Assessing the Monotonicity Assumption in IV and fuzzy RD designs

Mario Fiorini[1] and Katrien Stevens[2]

[1]University of Technology Sydney
[2]University of Sydney

February 15, 2016

Abstract

Whenever treatment effects are heterogeneous and there is sorting into treatment based on the gain, monotonicity is a condition that both Instrumental Variable and fuzzy Regression Discontinuity designs have to satisfy for their estimate to be interpretable as a LATE (Imbens and Angrist (1994), Hahn, Todd, and Van der Klaauw (2001)). However, applied economic work rarely discusses this important assumption. This is in stark contrast to the lengthy discussions dedicated to the other IV and fuzzy RD conditions. In this paper, we first use an extended Roy model to provide insights into the interpretation of IV and fuzzy RD estimates when monotonicity does not hold. We then extend our analysis to two applied settings to show that monotonicity can and should be investigated using a mix of economic insights, data patterns and formal tests. We point out that even in the absence of monotonicity interpretability of the estimate is not necessarily lost.

JEL classification: C1, C21, C26, I2, J01

Keywords: essential heterogeneity, monotonicity assumption, LATE, instrumental variable, regression discontinuity

We thank Jerôme Adda, Colin Cameron, Clément de Chaisemartin, Ben Edwards, Martin Huber, Susumu Imai, Markus Jäntti, Toru Kitagawa, Tobias Klein, Matthew Lindquist, Peter Siminski, Olena Stavrunova, Matthew Taylor and seminar participants at the Swedish Institute for Social Research (SOFI), Australasian Meetings of the Econometrics Society, University of Melbourne, University of Technology Sydney, University of Sydney, University of New South Wales, Monash University, University of Queensland and Australasian Labour Econometrics Workshop for helpful comments and suggestion.

1 Introduction

In the early 90's, work by Imbens and Angrist (1994), Angrist and Imbens (1995) and Angrist, Imbens, and Rubin (1996) provided the theoretical foundation for the identification of the Local Average Treatment Effect (LATE): the treatment effect for those individuals who are affected by the instrument. Heckman, Urzua, and Vytlacil (2006) stress that the LATE identification result is crucial when (i) the gain from treatment is heterogeneous across the population and (ii) there is sorting into treatment based on the gain from treatment. They define any context where both (i) and (ii) occur as *essential heterogeneity*. Under essential heterogeneity, the identification of the LATE requires the additional assumption of monotonicity: for a given change in the value of the instrument, it can not be that some individuals increase treatment intensity while others decrease treatment intensity. Hahn, Todd, and Van der Klaauw (2001) show that under essential heterogeneity, the assumption of monotonicity is also needed for the identification of a LATE in a fuzzy regression discontinuity (RD) approach. When monotonicity does not hold, the IV and fuzzy RD estimates are generally uninterpretable. Given the importance of the monotonicity condition in both IV and fuzzy RD designs, it is remarkable that this condition is seldom investigated in applied studies. This is in stark contrast to the lengthy discussions dedicated to the IV independence and rank conditions, and to the RD discontinuity (in the probability of treatment) and continuity (in the conditional regression function) conditions. A possible explanation for this missing step is the lack of a clear framework to think about monotonicity in practice.

The aim of this paper is to argue that applied work can and should investigate monotonicity by applying economic insights to the example at hand, and by data analysis, as it is generally done for other assumptions. When monotoncitiy is unlikely to hold, interpretation of the estimates should be adjusted accordingly. While IV and fuzzy RD designs can be incredibly powerful in dealing with endogeneity problems, a discussion of the monotonicity assumption is an extra step to validate the results. More specifically, we make two contributions.

First, we show how informative the IV and fuzzy RD estimates are under various degrees of heterogeneity in treatment effects, varying degrees of sorting on gain and the degree of violation of monotonicity. Using a simple model of selection into treatment, we provide general numerical examples. The originality here is to consider an extended Roy selection model, and to consider structural instead of reduced form parameters as inputs into the sensitivity analysis.

Since monotonicity is relevant for a wide range of applications, as a second contribution we go through two examples in different settings that adopt either the IV or fuzzy RD estimator: Black, Devereux, and Salvanes (2011) who use school entry age cutoffs as an instrument to investigate the effect of entering school older on IQ test scores and adult outcomes; Clark and Royer (2013) who use changes in compulsory schooling laws that generated discontinuities to investigate the effect of education on health. For each study we try to make a case for or against monotonicity. We show how the monotonicity assumption has sometimes been overlooked in the applied literature and how possible violations could have been detected and tested. For one of these studies, we also construct an extended Roy model to discuss interpretation of the estimates.

In doing so, we link the theoretical literatures on Instrument Variables, Regression

Discontinuity and Essential Heterogeneity with a focus on the applied-microeconomics profession.[1] Our paper complements the literature that focuses on deriving testable implications of instrument validity (Huber and Mellace (2015), Kitagawa (2015), Mourifie and Wan (2014)), and the literature that looks at replacing monotonicity with weaker assumptions while maintaining interpretable estimates (de Chaisemartin (2014), Huber and Mellace (2012), Klein (2010)).

The remaining of the paper is organized as follows. Section 2 discusses the monotonicity condition in an IV and RD setting, while section 3 illustrates interpretability of the estimates under a violation of monotonicity. Section 4 reviews a test of the monotonicity and independence conditions. We provide a thorough discussion of the monotonicity condition in several existing studies in section 5. Section 6 concludes.

