Jump or Kink? Identification of Binary Treatment Regression Discontinuity Design without the Discontinuity

Yingying Dong*

Department of Economics

University of California Irvine

First draft July 2011, Revised October 2014

Abstract

Standard Regression Discontinuity (RD) designs exploit a discontinuity (a jump) in the treatment probability to identify a local average treatment effect (LATE). RD identification fails or is weak when there is no jump or the jump is small. Unlike regression kink design (RKD), which requires a continuous treatment, this paper considers a binary treatment. This paper shows that without a jump, one can still identify a treatment effect utilizing a slope change (a kink) in the treatment probability. This paper provides weak and easily testable behavioral assumptions for identification based on a kink to be valid. While the standard RD model identifies a LATE for compliers, the kink identifies a limit form of the RD LATE, which can be viewed as a marginal treatment effect (MTE) or an average effect for marginal compliers. This paper further discusses a general model that utilizes either a jump, a kink or both for identification, and shows that a local two stage least squares (2SLS) estimator can be used regardless whether the treatment probability has a jump, a kink, or both. An empirical application is provided.

JEL Codes: C21, C25

Keywords: Regression discontinuity (RD), Fuzzy design, Local average treatment effect (LATE), Marginal treatment effect (MTE), Jump, Kink

---

*Correspondence: Department of Economics, 3151 Social Science Plaza, University of California Irvine, CA 92697-5100, USA. Email: yyd@uci.edu. http://yingyingdong.com/.

The author would like to thank Arthur Lewbel, Joshua Angrist, James Heckman, Han Hong and Damon Clark for very helpful comments, and thank Damon Clark and Emilia Del Bono for providing the data. Thanks also go to seminar participants at UCI, CSULB, CUHK, BC, and MIT, as well as conference participants at the SOLE, CEA, WEAI, and AMES annual conferences. Any errors are my own.

1 Introduction

Regression discontinuity (RD) designs for evaluating causal effects of interventions have been applied to many empirical areas of economics.[1] In a standard RD design, assignment of a binary treatment is determined by an observed covariate, the so-called running variable, exceeding a known threshold, and so there is a discrete change in the treatment probability at the threshold. Hahn, Todd, and van der Klaauw (2001) show that under weak conditions this discontinuity in the treatment probability, along with any observed corresponding discontinuity in the mean outcome, identifies a local average treatment effect (LATE). Intuitively, without any other discrete changes, any observed change in the mean outcome can be attributed to the change in the treatment probability.

Identification of a causal effect in RD designs relies on a discrete change or a jump in the treatment probability. Identification fails or is weak when there is no discontinuity, or when the discontinuity is small. See discussions in Imbens and Lemieux (2008), chapter 6 of Angrist and Pischke (2008), Imbens and Wooldridge (2009), and Lee and Lemieux (2010).

In this paper I show that one can still identify a causal effect when there is no discontinuity, but instead there is a kink, or a slope change in the treatment probability. Note that this is *not* the regression kink design (RKD), which I discuss in the next section. One key difference is that in RKD, treatment is required to be continuous (see Assumptions 3 and 3a and footnote 3 of Card et al. 2012), whereas here treatment is binary. For example, the primary RKD estimand (Proposition 1 of Card et al. 2012) depends on the derivative of the treatment variable, which is infeasible when treatment is binary.

Existing empirical applications of fuzzy design RD models almost always acknowledge slope changes in the treatment probability by allowing the slopes to differ between one side of the cutoff and the other. Existing RD plots also tend to show at least some change in slope at the cutoff. However, slope changes in the treatment probability are rarely exploited or even reported, even when they are relatively large compared to jumps. Typical RD applications allow slopes to change at the cutoff only to maintain flexible functional forms

This paper provides formal conditions under which identification based on kinks is valid. In particular, I show that a kink in the treatment probability can identify a causal effect under certain smoothness conditions. The required smoothness is satisfied given a weak behavioral assumption: in-

[1] See Lee and Lemieux (JEL, 2010) for an account of the history of the RD design and a list of recent applications.

dividuals can not precisely manipulate the running variable, so that there is no sorting in the neighborhood of the RD threshold. Not surprisingly, identification based on a kink in the treatment probability requires more smoothness than is required by standard RD models (requiring continuous differentiability instead of just continuity). However, this further smoothness is readily testable, and is already commonly assumed by standard RD estimators. It would be difficult to construct sensible models of behavior that would satisfy standard smoothness while not also satisfying the additional smoothness required here. See Dong and Lewbel (2014) and Dong (2014) for additional implications and discussions of this smoothness as well as applications where this further smoothness is shown to hold.

A kink or discrete change in slope can occur whenever the probability of being treated depends on one's distance from the threshold. This will happen, e.g., when an administrator's discretion or incentive to assign treatment depends on the distance from the threshold or when the costs or benefits of treatment depend on one's distance to the cutoff.

One example is Jacob and Lefgren (2004). They examine the effects of remedial education programs, including attending summer school and repeating a grade, on later academic performance. Repeating a grade, or grade retention is incurred by failing summer school tests. They note that "the probability of retention does not drop sharply (discontinuously) at the exact point of the cutoff , ...it rapidly decreases over a narrow range of values just below the cutoff." This lack of a true jump in the retention probability is consistent with schools and parents taking into account other factors in deciding whether to retain a student. Students who score just below the failing threshold are very likely to be socially promoted, and the further below the threshold a student scores, the more likely she will be retained. Their Figure 7 (reproduced in Figure 1 here) clearly shows a dramatic slope change but no discontinuity in the retention probability at the cutoff (normalized to zero), which keeps them from estimating an RD model. This is in contrast to attending summer school, which (see their Figure 4) clearly has a sharp discontinuity, so the standard RD model applies.

Sometimes the design of a policy rule itself creates kinks instead of or in addition to jumps at cutoffs. For example, Clark and Del Bono (2013) estimate the the effects of attending an elite school on adult outcomes in the UK. They show that the probability of attending an elite school is largely a kinked function of an assignment score. Students were almost all assigned to an elite school when their scores were above a certain high threshold, and assigned to a non-elite school when their assignment scores were below a certain low threshold. Students who score in between the two thresholds were ranked in order of merit, and so their probability of being assigned to an elite school increases

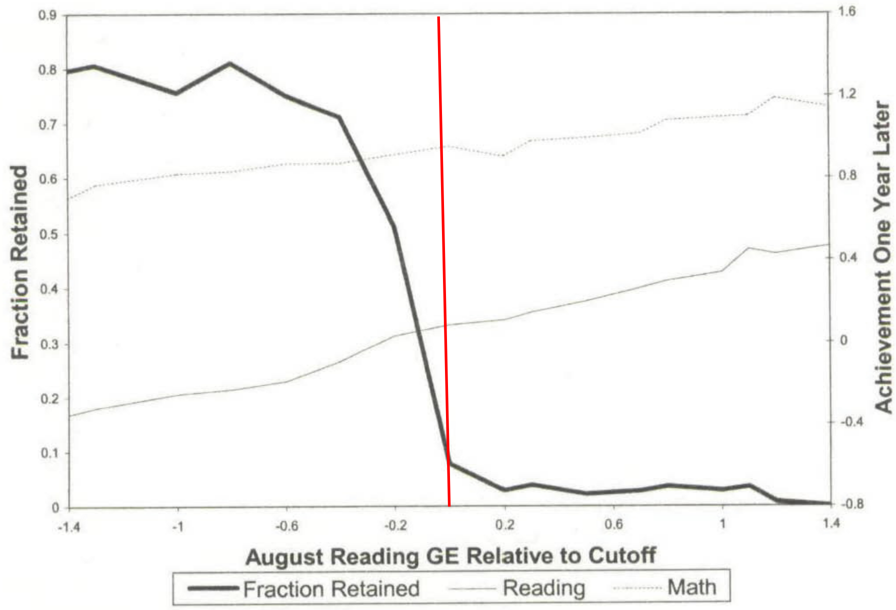

Figure 1: Retention rate and test score relative to the failing cutoff

with their assignment scores. Due to this particular policy rule, the probability of attending an elite school exhibits obvious kinks at the high and low thresholds.

While the standard RD model based on a jump identifies a LATE for compliers, I show that an RD model with a kink identifies a limit form of LATE, or a marginal treatment effect (MTE) as proposed by Heckman and Vytlacil (2005, 2007).

This paper further discusses a general model with either a jump, a kink or both. When both a jump and a kink exist, the kink estimand, the ratio of kinks, sometimes loses its causal interpretation. Easily interpretable conditions are provided under which the kink estimand and the jump estimand identify the same parameter.

A useful result is that a local two stage least squares (2SLS) estimator can be used regardless whether the observed change is a jump, a kink or both. It is asymptotically equivalent to the standard jump-based estimator if the true discontinuity in the treatment probability is nonzero; otherwise, the estimator reduces to the kink-based estimator. Identification based on changes in higher-order derivatives of the treatment probability is also briefly discussed. These results are applied to data in Clark and Del Bono (2013). I show that kinks in the treatment probability in this case provide strong identification when the true discontinuities may be small or insignificant.

For simplicity, this paper will mostly not deal with covariates other than the running variable in the analysis. The standard RD argument applies here, namely, that covariates are generally not needed for consistency in estimating the average (unconditional) treatment effect, though they may be useful for efficiency or for testing the validity of assumptions. However, if desired, additional covariates could be included in the analysis by letting all the assumptions hold conditional upon the values the covariates may take on. In applications, one could either partial out these covariates prior to analysis, or include them in the models as additional regressors.

The rest of the paper is organized as follows. Section 2 briefly reviews the related literature. Section 3 provides preliminary results. Section 4 discusses identification based on kinks. Section 5 discusses a general model with either a jump, a kink, or both. Section 6 relates the identification results to instrumental variables estimation and provides a local 2SLS estimator that can be applied regardless whether the treatment probability has a jump, a kink, or both, which is useful when the researcher is not certain whether a significant jump is present. Section 7 provides associated estimators. Section 8 presents an empirical application. Section 9 provides extensions to identification based on changes in higher order derivatives. Section 10 concludes.

2 Literature Review

This paper extends the standard FRD identification to allow for identification under more general conditions. The treatment being considered is binary. In contrast, the recent rising regression kink design (RKD) literature considers a continuous treatment that is a kinked function of the running variable. Below I give a brief review on the standard RD literature and the RKD literature, focusing mainly on theoretical developments. A list of recent empirical applications of RD models can be seen in Lee and Lemieux (2011).

The RD design is first introduced by Thistlethwaite and Campbell (1960) to estimate the effects of receiving a national merit award on educational aspiration and career plans among others. Eligibility for the award is based on a standard test score. Hahn, Todd and van der Klaauw (2001) provide formal identifying assumptions for RD models in the treatment model framework and propose local linear estimators. Porter (2003) proposes alternative partially linear and local polynomial estimators and discusses the optimal convergence rate. McCrary (2008) develops a formal density test to test the manipulation of the running variable. Lee (2008) discusses weak behavioral conditions to establish causal effects in SRD. Imbens and Kalyanaraman (2012) propose optimal bandwidth choices for RD models. Dong (2014) relax one of the key identifying assumption in Hahn, Todd and van der Klaauw

(2001) and propose a test on it. Calonico et al. (2014) propose alternative robust nonparametric bias-corrected inference in RD models.

