Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models[*]

Yingying Dong and Arthur Lewbel

University of California Irvine and Boston College

First version July 2010, revised December 2012

Abstract

   Regression discontinuity models, where the probability of treatment jumps discretely when a running variable crosses a threshold, are commonly used to nonparametrically identify and estimate a local average treatment effect. We show that the derivative of this treatment effect with respect to the running variable is nonparametrically identified and easily estimated. Then, given a local policy invariance assumption, we show that this derivative equals the change in the treatment effect that would result from a marginal change in the threshold, which we call the marginal threshold treatment effect (MTTE). We apply this result to Manacorda (2012), who estimates a treatment effect of grade retention on school outcomes. Our MTTE identifies how this treatment effect would change if the threshold for retention was raised or lowered, even though no such change in threshold is actually observed.

JEL Codes: C21, C25
Keywords: regression discontinuity, sharp design, fuzzy design, treatment effects, program evaluation, threshold, running variable, forcing variable, marginal effects.

1   Introduction

Consider a standard regression discontinuity (RD) model, where T is a treatment indicator, X is a

so-called running or forcing variable, c is the threshold for X at which the probability of treatment

changes discretely, and Y is some observed outcome that may be affected both by treatment and

smoothly by X. The goal in these models is to estimate the effect of treatment T on the outcome Y, and the main result in this literature is that under weak conditions this treatment effect can be nonparametrically identified and estimated at the point where $X = c$.

It is often useful to know how treatment effects would change if the threshold c were changed (examples are given below). But in most applications, we do not observe a change in threshold, and so cannot directly estimate the impact of a threshold change. This paper shows what can be nonparametrically identified about the impacts of a threshold change, when no such change is observed.

Let $\tau(C)$ denote the (local) average treatment effect that would be identified by standard RD methods if the threshold equaled C, and let $\tau'(C) = \partial \tau(C)/\partial C$. The true threshold is c, so the treatment effect that can be estimated is $\tau(c)$. We define the Marginal Threshold Treatment Effect (MTTE) at c to be $\tau'(c)$. The MTTE, if it were known, could be used to evaluate the impact of a small discrete change in c, exactly the way that, e.g., price elasticites are used to approximate the effects of small discrete changes in prices such as those arising from a small change in a sales tax or a value added tax. For example, the sign of the MTTE tells whether the effectiveness of treatment would be increased or decreased if the threshold for treatment were lowered. If we wanted to know what the treatment effect would be at a new threshold c_{new} that is close to c, given $\tau(c)$ and $\tau'(c)$, we could approximate this $\tau(c_{new})$ using the mean value expansion $\tau(c_{new}) \approx \tau(c) + (c_{new} - c)\tau'(c)$. We also provide conditions under which $\tau(c_{new})$ can be exactly rather than approximately identified nonparametrically.

In discussing RD methods Hahn, Todd, and van der Klaauw (2001) note that, "A limitation of the approach is that it only identifies treatment effects locally at the point at which the probability of receiving treatment changes discontinuously... It would be of interest, for example, if the policy change being considered is a small change in the program rules, such as lowering or raising the threshold for program entry, in which case we would want to know the effect of treatment for the subpopulation affected by the change." Our MTTE addresses this issue, by showing how the effect of treatment changes given a marginal change in the threshold. Our results may also be taken as an example of a marginal policy analysis of the sort advocated by Heckman (2010) and Carneiro,

Heckman, and Vytlacil (2010).

To illustrate the usefulness of the MTTE $\tau'(c)$, consider a few examples. Jacob and Lefgren (2004) consider the impact of attending summer school on later academic performance. In their application, the treatment T is summer school attendance, X is the test score on an exam, c is the threshold failing score below which one is required to attend summer school, and the outcome Y is a later test score. The MTTE in this case can be used to evaluate the effectiveness of summer school for students at the new lower (higher) cutoff if the cutoff score for determining summer school attendance is marginally lowered (raised).

A similar example is the original Thistlethwaite and Campbell (1960) RD paper, where T is receipt of a National Merit Award, X is the test score on the National Merit Award qualifying exam, c is the exam grade required to qualify for the award, and Y is receipt of other college scholarships (and other academic outcomes). In this application the treatment effect is the increase in the probability of receiving college scholarships resulting from winning the National Merit Award, and our paper's goal would be to evaluate what the probability of obtaining college scholarships would be for the new marginal winners if the National Merit Award were given at a lower or higher exam grade level.

Another example is Chay and Greenstone (2005), who examine impacts of the US Clean Air Act Amendments (CAAA) of 1970. The CAAA requires that a county be designated "nonattainment" status if its pollution concentrations exceed a federally determined ceiling. The CAAA imposes stringent pollution abatement regulations on "nonattainment" counties. Here T indicates nonattainment status so the treatment consists of stringent pollution regulations, X is the pollution concentration measure, c is the pollution ceiling, and Y is the subsequent pollution reduction or a side effect like housing value increases. In this sharp RD design, our MTTE would be used to address questions such as how the effectiveness of the regulations in reducing pollution would change, or how housing prices would be affected at the new pollution ceiling, if the pollution ceiling were marginally lowered.

A final example (which we empirically implement) is Manacorda (2012), who estimates the treatment effect of grade retention on school outcomes in Uruguay. This is a fuzzy design RD in which the running variable X is a student's course failure rate in junior high school, treatment is retention (that

is, repeating a grade in junior high school), and the outcome Y is the additional grades of schooling attained in the next five years. In this fuzzy design, compliers are defined as students that would be required to repeat a grade if and only if their measured X exceeds the threshold $c = 3$. Like Manacorda, we obtain a treatment effect $\tau(c)$ of more than $-3/4$, meaning that among compliers, repeating a grade results in at least three fourths of a year reduction in educational attainment in the next five years on average. However, our new result is that we find an MTTE $\tau'(c)$ of about one fourth. The fact that the MTTE is positive means that if the standards for promotion were made stricter by reducing the course failure cutoff (reducing c), then the compliers who are retained in this new regime would be more adversely affected than the compliers at the original threshold. For example if the course failure cutoff rate was reduced from $c = 3$ to $c_{new} = 2$, then using $\tau(c_{new}) \approx \tau(c) + (c_{new} - c)\tau'(c)$ the treatment effect at the new cutoff would be minus a full year of schooling or more on average.

Thresholds are often set by policy, and knowing the direction and magnitude of changes in effects resulting from a change in threshold can be important in practice. Many policy debates center precisely on these types of questions, e.g., what would be the effects on take-up or on various health and welfare measures if the qualifying standards for a certain welfare program were to change? Or, in the previously described applications, what are the impacts of changing the pollution regulation ceiling? Or the effects of changing grading thresholds for various educational interventions?

Let $Y(1)$ and $Y(0)$ denote the potential outcomes as in Rubin (1974), meaning the outcome when one is treated or not treated, so the observed outcome $Y = Y(1)T + Y(0)(1 - T)$. To simplify the discussion here, consider sharp designs first, where $T = I(X > c)$, meaning that $T = 1$ if $X > c$, otherwise $T = 0$. Define the function $D(X) = E[Y(1) - Y(0) \mid X]$ holding c fixed, and define its derivative $D'(X) = \partial D(X)/\partial X$. By definition, $D(X)$ is the average treatment effect of an individual having running variable equal to X. These definitions of the functions $D(X)$ and $D'(X)$ make the assumption, which is standard in RD analyses, that the threshold c is a fixed constant. But our interest is in analyzing the effects of changing the threshold.

To consider the possibility of changing the threshold, we let C denote a possible threshold value. The distributions of $Y(1)$ and $Y(0)$ can depend on both X and C. For example, a person's outcome

if treated, given by $Y(1)$, can depend both on what the threshold is for treatment, and on the value of that person's running variable. For example, in the case of national merit awards, the threshold C would affect the total the number of students who get such awards, which could in turn affect the difference in probability of getting a scholarship, and hence the treatment effect, of giving an award to an individual with running variable equal to X.