2 The Monotonicity Assumption

To illustrate the notion of essential heterogeneity and monotonicity in a simple way, let us define a latent utility model where the endogenous variable (D) is binary, and the instrumental variable (Z) has support in $\mathcal{Z}$. For each individual i:

$$\begin{aligned} Y_i &= \beta_{0i} + \beta_{1i} D_i \qquad \text{(Outcome)} \\ D_i^* &= \alpha_{0i} + \alpha_{1i} Z_i \qquad \text{(Latent Utility)} \end{aligned} \tag{1}$$

with

$$D_i = \begin{cases} 1 & \text{if } D_i^* > 0 \\ 0 & \text{if } D_i^* \leq 0 \end{cases} \qquad \text{(Observed Treatment)}$$

All the coefficients $\{\beta_0, \beta_1, \alpha_0, \alpha_1\}$ are random variables with non-degenerate distributions. Here β_1 is the gain from treatment that is heterogeneous across individuals, and α_1 is the heterogeneous response to the instrument. The case of essential heterogeneity arises when $E(\beta_1|D=1) \neq E(\beta_1|D=0)$, that is when the treatment is partly driven by the gain from treatment. Essential heterogeneity is different from sorting on levels $E(\beta_0|D=1) \neq E(\beta_0|D=0)$.[2] The model can be generalized with the inclusion of covariates (X), in which case essential heterogeneity arises when $E(\beta_1|D=1,X) \neq E(\beta_1|D=0,X)$. In the rest of the paper we keep the conditioning on X implicit. Unless helpful, we also drop the i subscript for notational simplicity.

[1] Throughout the paper our focus is on the identification of LATEs for selected types of individuals. Heckman and Vytlacil (2005) and Heckman, Urzua, and Vytlacil (2006) discuss the identification of the Marginal Treatment Effect (MTE), the average treatment effect of those individuals at a margin of indifference for selecting into treatment. Including the MTE in our analysis is outside the scope of this paper.

[2] The econometrician often estimates (1) without allowing for heterogeneity:

$$\begin{aligned} Y_i &= \overline{\beta_0} + \overline{\beta_1} D_i + \overbrace{(\beta_{0i} - \overline{\beta_0}) + (\beta_{1i} - \overline{\beta_1}) D_i}^{U_i^Y} \\ D_i^* &= \overline{\alpha_0} + \overline{\alpha_1} Z_i + \underbrace{(\alpha_{0i} - \overline{\alpha_0}) + (\alpha_{1i} - \overline{\alpha_1}) Z_i}_{U_i^D} \end{aligned}$$

Both sorting on gain and on levels are then a source of endogeneity.

2.1 Monotonicity in the IV design

Following the notation in Imbens and Angrist (1994), let $Y(0)$ be the response without treatment for a given individual. $Y(1)$ is the response with treatment for the same individual. Define for each $z \in \mathcal{Z}$, $D(z)$ as the individual's treatment value when $Z = z$. Imbens and Angrist (1994) show that, provided a random variable Z satisfying the three following conditions is available, β_1^{IV} identifies the average treatment effect for those individuals who are affected by the instrument (LATE).

IV1. $P(z) = E[D|Z = z]$ is a non trivial function of z (rank)

IV2. $[Y(0), Y(1), \{D(z)\}_{z\in\mathcal{Z}}]$ is jointly independent of Z (independence)

IV3. for any two points of support $z, w \in \mathcal{Z}$, Either $D_i(z) \geq D_i(w)\ \forall i$, Or $D_i(z) \leq D_i(w)\ \forall i$ (monotonicity)

Condition IV1 is the rank condition. Condition IV2 requires that the instrument is independent of both potential outcomes and potential treatment assignment. This is stronger than the standard IV exclusion restriction: in model (1) it not only implies that Z is independent of β_0 but also that Z is independent of β_1 and $\{D(z)\}_{z\in\mathcal{Z}}$. The latter independence between Z and the counterfactual treatment values under alternative values of the instrument $\{D(z)\}_{z\in\mathcal{Z}}$ is sometimes referred to as "type" independence.[3] The strengthening of the exclusion restriction to this independence condition together with IV3 are needed to identify a LATE whenever there is essential heterogeneity.[4] The monotonicity assumption requires that, for every individual i, a change in the value of the instrument from w to z must either leave the treatment unchanged or change the treatment in the same direction. Monotonicity is violated if, because of the same change in the value of the instrument from w to z, some individuals respond by getting the treatment ("switching in") while others stop getting it ("switching out").[5] Monotonicity is a condition on counterfactuals: it refers to an individual's treatment in two *alternative* states of the world $Z = w$ and $Z = z$.

The IV estimator for any two points of support z, w in $\mathcal{Z}$ is given by

$$\beta_1^{IV}(z, w) \equiv \frac{E[Y|Z = z] - E[Y|Z = w]}{E[D|Z = z] - E[D|Z = w]}$$

Following Angrist, Imbens, and Rubin (1996), and assuming both IV1 and IV2, we can interpret the IV estimate of β_1 as

$$\begin{aligned}\beta_1^{IV}(z, w) = \lambda \times E[Y(1) - Y(0)|D(z) - D(w) = 1] + \\ (1 - \lambda) \times E[Y(1) - Y(0)|D(z) - D(w) = -1]\end{aligned} \tag{2}$$

[3] Individuals are classified into types depending on the counterfactual treatment values. For instance, when both the treatment and the instrument are binary there are only four types: compliers, defiers, always-takers and never-takers. We define these types more formally later.

[4] With sorting on gain $Cov(Z, \beta_0) = 0$ is not sufficient for identification. See footnote 5 in Heckman, Urzua, and Vytlacil (2006) and the related discussion for a proof.

[5] A necessary condition for monotonicity to be violated in (1) is α_1 positive for some individuals and negative for some others.

where

$$\lambda = \frac{P[D(z) - D(w) = 1]}{P[D(z) - D(w) = 1] - P[D(z) - D(w) = -1]}$$

Equation (2) is very informative.

- First, $\beta_1^{IV}(z, w)$ does not "use" individuals who do not respond to changes in the value of the instrument: $D(z) - D(w) = 0$. This motivates the standard LATE interpretation.

- Second, since λ is of the form $\frac{a}{a-b}$ with $a \geq 0$ and $b \geq 0$, then $\lambda \leq 0$ or $\lambda \geq 1$. If monotonicity holds then either $\lambda = 0$ or $\lambda = 1$, and IV measures the LATE for a specific group of individuals. For instance if $D_i(z) - D_i(w) \geq 0 \; \forall i$, then $\lambda = 1$ and IV estimates the effect for those individuals that are induced to take the treatment because of the instrument (LATE-*in*): $\beta_1^{IV}(z, w) = E[Y(1) - Y(0)|D(z) - D(w) = 1]$. Alternatively, if monotonicity holds because $D(z) - D(w) \leq 0 \; \forall i$ then $\lambda = 0$, and IV estimates the effect for those individuals that stop getting the treatment because of the instrument (LATE-*out*): $\beta_1^{IV}(z, w) = E[Y(1) - Y(0)|D(z) - D(w) = -1]$.