In addition, Lee and Card (2008) consider the consequences of using a discrete running variable. They interpret deviations of the chosen approximating regression function from the true regression function as random specification errors and discuss the impact of this on inference. Dong (2014) discusses potential biases in using a rounded and hence discrete running variable, e.g., age in years, in empirical applications of RD models and proposes a simple bias correction and constructing bounds based on the moments of the rounding error. Papay et al. (2011) discuss estimation of RD models when treatment assignment is based on multiple variables. Frandsen, Frolich and Melly (2012) show identification and estimation of quantile treatment effects in RD models. Dong and Lewbel (2014) discuss nonparametric identification of the derivative of the RD LATE, its usage in extrapolation away from the RD threshold as well as identifying the change in the RD LATE resulting from a marginal change in the RD threshold. Gelman and Imbens (2014) discusses consequences of using higher-order polynomials in empirical applications of RD models and recommend against it.

The RKD is first introduced by Nielsen, Sorensen, and Taber (2010) to identify the causal effect of financial aid on college enrollment. The subsidy rate, or the subsidy amount as a function of family income, changes when family income crosses a threshold level. They therefore identify the effect of financial aid by the ratio of the kink in college enrollment to the kink in the subsidy amount as functions of family income. The fact that financial aid are a deterministic yet kinked function of family income helps identify the effect of financial aid separately from that of family income. Other empirical studies that use kinks to identify effects of continuous endogenous regressors include Guryan (2003), Dahlberg et al. (2008) and Simonsen et al. (2009).

Card, Lee, Pei, and Weber (2012) formally considers nonparametric identification of the average marginal effect of a continuous endogenous treatment variable in a generalized nonseparable model, where the treatment of interest is a known, deterministic, continuous but kinked function of an observed assignment variable, such as tax credit or subsidy rate as a function of family income. They then extend the sharp regression kink design to allow for omitted variables or measurement errors in the treatment function so that the exact kink is not deterministic or known, which they call fuzzy regression kink design. They show that the RKD estimand identifies an average effect of a marginal increase in the continuous treatment at the kink point, which can be viewed as the treatment effect on the treated (TT) or local average response (LAR) parameter that has been discussed in the nonseparable regression literature (Altonji and Matzkin, 2005, Florens et al. 2008).

Unlike RKD, this paper aims to identify the effect of a binary treatment as in standard RD design, but under more general conditions. The identification can be based either a jump, or a kink or both in the treatment probability. The purely kink-based estimand (Theorem 1) in this paper might appear to be similar to the RKD estimand. A key difference is that the RKD estimand depends on the derivative of the treatment variable when there is no measurement error, which would be infeasible when treatment is binary, while the estimand here depends on the derivative of the expected value of a binary treatment, i.e., the treatment probability.

3 Preliminary Results

Before discussing identification with a kink in the treatment probability, this section presents preliminary results that will facilitate such a discussion. In particular, I show that the standard RD identification result, the ratio of jumps identifies a LATE, as proved in Hahn, Todd and van der Klaauw (2001) can be re-produced under this paper's behavioral assumption. The result is then used to derive identification based on a kink or higher order derivative changes. Dong (2014) proves identification of both standard SRD and FRD under a similar but weaker behavioral assumption.

Let T be a binary indicator for some treatment such as participation in a social program or receiving a diploma. Let Y be some associated outcome of interest such as employment or wages, and let R be the so-called running or forcing variable that affects both T and Y. All the discussion in this section applies to R in a neighborhood of the RD cutoff, i.e., $R \in (r_0 - \varepsilon, r_0 + \varepsilon)$ for some small $\varepsilon > 0$. For example, R could be family income that affects eligibility for a social program, or an exam score affecting eligibility for a diploma.

Let Y_1 and Y_0 denote an individual's potential outcomes from being treated or not, respectively (Rubin, 1974). The observed outcome can then be written as $Y = Y_1 T + Y_0 (1 - T)$. Define $G(r) = E[Y \mid R = r]$ and $P(r) = E[T \mid R = r]$, so $G(r)$ and $P(r)$ are the expected outcome and expected probability of treatment when the running or forcing variable is $R = r$. In the standard RD model one would expect both $P(r)$ and $G(r)$ to have a jump (discontinuity) at the fixed threshold $r = r_0$.

Let $T(r)$ denote an individual's treatment when she has $R = r$. Further define an individual's potential status below the cutoff as $T_0(r) \equiv T(r)$ if the observed $r < r_0$, so the potential treatment below the cutoff is the same as the observed treatment status in this case, and $T_0(r) \equiv \lim_{\varepsilon \to 0} T(r_0 - \varepsilon)$

if the observed $r \geq r_0$ and the limit exists. Similarly, define an individual's potential treatment above the cutoff as $T_1(r) \equiv T(r)$ if the observed $r \geqslant r_0$ and $T_1(z) \equiv \lim_{\varepsilon \to 0} T(r_0 + \varepsilon)$ if the observed $r < r_0$ and the limit exists.[2] So $T_0(r)$ and $T_1(r)$ are well defined for any observed $r \in (r_0 - \varepsilon, r_0 + \varepsilon)$.

Given the above definitions, for an individual with $r < r_0$, her treatment status below the cutoff is observed and is given by $T_0(r) = T(r)$, and her counterfactual treatment just above the cutoff is $T_1(r) \equiv \lim_{\varepsilon \to 0} T(r_0 + \varepsilon)$. Similarly for an individual with $r \geq r_0$, her treatment status above the cutoff is observed and is given by $T_1(r) = T(r)$, and her counterfactual treatment just below the cutoff is $T_0(r) \equiv \lim_{\varepsilon \to 0} T(r_0 - \varepsilon)$.

Given any $r \in (r_0 - \varepsilon, r_0 + \varepsilon)$, we can then define the following four types of individuals as events in a common probability space $(\Omega, \mathcal{F}, P)$ (Angrist, Imbens and Rubin, 1996):

Always takers: $A = \{T_1(r) = T_0(r) = 1\}$

Never takers: $N = \{T_1(r) = T_0(r) = 0\}$

Compliers: $C = \{T_1(r) = 0, T_0(r) = 1\}$

Defiers: $D = \{ T_1(r) = 1, T_0(r) = 0\}$

For any individual, her potential treatment status and therefore type is allowed to depend on her value of the running variable. The standard RD models identify a LATE for compliers at $r = r_0$, i.e., individuals having $T_0(r_0) = 0$ and $T_1(r_0) = 1$. The identification requires $\Pr[C \mid R = r_0] \neq 0$, which would result in $P(r)$ having a discontinuity at r_0. For notational convenience, I will suppress the argument to simply use T_0 and T_1 but keep in mind they these are implicit functions of the running variable. Let Z be a dummy for crossing the threshold r_0, i.e., $Z = I\{R \geq r_0\}$, where $I\{\cdot\}$ is an indicator function equal to 1 if the expression in the bracket is true and 0 otherwise. The observed treatment can then be written as $T = T_0 + Z(T_1 - T_0)$, which is a function of the running variable.

Define the random vector $\mathbf{S} \equiv (Y_1, Y_0, T_1, T_0)$. An individual's observed outcome is fully determined by $\mathbf{S}$, given her draw of the running variable, so each individual can be seen as defined by $\mathbf{S}$. Let $f_{\mathbf{S}|R}(\mathbf{s} \mid r)$ denote the mixed joint density function of $\mathbf{S}$ conditional on $R = r$, if Y_0 and Y_1 are continuous, i.e., $f_{\mathbf{S}|R}(\mathbf{s} \mid r) = f_{Y_0,Y_1,R|T_0,T_1}(y_0, y_1, r \mid T_1 = t_0, T_0 = t_1) \Pr(T_1 = t_0, T_0 = t_1) / f_R(r)$ for $t_0 = 0, 1$ and $t_1 = 0, 1$, or it denotes the probability mass function if Y_0 and Y_1 are discrete.

ASSUMPTION A1 (No Defiers): There exists small $\varepsilon > 0$ such that $T_1(r) \geq T_0(r)$ for all $r \in (r_0 - \varepsilon, r_0 + \varepsilon)$.

[2] Note that defining these unobserved counterfactuals this way is without loss of generality, which just simplifies the derivation, since what matters are only those at the limit when z goes to z_0.

ASSUMPTION A2 (No Manipulation): $f_{\mathbf{S}|R}(\mathbf{s} \mid r)$ is continuously differentiable, and the density of R is continuously differentiable and strictly positive in the neighborhood of $r = r_0$ for all $\mathbf{s} \in supp(\mathbf{S})$.

Except for a first-stage, which requires a jump or kink in the treatment probability, A1 and A2 are this paper's main identifying assumptions. A1 is the standard LATE assumption assuming monotonicity or no defiers. Given A2, the conditional density of the running variable R for each individual defined by $\mathbf{S} = \mathbf{s}$, denoted as $f_{R|\mathbf{S}}(r \mid \mathbf{s}) = f_{\mathbf{S}|R}(\mathbf{s} \mid r) f_R(r)/f_{\mathbf{S}}(\mathbf{s})$ is continuously differentiable, or the derivative $\partial f_{R|\mathbf{S}}(r \mid \mathbf{s})/\partial r$ is continuous at $r = r_0$.

Dong (2014) utilizes a weaker version of no manipulation, in particular continuity instead of continuous differentiability of $f_{R|\mathbf{S}}(r \mid \mathbf{s})$ to identify standard sharp and fuzzy RD design. Continuity of the density $f_{R|\mathbf{S}}(r \mid \mathbf{s})$ means that individuals cannot sort to be just above the cutoff, so the probability of being above the threshold is in between 0 and 1. Continuity of the derivative of the density further requires that individuals do not have control to be even in an arbitrarily small neighborhood of the cutoff. Note that this stronger version of no-manipulation in A2 can be tested as discussed below. A2 also rules out other policies or programs using the same cutoff, which may result in a discrete change in the potential outcome distributions.

$\mathbf{S} \equiv (Y_0, Y_1, T_0, T_1)$ puts no restrictions on treatment effect heterogeneity and on selection into different types of individuals, so there could be arbitrary selection into treatment based on idiosyncratic gains and hence selection into types, say compliers. Therefore, this paper's assumptions allow for endogenous selection into compliers and self-selection into treatment based on idiosyncratic gains.

LEMMA 1: Given A1 and A2, $E[Y_t \mid R = r, \Psi]$ for $t = 0, 1$ and $\Pr[\Psi \mid R = r]$ for $\Psi \in \{A, N, C\}$ are continuously differentiable in the neighborhood of $r = r_0$.

Proofs are in the Appendix I. Lemma 1 says that given no manipulation as specified in A2, the conditional means of potential outcomes for each type of individuals $E[Y_t \mid R = r, \Psi]$ and the probabilities of selection into different types $\Pr[\Psi \mid R = r]$ are smooth in the sense of continuous differentiability. Smoothness of these conditional mean and probability functions are directly used for identification. The standard RD model requires only continuity for identification but not differentiability of these functions (See discussion in Hahn, Todd and van der Klaauw 2001 and Dong 2014). More smoothness is required here since this paper considers identification based on a discontinuity in slope. Assumption A2 is stronger than necessary compared with conditions in Lemma 1; however, A2 is more interpretable.