In contrast to the function D, the function $\tau(C)$ is defined as equaling what the RD treatment effect would be if the threshold were equal to C. At the actual, realized value of the threshold c we have $\tau(c) = D(c)$, but τ and D are in general different functions.[1]

We first show that with minimal regularity conditions, $D'(c)$ (the derivative of $D(X)$ evaluated at the point $X = c$) is nonparametrically identified and can be easily estimated. Call $D'(c)$ the Treatment Effect Derivative, or TED. The function $D'(c)$ has some interesting uses by itself, e.g., Abdulkadiroglu, Angrist, and Pathik (2011) apply our TED estimator in a context where the treatment is exam school attendance and the outcome is scores on standardized achievement tests. The TED can also be used to test if treatment effects are locally constant, since a necessary condition for locally constant treatment effects is that $D'(c) = 0$.

However, what we mainly use the TED for is to show is that if (and only if) a local policy invariance assumption holds, then the TED $D'(c)$ will also equal the MTTE $\tau'(c)$, so in that case the MTTE is nonparametrically identified and easily esimated.

Abbring and Heckman (2007) define policy invariance as an "assumption that an agent's outcome only depends on the treatment assigned to the agent and not separately on the mechanism used to assign treatments. This excludes (strategic) interactions between agents and equilibrium effects of the policy." Example applications of policy invariance assumptions and marginal policy analyses include

---

[1]Formally, we can define the function $S(X, C) = E(Y(1) - Y(0) \mid X, C)$, which is the expected difference in treated vs untreated outcomes of a hypothetical individual having a running variable equal to X, living in an environment in which everybody (else) is treated if and only if their running variable exceeds C. The functions D and τ are defined by $D(X) = S(X, c)$ and $\tau(C) = S(C, C)$. The standard sharp design RD treatment effect is then $D(c) = \tau(c)$, which is identified by $D(c) = \lim_{x \downarrow c} E(Y \mid X = x, C = c) - \lim_{x \uparrow c} E(Y \mid X = x, C = c)$. See the Appendix for details, where we further generalize these definitions to handle fuzzy designs by conditioning on compliers.

Heckman (2010) and Carneiro, Heckman, and Vytlacil (2010). What we require to make $\tau'(c) = D'(c)$ is a limited version of policy invariance that only applies to infinitesimal changes in the assignment mechanism (specifically, changes in the threshold c) and which we therefore refer to as, "local policy invariance. We show that this local policy invariance, which we discuss in detail in the next section, is a necessary and sufficient condition to make $D'(c) = \tau'(c)$. Note that this equality only needs to hold at the true threshold c, so in general we can have $\tau'(C) \neq D'(C)$ for $C \neq c$."

The above discussion focused on the sharp design case, but we show later that the above definitions and results can all be generalized to identify the treatment effect and MTTE in fuzzy designs as well.

In many RD applications the function $E(Y \mid X = x)$ is parameterized, e.g., it is assumed to be a polynomial in chapter 6 of Angrist and Pischke (2008). When this expectation is parameterized, the function $D(x)$ can be directly estimated for all values of x. For example, in a linear sharp design model where $Y = a + Xb + T\beta + XT\gamma + e$ and $E(e \mid X = x) = 0$, we would have $D(x) = \beta + x\gamma$, so $D'(x) = \gamma$, and so the TED (and, given local policy invariance, the MTTE) would be identified. But is this identification due to the assumed functional form, or can the TED be nonparametrically identified?

In parametric models $D(x)$ is identified both at $x = c$ and for values $x \neq c$, permitting identification of $D'(c)$, only because the functional form allows us to evaluate counterfactual objects like $E(Y \mid T = 1, X = x < c)$, even though in the data with a sharp design we could never see any observations having both $T = 1$ and $x < c$. One might more generally think that nothing regarding changes in c can be identified nonparametrically, because we only observe treatment at $x = c$ itself.

However, what we show is that, analogous to the way $D(c)$ in sharp designs is nonparametrically identified by differencing the left and right limits of $E(Y \mid X = x)$ as $x \to c$, one can similarly nonparametrically identify $D'(c)$ by differencing the left and right derivatives of $E(Y \mid X = x)$ as $x \to c$. We also show that the TED can be identified and easily estimated in fuzzy designs as well (though with a somewhat more complicated formula). In the fuzzy design case we will also still be able to provide a local policy invariance assumption that makes the TED equal the MTTE.

A few other papers exist that appeal to derivative conditions in RD analyses. Card, Lee, Pei,

and Weber (2012) analyze a regression kink design model, where the treatment is a continuous but kinked function of the running variable, so the kink, which is a change in the derivative, identifies the treatment effect. Dong (2012) uses changes in the derivative or slope of the treatment probability function at the threshold to identify standard fuzzy RD design treatment effects in applications where there is a kink instead of, or in addition to, a discontinuity at the threshold. Perhaps the closest result to ours is a few paragraphs in a survey article by Dinardo and Lee (2011), in which they informally propose using a Taylor expansion at the threshold to identify an average treatment effect on the treated (ATT) parameter. In contrast, we use a similar expansion to estimate a different object, that is, we consider the impact of changing the threshold, and we provide formal results for both fuzzy and sharp designs.

In the next section we show nonparametric identification of the TED in sharp RD designs, and describe in detail the local policy invariance condition required to make the MTTE equal the TED. Later sections consider identification when changes in the threshold are larger than marginal and identification in fuzzy designs. We then discuss estimation based on these identification results, and provide empirical estimates based on Manacorda's (2012) study of grade retention effects on school outcomes in Uruguay. This is followed by conclusions, and an Appendix containing proofs.

2 Sharp Design TED and MTTE

We present our results for sharp design RD models first, and later consider the extension to fuzzy designs. For simplicity we give assumptions and results without consideration of covariates other than the running variable X. We later discuss how additional covariates Z could be included in the regressions.

ASSUMPTION A1: For each unit (individual) i we observe Y_i, T_i, X_i where T_i is a binary treatment indicator, X_i is a running variable, and $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$ for potential outcomes $Y_i(1)$ and $Y_i(0)$.

For ease of notation we will drop the i subscript when refering to the random variables $Y(1)$,

$Y(0)$, Y, T, and X.

ASSUMPTION A2 (sharp design): $T = I(X > c)$ for some known constant threshold c. The support of X includes a neighborhood of c. For $t = 0$ and $t = 1$, $E(Y(t) \mid X = x)$ is continuously differentiable in x in a neighborhood of $x = c$.

Define $g(x) = E(Y \mid X = x)$. The main identification result in the literature for sharp design RD is that the value of the treatment effect $\tau(c)$ is identified by[2]

$$\tau(c) = \lim_{x \downarrow c} g(x) - \lim_{x \uparrow c} g(x). \tag{1}$$

based on

$$E(Y(0) \mid X = c) = \lim_{x \uparrow c} E(Y(0) \mid X = x) = \lim_{x \uparrow c} E(Y(0) \mid X = x, T = 0) = \lim_{x \uparrow c} E(Y \mid X = x) \tag{2}$$

and similarly for $Y(1)$ using the limit $x \downarrow c$.

Assumptions A1 and A2 are identical to those used in the literature to prove equation (1) via equation (2) arguments, except that A2 assumes not just continuity, but differentiability. However, virtually all empirical implementations of RD models satisfy this stronger smoothness assumption. In particular, parametric models generally assume polynomials or other differentiable functions, while most nonparametric estimators, including local linear regressions, assume (for establishing asymptotic theory) at least continuous differentiability of $E(Y(1) \mid X = x)$ for $x \geq c$ and of $E(Y(0) \mid X = x)$ for $x \leq c$.[3] We do not know of any application of RD methods where the usual continuity assumptions hold but differentiability does not.

---

[2]See Hahn, Todd, and van der Klaauw (2001), or for recent surveys see Lee and Lemieux (2010), Imbens and Wooldridge (2009) and Imbens and Lemieux (2008).