- Third, if monotonicity does not hold, then either $\lambda < 0$ or $\lambda > 1$, and the IV estimate is neither a LATE nor a weighted average of LATEs. In this case, individuals respond differently to a given change in the value of Z, such that for some individuals $D(z) - D(w) = 1$ while for some others $D(z) - D(w) = -1$.

- Fourth, if monotonicity does not hold, it is close to impossible to recover a reliable estimate of the LATE for the group of interest. Equation (2) clearly shows that the interpretation of the IV estimate depends on four unknowns: the proportion of switchers-in, the proportion of switchers-out, and the LATEs for each of these groups. Therefore, one needs to make assumptions on three out of four unknowns to confidently back-out the fourth one. For example, even if one has a very good guess on both proportions such that λ is "known", recovering one of the LATEs is still not possible.

- Finally, if the return to treatment is heterogeneous but there is no sorting on gain, then monotonicity is not required. Without sorting on gain the expected return from treatment is the same among those who switch in and out, and equal to the average treatment effect (ATE):

$$\beta_1^{IV}(z, w) = \text{LATE-}in = \text{LATE-}out = E[Y(1) - Y(0)]$$

2.2 Monotonicity in the fuzzy RD design

Let the setting be similar to equation (1), but with Z taking on a continuum of values ($v \in \mathcal{Z}$) and $D_i^* = \alpha_{0i} + \alpha_{1i}\mathbb{1}[Z_i > v_0]$. Also let

$$\lim_{v \downarrow v_0} P[D = 1|Z = v] \neq \lim_{v \uparrow v_0} P[D = 1|Z = v]$$

such that the probability of treatment jumps at the threshold v_0, without requiring the jump to be equal to 1.[6] Hahn, Todd, and Van der Klaauw (2001) point out that there is a very close analogy between the fuzzy Regression Discontinuity and the IV estimators since both can be expressed as a Wald estimator:

$$\beta_1^{RD} \equiv \frac{\lim_{v \downarrow v_0} E[Y|Z=v] - \lim_{v \uparrow v_0} E[Y|Z=v]}{\lim_{v \downarrow v_0} E[D|Z=v] - \lim_{v \uparrow v_0} E[D|Z=v]}$$

which measures a similar LATE to $\beta_1^{IV}(z,w)$ if $z = v_0 + \epsilon$, $w = v_0 - \epsilon$ and ϵ is arbitrarily small. Hahn, Todd, and Van der Klaauw (2001) show that under essential heterogeneity, regression discontinuity estimation identifies β_1 for those individuals with $Z = v_0$ who are affected by the threshold (LATE at v_0) under the following conditions

RD1. $\lim_{v \downarrow v_0} P[D=1|Z=v] \neq \lim_{v \uparrow v_0} P[D=1|Z=v]$ (RD)

RD2. $E[Y(0)|Z=v]$ is continuous in v at v_0 (continuity)

There exists a small number $\xi > 0$ such that for all $0 < e < \xi$

RD3. $[\beta_1, D(v-e), D(v+e)]$ is jointly independent of Z near v_0 (independence)

RD4. Either $D_i(v_0+e) \geq D_i(v_0-e)\ \forall i$, Or $D_i(v_0+e) \leq D_i(v_0-e)\ \forall i$ (monotonicity)

Condition RD1 is the RD equivalent of the rank condition in the IV setting. Condition RD2 implies that in the absence of treatment, individuals close to the threshold v_0 are similar. These first two conditions must hold whether there is essential heterogeneity or not. Similar to IV estimation, whenever there is essential heterogeneity a stronger independence condition together with monotonicity are needed to identify a LATE at v_0. Note again that monotonicity is a condition on counterfactuals: for every individual i, crossing the threshold must either leave the treatment unchanged or change the treatment in the same direction. Invoking the reasoning in Angrist, Imbens, and Rubin (1996), we can interpret the RD estimate of β_1 as[7]

$$\begin{aligned} \beta_1^{RD} = \lim_{e \to 0} \Big\{ & \lambda \times E[Y(1) - Y(0) | D(v_0+e) - D(v_0-e) = 1] + \\ & (1-\lambda) \times E[Y(1) - Y(0) | D(v_0+e) - D(v_0-e) = -1] \Big\} \end{aligned} \tag{3}$$

where

$$\lambda = \frac{P[D(v_0+e) - D(v_0-e) = 1]}{P[D(v_0+e) - D(v_0-e) = 1] - P[D(v_0+e) - D(v_0-e) = -1]}$$

Equation (3) is the equivalent of equation (2) in an RD setting and provides the same insights. Thus, if monotonicity holds, λ is equal to either 0 or 1, and the RD estimate can be interpreted as a LATE at the threshold. When monotonicity is violated either $\lambda < 0$ or $\lambda > 1$, and the RD estimate measures neither a LATE for any particular group of individuals nor a weighted average of LATEs.

[6] When the jump in treatment probability is equal to 1, monotonicity is satisfied by definition. This case is defined in the literature as a *sharp* regression discontinuity design.

[7] Proof in appendix A.

2.3 Monotonicity when either treatment or instrument are multivalued

When either the treatment or instrument take multiple values, monotonicity remains an important assumption to allow interpretation of IV and fuzzy RD estimates. With a multivalued treatment, monotonicity implies that a change in instrument value w to z causes all individuals to either increase treatment intensity, or be unaffected. If monotonicity is satisfied, the IV or fuzzy RD estimate measures an "average causal response" (ACR, Angrist and Imbens (1995)). A violation of monotonicity renders the estimate uninterpretable. With a multivalued instrument, Imbens and Angrist (1994) supplement monotonicity with another condition to show that β^{IV} is a weighted average of LATEs (if the treatment is binary) or ACRs (if the treatment is multivalued). However, if monotonicity is violated, this LATE or ACR interpretation is lost. We refer the reader to appendix B for a thorough discussion in both the IV and fuzzy RD setting.