Assumption A2 and the resulting Lemma 1 suggest two types of validity tests for this paper's identifying assumptions: 1) testing smoothness, or no jumps or kinks, in the conditional means of predetermined covariates, and 2) testing smoothness of the density of the running variable at the RD cutoff . Standard RD validity tests test only continuity or no jumps (see e.g., Lee 2008 and McCrary 2008). The additional smoothness here requires ruling out significant kinks as well as any jumps, so for example, to test the smoothness of the density of the running variable, one can include not only Z but also $Z(R - r_0)$ in the local linear regression for the empirical density (fraction of observations in equally spaced bins) and test significance of both coefficients. To test smoothness of covariate means one can estimate the treatment effect on a covariate including both a a jump and a kink in the first stage. A false significant treatment effect implies that a significant jump and/or kink exist(s) in the covariate mean.

This paper's discussion utilizes one-sided limits and one-sided derivatives at $r = r_0$. For any function $H(r)$, define one-sided limits $H_+ = \lim_{\varepsilon \to 0} H(r_0 + \varepsilon)$ and $H_- = \lim_{\varepsilon \to 0} H(r_0 - \varepsilon)$ when they exist. Also define one-sided derivatives when exist $H'_+ = \lim_{\varepsilon \to 0} \frac{H(r_0+\varepsilon)-H_+}{\varepsilon}$ and $H'_- = \lim_{\varepsilon \to 0} \frac{H_- - H(r_0-\varepsilon)}{\varepsilon}$ for some small $\varepsilon > 0$.

DEFINITION 1: At the point $r = r_0$, a jump in the function $P(r)$ (or simply a jump) is defined as $P_+ - P_- \neq 0$.

Denote the treatment of interest as $\tau \equiv E[Y_1 - Y_0 \mid R = r_0, C]$. Lemma 2 below shows that a jump exists when $P_+ - P_- = \Pr[C \mid R = r_0]$ is not zero, i.e., a positive fraction of compliers take up treatment immediately when they cross the threshold at r_0.

LEMMA 2: If A1 and A2 hold, then $G_+ - G_- = \tau (P_+ - P_-)$ and $\Pr[C \mid R = r_0] = P_+ - P_-$. Further assume there is a jump, then

$$\tau = \frac{G_+ - G_-}{P_+ - P_-}. \tag{1}$$

Lemma 2 shows that given a positive fraction of compliers (a first stage) and no defiers, smoothness in A2 is sufficient to reproduce the standard identification result, i.e., the ratio of the mean outcome discontinuity to the treatment probability discontinuity identifies a LATE for compliers at the threshold r_0. In fact, as shown in the proof of Lemma 2, except for existence of compliers and no defiers, establishing the above result only requires the conditional means and probabilities listed in Lemma 1 to be continuous. A similar result is derived in Dong (2014).

4 Identification with a Kink

I now consider identifying the treatment effect at $r = r_0$ when there is no jump, but instead there is a kink or a slope change in the treatment probability. Formally define a kink as follows.

DEFINITION 2: At the point $r = r_0$, a kink in the function $P(r)$ (or simply a kink) is defined as $P'_+ - P'_- \neq 0$.

THEOREM 1: If A1 and A2 hold, then $\partial \Pr[C \mid R = r] / \partial r \mid_{r=r_0} = P'_+ - P'_-$. Assume there is no jump, but a kink at r_0. Then $\Pr[C \mid R = r_0] = 0$, $\partial \Pr[C \mid R = r] / \partial r \mid_{r=r_0} \neq 0$, and

$$\tau = \frac{G'_+ - G'_-}{P'_+ - P'_-}. \tag{2}$$

Theorem 1 shows that given A1 and A2, a kink exists if and only if $\partial \Pr[C \mid R = r] / \partial r \mid_{r=r_0} = P'_+ - P'_-$ is not zero, so there is a 'marginal' change in the probability of compliers. Further when there is a kink but no jump, the ratio of kinks can identify a causal effect of the treatment. Recall $\tau \equiv E[Y_1 - Y_0 \mid R = r_0, C]$. Here C is such that $\Pr[C \mid R = r_0] = 0$, and $\partial \Pr[C \mid R = r] / \partial r \mid_{r=r_0} \neq 0$, which means a zero compliance propensity at $r = r_0$ but non-zero at $r_0 + \varrho$ for an arbitrarily small ϱ. Both conditions together essentially imply that these are individuals at the margin of indifference between being treated or not at $r = r_0$. Formally define marginal compliers at $r = r_0$ as individuals with $\Pr[T_0 = 0, T_1 = 1 \mid R = r_0] = 0$ and $(\partial \Pr[T_0 = 0, T_1 = 1 \mid R = r] / \partial r \mid r = r_0) \neq 0$. Here marginal compliers are defined in terms of the derivative of the complying probability, consistent with marginal treatment effects expressed as derivatives. Note that the probability of compliers $\Pr[C \mid R = r]$ is still assumed to be smooth.

In the Appendix II, I set up an equivalent structure with a latent index model for treatment and show that in this case $\left(G'_+ - G'_-\right) / \left(P'_+ - P'_-\right) = E\left[Y_1 - Y_0 \mid R = r_0, U_T = P_- = P_+\right]$, where U_T is a normalized uniformly distributed latent error in the treatment model. It is sometimes viewed as a willingness-to-pay measure (Heckman and Vytlacil 2005, 2007, Carneiro et al. 2010). Since a generic definition for a marginal person at a probability margin (propensity score) p is an individual with exactly $U_T = p$, this result confirms that the identified effect is indeed an average effect for marginal persons at the kink point, or an MTE at the probability margin $p = P_- = P_+$.

Intuitively, the kink estimand in eq. (2) identifies a limit of the standard RD LATE when $P_+ \to P_-$. Heckman and Vytlacil (2005, 2007) note that MTE is equal to the limit of LATE when the treatment probability given the IV $Z = 1$ gets arbitrarily close to that given $Z = 0$. Here Z is the binary indicator for being above the threshold. Just as existence of compliers identifies the standard RD model, existence of marginal compliers provides identification here. Detailed discussion is provided in the Appendix.

With a jump, the compliers are the largest population for which treatment effects can be point identified. With no jump, but instead a kink, average effects for marginal compliers are what can be identified nonparametrically. The usefulness of such an identified effect depends on context. In the next section, I discuss under what conditions the treatment effect identified by a kink is the same as that identified by a jump.

5 Models with a Jump, a Kink or Both

A jump identifies a LATE. When there is no jump then a kink identifies an MTE, a limit form of the LATE. What happens when both a jump and an kink are present? Importantly, when both a jump and kink are present, the kink estimand sometimes loses its causal interpretation.

Below I provide two ways to address this issue. I first show how to identify a causal effect when the researcher does not know if there is a jump or not. I then list the condition under which a kink and a jump will both identify the same treatment effect parameter, and show how both can be exploited for estimation in that case. Note that even when a jump is present, and when the kink and jump do not identify the same parameter, there can still be an advantage (in root mean square error) to exploiting kink estimation, when equation (2) is precisely estimated by a large kink and when the bias $(P_+ - P_-)\,\tau'/\left(P'_+ - P'_-\right)$ is numerically small. This is discussed in more detail in the empirical application later.

The proof of Theorem 1 in the previous section shows that $G'_+ - G'_- = \tau'(P_+ - P_-) + \tau\left(P'_+ - P'_-\right)$, so in general the kink based estimator equals

$$\frac{G'_+ - G'_-}{P'_+ - P'_-} = \tau + \frac{P_+ - P_-}{P'_+ - P'_-}\tau', \tag{3}$$

where $\tau' = \partial E(Y_1 - Y_0 \mid R = r, C)/\partial r \mid_{r=r_0}$. It measures how the treatment effect varies with the running variable locally near the RD threshold r_0. Dong and Lewbel (2014) refer to τ' as the treatment effect derivative (TED), and show that TED is nonparametrically identified given the smoothness conditions listed in this paper. Detailed discussion of τ' and its applications can be found in Dong and Lewbel (2014). If both a jump and kink are present, both $P_+ - P_-$ and $P'_+ - P'_-$ are nonzero. Then the kink estimand may no longer equals the causal parameter τ.

THEOREM 2: Let A1 and A2 hold. Assume there is either a jump or a kink (or both) at r_0. Given any sequence of nonzero weights w_n such that $\lim_{n\to\infty} w_n = 0$,

$$\tau = \frac{G_+ - G_- + w_n\left(G'_+ - G'_-\right)}{P_+ - P_- + w_n\left(P'_+ - P'_-\right)}. \tag{4}$$

Theorem 2 uses a weight w_n to combine the standard RD estimand (1) and the new kink based estimand (2). Here w_n is assumed to go to zero as the sample size goes to infinity. In the next section I show that the weights in the local 2SLS estimator, a special case of the proposed estimator here, have this property. These results together than show that local 2SLS estimators utilizing both the jump and kink as IV's are valid estimators when one is not sure whether there is a jump, a kink or both.

When there is no jump, i.e., $P_+ - P_- = 0$ and $G_+ - G_- = 0$, then equation (4) will equal the kink based estimator (2). However, when there is a jump, then regardless of whether there is a kink or not, as $n \to \infty$ equation (4) becomes equal to the jump based estimator. This is because, as equation (3) shows, in general only the jump estimator is valid when both a jump and kink are present. So equation (4) (and its special case, the local 2SLS estimator discussed later) has the advantage that it can be used to estimate a valid causal effect if one is unsure whether a jump is present or not.

The following Corollary 1a shows the conditions under which the kink based estimand and the standard jump based estimand identify the same parameter, and so in this case both could be used for estimation.

COROLLARY 1a: Assume that A1 and A2 hold, and $\tau' = 0$. When both a jump and a kink exist,

$$\tau = \frac{G_+ - G_-}{P_+ - P_-} = \frac{G'_+ - G'_-}{P'_+ - P'_-}.$$

Corollary 1a follows immediately from equations (2) and (3).

$\tau' = 0$ means that the treatment effect does not vary linearly with the running variable R, as in the case where the treatment effect is locally constant.[3] So e.g., Y cannot be a nontrivial function of $T(R - r_0)$, but can be a function of T or $T(R - r_0)^2$. Intuitively, when the treatment effect is (locally) constant around the threshold, then the LATE equals its limiting value, the MTE, as have noted in, e.g., Heckman and Vytlacil (2005).

Note that $\tau' = 0$ is a strong restriction that is not required anywhere else in this paper; however, it is testable given A1 and A2. For example, one can estimate τ' as the coefficient of $T(R - r_0)$ in a local linear regression of Y on T, $(R - r_0)$ and $T(R - r_0)$ and then test its significance.[4]

Based on Corollary 1a, if $\tau' = 0$ then one could estimate τ by taking any weighted average of the jump and kink estimators. However, that would require that both a jump and a kink be present. One may instead use the following Corollary, which can be applied if either a jump, or a kink, or both exist. Unlike Theorem 2, Corollary 1b employs fixed weights w, and so exploits both the jump and the kink when both are present.

COROLLARY 1b: Assume that A1 and A2 hold, and $\tau' = 0$. If either a jump, or a kink, or both exist then

$$\tau = \frac{G_+ - G_- + w\left(G'_+ - G'_-\right)}{P_+ - P_- + w\left(P'_+ - P'_-\right)} \tag{5}$$

for any $w \neq -\left(P_+ - P_-\right) / \left(P'_+ - P'_-\right)$.