[3]Local linear or higher order local polynomial regressions are used to estimate RD models to mitigate boundary bias issues as discussed by, e.g., Porter (2003). The asymptotic theory for local linear or local polynomial estimation (see Fan and Gijbels 1996) requires not just continuity but continuous differentiability. Generally twice differentiability or more smoothness is assumed. This differentiability is not required for consistent estimation of RD models, but is always assumed in practice to reduce bias and thereby increase estimation precision. Similarly, still greater smoothness would be useful for bias mitigation in derivative estimation, but is not required for consistency of MTTE estimation.

Recall from the introduction that $D(x) = E(Y(1) - Y(0) \mid X = x)$ is the average treatment effect across individuals having running variable equal to x and $D'(x) = \partial D(x)/\partial x$. For example, in the summer school application $D(x)$ is the expected effect of attending summer school for an average individual having test score x. Sharp design RD identifies $D(c)$, the value of the function $D(x)$ just at the point $x = c$. We now similarly show nonparametric identification of the TED $D'(c)$.

As discussed earlier, the function $D(x)$ depends on the distributions of $Y(1)$ and $Y(0)$, which can in turn depend on the threshold c. The definitions of $D(x)$ and $D'(x)$ hold c fixed (see the Appendix for more details and more precise definitions). It will be convenient later to use the notation $h_+(x) = \lim_{\varepsilon \downarrow 0} h(x + \varepsilon)$ and $h_-(x) = \lim_{\varepsilon \uparrow 0} h(x + \varepsilon)$ for any function h. In this notation we have

$$D(c) = g_+(c) - g_-(c). \tag{3}$$

To show identification of the TED we require one-sided derivatives. The right and left derivatives of a function $h(x)$ at the point x, which we denote as $h'_+(x)$ and $h'_-(x)$ respectively, are defined by

$$h'_+(x) = \lim_{\varepsilon \downarrow 0} \frac{h_+(x + \varepsilon) - h_+(x)}{\varepsilon} \quad \text{and} \quad h'_-(x) = \lim_{\varepsilon \uparrow 0} \frac{h_-(x + \varepsilon) - h_-(x)}{\varepsilon}$$

A property of right and left derivatives is that if a function $h(x)$ is differentiable at a point x, then $h'_+(x) = h'_-(x) = \partial h(x)/\partial x = h'(x)$.

THEOREM 1: If Assumptions A1 and A2 hold then the treatment effect $D(c) = \tau(c)$ is identified by equation (3) and the function $D'(x)$ evaluated at the point $x = c$ (the TED) is identified by

$$D'(c) = g'_+(c) - g'_-(c). \tag{4}$$

Proofs are in the appendix. Estimation based on the identification in Theorem 1 will be discussed in more detail later, but for now note that local polynomial regressions can be used to estimate the function $g(x)$ and its derivatives separately on either side of the threshold. Evaluting these regression derivatives in the limits as $x \to c$ provides consistent estimates of $g'_+(c)$ and $g'_-(c)$, and hence consistent estimates of $D'(c)$. In short, $D'(c)$ equals the difference between the left and right derivatives of $g(x)$ around $x = c$, just as the local treatment effect $D(c)$ equals the difference between the left and right limits of $g(x)$ around $x = c$.

The TED $D'(c)$ provides a measure of the impact of a marginal change in the running variable x on the treatment effect, at $x = c$, holding c fixed. This estimate can be used for a variety of purposes as discussed in the introduction, but we will now focus on its application to the MTTE.

Now consider policy invariance. Roughly speaking, we say that "local policy invariance" holds if an infinitesimal change in the true threshold c would not change the function $D(x)$ at points x in an arbitrarily small neighborhood of $x = c$. [4] It is important to note that local policy invariance does not imply, and is not implied by, locally constant treatment effects. In fact, local policy invariance places no restriction at all on the shape of the function $D(x)$. It only restricts how this function could change if the threshold changed marginally.

To further explain local policy invariance, consider first the examples of RD discussed in the introduction where treatment is attending summer school or winning a National Merit Award. Consider a group of compliers who have test scores x that equal a value c_{new} that is infinitesimally larger or smaller than c. Policy invariance means that the average treatment effect for this group would not change if the threshold test score used for determining treatment were changed infinitesimally from c to c_{new}. In the summer school example, this local policy invariance might not hold if a marginal change in the qualifying score affects the curriculum or quality of instruction in summer school, or if the resulting small compositional change in the population of students attending summer school affects learning through peer effects. In the National Merit Award example, local policy invariance might be violated if the increase in the number of award winners resulting from an infinitesimal lowering of the qualifying score c resulted in increased competition for outside scholarships, or in a lowering in the perceived prestige of an award.

In the pollution application, a stronger than necessary condition for local policy invariance to hold is if the impact of the pollution regulation associated with non-attainment status depends only on a county's pollution level x, but not on the actual pollution ceiling for assigning the non-attainment

---

[4]More precisely, consider the function $S(X, C) = E(Y(1) - Y(0) \mid X, C)$ discussed in footnote 1. For sharp design RD, Local policy invariance says that $\partial S(X, C) / \partial C$ equals zero when $X = C = c$. See the Appendix for more formal details.

status, which is likely to be the case.

As these examples show, local policy invariance is essentially a ceteris paribus assumption of the type commonly employed in partial equilibrium analyses. It can also be interpreted as an external validity assumption, since it assumes that a function we can identify does not change when one feature of the environment (specifically, the value of c) changes infinitesimally. Analogous ceteris paribus or external validity assumptions are required to apply almost any reduced form program evaluation calculation to a change in context or environment.

COROLLARY 1: Let Assumptions A1 and A2 hold. Then the MTTE $\tau'(c)$ is nonparametrically identified by $\tau'(c) = D'(c)$ if and only if local policy invariance holds.

Without futher assumptions (which we discuss in the next section), the functions $\tau(C)$ and $\tau'(C)$ are only identified nonparametrically at the observed true threshold level $C = c$, and not at other values of C. However, given the MTTE $\tau'(c)$, we can use the mean value theorem to obtain an approximate estimate of the effect of a small discrete change in the threshold. For example, an estimate of what the treatment effect $\tau(c_{new})$ would be if the threshold were changed a small amount from c to c_{new} is

$$\tau(c_{new}) \approx \tau(c) + (c_{new} - c)\tau'(c). \qquad (5)$$

3 Extensions to Nonmarginal Changes

Equation (5) gives an estimate of the treatment effect $\tau(c_{new})$ when c_{new} is a marginal change in c. Here we discuss what can be done to consider larger changes in c. One possibility is to refine equation (5) using higher order derivatives.

Theorem 1 can be immediately extended to identify and estimate higher order derivatives. Let $D''(x) = \partial^2 D(x)/\partial^2 x$. If we replace the continuous differentiability in Assumption A2 with twice continuous differentiability for x in a neighborhood of c, then by taking second order instead of first order derivatives in the proof of Theorem 1 we obtain $D''(c) = g''_+(c) - g''_-(c)$. Given local policy invariance for second order derivatives (see the Appendix for details), we would have $\tau''(c) =$

11

$D''(c)$ is identified. If sufficient data are available to precisely estimate second derivatives in the neighborhood of $x = c$, then these could be used to further refine estimates of the effects of small discrete changes in c. For c_{new} close to c, a second order Taylor expansion gives

$$\tau(c_{new}) \approx \tau(c) + (c_{new} - c)\tau'(c) + \frac{(c_{new} - c)^2}{2}\tau''(c). \tag{6}$$

In the same way, if $E(Y(t) \mid X = x)$ for $t = 0$ and $t = 1$ are analytic functions, then derivatives $D^k(x) = \partial^\kappa D(x)/\partial^\kappa x$ at $x = c$ for all integers κ exist and can be correspondingly identified. Given a strong form of local policy invariance, the following corollary shows that, in theory at least, $\tau(c_{new})$ can be exactly rather than approximately identified, making use of an infinite order Taylor expansion.