3 Interpretation of IV and RD estimates: Sensitivity to Key Assumptions

The discussion in section 2 highlights that in a world with essential heterogeneity, monotonicity is an important condition to interpret the IV and RD estimates. However, the researcher might be interested in knowing to what degree the presence of essential heterogeneity and a violation of monotonicity are a problem. What if the treatment effects are "roughly" homogeneous? What if there is just a "bit" of sorting on gain? What if only a "few" individuals violate monotonicity? To answer this question we set up a extended Roy model with an exogenous variable Z affecting the treatment decision. The key characteristics of this model are i) treatment effects are heterogeneous, ii) selection into treatment is based on the gain, and iii) the impact of Z is heterogeneous across individuals. The goal of this Roy model is to investigate the extent to which the IV and fuzzy RD estimates can be interpreted as a LATE of interest as i), ii) and iii) are strengthened or weakened, by changing the structural parameters of the model.

3.1 A Roy model with violation of monotonicity

For each individual i let Y_{0i} be the outcome if she does not take the treatment, and let Y_{1i} be the outcome if she does:

$$Y_{1i} = \alpha + \overline{\beta} + U_{1i}$$
$$Y_{0i} = \alpha + U_{0i}$$

Thus the gain from treatment is heterogeneous and given by $\beta_i = Y_{1i} - Y_{0i} = \overline{\beta} + U_{1i} - U_{0i}$. Individuals decide whether or not to take treatment partly on the basis of the idiosyncratic gain (sorting on gain) and partly on the basis of an exogenously determined variable Z (instrument).[8] Dropping the i subscript for notational simplicity, let the treatment D be determined as follows:

[8] We illustrate the Regression Discontinuity case in section 3.4.

$$D = \begin{cases} 1 & \text{if } Y_1 - Y_0 + \gamma Z > 0 \Leftrightarrow \beta > -\gamma Z \\ 0 & \text{if } Y_1 - Y_0 + \gamma Z \leq 0 \Leftrightarrow \beta \leq -\gamma Z \end{cases}$$

Here γZ could be interpreted as the cost or taste for treatment.[9] To keep things simple, consider a binary instrument Z and a coefficient γ that can take two values: $z \in \{0, 1\}$, and $\gamma = \gamma_L$ or $\gamma = \gamma_H$. Importantly, for monotonicity to be violated, we need to set $\gamma_L < 0$ and $\gamma_H > 0$. Thus, the instrument "pulls" some individuals out of treatment ($\gamma < 0$) while it "pushes" other individuals intro treatment ($\gamma > 0$). The proportion of individuals with γ_L and γ_H are given by p_{γ_L} and $p_{\gamma_H} = 1 - p_{\gamma_L}$.

Now we can define how individuals make their treatment decision based on their realization of Z, γ, U_1 and U_0. Given sorting on gain, there is a cut-off value of β above which individuals take treatment. That cut-off value is affected by the instrument, as indicated in table 1. The treatment decisions under alternative values of Z allow us to distinguish between 4 types based on their counterfactual choices: always-takers (AT), never-takers (NT), compliers (CM) and defiers (DF).

Table 1: Counterfactual Choices $\big(D(Z=1), D(Z=0)\big)$

$\gamma = \gamma_L$			
	$\beta \leq 0$	$0 < \beta \leq -\gamma_L$	$\beta > -\gamma_L$
$Z = 0$	$D = 0$	$D = 1$	$D = 1$
$Z = 1$	$D = 0$	$D = 0$	$D = 1$
type	NT	DF	AT

$\gamma = \gamma_H$			
	$\beta \leq -\gamma_H$	$-\gamma_H < \beta \leq 0$	$\beta > 0$
$Z = 0$	$D = 0$	$D = 0$	$D = 1$
$Z = 1$	$D = 0$	$D = 1$	$D = 1$
type	NT	CM	AT

We maintain that conditions IV1 (rank) and IV2 (independence) are satisfied. The rank condition requires that $P(D = 1|Z = 1) \neq P(D = 1|Z = 0)$. Under IV2, $P(D = 1|Z = 1) = p_{AT} + p_{CM}$ and $P(D = 1|Z = 0) = p_{AT} + p_{DF}$ where p_{AT}, p_{CM} and p_{DF} are the proportions of always-takers, compliers and defiers respectively. Thus the rank condition implies that $p_{CM} \neq p_{DF}$. Moreover, since in our model types are defined by the pair (β, γ), type independence requires that these parameters are independent of Z. To simplify our discussion below, we also constrain β and γ to be uncorrelated though our results generalize to the case where they are correlated. Finally, we assume that the treatment effects are distributed normally: $\beta \sim N(\overline{\beta}, \sigma_\beta)$.

We can then define the probability of observing each type as a function of γ_L, γ_H, p_{γ_L} and the distribution of the gain from treatment $f(\beta)$. Figure 1 illustrates where the different types are located along the β distribution.[10]

- $p_{AT} = p_{\gamma_L} \times P[\beta > -\gamma_L] + p_{\gamma_H} \times P[\beta > 0]$. Always-takers (ATs) take treatment irrespective of the instrument. Therefore, they must have a high return to treatment. In particular, ATs with γ_L must have an especially large β. Since $\gamma_L < 0$,

[9] The model could be extended to include an explicit cost or taste parameter, in which case the γZ term could be interpreted as a change in the cost or taste for treatment.

[10] Figure 1 is obtained using the parametrization described in table 2a.

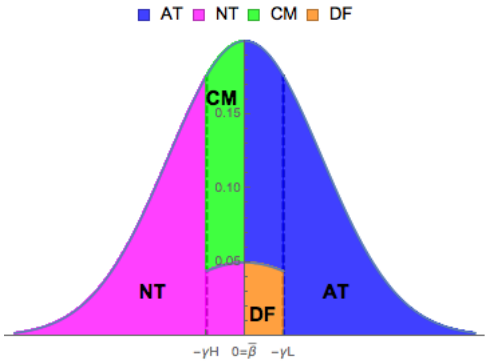


Figure 1: Types over the β distribution

the instrument tends to pull them out of treatment, hence they need a large and positive β to compensate for it and remain treated.