The weight w could be chosen to maximize efficiency, i.e., choosing the value of w that minimizes the estimated standard error of the corresponding estimate of τ. The following Section provides a local two stage least squares estimator (2SLS) that uses weights based on a measure of the relative strength of the two possible sources of identification, the jump and kink.[5]

[3] $\tau' = 0$ is a strictly weaker condition than assuming a locally constant treatment effect, because the latter would imply that all derivatives of τ were zero, not just the first derivative τ'.

[4] Dong and Lewbel (2014) shows that τ' is given by $\frac{G'_+ - G'_-}{P_+ - P_-} - \frac{(G_+ - G_-)(P'_+ - P'_-)}{(P_+ - P_-)^2}$ or $\frac{(G'_+ - G'_-) - \tau(P'_+ - P'_-)}{P_+ - P_-}$. All the terms involved can be nonparametrically identified given this paper's smoothness conditions.

[5] The result is not surprising. It is in fact a generic feature of 2SLS that when there exit more than one instrumental

6 Instrumental Variable Interpretation

This section provides an instrumental variables (IV) interpretation for the identification results of the previous section. The IV (particularly 2SLS) estimator discussed in this section has a varying bandwidth shrinking to zero as the sample size goes to infinity. It is therefore referred to as a local 2SLS estimator. More importantly I show that when there is either a jump, or a kink, or both, local 2SLS estimators automatically provide the type of weights satisfying the property required in Theorem 2.

Suppose that for observations in the neighborhood of r_0, one has the local linear outcome model

$$Y = \alpha_0 + \alpha_1(R - r_0) + \tau T + e, \tag{6}$$

where α_0, α_1, and τ are coefficients, and the error e may be correlated with the treatment indicator T. In general, e might also be correlated with R and hence Z for strictly positive ε.

Hahn, Todd, and van der Klaauw (2001) show that the standard fuzzy design RD estimator given by equation (1) is numerically equivalent to the IV estimator of τ in equation (6), using $(R - r_0)$ and Z as instruments for any given ε, even though the IV zero correlation assumption is violated for any $\varepsilon > 0$. Continuity of potential outcomes (essentially continuity of R and e in this case) at the threshold and having the bandwidth $\varepsilon \to 0$ as the sample size $n \to \infty$ establishes the consistency of the standard RD estimator.

Similarly, a kink in the treatment probability at the threshold, or the interaction term $(R - r_0)\,Z$, might also be an instrument for T. So when there is no jump, one can then still identify the effect of the treatment.

To include either Z, or $(R - r_0)\,Z$, or both as possible instruments for T, write the reduced-form treatment as

$$T = \beta_1 Z + \beta_2\,(R - r_0)\,Z + \beta_3\,(R - r_0) + \beta_4 + \upsilon, \tag{7}$$

where β's are the coefficients in this equation.

Substituting equation (7) into equation (6) yields the reduced form Y equation

$$Y = \gamma_1 Z + \gamma_2(R - r_0)Z + \gamma_3(R - r_0) + \gamma_4 + u, \tag{8}$$

variables, 2SLS uses efficient weights in combining these instrumental variables (see, e.g., Davidson and MacKinnon, 1993, and Chapter 4 of Angrist and Pischke, 2008).

where $\gamma_1 = \beta_1 \tau$, $\gamma_2 = \beta_2 \tau$, $\gamma_3 = \alpha_1 + \tau \beta_3$, $\gamma_4 = \alpha_0 + \tau \beta_4$ and $u = \tau v + e$.

Given equations (7) and (8), one has

$$P_+ - P_- = \beta_1, \quad P'_+ - P'_- = \beta_2, \tag{9}$$

$$G_+ - G_- = \gamma_1, \quad G'_+ - G'_- = \gamma_2. \tag{10}$$

Since the coefficients in local linear regressions equal conditional means and derivatives of conditional means regardless of their true functional forms (as long as they are sufficiently smooth), equations (9) and (10) would hold regardless of the true functional forms of Y and T.

Let Y^*, T^*, Z_1^*, and Z_2^* be Y, T, Z, and $(R - r_0)\, Z$ after partialling out $(R - r_0)$, respectively, i.e., they are the residuals from local linear regressions of Y, T, Z, and $(R - r_0)\, Z$ on a constant and $(R - r_0)$. Then the first and second stage regression equations can be rewritten as

$$\begin{aligned} T^* &= \beta_1 Z_1^* + \beta_2 Z_2^* + v, \\ Y^* &= \tau T^* + e, \end{aligned}$$

and the reduced form for Y as

$$Y^* = \gamma_1 Z_1^* + \gamma_2 Z_2^* + u.$$

The 2SLS estimator in this case is then

$$\begin{aligned} \tau &= \frac{cov\left(Y^*, \beta_1 Z_1^* + \beta_2 Z_2^*\right)}{cov\left(T^*, \beta_1 Z_1^* + \beta_2 Z_2^*\right)} = \frac{cov\left(\gamma_1 Z_1^* + \gamma_2 Z_2^*, \beta_1 Z_1^* + \beta_2 Z_2^*\right)}{cov\left(T^*, \beta_1 Z_1^* + \beta_2 Z_2^*\right)} \\ &= \frac{cov\left(\gamma_1 Z_1^* + \gamma_2 Z_2^*, T^*\right)}{cov\left(T^*, \beta_1 Z_1^* + \beta_2 Z_2^*\right)} \\ &= \frac{cov\left(T^*, Z_1^*\right)\gamma_1 + cov\left(T^*, Z_2^*\right)\gamma_2}{cov\left(T^*, Z_1^*\right)\beta_1 + cov\left(T^*, Z_2^*\right)\beta_2} \end{aligned}$$

which is the same as

$$\tau = \frac{w_1 \gamma_1 + w_2 \gamma_2}{w_1 \beta_1 + w_2 \beta_2}$$

where the weights are given by $w_1 = cov\left(T^*, Z_1^*\right)$ and $w_2 = cov\left(T^*, Z_2^*\right)$, so these weights reflect the relative strength of the two IVs, Z_1^* and Z_2^*. Plugging in γ_1, γ_2, β_1, and β_2, gives

$$\tau = \frac{w_1\gamma_1 + w_2\gamma_2}{w_1\beta_1 + w_2\beta_2} = \frac{w_1 (G_+ - G_-) + w_2 \left(G'_+ - G'_-\right)}{w_1 (P_+ - P_-) + w_2 \left(P'_+ - P'_-\right)}. \tag{11}$$

When there is no jump, but a kink, then $\beta_1 = \gamma_1 = 0$ and $w_1 = 0$, equation (11) reduces to equation (2), i.e., the IV estimand for τ equals the ratio of the coefficients for $Z(R - r_0)$ in the reduced-form Y and T equations, which confirms that the kink or slope change β_2 provides identification.

The 2SLS estimand in equation (11) is numerically equivalent to a special case of the estimand in Theorem 2. The weights here satisfy the property specified in Theorem 2. In particular, asymptotically this local 2SLS puts a zero weight on the kink as long as there is a jump. This is because this local 2SLS estimator has a variable bandwidth $\varepsilon \to 0$ as the sample size $n \to \infty$.

Asymptotically when the bandwidth $\varepsilon \to 0$, $R - r_0$ and hence the kink $Z(R - r_0)$ goes to zero, which makes Z_2^* go to zero. It follows that $w_2 = cov\left(T^*, Z_2^*\right)$, and hence w_2/w_1 goes to zero. Therefore with the local 2SLS if there is a jump, i.e., $\beta_1 = P_+ - P_- \neq 0$, the 2SLS weight $w_2/w_1 = w_n \to 0$ as $n \to \infty$, and hence

$$\frac{w_1\gamma_1 + w_2\gamma_2}{w_1\beta_1 + w_2\beta_2} = \frac{\gamma_2}{\beta_2} = \frac{G_+ - G_-}{P_+ - P_-}.$$

Alternatively, if there is no jump, $\beta_1 = P_+ - P_- = 0$ and hence $\gamma_1 = G_+ - G_- = 0$, then the weights are asymptotically irrelevant. In this case one has

$$\frac{w_1\gamma_1 + w_2\gamma_2}{w_1\beta_1 + w_2\beta_2} = \frac{w_2\gamma_2}{w_2\beta_2} = \frac{\gamma_2}{\beta_2} = \frac{G'_+ - G'_-}{P'_+ - P'_-},$$

which is the kink estimand in Theorem 2.

7 Estimation

In this section I briefly describe how to implement the proposed estimators. The estimation methods provided here are standard ones for RD analyses. All that is new here is taking the coefficients from these existing estimators, and using them to obtain estimates of the treatment effects defined by the Theorems in this paper.

One convenient way to implement the proposed estimators is to estimate local linear or polynomial regressions. The proposed estimators are simple functions of the coefficients in these local linear or polynomial regression. For example, based on a uniform kernel and observations near the threshold r_0, one could estimate $Y = \gamma_1 Z + \gamma_2 (R - r_0) Z + \gamma_3 (R - r_0) + \gamma_4 + \upsilon$ and $T = \beta_1 Z + \beta_2 (R - r_0) Z + \beta_3 (R - r_0) + \beta_4 + u$ by ordinary least squares (OLS). The bandwidth might be chosen using methods proposed by Imbens and Kalyanaraman (2012) or Ludwig and Miller (2007). See also discussion on bandwidth choices in Imbens and Lemieux (2008) and Lee and Lemieux (2010).

With these estimates the standard RD treatment effect estimator given a jump (Lemma 2) is

$$\widehat{\tau}(c) = \frac{\widehat{\gamma}_1}{\widehat{\beta}_1}. \tag{12}$$

This estimator can also be implemented as the estimated coefficient of T using IV estimation, regressing Y on a constant, $R - r_0$, and T, using $(R - r_0)$ and Z as instrumental variables.

The RD treatment effect estimator given a kink but no jump at the threshold r_0 (Theorem 1) can be estimated by

$$\widehat{\tau} = \frac{\widehat{\gamma}_2}{\widehat{\beta}_2}. \tag{13}$$

Equivalently, one could take $\widehat{\tau}$ to be the estimated coefficient of T in an IV estimation, regressing Y on a constant, $R - r_0$, and T, using $(R - r_0)$ and $(R - r_0)Z$ as instrumental variables.

The RD treatment effect estimator proposed in Corollary 1 can be implemented as

$$\widehat{\tau} = \frac{\widehat{\gamma}_1 + \widehat{w}\widehat{\gamma}_2}{\widehat{\beta}_1 + \widehat{w}\widehat{\beta}_2}. \tag{14}$$

where the weight $\widehat{w}$ can be chosen to minimize the bootstrapped standard error for $\widehat{\tau}$. Alternatively, equation (14) could be estimated by a 2SLS regression of Y on a constant, $R - r_0$, and T, using as instruments $(R - r_0)$, Z, and $(R - r_0)Z$. The resulting estimator corresponds to the one in Theorem 2, and the estimated weights will then be as described in Section 4.

For all the estimators in the above, one could use the Delta method to calculate standard errors. Alternatively, IV estimation provides robust standard errors along with the point estimate of the local average treatment effect.

8 Empirical Application

This section applies results in the previous sections to estimate the effect of elite school attendance on educational attainment in the UK. The data are from Clark and Del Bono (2013). Students in the sample were assigned to either an elite or a non-elite secondary school. The school assignment formula is largely a kinked function of the assignment score, with small or possibly nonexistant jumps. These jumps and kinks are explored for identification. Performance of the kink estimator is compared to that of the standard RD estimator based on a jump and the general estimator that uses both a jump and a kink.