COROLLARY 2: Let Assumptions A1 and A2 hold. Assume that $E(Y(t) \mid X = x)$ for $t = 0$ and $t = 1$ are analytic functions and that strong local policy invariance holds. Then for any threshold c_{new} the local average treatment effect $\tau(c_{new})$ is identified and is given by the Taylor expansion

$$\tau(c_{new}) = \tau(c) + \sum_{k=1}^{\infty} \frac{(c_{new} - c)^k}{k}D^k(c). \tag{7}$$

The proof Corollary 2, and a formal definition of strong local policy invariance, is provided in the Appendix. Given Corollary 2 one might construct a consistent estimator of $\tau(c_{new})$ based on the method of sieves, using a K'th order Taylor expansion to estimate $\tau(c_{new})$ via equation (7), with each $D^k(c)$ for $k = 0, ..., K$ estimated by K order polynomial regressions, and letting $K \to \infty$ as the sample size goes to infinity.

Another way to obtain a consistent estimator of $\tau(c_{new})$ would be to assume a functional form for the treatment effect. For example, if we assume the function $\tau(c)$ is linear, then equation (5) will hold exactly rather than approximately and could then be applied to any size change to c_{new} from c. Similarly, if $\tau(c)$ is quadratic then equation (6) becomes exact. These assumptions would still be less restrictive than the requirement that one have a complete, correctly specified parametric model for Y.

4 Fuzzy Design TED and MTTE

We now extend our previous results to fuzzy designs. Let T continue to indicate whether an individual is treated, but now let $T^* = I\,(X > c)$, so T would be the same as T^* for all individuals if the design were sharp. Define $f\,(x) = E\,(T \mid X = x)$, so $f\,(x)$ is the conditional probability of being treated, conditional upon $X = x$. The counterfactuals $Y\,(0)$ and $Y\,(1)$ are defined in terms of T, not T^*, so we still have $Y = Y\,(1)\,T + Y\,(0)\,(1 - T)$. Analogous to $Y\,(0)$ and $Y\,(1)$ in the sharp design, define the potential treatment status variables $T\,(0)$ and $T\,(1)$, where $T\,(t^*)$ is what an individual's treatment status would be if $T^* = t^*$, so $T = T\,(1)\,T^* + T\,(0)\,(1 - T^*)$. Compliers can then be defined as individuals i having $T_i\,(0) < T_i\,(1)$, while defiers have $T_i\,(0) > T_i\,(1)$ (always takers and never takers have $T_i\,(0) = T_i\,(1)$ equal to one or zero, respectively).

Analogous to $D\,(x)$ in sharp designs, define

$$D_f\,(x) = E\,[Y\,(1) - Y\,(0) \mid X = x, T\,(0) < T\,(1)] \tag{8}$$

(see the Appendix for details). The only difference between the functions $D_f\,(x)$ and $D\,(x)$ is that the definition of $D_f\,(x)$ conditions explicitly on compliers, $T\,(0) < T\,(1)$. Both $D_f\,(x)$ and $D\,(x)$ implicitly condition on the true threshold equaling c.

ASSUMPTION B2 (Fuzzy design): The support of X includes a neighborhood of a known threshold constant c. $T\,(0) \leq T\,(1)$. The functions $E\,(Y\,(t) \mid X = x)$ and $E\,(T\,(t^*) \mid X = x)$ are continuously differentiable in x in a neighborhood of $x = c$.

The monotonicity assumption $T\,(0) \leq T\,(1)$ in Assumption B2 rules out defiers. Based on Assumptions A1 and B2, $D_f\,(x)$ equals the treatment effect that is identified by fuzzy design RD, using the formula

$$D_f\,(c) = \frac{g_+\,(c) - g_-\,(c)}{f_+\,(c) - f_-\,(c)} \tag{9}$$

Assumptions A1 and B2 are equivalent to standard assumptions used in Fuzzy RD design identification to obtain equation (9) (see, e.g., Hahn, Todd, and van der Klaauw (2001) or Imbens and

Lemieux (2008)) except that the functions in B2 are assumed to be differentiable instead of just continuous. As noted in discussing the sharp design, the estimators that are always used in practice to estimate RD models implicitly or explicitly assume this differentiability. Given the differentiability in Assumption B2, define $D'_f(x) = \partial D_f(x)/\partial x$. The TED for compliers is then defined by $D'_f(c)$.

The following Theorem shows identification of the TED $D'_f(c)$. As in sharp designs, a necessary condition for local constant treatment effects is $D'_f(c) = 0$, so estimation of the TED can be used for testing that condition.

THEOREM 2: If Assumptions A1 and B2 hold then the treatment effect $D_f(c) = \tau_f(c)$ is identified by equation (9) and the TED $D'_f(c)$ is identified by

$$D'_f(c) = \frac{g'_+(c) - g'_-(c) - \left[f'_+(c) - f'_-(c)\right] D_f(c)}{f_+(c) - f_-(c)} \tag{10}$$

Let $p(c)$ denote the fraction of the population that are compliers when the threshold equals c. The proof of Theorem 2 shows that $p(c) = f_+(c) - f_-(c)$, the denominator of equation (10). Applying the mean value theorem as before shows that we can approximate the proportion of the population who would be compliers at a new threshold c_{new} by

$$p(c_{new}) \approx p(c) + (c_{new} - c) p'(c). \tag{11}$$

The proof of Theorem 2 also shows that $p'(c) = f'_+(c) - f'_-(c)$, so Equation (10) can be written as the sum of two terms, $\left[g'_+(c) - g'_-(c)\right]/p(c)$ and $p'(c) D_f(c)/p(c)$. The first of these terms is what the TED would equal if the probability of compliance $p(c)$ was held fixed, while the second term, which is proportional to $p'(c)$, accounts for the effect on the TED of changes in the probability of compliance that can occur when c marginally changes.

Define $\tau_f(C)$ the same as $\tau(C)$ but just for compliers, i.e., conditioning on the additional argument that $T(0) < T(1)$, in exactly the same way that $D_f(x)$ is defined relative to $D(x)$ (see the Appendix for details). It follows that at the observed threshold c we will have $\tau_f(c) = D_f(c)$. Let $\tau'_f(C) = \partial \tau_f(C)/\partial C$.

14

The fuzzy MTTE is defined to be $\tau'_f(c)$. To identify the fuzzy MTTE, we again need a policy invariance assumption. Extending the sharp design definition, we say that local policy invariance holds in the fuzzy design if an infinitesimal change in c does not change the function $D_f(x)$ at points x in an arbitrarily small neighborhood of $x = c$ (see the Appendix for details). This definition shows that local policy invariance only restricts how outcomes of compliers might change if the threshold c changes. Local policy invariance allows the composition of complier status to change when c changes. As in the sharp design case, local policy invariance does not place any restriction on how the treatment effect depends on the running variable.

COROLLARY 3: If Assumptions A1 and B2 hold and local policy invariance holds, then the fuzzy MTTE $\tau'_f(c)$ is nonparametrically identified by $\tau'_f(c) = D'_f(c)$.

Corollary 1 shows that if policy invariance holds, then analogous to the sharp design case, an estimate of the TED $D'_f(c)$ is also an estimate of the MTTE $\tau'_f(c)$. By the mean value theorem, the fuzzy design analog to equation (5),

$$\tau_f(c_{new}) \approx \tau_f(c) + (c_{new} - c)\,\tau'_f(c) \tag{12}$$

can be used as before to approximate the effect of treatment on compliers if the threshold is changed a small amount from c to c_{new}. Recall that $p(c) = f_+(c) - f_-(c)$ is the fraction of the population that are compliers when the threshold equals c. Even though the set of compliers can change in unknown ways when the threshold changes, we can approximate both $p(c_{new})$, the fraction of the population who would be compliers at the new threshold c_{new}, and the treatment effect $\tau_f(c_{new})$ on those compliers, using equations (11) and (12).

Fuzzy design analogs to Corollary 2 can also be immediately constructed, allowing exact rather than approximate identification of $\tau_f(c_{new})$ and $p(c_{new})$ if the conditional expectation functions in Assumption B2 are analytic. As in the sharp design case exact estimation of these objects would require increasing the order of derivatives one estimates as the sample size grows.