- $p_{NT} = p_{\gamma_L} \times P[\beta \leq 0] + p_{\gamma_H} \times P[\beta \leq -\gamma_H]$. Never-takers (NTs) do not take treatment irrespective of the instrument. Therefore, they must have a low return to treatment. In particular, NTs with γ_H must have an especially low β. Since $\gamma_H > 0$, the instrument tends to push them into treatment, in which case they need a very negative β to compensate for it and remain untreated.

- $p_{CM} = 0 + p_{\gamma_H} \times P[-\gamma_H < \beta \leq 0]$. Compliers (CMs) are on the margin of taking treatment. They switch into treatment only when the instrument changes from 0 to 1. Therefore they must have $\beta \leq 0$ and $\gamma = \gamma_H$ such that $Z = 1$ pushes them into treatment, but they must also have $\beta > -\gamma_H$ otherwise the push is not sufficient to compensate for the negative treatment effect.

- $p_{DF} = p_{\gamma_L} \times P[0 < \beta \leq -\gamma_L] + 0$. Defiers (DFs) are also on the margin of taking treatment. However, they switch out of treatment if the instrument changes from 0 to 1. Therefore they must have $\beta > 0$ and $\gamma = \gamma_L$ such that $Z = 1$ pulls them out of treatment, but they must also have $\beta < -\gamma_L$ otherwise the push is not sufficient to compensate for the positive treatment effect.

Given the location of types along the β distribution, one can easily derive the Average Treatment Effects of the various types. See appendix C for the formal expressions.

3.2 A parametrized baseline model

In this section we investigate the effect of violating the monotonicity condition. We do so under our baseline parametrization, while in the next section we alter the baseline. Figure 1 was obtained using the parametrization in table 2a, and letting $U_1 \sim N(\mu_1, \sigma_1)$, $U_0 \sim N(\mu_0, \sigma_0)$ and $Cov(U_1, U_0) = \sigma_{01}$. The baseline parametrization is chosen to be simple while ensuring that β is heterogeneous. Since $\beta = \overline{\beta} + U_1 - U_0$ then $\sigma_\beta = \sqrt{\sigma_1^2 + \sigma_0^2 - 2\sigma_{01}}$. In the baseline we set $\sigma_0 = \sigma_1 = 1$ and $\sigma_{01} = -1$ in order to get a positive standard deviation of $\sigma_\beta = 2$. We also set $p_{\gamma_L} = 0.25$ to ensure that the rank condition is satisfied: $p_{CM} \neq p_{DF}$. The resulting proportions of types is described in table 2b. Most of the population is given by always-takers, followed by never-takers. However, there are also

both compliers and defiers. In this Roy model, with a binary treatment and a binary instrument

$$\beta^{IV} = \lambda \times LATE_{CM} + (1 - \lambda) \times LATE_{DF} \tag{4}$$

where

$$\lambda = \frac{p_{CM}}{p_{CM} - p_{DF}}$$

Whenever both $p_{CM} \neq 0$ and $p_{DF} \neq 0$ monotonicity is violated. Thus, in our example, no economic information can be recovered from β^{IV}. It is neither a LATE nor a weighted average of the LATE for compliers and defiers, but it is more extreme since $\lambda > 1$ (see table 2c).[11]

Table 2: Baseline

(a) Parametrization

γ_L	γ_H	p_{γ_L}	$\overline{\beta}$	μ_1	μ_0	σ_1	σ_0	σ_{10}
-1	1	0.25	0	0	0	1	1	-1

(b) Types

p_{AT}	p_{NT}	p_{CM}	p_{DF}	$p_{AT} + p_{NT} + p_{CM} + p_{DF}$	λ
0.452134	0.356403	0.143597	0.0478656	1	**1.5**

(c) ATE, LATEs and IV estimate

ATE	ATE_{AT}	ATE_{NT}	$LATE_{CM}$	$LATE_{DF}$	β^{IV}
0	1.71287	-2.04142	-0.489673	0.489673	**-0.979345**

3.3 Altering the baseline: interpretation of β^{IV}

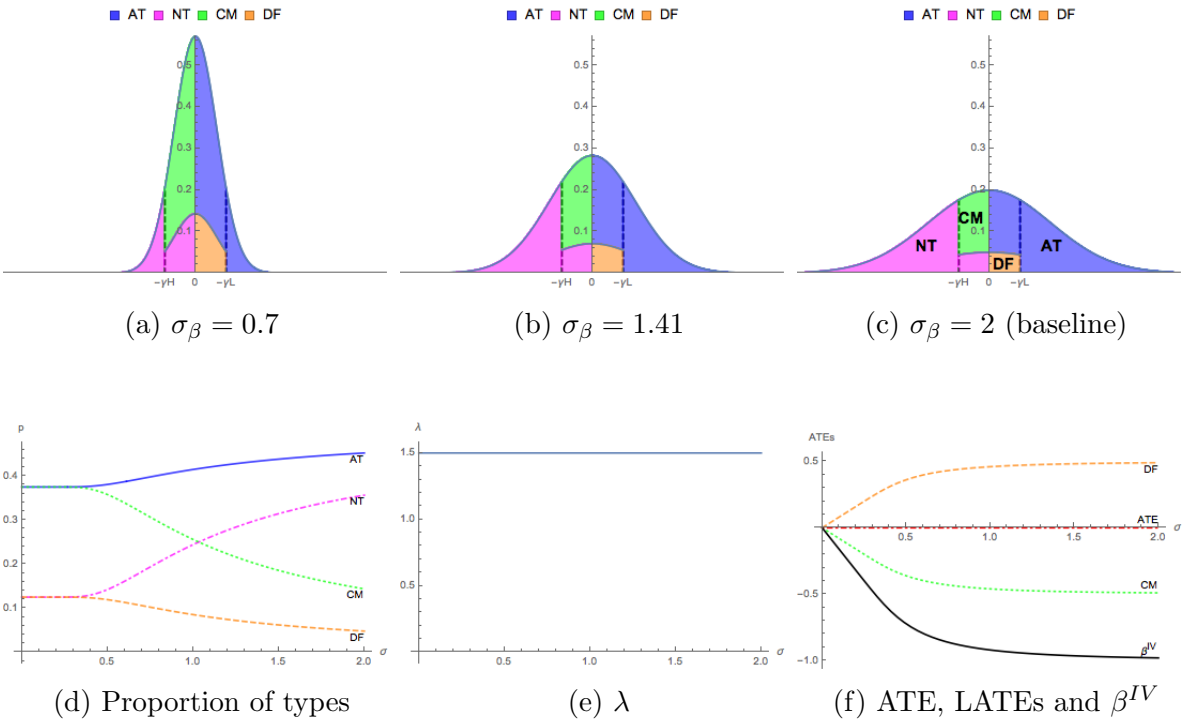
In this section we alter the parameters in the baseline scenario and show how the decisions of individuals are affected. The goal is to illustrate whether β^{IV} becomes more/less interpretable as we change the degree of heterogeneity in the treatment effect (σ_β), the heterogeneity in the impact of the instrument (p_{γ_L}) and the degree of sorting on gain.