One disadvantage of this application is that the sample size is rather small. However, the kinks in the probability of attending an elite school are large and highly significant, so even with a small sample, they still provide strong identifying power as shown below.

The treatment T here is a binary indicator for attending an elite secondary school. The outcome Y is years of post-secondary education. Secondary education is compulsory in the UK for the cohort (those born in the 1950's) considered here. The running variable, the assignment score R is a sum of four test scores and a teacher assessment component. Two of the test scores are from Verbal Reasoning Quotient or VRQ tests, one from an English attainment test and the other from an arithmetic attainment test. Each component is standardized to have mean 100 and standard deviation 15. Students with assignment scores below 540 were almost all allocated to a non-elite school. Students with assignment scores above 560 were allocated to an elite school unless 1) they were assessed by their head teacher as "unsuitable," or 2) one of their VRQ test scores was below 112 and their overall assignment score was below 580. Students with scores between 540 and 560 were assigned to an elite school partly based on their order of merit, and so the higher their scores, the more likely they were assigned to an elite school. Details of the construction of the assignment score and the school assignment procedure can be found in Clark and Del Bono (2013).

Figure 2 shows the probability of attending an elite school as a function of the assignment score for males. Consistent with the assignement rule, it obviously kinks at the two thresholds $R = 540$ and $R = 560$. Figure 3 shows years of post-secondary education. The pattern in Figure 3 largely mimics that in Figure 2 but in a smaller magnitude. Plots for females are very similars. They are omitted to save space.

I focus on the sample of males to estimate the impact of elite school attendance on educational

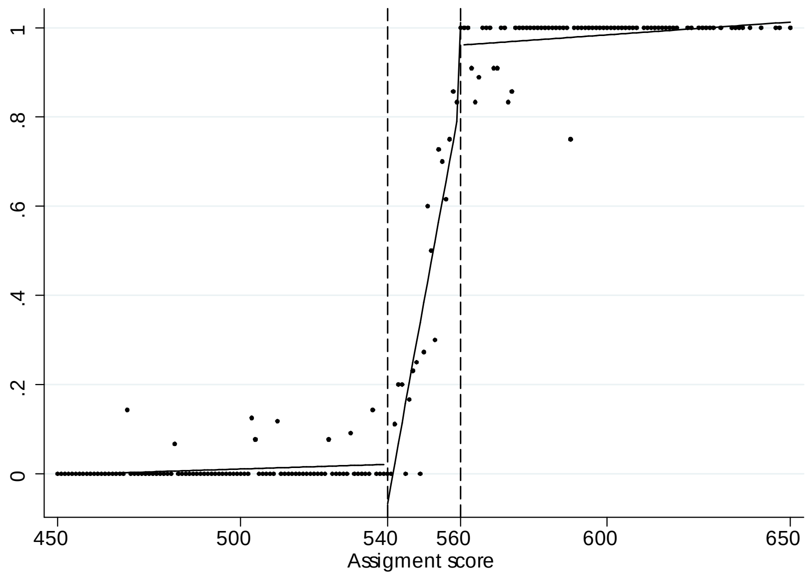

Figure 2: Assignment score and the probability of attending an elite school for males

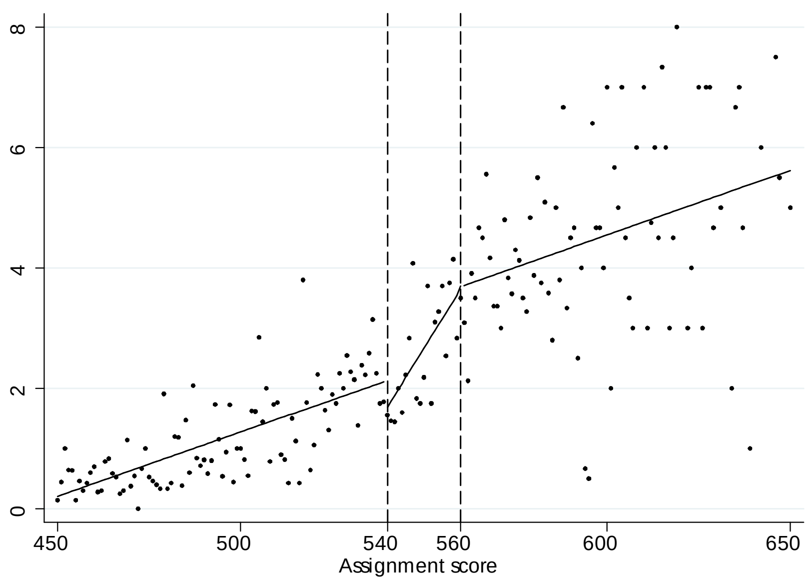

Figure 3: Assignment score and years of post-secondary education for males

outcome. Females in general show more heterogeneity in schooling choices, which leads to much less precise estimates, especially given the small sample size. I separately estimate a local effect at each threshold, using the “high" and “low" assignment score samples. The former consists of assignment scores between 540 and 580 and the latter between 520 and 560. The sample sizes are 361 and 409, respectively. I estimate local linear regressions using a uniform kernel. The small sample size and the narrow distance between the cutoffs greatly limit the range of feasible alternative estimators; however, estimates using different bandwidths are reported in the Appendix. The following conclusions largely hold with different bandwidth choices.

Table 1 presents the estimated jumps and kinks in the first-stage and the reduced-form outcome regressions. For each, the first column presents estimates without controlling for any additional covariates except for a linear term in the assignment score, while the second column presents estimates controlling for an extensive set of covariates, including fixed effects for the school and grade attended in 1962, father's occupation, and mother's socioeconomic status, age within grade and linear and quadratic terms of test scores at ages 7 and 9. Consistent with Figure 2, estimates in Table 1 show significant kinks at both thresholds, no significant jump at the upper threshold and a somewhat imprecisely estimated jump of 10-11% at the lower threshold. Estimates remain similar regardless of whether one controls for the extensive set of covariates or not.

Table 1 Elite school attendance and reduced-form post-compulsory education for males

Dependent var.	Elite school	Elite school	Education	Education
		$R \in [540, 580]$		
1(R>560)	0.074	0.089	-0.050	-0.045
	(0.057)	(0.065)	(0.478)	(0.555)
(R-560)1(R>560)	-0.044***	-0.043***	-0.077**	-0.066*
	(0.004)	(0.005)	(0.035)	(0.038)
Covariates	N	Y	N	Y
R-squared	0.549	0.622	0.086	0.283
		$R \in [520, 560]$		
1(R<540)	0.098**	0.111**	0.689*	0.968**
	(0.044)	(0.047)	(0.361)	(0.412)
(R-540)1(R<540)	-0.046***	-0.044***	-0.062**	-0.059*
	(0.004)	(0.004)	(0.029)	(0.032)
Covariates	N	Y	N	Y
R-squared	0.439	0.524	0.053	0.259

Note: Robust standard errors are in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 2 presents the estimated impact of attending an elite school on educational attainment. These estimates are generated by 2SLS using either a jump, or a kink or both as IV(s), corresponding to the standard jump estimator, the kink estimator and the general estimator in equations (12), (13) and (14), respectively. For the high assignment score sample, the jump is insignificant and hence identification based on a jump fails in this case. As a result, the standard jump estimator gives an economically implausible negative sign. In contrast, the kink estimator generates an estimated effect of about 1.5 years, so attending an elite school increase completed post-secondary education by 1.5 years for males. The estimated effect is comparable to the estimates reported in Clark and Del Bono (2013). They implement an IV estimator using the nonlinearity shape (the jumps and the kinks) of the treatment function as IVs based on the full range of data from 350 to 650. The combined estimator generates similar estimates to those of the kink estimator. This is consistent with Theorem 2 and Corollary 1b in that the 2SLS estimator using both a jump and a kink as IVs reduces to the kink estimator when there is no jump. Those estimates do not change much when excluding covariates.

For the low assignment score sample, the estimated jump is significant at the 5% level. Recall that when a jump exists, the kink estimator is biased unless TED or $\tau' = 0$. More precisely, based on equation (3), The bias in the kink estimator equals $\tau'(P_+ - P_-)/(P'_+ - P'_-)$. Here the jump appears to be much less precisely estimated than the kink, having a standard error that is an order of magnitude larger than the standard error of the kink estimate (see Table 2). So unless the bias in the kink estimator is very large, estimation based on the kink will be much more accurate than the jump estimator, at least in terms of root mean squared error.

Table 2 Impact of elite school attendance on post-compulsory education for males

	Jump Estimator		Kink Estimator		Jump&Kink Estimator	
			$R \in [540, 580]$			
	-0.678	-0.509	1.735**	1.546*	1.703**	1.503*
	(6.330)	(5.709)	(0.805)	(0.824)	(0.853)	(0.888)
Covariates	N	Y	N	Y	N	Y
			$R \in [520, 560]$			
	6.995	8.696*	1.335**	1.332**	1.536**	1.705**
	(4.563)	(4.516)	(0.604)	(0.646)	(0.647)	(0.731)
Covariates	N	Y	N	Y	N	Y

Note: Robust standard errors clustered at the assignment score level are in parentheses;
*** p<0.01, ** p<0.05, * p<0.1

In fact, the bias in the kink estimator is likely to be very small, both because $P'_+ - P'_-$ is relatively quite large (this is visible in Figure 2) and because $\tau'(P_+ - P_-)$ is likely to be quite small. $\tau'(P_+ - P_-) = 0$ can be formally tested. I estimate $\tau'(P_+ - P_-)$ using the equation $\tau'(P_+ - P_-) = G'_+ - G'_- - \tau\left(P'_+ - P'_-\right)$, where $G'_+ - G'_-$ and $P'_+ - P'_-$ are slope changes in the reduced-form outcome and the first-stage treatment probability equation, and τ is estimated by the general estimator. Note that using the general estimator to estimate τ is valid under the null $\tau' = 0$ or $(P_+ - P_-) = 0$. This is also preferred because the true jumps may be small or insignificant, which yields largely misleading estimates of τ, as shown by the estimates in first columns in Table 2. Estimates of $\tau'(P_+ - P_-)$ are reported in Table 3. For both the high and low assignment score samples, the estimates are both numerically very small and are statistically insignificant, regardless whether one controls for covariates or not.

These results strongly suggest that, by greatly decreasing variance while adding little bias, the kink estimator provides a much more accurate estimate of the treatment effect than the jump estimator at the lower threshold, despite the possible presence of a nonzero jump there. Consistent with this, the standard jump estimator yields implausibly large and imprecise positive estimate. In contrast, the kink estimator and the general estimator produce estimates close to the estimates from the high assignment score sample.