5 Estimation

Here we describe estimators for the TED and hence for the MTTE. The estimators we provide here are not themselves new, being equivalent to estimators for treatment effects like those summarized in surveys such as Imbens and Wooldridge (2009) and Lee and Lemieux (2010). We therefore do not provide associated limiting distribution theory, since it corresponds to the standard theory of estimation of local polynomial estimators, albeit at boundary points. See, e.g., Fan and Gijbels (1996) and Porter (2003). What is new here is not the estimators themselves, but rather their application to estimation of the TED.

First consider local linear estimation of sharp design RD models with a uniform kernel. This is equivalent to selecting observations i such that $-\varepsilon \leq X_i - c \leq \varepsilon$ for a chosen bandwidth ε, and then using just those observations to estimate the model

$$Y_i = \alpha + (X_i - c)\,\delta + T_i^*\beta + (X_i - c)\,T_i^*\gamma + e_i \tag{13}$$

by ordinary least squares, where $T_i^* = I(X_i > c)$. Let $\widehat{\alpha}$, $\widehat{\delta}$, $\widehat{\beta}$, and $\widehat{\gamma}$, be these ordinary least squares estimates of the coefficients α, β, δ, and γ. The line $(\alpha + \beta) + (\delta + \gamma)(X - c)$ is a linear approximation to $E(Y \mid X, c \leq X \leq c + \varepsilon)$, so by the standard theory of local linear estimation, in the limit as $\varepsilon \to 0$ (at an appropriate rate) this approximation becomes equal to the tangent line to the function $g(x)$ as $x \downarrow c$, making $plim\left(\widehat{\alpha} + \widehat{\beta}\right) = g_+(c)$ and $plim\left(\widehat{\delta} + \widehat{\gamma}\right) = g'_+(c)$. Similarly, the line $\alpha + \delta(X - c)$ is a linear approximation to $E(Y \mid X, c - \varepsilon \leq X < c)$, making $plim\left(\widehat{\alpha}\right) = g_-(c)$ and $plim\left(\widehat{\delta}\right) = g'_-(c)$.

It follows immediately that

$$plim\left(\widehat{\beta}\right) = g_+(c) - g_-(c) \quad \text{and} \quad plim\left(\widehat{\gamma}\right) = g'_+(c) - g'_-(c) \tag{14}$$

and therefore, by Theorem 1, $\widehat{\tau}(c) = \widehat{D}(c) = \widehat{\beta}$ is a consistent estimator of the sharp design treatment effect and $\widehat{D}'(c) = \widehat{\gamma}$ is a consistent estimator of the sharp design TED. Given local policy invariance, by Corollary 1 $\widehat{\tau}'(c) = \widehat{\gamma}$ is a consistent estimator of the sharp design MTTE.

This local linear estimator of $\widehat{\tau}(c)$ is standard in the literature and commonly used. What we are

adding here is just the observation that these same local linear estimators also provide estimates of derivatives, which is what we need to recover the TED and hence the MTTE.

Porter (2003) suggests using local linear regressions rather than ordinary kernel (locally constant) estimators because inclusion of the linear terms $(X_i - c)$ and $(X_i - c) T_i^*$ reduces small sample boundary bias in $\widehat{\alpha}$ and $\widehat{\beta}$. For the same reason, to reduce bias in $\widehat{\delta}$ and $\widehat{\gamma}$, it could be advantageous to use local quadratic rather than local linear estimation, which with a uniform kernel corresponds to ordinary least squares estimation of the model

$$Y_i = \alpha + (X_i - c)\,\delta + (X_i - c)^2\,\zeta + T_i^*\beta + (X_i - c)\,T_i^*\gamma + (X_i - c)^2\,T_i^*\xi + \widetilde{e}_i \quad (15)$$

Equation (14) only depends on consistency of local (quadratic) polynomial estimation, and so will continue to hold if equation (15) is used instead of equation (13).

For fuzzy designs, we can still use equation (13) or (15), but in addition, we can use local linear or local quadratic estimation on the same observations i such that $-\varepsilon \leq X_i - c \leq \varepsilon$ to estimate

$$T_i = \alpha_T + (X_i - c)\,\delta_T + T_i^*\beta_T + (X_i - c)\,T_i^*\gamma_T + e_{Ti} \quad (16)$$

or

$$T_i = \alpha_T + (X_i - c)\,\delta_T + (X_i - c)^2\,\zeta_T + T_i^*\beta_T + (X_i - c)\,T_i^*\gamma_T + (X_i - c)^2\,T_i^*\xi_T + \widetilde{e}_{Ti} \quad (17)$$

by ordinary least squares, yielding estimated coefficients including $\widehat{\beta}_T$ and $\widehat{\gamma}_T$. By the exact same derivations as above, we obtain consistent estimators

$$plim\left(\widehat{\beta}_T\right) = p\,(c) = f_+\,(c) - f_-\,(c) \quad \text{and} \quad plim\left(\widehat{\gamma}_T\right) = p'\,(c) = f'_+\,(c) - f'_-\,(c) \quad (18)$$

Then by Theorem 2 we have consistent estimators of the fuzzy design treatment effect and TED given by

$$\widehat{D}_f\,(c) = \widehat{\tau}_f\,(c) = \widehat{\beta}/\widehat{\beta}_T \quad \text{and} \quad \widehat{D}'_f\,(c) = \left(\widehat{\gamma} - \widehat{\gamma}_T\widehat{D}_f\,(c)\right)/\widehat{\beta}_T \quad (19)$$

and, given policy invariance, the estimated fuzzy design MTTE is $\widehat{\tau}'_f\,(c) = \widehat{D}'_f\,(c)$. Note that the ratio of ordinary least squares estimates $\widehat{\beta}/\widehat{\beta}_T$ is numerically equivalent to linear instrumental variables

17

estimation of β in equation (13) using T^* as an instrument for T, which is how fuzzy design estimation is often implemented. See, e.g., Hahn, Todd, and van der Klaauw (2001).

In empirical practice, it is common to add other covariates Z_i as additional controls in these regressions. Adding or omitting these additional terms can be helpful for estimation precision, but does not affect the consistency of the above described estimators. All of our results extend immediately to the inclusion of covariates.[5]

6   The Cost of Grade Retention

Here we follow Manacorda (2012) in estimating the impact of retaining underperforming students (i.e, making them repeat a grade school) in Uruguay. The running variable X is the number of subjects that students are observed to fail in junior high school (grades 7 to 9) in 1996 and 1997. Students who have more than three accumulated subject failures by the beginning of the following school year are required to repeat a grade, so treatment T equals one if a student fails a grade and hence is required to repeat that grade, and zero otherwise. Students are followed until 2001. The outcome Y is defined as the additional grades attended by 2001 for students who were retained, and by 2000 for students who were not (the difference in years is to make up for the year lost due to grade repetition by the retained students). Not all students who have $X > c$ are retained, because students who fail a subject may retake the exam and pass before the following school year starts. Also, students with fewer than three failed subjects can still be retained if they have more than 25 missed school days, Still, the administrative rule results in a significant discontinuity. The retention rate changes by close to 60% when the observed failed subjects X exceeds $c = 3$, so this is a fuzzy design RD.