3.3.1 Heterogeneity in the treatment effect: σ_β

In the baseline we set $\sigma_\beta = 2$ in order to get heterogeneous treatment effects. In figure 2a-c we illustrate the types over the β distribution as a function of σ_β. As σ_β decreases the distribution narrows around the Average Treatment Effect ($\overline{\beta} = 0$). Figure 2d shows the proportion of types: as σ_β goes towards zero all observations get concentrated within

[11] In table 2c $\beta^{IV} = 2LATE_{CM} = -2LATE_{DF}$. This occurs because $LATE_{CM} = -LATE_{DF}$ and $\lambda = 1.5$. It might be tempting to conclude that if we had a good guess of λ then it would be possible to recover the $LATE_{CM}$ and $LATE_{DF}$ from the β^{IV}. However, $LATE_{CM} = -LATE_{DF}$ only because $f(\beta)$ is centred around 0 and $\gamma_H = -\gamma_L$. More generally, $LATE_{CM}$ and $LATE_{DF}$ are not related and even knowledge of λ would not allow us to recover any LATE.

the $[-\gamma_H, -\gamma_L]$ interval. Thus, the proportion of always-takers and never-takers falls. Moreover, the proportions of compliers and always-takers converge. The same is true for the proportions of never-takers and defiers. Because of the symmetry around the threshold, λ is constant. Figure 2f shows the ATE, the LATEs for compliers and defiers, and β^{IV}. All the LATEs and the β^{IV} tend to the ATE as σ_β goes towards zero: in the absence of treatment effect heterogeneity there is no monotonicity requirement. However, for positive σ_β the β^{IV} is always more extreme and far from the LATEs (since $\lambda > 1$).



(a) $\sigma_\beta = 0.7$ (b) $\sigma_\beta = 1.41$ (c) $\sigma_\beta = 2$ (baseline)

(d) Proportion of types (e) λ (f) ATE, LATEs and β^{IV}

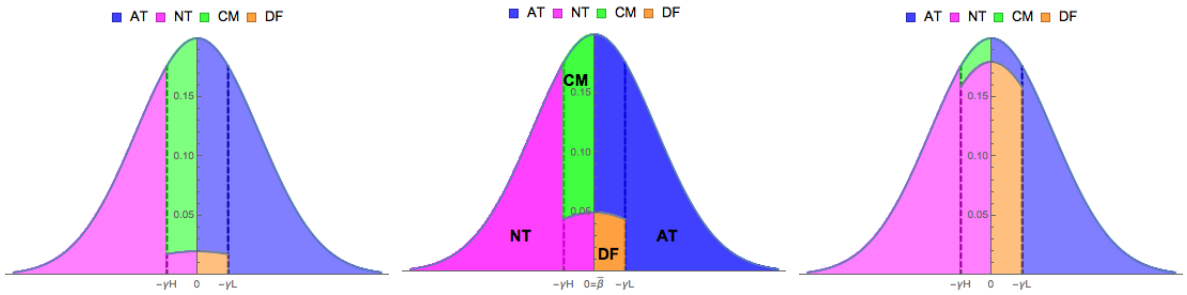
Figure 2: Sensitivity to σ_β

3.3.2 Heterogeneity in the response to the instrument: p_{γ_L}

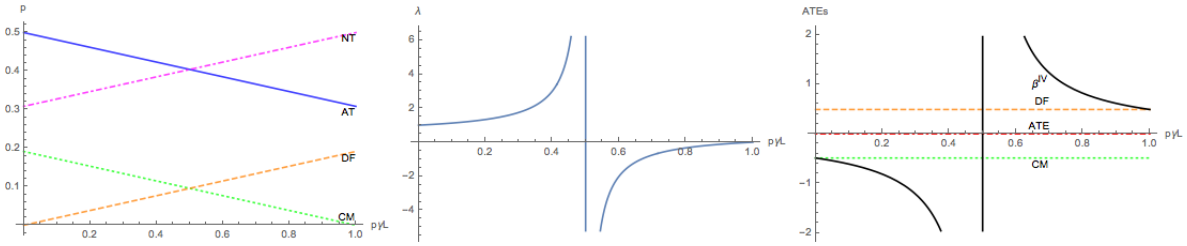
In this section we investigate what happens when we vary the degree of heterogeneity in the response to the instrument. In our model, this heterogeneity exists since γ can be either positive (γ_H) or negative (γ_L): individuals with positive γ are needed to generate compliers, while individuals with negative γ are needed to generate defiers.[12] As long as $0 < p_{\gamma_L} < 1$ there are both compliers and defiers causing monotonicity to be violated. Figure 3a-c illustrates the types over the β distribution as a function of p_{γ_L}. Note that the case with $p_{\gamma_L} > 0.5$ is a mirror image of $p_{\gamma_L} < 0.5$, with defiers rather than compliers being the dominant group affected by the instrument. As p_{γ_L} rises, there are fewer individuals positively affected by the instrument: thus the proportion of compliers decreases while the proportion of defriers increases. In the limit case where $p_{\gamma_L} = 0$ (no heterogeneity in the response to the instrument) there are no defiers. The opposite is true if $p_{\gamma_L} = 1$:

[12]Strictly speaking, there is heterogeneity in the response to treatment also if γ_L and γ_H have the same sign. But in that case, heterogeneity does not create a violation of monotonicity: there are either no compliers or no defiers.

there are no more compliers. Figure 3f shows that once again β^{IV} cannot be interpreted as a LATE of interest unless we are in the limit cases of $p_{\gamma_L} = 0$ (leading to $\lambda = 1$) or $p_{\gamma_L} = 1$ (leading to $\lambda = 0$): no heterogeneity in the response to the instrument. As soon as we depart from either of the limit cases β^{IV} becomes uninformative rapidly. An especially problematic situation occurs when p_{γ_L} is close to 0.5 in which case $p_{CM} \approx p_{DF}$ and $|\lambda| \to \infty$.[13]



(a) $p\gamma_L = 0.1$ (b) $p\gamma_L = 0.25$ (baseline) (c) $p\gamma_L = 0.9$



(d) Proportion of types (e) λ (f) ATE, LATEs and β^{IV}

Figure 3: Sensitivity to p_{γ_L}

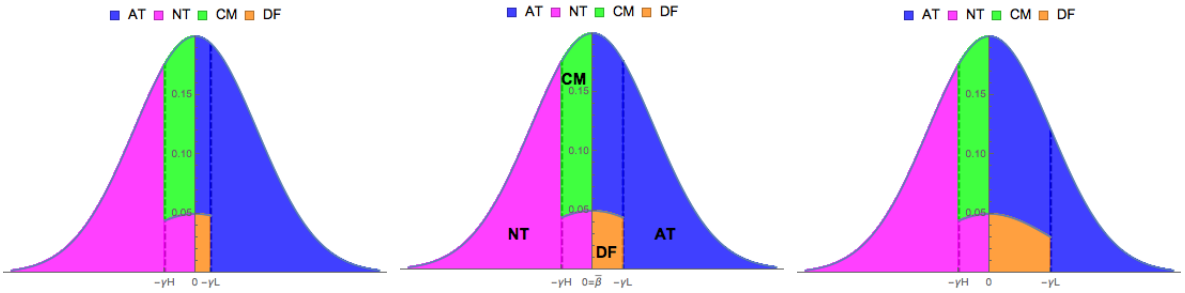
3.3.3 Other forms of Heterogeneity: γ_L

There is another way to alter the degree of heterogeneity in this model aside from shifting p_{γ_L}. In our baseline $\gamma_H = -\gamma_L$, that is the magnitude of the response to the instrument is the same between individuals with different γ, albeit in opposite directions. However, it could be that $\gamma_H \neq -\gamma_L$. Thus, holding γ_H constant, shifting the value of γ_L leads to different proportions of defiers (for a given proportion of compliers). Figure 4a-c illustrates the types over the β distribution as a function of γ_L: the closer γ_L is to zero the smaller the proportion of defiers. A very negative γ_L has the opposite effect. Figure 4d shows the proportion of types as a function of ψ_L, a parameter re-scaling the baseline: $\gamma_L = \psi_L \gamma_L^B$ (with $\gamma_L^B = -1$). For $\psi_L = 0$ then $\gamma_L = 0$: this is a limit case with heterogeneity in the response to the instrument but no defiers since no individual is pushed out of treatment by the instrument.[14] If $\psi_L > 0$ then γ_L becomes a negative number, $\lambda > 1$ and β^{IV} cannot be interpreted as a LATE of interest. In our example λ

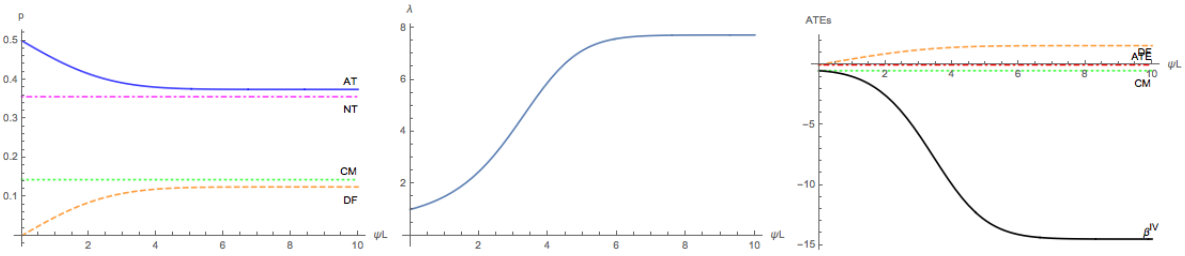
[13]The limit case itself ($p_{\gamma_L} = 0.5$) is uninteresting since the instrument does not satisfy the rank condition ($p_{CM} = p_{DF}$). More generally, the closer pCM and pDF are, the weaker the instrument.

[14]This case is equivalent to a situation where γ is heterogeneous but only takes positive values.

has an horizontal asymptote since, even for very negative γ_L, the proportion of defiers is always smaller than the proportion of compliers. Note that a shift in the value of γ_L also has implications for the impact of the instrument. As we discussed earlier, the rank condition demands that $p_{CM} \neq p_{DF}$. Figure 4d clearly shows that the difference in the proportion of compliers and defiers is largest for $\gamma_L = 0$, right because in that case $p_{DF} = 0$.



(a) $\gamma_L = -0.5$ (b) $\gamma_L = -1$ (baseline) (c) $-\gamma_L = -2$



(d) Proportion of types (e) λ (f) ATE, LATEs and β^{IV}

Figure 4: Sensitivity to γ_L

3.3.4 Sorting on gain

In the model above, we assume that individuals sort into treatment based on the gain, and treatment decisions are potentially affected by the instrument Z. Theoretically we know that monotonicity is not a problem in the limit case of no sorting on gain. However, is a small degree of sorting on gain sufficient to render β^{IV} uninformative?