Table 3 Biases in the kink estimator

	$R \in [540, 580]$		$R \in [520, 560]$	
	0.001	0.001	-0.008	-0.008
	(0.010)	(0.009)	(0.010)	(0.009)
Covariates	N	Y	N	Y

Note: Boostrapped standard errors based on 1000 simulations are in parentheses; *** p<0.01, ** p<0.05, * p<0.1

To evaluate the plausibility of the this paper's identifying assumptions, I examine whether the density of the running variable has any unusual jumps or kinks at the cutoff. The estimates are based on a similar local linear regression as the first-stage regression, where the dependent variable is replaced by the empirical density of the running variable, i.e., the fraction of observations at each assignment score. The results are presented in Table 4. No significant jumps or kinks in the density appear at either threshold.[6]

[6]Another test is to evaluate whether the baseline covariates are smooth. This can be done by estimating the general model after replacing the outcome variable with a covariate, and then testing whether. Among the 46 covariates (43 binary and 3 continuous covariates) included in the outcome equations, only one out of eight included indicators for mother's SES categories shows a weakly significant coefficient, which is reasonable given the large set of covariates. So the overall evidence suggests that the smoothness conditions required for identification hold.

Table 4 Tests for the smoothness of the density of the running variable

$R \in 540, 580$		$R \in 520, 560$	
1(R>560)	0.003	1(R<540)	-0.007
	(0.004)		(0.004)
(R-560)1(R>560)	0.000	(R-540)1(R<540)	-0.000
	(0.000)		(0.000)

Note: obust standard errors clustered at the assignment score level are in parentheses

Lastly, Table A1 in the Appendix III provides additional estimates extending the low assignment score sample to be between 450 and 560 and the high assignment score sample to be between 540 and 650. Again one can see that the kink estimator and the general estimator using both the jump and the kink produce estimates reasonably close to those reported in Table 2, while the standard jump estimator yields estimates either with an improbable negative sign or in implausibly large magnitude.

9 Extensions

The estimand in Corollary 1 requires $\tau' = 0$ to use both a jump and a kink for identification. As mentioned, having $\tau' = 0$ means that the treatment effect does not vary linearly with the running variable R. For example, in the true parametric form, Y cannot be a function of $(R - r_0)T$.

The following Corollary 2 provides extensions of Corollary 1 to allow $\tau' \neq 0$ while still exploiting information in a kink in addition to a jump. For example, if the treatment is grade retention, the running variable is test score, and the outcome is later academic performance, then $\tau' \neq 0$ would mean that the effect of repeating a grade on later performance depends on the pre-treatment test score, and in this case one still could use both jump and kink information to estimate the treatment effect.

For convenience of notation, formally define $\gamma_1 \equiv G_+ - G_-$, $\gamma_2 \equiv G'_+ - G'_-$, $\gamma_3 \equiv G''_+ - G''_-$, $\beta_1 \equiv P_+ - P_-$, $\beta_2 \equiv P'_+ - P'_-$ and $\beta_3 \equiv P''_+ - P''_-$. So γ_1, γ_2, and γ_3 (β_1, β_2, and β_3) as the intercept, slope, and second-order derivative changes, respectively, in the conditional means of outcome (treatment) at r_0 when the derivatives exist.

COROLLARY 2: Assume A1 and A2 hold. Further assume that $f_{\mathbf{S}|R}(\mathbf{s} \mid r)$ in A1 is continuously differentiable in a neighborhood of $r = r_0$. When both a jump and a kink exist, assume $\tau'' = 0$, then

$$\tau = \frac{\gamma_1 + \omega\left(2\beta_2\gamma_2 - \gamma_3\beta_1\right)}{\beta_1 + \omega\left(2\beta_2^2 - \beta_3\beta_1\right)} \tag{15}$$

for any weights $w \neq -\beta_1 / \left(2\beta_2^2 - \beta_3\beta_1\right)$.

If $f_{\mathbf{S}|R}(\mathbf{s} \mid r)$ in A1 is continuously differentiable, then the conditional means of potential outcomes for each type of individuals $E(Y_t \mid R = r, \Psi)$ for $t = 0, 1$ and the probabilities of different types $\Pr(\Psi \mid R = r)$ for $\Psi \in \{A, N, C\}$ are continuously twice differentiable. It follows that all the involved limits and derivatives in γ_1, γ_2, γ_3, β_1, β_2 and β_3 exist.

Analogous to Corollary 1, the assumption for Corollary 2 that $\tau'' = 0$ will hold if the treatment effect is locally linear or locally constant. However, while a locally linear or constant treatment effect is sufficient for $\tau'' = 0$, it is stronger than necessary, because it implies that all derivatives higher than the first are zero, instead of just the second derivative being zero. With the assumption $\tau'' = 0$, the corresponding estimator does not allow the treatment effect to vary with $(R - r_0)^2$, because in this case the estimator using the second derivative changes $\left(2\beta_2\gamma_2 - \gamma_3\beta_1\right) / \left(2\beta_2^2 - \beta_3\beta_1\right)$ would not be a valid estimator for the local treatment effect at r_0. So for example, Y cannot be a function of $T(R - r_0)^2$, but can be a function of T or $T(R - r_0)$ or both.

When the treatment effect is locally linear instead of locally constant, the above estimand uses this information, while the estimand in Theorem 2 does not. In particular, when the treatment effect τ is either locally constant or locally linear, given a kink ($\beta_2 \neq 0$), $\tau = \left(2\beta_2\gamma_2 - \gamma_3\beta_1\right) / \left(2\beta_2^2 - \beta_3\beta_1\right)$ regardless whether there is a jump or not, while $\tau \neq \gamma_2/\beta_2$ unless the treatment effect is locally constant or there is no jump ($\beta_1 = 0$).

Similar estimands can be constructed if the d-th derivative $\tau^{(d)} = 0$, as is the case if the treatment effect is up to a polynomial of degree $d - 1$ in $(R - r_0)$, for any positive integer d. Construction of the estimand when $\tau^{(d)} = 0$ for any finite d is briefly discussed in the Appendix. In this case, the treatment effect can be an arbitrarily high-order (e.g., up to the $(d - 1)$-th order) polynomial of $(R - r_0)$, as long as the order is finite.

So far, all the estimators have been discussed without considering other covariates except for R. The usual argument applies that if one is interested in estimating the average treatment effect, or the unconditional treatment effect, covariates are not necessary for consistency, but may be useful to increase efficiency or for robustness check. If desired, one can directly include covariates in the treatment and outcome equations, or partial covariates out by first regressing Y and T on covariates both above and below the threshold, and then use the residuals from those regressions in place of Y and T in estimation.

10 Conclusions

The standard RD design identifies a causal effect of treatment by associating a discrete change in the mean outcome with a corresponding discrete change (a jump) in the treatment probability. Identification fails or is weak if there is no discontinuity or the discontinuity is small. This paper shows that it is possible to identify a causal effect under more general conditions, i.e., from a kink (a discrete change in slope) rather than, or in addition to, a jump (a discrete change in level) in the treatment probability.

When there is no jump but a kink in the treatment probability, the ratio of kinks in the mean outcome and the treatment probability identifies a causal effect of the treatment. The identified effect is a limit form of the RD local average treatment effect, which can be viewed as a marginal treatment effect.

Identification based on kinks is valid given smoothness, or continuous differentiability, of the conditional means of potential outcomes for each type of individuals (always takers, never takers and compliers) and smoothness of the probabilities of individual types. The required smoothness is satisfied under a weak behavioral assumption, i.e., individuals cannot sort in the neighborhood of the threshold, which results in the smoothness of the conditional density of the running variable. The identifying assumptions have easily testable implications – one can test smoothness (no jumps and kinks) of covariate means and the density of the running variable.

This paper also discusses a general model with either a jump, a kink or both. When both a jump and a kink exist, I show that the kink estimand may not identify a causal effect and that the jump estimand and the jump estimand identify the same parameter if and only if the treatment effect derivative is zero. This condition is testable given this paper's identifying assumption. When the derivative is nonzero but small, it can be beneficial in a mean squared error sense to consider kink based estimation, as in the case of this paper's empirical application.

A useful result is that the local two stage least squares (2SLS) estimator that uses both a jump and a kink as IVs is valid when there is either a jump or a kink or both. This 2SLS asymptotically reduces to the jump estimator if there is a jump, and otherwise reduces to the kink estimator if there is no jump. An empirical application shows usefulness of these results. Extensions to identification based on discrete changes in higher order derivatives are also briefly discussed.

As with standard RD, this paper discusses kink based identification in a static model. When the running variable is time, kinks in outcomes may result from dynamic effects, such as variable delays in response to treatment. Extending the results in this paper to a dynamic context would be an interesting direction for future research.

11 Appendix I: Proofs

PROOF of LEMMA 1:

The following applies to $R \in (r_0 - \varepsilon, r_0 + \varepsilon)$ for some small $\varepsilon > 0$ and assumes that Y_0 and Y_1 are continuous. For discrete Y_0 and Y_1, replace with summation the integration over the conditional distribution of Y_0 and Y_1.

By Assumption A2, $f_{\mathbf{S}|R}(\mathbf{s} \mid r)$ is continuously differentiable in r at $r = r_0$, where $\mathbf{S} \equiv (Y_0, Y_1, T_0, T_1)$, then probability of each type of individual $Pr(T_0 = t_0, T_1 = t_1 \mid R = r) = \int_{\Omega_1} \int_{\Omega_0} f_{\mathbf{S}|R}(\mathbf{s} \mid r) dy_0 dy_1$ for $t_0 = 0, 1$ and $t_1 = 0, 1$ is continuously differentiable in r at $r = r_0$, where $\mathbf{\Omega}_t$ for $t = 0, 1$ denotes the conditional support of Y_t conditional on $R = r$. Assumption A1 rules out defiers, so $\Pr(\Psi \mid R = r)$ for $\Psi \in \{A, N, C\}$ are continuously differentiable in a neighborhood of $r = r_0$

Again by Bayes' Rule, $f_{Y_0,Y_1|T_0,T_1,R}(y_0, y_1 \mid t_0, t_1, r) = f_{\mathbf{S}|R}(\mathbf{s} \mid r)/Pr(T_0 = t_0, T_1 = t_1 \mid R = r)$. Both $f_{\mathbf{S}|R}(\mathbf{s} \mid r)$ and $Pr(T_0 = t_0, T_1 = t_1 \mid R = r)$ are continuously differentiable in r at $r = r_0$, so $f_{Y_0,Y_1|T_0,T_1,R}(y_0, y_1 \mid t_0, t_1, r)$ for $t_0 = 0, 1$ and $t_1 = 0, 1$ is continuously differentiable in r at $r = r_0$. It follows that type-specific conditional means of potential outcome $E(Y_t \mid T_0 = t_0, T_1 = t_1, R = r)$ for $t = 0, 1$, $t_0 = 0, 1$ and $t_1 = 0, 1$ or $E(Y_t \mid R = r, \Psi)$ for $t = 0, 1$ and $\Psi \in \{A, N, C\}$ are continuously differentiable in r in the neighborhood of $r = r_0$.