---

[5]A conditional on covariates MTTE could be obtained by including covariates interacted with T^* in the estimating regressions, and an unconditional MTTE would then be obtained by averaging the conditional MTTE over covariate values. Conditioning on covariates could help in making local policy invariance hold. For example, in applications where treatment is summer school, if the only source of potential violation of local policy invariance is that a change in threshold could change summer school class sizes, then including class sizes as an additional covariate could suffice to satisfy a conditional local policy invariance restriction.
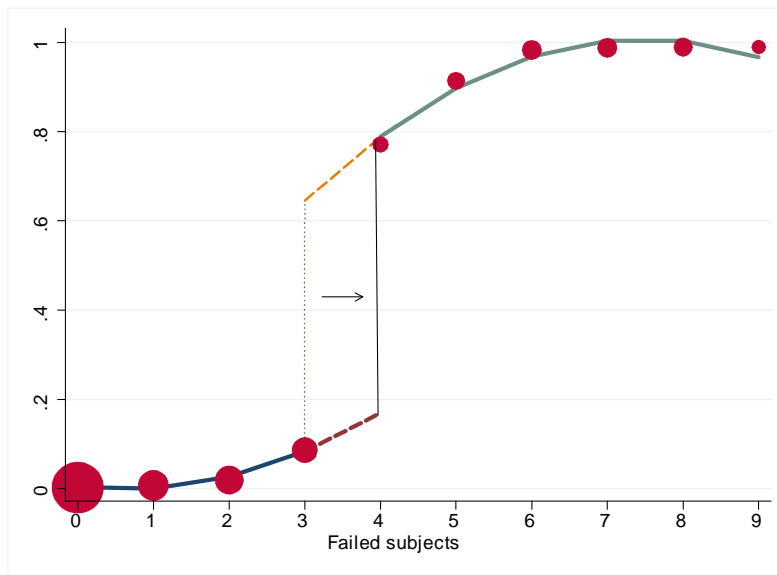
Figure 1: Grade failure by number of failed subjects in junior high school, Uruguay, 1996-1997

We use the full sample from Manacorda (2012). Details of the data can be found therein. The data contains 99,729 observations on 73,621 individual students. In this data roughly one in four students fail a grade in 1996 or 1997 and are therefore retained in those years, so the number of students affected by the treatment is large. While passers on average attended around 2.4 additional grades, failers only attended about extra 0.6 years of school. Manacorda reports estimated treatment effects ranging from minus three fourths to more than minus one, which means that among compliers, retention on average causes them to drop out of school at least 3/4 of a year earlier within the next 5 years, so grade retention is associated with large negative outcomes in terms of years of schooling later. It is therefore of substantive policy interest to know what the treatment effect would be at the marginally higher or lower threshold. For example, would the treatment effect be positive or negative (and larger or smaller) if the threshold were raised or lowered slightly? This is exactly what the sign and magnitude of the MTTE tell us.

Figures 1 and 2 show the rates of grade failure or grade retention and additional grades attended in five years for the students in the sample, corresponding to the functions f and g respectively. Both appear close to quadratic with obvious discontinuities after $c = 3$. In figures 1 and 2 we also illustrate the impact of changing the threshold c from three to four, analogous to what the MTTE estimates in
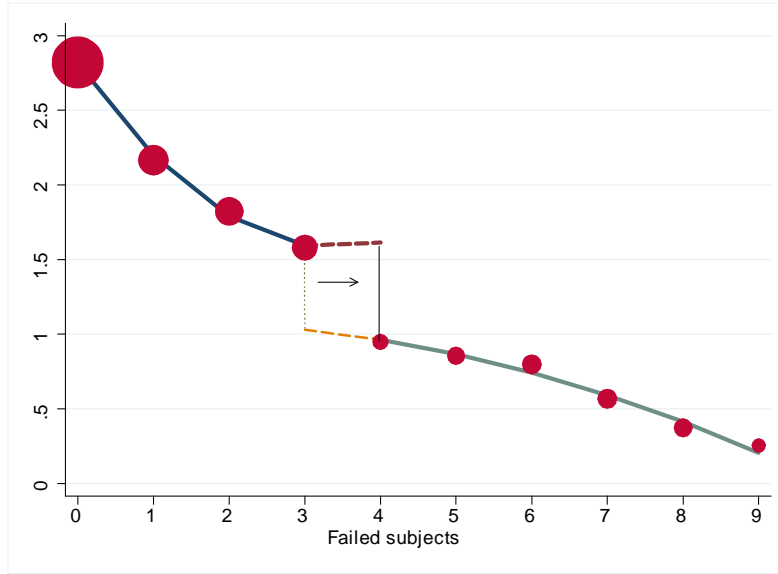
Figure 2: Additional grades attained by 2001 (2000 for nonfailers) by number of failed subjects in junior high school, Uruguay, 1996-1997

this case. Similarly one could easily envision the impact of changing the threshold from three to two instead, by extrapolating the upper curve in Figure 1 and the lower curve in Figure 2, which we do not depict in these figures.

Based on figures 1 and 2, we estimate treatment effects using the quadratic specification of equations (15) and (17). Manacorda reports estimates with second, third and fourth order polynomials. Higher order polynomials do not make much difference, consistent with the figures looking close to quadratic. In Table 1, column 1 gives the results of estimating these quadratic equations without controlling for additional covariates, which corresponds to the specification in Table 2 column 1 in Manacorda. We obtain the same treatment effect estimate as Manacorda. Columns 2 and 3 in Table 1 give estimates of the same equations after adding controls Z. In column 2 the controls are school, year, and grade fixed effects, and column 3 includes these fixed effects along with additional dummies for age, missed school days, subject scores, and grade-age distortions (see Manacorda for details regarding these controls).

The first panel of numbers in Table 1 are the estimates $\widehat{\beta}_T$ and $\widehat{\gamma}_T$, which are the coefficients of T_i^* and $(X_i - 3) T_i^*$ in equation (17), where $T_i^* = I(X_i > 3)$. The second panel of numbers are the

estimates $\widehat{\beta}$ and $\widehat{\gamma}$, which are the coefficients of T_i^* and $(X_i - 3) T_i^*$ in equation (15). The third panel of numbers gives the estimated treatment effect and treatment effect derivative $\widehat{D}_f(c)$ and $\widehat{D}'_f(c)$ obtained by equation (19).

TABLE 1 Estimated Treatment Effects and Derivatives of Grade Retention on Additional Grades Attended

	(1)	(2)	(3)
Dependent Variable: Grade Retention T			
I(X>3)	0.562	0.574	0.562
	(0.011)***	(0.010)***	(0.009)***
I(X>3)·(X−3)	0.091	0.084	0.076
	(0.033)***	(0.005)***	(0.006)***
Dependent Variable: Additional Grades Attained Y			
I(X>3)	-0.565	-0.472	-0.380
	(0.021)***	(0.031)***	(0.031)***
I(X>3)·(X−3)	0.034	0.039	0.111
	(0.007)***	(0.020)***	(0.023)***
Treatment Effect	-1.005	-0.822	-0.676
	(0.054)***	(0.051)***	(0.052)***
Treatment Effect Derivative	0.223	0.188	0.289
	(0.043)***	(0.039)***	(0.045)***

Note: Based on data from Manacorda (2012); both the reduced outcome and the treatment equations are quadratic regressions, fully interacted with the dummy indicating crossing the threshold, 1(X>3); (1)no controls; (2)controls for school, year and grade fixed effects; (3) controls for school, year and grade fixed effects and additional dummies for age, missed school days, subject scores, and grade-age distortion. Bootstrap standard errors are in parentheses. * significant at the 10% level, ** significant at the 5% level, *** significant at the 1% level.

Table 1 shows treatment effect estimates in the range -.68 to -1.01. This means that among compliers who fail more than three subjects, repeating a grade on average lead them to drop out of school around 2/3 of a year to a full year earlier within the next 5 years, which again confirms Manacorda's finding that the retention treatment has a substantial negative impact on the outcome of grades of schooling attained.

The TED estimates in Table 1 are in the range of .19 to .29, with an average of .23. This value differs substantially and significantly from zero, so we can strongly reject the hypothesis of locally constant treatment effects.

In order to interpret the estimated TED as MTTE, we need the policy invariance assumption to hold. Policy invariance in this case requires that students at the new threshold would not change their later schooling decision if there were a marginal change in the threshold, so their dropout decision (conditional on being retained or not) would not depend on the change in threshold, though it can depend on how many subjects they fail. Local policy invariance might be violated if, for example, a marginal change in the cutoff c leads to a change in the number or the composition of students in each grade that in turn impacts the later decisions of students to drop out, either by affecting the quality of instruction or through peer effects.

Assume that local policy invariance is satisfied, so the TED equals the MTTE. The fact that we find the MTTE is positive and large means that if the threshold c were lowered, so more students were retained, then the effects of this treatment would on average be substantially worse. Specifically knowing that the sign of the MTTE is positive tells policy makers that raising the standard required to pass a grade will cause new compliers to drop out even earlier on average.

As discussed after equation (10), by equation (19) we can decompose the estimated MTTE into the sum of the terms $\widehat{\gamma}/\widehat{\beta}_T$ and $-\left(\widehat{\gamma}_T/\widehat{\beta}_T\right)\widehat{\tau}_f\left(c\right)$. The first term estimates the MTTE or treatment effect change due to a change in the mean outcome difference for compliers, while holding the compliance rate constant, while the second term accounts for the change in the treatment effect due to a change in the compliance rate when c marginally changes. Our empirical estimates suggest that both terms contribute substantially to the final MTTE. So our estimated negative impact of raising standards is due both to a change in the behavior of compliers and a change in the compliance rate that would result from the change in threshold.

We can use equations (11) and (12) to evaluate the impacts of a specific marginal change in c. For convenience we will average results across the three specifications in Table 1, but note that similar results are obtained regardless of which specification is used.

First we consider changing the threshold from $c = 3$ to $c_{new} = 2$, corresponding to a raising of standards, so that students are required to repeat a grade if they fail more than two subjects instead of the existing standard of three subjects. Consider compliance rates first. By equation (18), the

estimated probability of compliance $p(c)$ is given by β_T, which in the first row of Table 1 is very precisely estimated at 0.57 (averaging across the three specifications). The derivative $p'(c)$ is estimated in the second row of Table 1 as around 0.08 which using equations (11) indicates that compliance rate at the new margin of $c_{new} = 2$ drops to 0.49 from 0.57 at the original $c = 3$ margin.

The estimated local average treatment effect at the observed $c = 3$ threshold is -0.83 (again averaging across the columns of Table 1) and the estimated MTTE is 0.23. This MTTE says that, on average, the new compliers at $c = 2$ would drop out of school a fourth of a year earlier than the old compliers if the standard for passing a grade were made more difficult.

By a similar analysis, if the threshold were moved from $c = 3$ to $c_{new} = 4$, the compliance rate would inrease from 0.57 to 0.65, and here the MTTE equal to 0.23 implies that the new compliers at $c_{new} = 4$ would be less negatively affected. In particular, they would drop out of school a quarter of a year later compared with those compliers at the orginal threshold $c = 3$.

7   Conclusions

We have shown nonparametric identification of the treatment effect derivative (TED), which we show equals the marginal threshold treatment effect (MTTE) given a local policy invariance assumption. We also provide simple nonparametric estimators for these objects in both sharp and fuzzy RD designs, and discuss the use of these estimates for policy analysis.

Without the local policy invariance assumption, the MTTE will generally not be identified. Local policy invariance is a weak version of a specific external validity assumption for applying program evaluation results when the environment changes. In any particular application, one may assess from the institutional setting whether this assumption is likely to hold.

We apply our results to a model of the impacts of junior high school grade retention on the number of additional years (up to five) that students stay in school. In this empirical application, there is a discontinuity in the retention rate of students who fail more than three subjects. We use our estimate of the MTTE to see how the effect of the retention treatment on average additional years of schooling

23

would change if, hypothetically, the threshold were lowered. We find that retaining students who fail more than two subjects instead of three would substantially worsen the treatment effect, making the compliers at this new threshold drop out of school an average of .23 years earlier than the compliers at the old threshold.

8 Appendix: Proofs

Define the functions $G_{ct}(X)$, $s(X, C)$, $r(X, C)$ and $d(X, C)$ by

$$G_{ct}(x) = E(Y(t) \mid X, C = c), \quad s(X, C) = E(Y(1) - Y(0) \mid X, C, T(0) < T(1))$$

$$r(X, C) = \partial s(X, C) / \partial C, \quad d(X, C) = \partial s(X, C) / \partial X$$

The same definitions for s, r, and d can be applied in both sharp and fuzzy designs, since sharp designs are just the special case in which $T(0) < T(1)$ always holds. By these definitions, $s(X, C)$ is the expected difference in treated vs untreated outcomes of a hypothetical individual who is a complier and so has $T(0) < T(1)$, has a running variable equal to X, and is in an environment in which all other compliers are treated if and only if their running variable exceeds C.

For the special case of sharp designs, the function s reduces to equaling the function S in footnote 1, making the functions $D(X)$ and $D'(X)$ equal $s(X, c)$ and $d(X, c)$ respectively, and $\tau(C) = s(C, C)$. More generally, the functions $D_f(X)$, $D'_f(X)$ and $\tau_f(C)$ defined in the section on fuzzy designs equal $s(X, c)$, $d(X, c)$, and $s(C, C)$, respectively.

In this notation the formal definition of local policy invariance is the following:

Definition: Local policy invariance holds if and only if $r(c, c) = 0$.

Local policy invariance therefore says that, when holding the first argument of $s(x, c)$ fixed at $x = c$, infinitesimal changes in the second element of $s(x, c)$ have no effect on the value that $s(x, c)$ takes on. Local policy invariance refers only to impacts of changing the second argument of s in the neighborhood of c. In contrast, local constant treatment effects would mean that small changes in the first argument of s had no effect on the value of s.

In the text we followed the standard RD convention in defining the observable function $g(x) = E(Y \mid X = x)$. But, even conditioning on a given level of x, outcomes could change if the threshold were changed, and to account for this we can define $g_c(x) = E(Y \mid X = x, C = c)$. The function $g_c(x)$ is observable only for the threshold level c that generated the data. Then $g_c(x)$ as a function of x has one sided limits $g_{c-}(x)$ and $g_{c+}(x)$ that by continuity are equal to each other except at $x = c$, and similarly has one sided derivatives $g'_{c-}(x)$ and $g'_{c+}(x)$.

In this notation

$$\tau(c) = g_{c+}(c) - g_{c-}(c). \tag{20}$$

based on

$$G_{c0}(x) = E(Y(0) \mid X = c, C = c) = \lim_{x \uparrow c} E(Y(0) \mid X = x, C = c)$$

$$= \lim_{x \uparrow c} E(Y(0) \mid X = x, T = 0, C = c) = \lim_{x \uparrow c} E(Y \mid X = c, C = c) \tag{21}$$

and similarly for $G_{c1}(x) = \lim_{x \downarrow c} E(Y \mid X = c, C = c)$. As this example shows, the reference to $E(Y(t) \mid X = x)$ in Assumption A2 refers more precisely to the function $G_{ct}(x)$.

We can now restate Theorem 1 as saying that if Assumptions A1 an A2 hold then $s(c, c)$ and $d(c, c)$ are identified by $s(c, c) = g_{c+}(c) - g_{c-}(c)$ and

$$d(c, c) = g'_{c+}(c) - g'_{c-}(c). \tag{22}$$

PROOF of Theorem 1: Equation (20) or equivalently $s(c, c) = g_{c+}(c) - g_{c-}(c)$, is the standard sharp design identification result which was derived as above. To obtain equation (22), by the definition of a sharp design we first have $s(x, c) = G_{c1}(x) - G_{c0}(x)$. By assumption, $G_{ct}(x)$ is differentiable at $x = c$, so $d(c, c) = G'_{c1}(c) - G'_{c0}(c)$. By equation (21), $G_{c0}(c) = g_-(c)$, and differentiability of $G_{c0}(c)$ means that $G'_{c0}(c)$ equals its own one sided derivative $G'_{c0-}(c)$, giving

$$
\begin{aligned}
G'_{c0}(c) &= G'_{c0-}(c) = \lim_{\varepsilon \uparrow 0} \frac{G_{c0}(c+\varepsilon) - G_{c0-}(c)}{\varepsilon} = \lim_{\varepsilon \uparrow 0} \frac{E(Y(0) \mid X = c+\varepsilon, C = c) - g_{c-}(c)}{\varepsilon} \\
&= \lim_{\varepsilon \uparrow 0} \frac{E(Y(0) \mid X = c+\varepsilon, T = 0, C = c) - g_-(c)}{\varepsilon} = \lim_{\varepsilon \uparrow 0} \frac{E(Y \mid X = c+\varepsilon, C = c) - g_{c-}(c)}{\varepsilon} \\
&= \lim_{\varepsilon \uparrow 0} \frac{g_c(c+\varepsilon) - g_{c-}(c)}{\varepsilon} = g'_{c-}(c)
\end{aligned}
$$

An analogous derivation sending $\varepsilon \downarrow 0$ gives $G'_{c1}(c) = g'_{c+}(c)$, so equation (22) holds.