PROOF of LEMMA 2:

Given Assumption A1 as well as definitions of individual types, for observations above the cutoff, we have

$$
\begin{aligned}
&G(r_0+\varepsilon)\equiv E\left[Y \mid R=r_0+\varepsilon\right]=E\left[Y_0+(Y_1-Y_0)T \mid R=r_0+\varepsilon\right]\\
&=E\left[Y_0 | \; R=r_0+\varepsilon, T=0\right]\Pr\left[T=0 \mid R=r_0+\varepsilon\right]\\
&+E\left[Y_1 | \; R=r_0+\varepsilon, T=1\right]\Pr\left[T=1 \mid R=r_0+\varepsilon\right]\\
&=E\left[Y_0 | \; R=r_0+\varepsilon, T_1=0\right]\Pr\left[T_1=0 \mid R=r_0+\varepsilon\right]\\
&+E\left[Y_1 | \; R=r_0+\varepsilon, T_1=1\right]\Pr\left[T_1=1 \mid R=r_0+\varepsilon\right]\\
&=E\left[Y_0 | \; R=r_0+\varepsilon, T_0=0, T_1=0\right]\Pr\left[T_0=0, T_1=0 \mid R=r_0+\varepsilon\right]\\
&+E\left[Y_1 | \; R=r_0+\varepsilon, T_0=0, T_1=1\right]\Pr\left[T_0=0, T_1=1 \mid R=r_0+\varepsilon\right]\\
&+E\left[Y_1 | \; R=r_0+\varepsilon, T_0=1, T_1=1\right]\Pr\left[T_0=1, T_1=1 \mid R=r_0+\varepsilon\right]\\
&=E\left[Y_0 | \; R=r_0+\varepsilon, N\right]\Pr\left[N \mid R=r_0+\varepsilon\right]\\
&+E\left[Y_1 | \; R=r_0+\varepsilon, C\right]\Pr\left[C \mid R=r_0+\varepsilon\right]\\
&+E\left[Y_1 | \; R=r_0+\varepsilon, A\right]\Pr\left[A \mid R=r_0+\varepsilon\right]
\end{aligned}
$$

Similarly for observations below the cutoff we have

$$
\begin{aligned}
&G(r_0-\varepsilon)=E\left[Y_0 | \; R=r_0+\varepsilon, T_0=0, T_1=0\right]\Pr\left[T_0=0, T_1=0 \mid R=r_0+\varepsilon\right]\\
&+E\left[Y_0 | \; R=r_0+\varepsilon, T_0=0, T_1=1\right]\Pr\left[T_0=0, T_1=1 \mid R=r_0+\varepsilon\right]\\
&+E\left[Y_1 | \; R=r_0+\varepsilon, T_0=1, T_1=1\right]\Pr\left[T_0=1, T_1=1 \mid R=r_0+\varepsilon\right],\\
&=E\left[Y_0 | \; R=r_0+\varepsilon, N\right]\Pr\left[N \mid R=r_0+\varepsilon\right]\\
&+E\left[Y_0 | \; R=r_0+\varepsilon, C\right]\Pr\left[C \mid R=r_0+\varepsilon\right]\\
&+E\left[Y_1 | \; R=r_0+\varepsilon, A\right]\Pr\left[A \mid R=r_0+\varepsilon\right],
\end{aligned}
$$

By Lemma 1, continuity of probabilities of types and type-specific conditional means of potential outcomes, we have

$$
\begin{aligned}
&G_+=\lim_{\varepsilon\to 0} G(r_0+\varepsilon) \qquad (16)\\
&=E\left[Y_0 | \; R=r_0, N\right]\Pr\left[N \mid R=r_0\right]+E\left[Y_1 | \; R=r_0, C\right]\Pr\left[C \mid R=r_0\right]\\
&+E\left[Y_1 | \; R=r_0, A\right]\Pr\left[A \mid R=r_0\right]
\end{aligned}
$$

and

$$G_- = \lim_{\varepsilon \to 0} G(r_0 - \varepsilon) \tag{17}$$
$$= E[Y_0 | \ R = r_0, N] \Pr[N \mid R = r_0] + E[Y_0 | \ R = r_0, C] \Pr[C \mid R = r_0]$$
$$+ E[Y_1 | \ R = r_0, A] \Pr[A \mid R = r_0].$$

It follows that

$$G_+ - G_- = E[Y_1 - Y_0 | \ R = r_0, C] \Pr[C \mid R = r_0] = \tau \Pr[C \mid R = r_0],$$

The treatment probabilities above or below the threshold are given by

$$\begin{aligned} P(r_0+\varepsilon) &= E\left[T \mid R = r_0+\varepsilon\right] \\ &= \Pr\left[T_1 = 1, T_0 = 0 \mid R = r_0 + \varepsilon\right] + \Pr\left[T_1 = 1, T_0 = 1 \mid R = r_0 + \varepsilon\right] \\ &= \Pr\left[C \mid R = r_0 + \varepsilon\right] + \Pr\left[A \mid R = r_0 + \varepsilon\right] \end{aligned}$$

and

$$\begin{aligned} P(r_0-\varepsilon) &= ET \mid R = r_0-\varepsilon = \Pr\left[T_1 = 1, T_0 = 1 \mid R = r_0 - \varepsilon\right] \\ &= \Pr\left[A \mid R = r_0 - \varepsilon\right] \end{aligned}$$

If follows that

$$\begin{aligned} P_+ - P_- &\equiv \lim_{\varepsilon \to 0} P(r_0 + \varepsilon) - \lim_{\varepsilon \to 0} P(r_0-\varepsilon) \\ &= \Pr\left[C \mid R = r_0\right] + \Pr\left[A \mid R = r_0\right] - \Pr\left[A \mid R = r_0\right] \\ &= \Pr[C \mid R = r_0], \end{aligned}$$

PROOF of THEOREM 1:

By Lemma 1, we have that each term involved in the expressions for $G(r_0 + \varepsilon)$ and $G(r_0 - \varepsilon)$ is continuously differentiable. Continuous differentiability means that the right derivative is equal to the left derivative so terms involving always takers and never taker are canceled out.

Define $g_1 \equiv E[Y_1 | \ R = r_0 + \varepsilon, C] \Pr[C \mid R = r_0 + \varepsilon]$ and $g_0 \equiv E[Y_0 | \ R = r_0 - \varepsilon, C] \Pr[C \mid R = r_0 - \varepsilon]$. Note that individual type C is implicitly a function of $R = r$. Let g'_{1+} and g'_{0-} be the right and left derivatives of g_1 and g_0 at $r = r_0$, respectively. For simplicity, further let E' and Pr' to denote derivatives of the conditional expectation and probability function, respectively. Then we have

$$g'_{1+} = \lim_{\varepsilon\to 0} E' [Y_1 | \ R = r_0 + \varepsilon, C] \Pr [C \mid R = r_0 + \varepsilon]$$
$$+ \lim_{\varepsilon\to 0} E [Y_1 | \ R = r_0 + \varepsilon, C] \, Pr' [C \mid R = r_0 + \varepsilon]$$
$$= E' [Y_1 | \ R = r_0, C] \Pr [C \mid R = r_0]$$
$$+ E [Y_1 | \ R = r_0, C] \, Pr' [C \mid R = r_0]$$

and

$$g'_{1-} = \lim_{\varepsilon\to 0} E' [Y_0 | \ R = r_0 - \varepsilon, C] \Pr [C \mid R = r_0 - \varepsilon]$$
$$+ \lim_{\varepsilon\to 0} E [Y_0 | \ R = r_0 - \varepsilon, C] \, Pr' [C \mid R = r_0 - \varepsilon]$$
$$= E' Y_0 | \ R = r_0, C \Pr [C \mid R = r_0]$$
$$+ E [Y_0 | \ R = r_0, C] \, Pr' [C \mid R = r_0] .$$

It follows that

$$\begin{aligned} G'_+ - G'_- &= g'_{1+} - g'_{0-} \\ &= \tau' \Pr [C \mid R = r_0] + \tau \left[\frac{\partial \Pr (C \mid R = r)}{\partial r} \mid r = r_0 \right] \end{aligned} \tag{18}$$

where $\tau = E [Y_1 - Y_0 | \ R = r_0, C]$ and $\tau' = \frac{\partial E[Y_1 - Y_0 | R = r, C]}{\partial r} \mid r = r_0$.

Similarly, each term in $P(r_0 + \varepsilon)$ and $P(r_0 - \varepsilon)$ is continuous differentiable, so right derivative is equal to left derivative for terms involving always takers, then we have

$$P'_+ - P'_- = \frac{\partial \Pr [C \mid R = r]}{\partial r} \mid r = r_0. \tag{19}$$

Plug eq. (19) into eq. (18). Also by Lemma 2 $\Pr [C \mid R = r_0] = P_+ - P_-$, then we have

$$G'_+ - G'_- = \tau'(P_+ - P_-) + \tau \left(P'_+ - P'_- \right). \tag{20}$$

When there is no jump, meaning $P_+ - P_- = 0$, but a kink, meaning $P'_+ - P'_- \neq 0$, eq. (20) can be rewritten as

$$\tau = \frac{G'_+ - G'_-}{P'_+ - P'_-}.$$

Recall by definition $\tau \equiv EY_1 - Y_0 \mid R = r_0, C$, and C here is such that $\Pr[C \mid R = r_0] = 0$ and $\left(\frac{\partial \Pr[C|R=r]}{\partial r} \mid r = r_0\right) \neq 0$. This proves that the ratio of kinks identifies the average treatment effect for marginal compliers defined by $\Pr[T_0 = 0, T_1 = 1 \mid R = r_0] = 0$ and $\left(\frac{\partial \Pr[T_0=0,T_1=1|R=r]}{\partial r} \mid r = r_0\right) \neq 0$.

PROOF of THEOREM 2:

When there is a jump, asymptotically w_n goes to zero, so eq. (4) will reduce to $\tau = (G_+ - G_-)/(P_+ - P_-)$, which is valid and is simply the standard RD estimand. When there is no jump, by the assumption there is a kink, eq. (4) will reduce to $\tau = \left(G'_+ - G'_-\right)/\left(P'_+ - P'_-\right)$, the kink based estimand, which is valid by Theorem 1.

PROOF of COROLLARYs 1a and 1b:

By Lemma 2 $G_+ - G_- = \tau(P_+ - P_-)$. Then given a jump one has $\tau = (G_+ - G_-)/(P_+ - P_-)$. By eq. (20) in the proof of Theorem 1, $G'_+ - G'_- = \tau'(P_+ - P_-) + \tau\left(P'_+ - P'_-\right)$. If and only if $\tau' = 0$, $G'_+ - G'_- = \tau\left(P'_+ - P'_-\right)$, so $\tau = \left(G'_+ - G'_-\right)/\left(P'_+ - P'_-\right)$. That is, when $\tau' = 0$, jump and kink identify the same parameter, i.e.,

$$\tau = \frac{G_+ - G_-}{P_+ - P_-} = \frac{G'_+ - G'_-}{P'_+ - P'_-}.$$

Then by the rule of fraction

$$\tau = \frac{G_+ - G_- + w\left(G'_+ - G'_-\right)}{P_+ - P_- + w\left(P'_+ - P'_-\right)},$$

for any $w \neq -(P_+ - P_-)/\left(P'_+ - P'_-\right)$. This estimand works when there is either a jump, or a kink, or both. In the first case it reduces to the jump estimand; in the second case, it reduces to the kink estimand; and in the third case, it is valid by the rule of fraction.