PROOF of Corollary 1: By definition in the sharp design $\tau'(c) = \partial s(c, c)/\partial c = d(c, c) + r(c, c)$, and hence equals $d(c, c)$ if and only if the local policy invariance condition $r(c, c) = 0$ holds.

Analogous to the local policy invariance definition that $\partial s(X, C)/\partial C$ equals zero when $X = C = c$, define local policy invariance for second order derivatives to be that both $\partial s(X, C)/\partial C$ and $\partial^2 s(X, C)/\partial C^2$ equal zero when $X = C = c$. It then follows immediately that if this condition holds then $\tau''(c) = D''(c)$ in addition to Theorem 1 and Corollary 1, which can then be used to apply equation (6).

PROOF of Corollary 2: Define strong policy invariance to be the assumption that $s(c, C) = s(c, c)$ for all C in a neighborhood of c. The assumptions of Corollary 2 imply that $s(X, C)$ is analytic, and by strong policy invariance, $\partial^k s(c, C)/\partial C^k$ equals zero when $C = c$ for all integers k. This then makes $\tau^k(c) = D^k(c)$ and equation (7) then holds by a Taylor expansion of $\tau(c_{new})$, replacing each $\tau^k(c)$ term with $D^k(c)$.

PROOF of Theorem 2: Define $\pi_c(x) = \Pr(T(0) < T(1) \mid X = x, C = c)$. Beginning with the definition of $G_{ct}(x)$,

$$
\begin{aligned}
G_{c1}(x) - G_{c0}(x) &= E(Y(1) - Y(0) \mid X = x, C = c) \\
&= E(Y(1) - Y(0) \mid X = x, C = c, T(0) < T(1))\Pr(T(0) < T(1) \mid X = x, C = c) \\
&\quad + E(Y(1) - Y(0) \mid X = x, C = c, T(0) = T(1))\Pr(T(0) = T(1) \mid X = x, C = c) \\
&\quad + E(Y(1) - Y(0) \mid X = x, C = c, T(0) > T(1))\Pr(T(0) > T(1) \mid X = x, C = c) \\
&= E(Y(1) - Y(0) \mid X = x, C = c, T(0) < T(1))\Pr(T(0) < T(1) \mid X = x, C = c) \\
&= s(x, c)\pi_c(x)
\end{aligned}
$$

where the penultimate equality above holds because $T(0) = T(1)$ makes $Y(1) = Y(0)$ and by the no defiers assumption $T(1) \geq T(0)$. Therefore $s(x, c) = [G_{c1}(x) - G_{c0}(x)]/\pi_c(x)$. Take the

derivative of this expression with respect to x, and evaluate the result at $x = c$ to get

$$\widetilde{d}(c) = d(c, c) = \frac{G'_{c1}(c) - G'_{c0}(c)}{\pi_c(c)} - s(c, c)\frac{\pi'_c(c)}{\pi_c(c)} \tag{23}$$

By Theorem 1, $s(c, c) = g_+(c) - g_-(c)$ and $G'_{c1}(c) - G'_{c0}(c) = g'_+(c) - g'_-(c)$. Next note that $\pi_c(x) = E(T(1) - T(0) \mid X = x, C = c)$ because the probability that $T(0) < T(1)$ is the probability that $T(1) = 1$ and $T(0) = 0$, which equals the expected value of $T(1) - T(0)$, again exploiting the no defiers assumption. Applying Theorem 1 replacing Y with T gives $\pi_c(c) = p(c) = f_+(c) - f_-(c)$ and $\pi'_c(c) = p'(c) = f'_+(c) - f'_-(c)$. Substituting these expressions in equation (23) proves equation (10).

PROOF of Corollary 3: By definition in the fuzzy design $\tau'_f(c) = \partial s(c, c)/\partial c = d(c, c) + r(c, c)$, and hence equals $d(c, c)$ if and only if local policy invariance $r(c, c) = 0$ holds.

References

[1] Abbring, J. H. and Heckman, J. J., (2007) "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in J.J. Heckman & E.E. Leamer (ed.), Handbook of Econometrics, volume 6, chapter 72, Elsevier.

[2] Abdulkadiroglu, A., J. D. Angrist, and P. A. Pathik (2011), "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," NBER working paper 17264.

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

[4] Battistin, E., A. Brugiavini, E. Rettore and G. Weber (2009) "The Retirement Consumption Puzzle: Evidence from a Regression Discontinuity Approach," American Economic Review, 99, 2209-2226.

[5] Card, D., C. Dobkin, and N. Maestas (2008) "The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare" American Economic Review, 98, 2242–2258.

[6] Card, D., D. S. Lee, Z. Pei, and A. Weber (2012), "Nonlinear Policy Rules and the Identification and Estimation of Causal Effects in Generalized Regression Kink Design," Working Paper 1206, Princeton University, Department of Economics, Industrial Relations Section.

[7] Carneiro, P., J. J. Heckman, and E. Vytlacil, (2010), "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," Econometrica, 78, 377–394.

[8] Chay, K. Y. and M. Greenstone (2005), "Does air quality matter? Evidence from the housing market," Journal of Political Economy, 113, 376–424.

[9] Deaton A., (2009), "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development," NBER Working Paper No. 14690.

[10] Dinardo, J. and D. S. Lee (2011), "Program Evaluation and Research Designs," in Handbook of Labor Economics, Ashenfelter and Card, eds., vol. 4a, Chap. 5, 463-536.

[11] Dong, Y. (2012), "Jumpy or Kinky? Regression Discontinuity Without the Discontinuity," Unpublished Manuscript.

[12] Fan, J. and Gijbels, I. (1996), Local Polynomial Modelling and its Applications. London: Chapman and Hall.

[13] Hahn, J., P. E. Todd, and W. van der Klaauw (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," Econometrica, 69, 201–09.

[14] Heckman, J. J. (2010), "Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy," Journal of Economic Literature, 48, 356-398

[15] Heckman, J., J. and S. Urzua, (2010), "Comparing IV With Structural Models: What Simple IV Can and Cannot Identify," Journal of Econometrics, 156, 27-37

[16] Imbens, G. W. (2010), "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)" Journal of Economic Literature, 48, 399–423.

[17] Imbens, G. W. and K. Kalyanaraman (2009), "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," NBER working paper number 14726.

[18] Imbens, G. W. and T. Lemieux (2008), "Regression Discontinuity Designs: A Guide to Practice," Journal of Econometrics, 142, 615–35.

[19] Imbens, G. W. and J. M. Wooldridge (2009), "Recent Developments in the Econometrics of Program Evaluation," Journal of Economic Literature 47, 5–86.

[20] Ludwig J., and D. L. Miller (2007), "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," The Quarterly Journal of Economics, 122, 159-208

[21] Jacob, B. A., and L. Lefgren, (2004) "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," Review of Economics and Statistics, 86, 226–244.

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

[23] Manacorda, M. (2012), "The Cost of Grade Retention," The Review of Economics and Statistics, 94, 596-606.

[24] Porter, J. R. (2003) "Estimation in the Regression Discontinuity Model," Unpublished Manuscript.

[25] Rubin, D. B. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology, 66, 688–701.