PROOF of COROLLARY 2:

By eq. (20) $G'_+ - G'_- = \tau'(P_+ - P_-) + \tau\left(P'_+ - P'_-\right)$. When $f_{\mathbf{S}|R}(\mathbf{s} \mid r)$ is continuous differentiable, then we have the conditional means of potential outcomes for each type of individuals $E(Y_t \mid R = r, \Psi)$ for $t = 0, 1$ and the probabilities of different types $\Pr(\Psi \mid R = r)$ for $\Psi \in \{A, N, C\}$ are continuously twice differentiable in a neighborhood of $r = r_0$. Then analogous to the derivation in the proof of Theorem 2, we have

$$G''_+ - G''_- = \tau''(P_+ - P_-) + 2\tau'(P'_+ - P'_-) + \tau\left(P''_+ - P''_-\right)$$

The above along with $G_+ - G_- = \tau(P_+ - P_-)$ in Lemma 2 and eq. (20) give the following system of eq.s:

$$\begin{aligned} G_+ - G_- &= \tau(P_+ - P_-), \\ G'_+ - G'_- &= \tau'(P_+ - P_-) + \tau\left(P'_+ - P'_-\right), \\ G''_+ - G''_- &= \tau''(P_+ - P_-) + 2\tau'(P'_+ - P'_-) + \tau\left(P''_+ - P''_-\right). \end{aligned}$$

Rewrite the above system of eq.s using notations given in the text, we have

$$\begin{aligned} \gamma_1 &= \tau\beta_1, \\ \gamma_2 &= \tau'\beta_1 + \tau\beta_2, \\ \gamma_3 &= \tau''\beta_1 + 2\tau'\beta_2 + \tau\beta_3. \end{aligned}$$

Given $\tau'' = 0$, and there is a kink, one can solve for τ from the second and third eqs. in the above to have $\tau = \left(2\beta_2\gamma_2 - \gamma_3\beta_1\right) / \left(2\beta_2^2 - \beta_3\beta_1\right)$. Given a jump we also have $\tau = \gamma_1/\beta_1$. Then by the rule of fraction

$$\tau = \frac{\gamma_1}{\beta_1} = \frac{2\beta_2\gamma_2 - \gamma_3\beta_1}{2\beta_2^2 - \beta_3\beta_1} = \frac{\gamma_1 + w\left(2\beta_2\gamma_2 - \gamma_3\beta_1\right)}{\beta_1 + w\left(2\beta_2^2 - \beta_3\beta_1\right)},$$

for any $w \neq -\beta_1/\left(2\beta_2^2 - \beta_3\beta_1\right)$.

The same procedure can be applied to cases where the d-th derivative $\tau^{(d)} = 0$ for any finite positive integer d. Keep taking derivatives on both sides of $\gamma_1 = \tau\beta_1$, until the d-th derivative. With the system of d equations and $\tau^{(d)} = 0$, one can back out τ, as the system of equations are recursive in nature.

12 Appendix II: Discussion of the Kink Estimand as an MTE

The analysis here shows that the treatment effect identified by the ratio of kinks $\tau = \frac{G'_+ - G'_-}{P'_+ - P'_-}$ in Theorem 1 is a marginal treatment effect (MTE) at the probability margin $p = P_- = P_+$, or an average effect for marginal persons at the indifference margin to treatment at the threshold.

To see this, one needs to introduce a latent index model for the treatment. Given $P(r) = E[T \mid R = r]$, the LATE framework in this paper implies a treatment model $T = I\{U_T \leq P(r)\}$, where U_T is a transformation of the latent error, say V in the binary treatment model, so $U_T = F_{V|R}(v)$, where $F_{\cdot|\cdot}$ is the conditional cumulative density function, and hence $U_T \sim U(0,1)$ (Discussion of the equivalence between the latent index model with the LATE framework can be found in Vytlacil 2002).

A marginal person who is just indifferent between being treated or not at the probability margin p is the person with $U_T = p$. MTE at a probability margin p is then defined as $EY_1 - Y_0 \mid R = r, U_T = p$, i.e. the average treatment effect for marginal persons at the probability margin p. LATE identified by the standard RD is then $EY_1 - Y_0 \mid R = r_0, P_- \leq U_T < P_+$, and compliers are individuals with $P_- \leq U_T < P_+$. We then have

$$\begin{aligned} E\left[Y \mid R = r, P(r) = p\right] &= E\left[Y_0 + T(Y_1 - Y_0) \mid R = r, P(r) = p\right] \\ &= E[Y_0 \mid R = r] + E\left[T(Y_1 - Y_0) \mid R = r, P(r) = p\right] \\ &= E[Y_0 \mid R = r] + E[Y_1 - Y_0 \mid R = r, T = 1]\, p \\ &= EY_0 \mid R = r + \int_0^p E[Y_1 - Y_0 \mid R = r, U_T = u_T]\, du_T \end{aligned}$$

where the last equality follows from that $T = 1$ when $U_T \in [0, p]$ and $T = 0$ when $U_T \in (p, 1$.

Take derivative of the above w.r.t. p, then we have

$$E\left[Y_1 - Y_0 \mid R = r, U_T = p\right] = \frac{\partial E\left[Y \mid R = r, P(r) = p\right]}{\partial p},$$

where the right-hand side of the above is identified if we have a valid IV for $P(r)$ – a standard result of the MTE literature, i.e., MTE can be identified by a local instrumental variable (LIV) estimator. Therefore, MTE represents the impact of a marginal change in the treatment probability on the outcome. Note that MTE is invariant to different IVs.

Recall $G(r) = EY \mid R = r = E[E[Y \mid R = r, T] \mid R = r] = EY \mid R = r, P(r)$. $P(r)$ is a kinked function of r at $r = r_0$; $P(r)$ is also continuous at $r = r_0$, so one can use $Z = 1\{R \geq r_0\}$ as an IV to identify MTE at one point $p = P_- = P_+$. That is

$$\begin{aligned}
E\left[Y_1 - Y_0 \mid R = r, U_T = P_-\right] &= \frac{\partial E\left[Y \mid R = r_0, P(r_0) = p\right]}{\partial p} \mid_{p=P_-} \\
&= \lim_{\varepsilon \to 0} \frac{G(r_0 + \varepsilon) - G(r_0 - \varepsilon)}{P(r_0 + \varepsilon) - P(r_0 - \varepsilon)} \\
&= \lim_{\varepsilon \to 0} \frac{G'(r_0 + \varepsilon) - G'(r_0 - \varepsilon)}{P'(r_0 + \varepsilon) - P'(r_0 - \varepsilon)} \\
&= \frac{G'_+ - G'_-}{P'_+ - P'_-}.
\end{aligned}$$

where the second equality follows essentially from the Chain rule by replacing derivatives in the numerator and the denominator with differences, and the third equality follows from L'Hôpital's rule and the fact that the differences $G(r_0 + \varepsilon) - G(r_0 - \varepsilon)$ and $P(r_0 + \varepsilon) - P(r_0 - \varepsilon)$ are differentiable w.r.t. ε by assumption.

The above proves that the ratio of kinks identifies an average treatment effect for marginal persons with $U_T = P_- = P_+$.

Intuitively, $P(r)$ is continuous at $r = r_0$, so when r changes marginally from just below r_0 to just above, the induced change in the treatment probability $P(r)$, $P_+ - P_-$, represents a marginal change rather than a finite (non-zero) change. The resulting change in the mean outcome $G_+ - G_-$ is then the change in outcome given a marginal change in the treatment probability at $p = P_- = P_+$.

13 Appendix III: Additional Estimation Results

Table A1 Impact of elite school attendance on post-compulsory education for males

	Jump		Kink		Jump&Kink	
			$R \in [540, 580]$			
	-0.730	-0.416	1.571***	1.582***	1.335***	1.343***
	(3.547)	(2.881)	(0.442)	(0.522)	(0.413)	(0.420)
Covariates	N	Y	N	Y	N	Y
			$R \in [520, 560]$			
	4.100	4.495	1.663***	1.719***	1.364***	1.378***
	(3.492)	(3.520)	(0.422)	(0.428)	(0.349)	(0.339)
Covariates	N	Y	N	Y	N	Y

Note: Robust standard errors clustered at the assignment score level are in parentheses; *** p<0.01, ** p<0.05, * p<0.1

References

[1] Altonji, J. G. and R. L. Matzkin (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," Econometrica, 73(4), 1053-1102

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

[3] Sebastian Calonico, M.Cattaneo and R. Titiunik (2014), "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs", Econometrica, forthcoming.

[4] Card, D., D. S. Lee, Z. Pei and A. Weber (2012), "Nonlinear Policy Rules and The Identification and Estimation of Causal Effects in a Generalized Regression Kink Design," NBER Working Paper 18564.

[5] 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.

[6] Dahlberg, M., E. Mork, J. Rattso and H. Agren (2008), "Using a Discontinuous Grant Rule to Identify the Effect of Grants on Local Taxes and Spending," Journal of Public Economics, 92 (12), 2320–2335.

[7] Clark, D. and E. Del Bono (2014) "The Long-Run Effects of Attending an Elite School: Evidence from the UK," Unpublished manuscript.

[8] Davidson, R. and J. G. MacKinnon (1993), Estimation and Inference in Econometrics, Oxford University Press.

[9] Dong, Y. and A. Lewble (2014), “Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," Boston College working paper.

[10] Dong, Y. (2014), “Regression Discontinuity Applications with Rounding Errors in the Running Variable," Forthecoming, Journal of Applied Econometrics.

[11] Dong, Y. (2014), “An Alternative Assumption to Identify LATE in Regression Discontinuity Designs," Unpublished manuscript.

[12] Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008), “Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects,” Econometrica, 76 (5), 1191-1206.

[13] Gelman and Imbens (2014), “Why High-order Polynomials Should not be Used in Regression Discontinuity Designs,” National Bureau of Economic Research, WP 20405

[14] Guryan, J. (2003), “Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts," National Bureau of Economic Research, WP8269.

[15] 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.

[16] Heckman, J. J. and E. Vytlacil (2005),“Structural Equations, Treatment Effects, and Econometric Policy Evaluation," Econometrica, 73(3), 669-738.

[17] Heckman, J. J. and E. J. Vytlacil (2007), “Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation," Handbook of Econometrics 6, in: J.J. Heckman & E.E. Leamer (ed.), 4779-4874.

[18] Imbens, G. W. and K. Kalyanaraman (2012), “Optimal Bandwidth Choice for the Regression Discontinuity Estimator," Review of Economic Studies 79(3), 933-959

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

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

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

[22] Lee, D. S. (2008), “Randomized Experiments from Non-random Selection in U.S. House Elections," Journal of Econometrics, 142(2), 675–697.

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

[24] Lemieux, T. and K. Milligan (2008), "Incentive effects of social assistance: A regression discontinuity approach," Journal of Econometrics 142, 807–828.

[25] Ludwig, J., and D. L. Miller (2007), “Does Head Start Improve Children’s Life Chances? Evidence from a Regression Discontinuity Design,” Quarterly Journal of Economics, 122, 159–208.

[26] McCrary, J. (2008), “Manipulation of the running variable in the regression discontinuity design: A density test," Journal of Econometrics 142,698–714.

[27] Nielsen, H. S., T. Sørensen, and C. Taber (2010), “Estimating the Effect of Student Aid on College Enrollment: Evidence from a Government Grant Policy Reform." American Economic Journal: Economic Policy, 2(2), 185-215.

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

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

[30] Simonsen, M., L. Skipper, and N. Skipper (2009), “Price Sensitivity of Demand of Prescription Drugs: Exploiting a Kinked Reimbursement Scheme,” University of Aarhus Economics Working Papers no. 2010-03.

[31] Thistlethwaite, D. L. and D. T. Campbell (1960), “Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment," Journal of Educational Psychology 51, 309–317.

[32] Vytlacil, E. (2002) , “Independence, monotonicity, and latent index models: An equivalence result," Econometrica, 70 (1), 331-341