To address this question, we adjust our model slightly by multiplying the gain from treatment with a homogenous parameter $\theta \geq 0$:

$$D = \begin{cases} 1 & \text{if } \theta\ (Y_1 - Y_0) + \gamma Z > 0 \Leftrightarrow \beta > \frac{-\gamma}{\theta} Z \\ 0 & \text{if } \theta\ (Y_1 - Y_0) + \gamma Z \leq 0 \Leftrightarrow \beta \leq \frac{-\gamma}{\theta} Z \end{cases}$$

Individuals make a treatment decision based on a signal about β. In economic terms, $\theta \neq 1$ can be interpreted as individuals having incorrect information about the actual return to treatment β. θ can be either smaller or larger than one. For instance, this could occur because the media under or overestimate the true return to treatment by not considering the endogeneity problem. θ close to zero can also reflect the existence of

some constraints that preclude β from being used in the treatment decision.[15] From this adjusted treatment equation, it is clear that the ratio $\frac{\gamma}{\theta}$ matters for treatment decisions (not just γ as earlier). This ratio indicates the relative strength of the instrument versus selection in terms of the gain in determining treatment decisions. Figures 5a-c show what happens when θ rises (from 0.5 to 2). This implies the strength of the instrument falls relative to the sorting on gain channel. The β-thresholds inbetween which compliers and defiers exist move closer to zero. Hence, the fraction of compliers and defiers falls and converges to 0, while the fraction of always-takers and never-takers rises and converges to 0.5 (see figure 5d). This again results in a constant λ for any value of θ, as can be seen in figure 5e. As the β-thresholds move towards zero, the magnitude of the LATEs on compliers and defiers fall with θ (see figure 5f). With λ constant this results in a β^{IV} that is always more extreme, and converges to the LATEs and ATE. A violation of monotonicity is thus a problem even with little sorting on the gain.

Intuitively, as long as θ is strictly non-zero, individuals have some information on their β, and thus take the gain into account in their treatment decision. Hence there are always-takers, never-takers, compliers and defiers (as shown in figure 5). As θ goes to ∞ (extreme sorting on gain), the LATEs and β^{IV} tend towards the ATE: this is an uninteresting limit case since it implies the instrument has no impact at all. In the other extreme of no sorting on gain ($\theta = 0$) the instrument is the only factor driving the decision. The treatment decision now simplifies to

$$D = \begin{cases} 1 & \text{if } \gamma Z > 0 \\ 0 & \text{if } \gamma Z \leq 0 \end{cases}$$

There is no more threshold value of β above or below which individuals change treatment decision. If only Z matters, no one takes treatment under Z=0, thus there cannot be always-takers or defiers. Individuals with a negative impact of the instrument ($\gamma < 0$) will never take treatment (never-takers), while a positive γ will always push individuals into treatment: compliers. In this limit case $\beta^{IV} = LATE_{CM} = LATE_{NT} = ATE$.

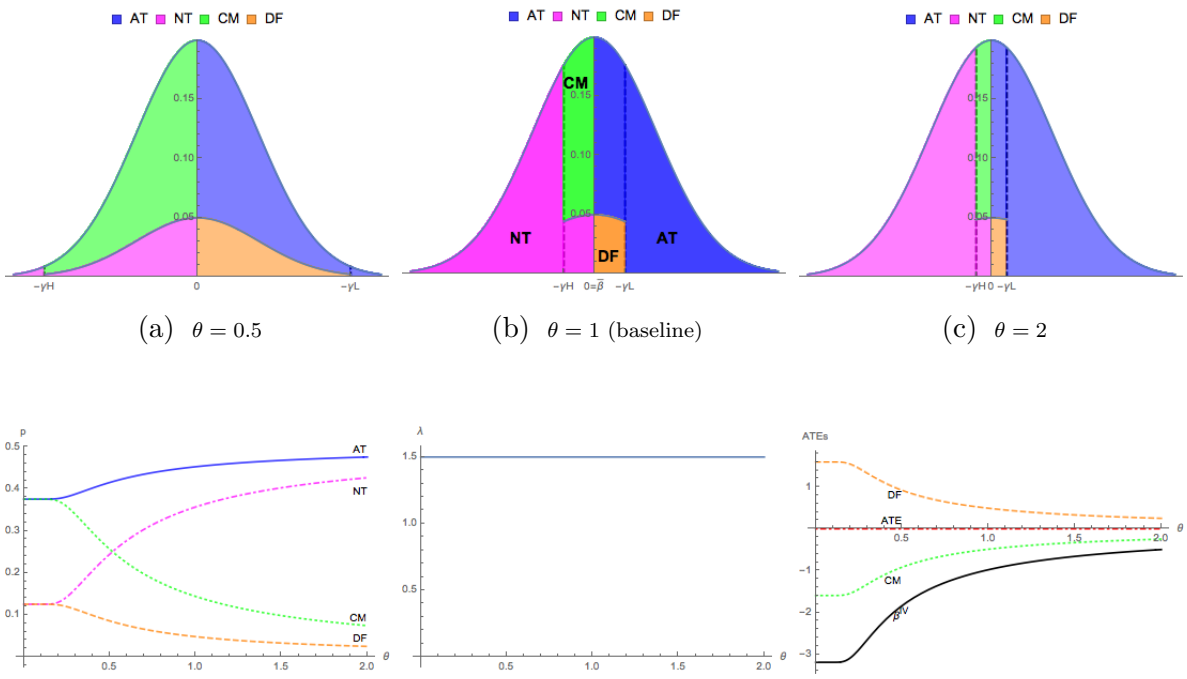
3.4 Regression Discontinuity

The same Roy model can be used to explain the importance of monotonicity in a fuzzy Regression Discontinuity Design. Let the selection equation be

$$D = \begin{cases} 1 & \text{if } Y_1 - Y_0 + \gamma \mathbb{1}[Z > v_0] > 0 \Leftrightarrow \beta > -\gamma \mathbb{1}[Z > v_0] \\ 0 & \text{if } Y_1 - Y_0 + \gamma \mathbb{1}[Z > v_0] \leq 0 \Leftrightarrow \beta \leq -\gamma \mathbb{1}[Z > v_0] \end{cases}$$

Thus treatment now depends on Z in a discontinuous way: if Z is larger than a cut-off value v_0 there is an additional effect determined by γ. All other elements of the model stay unchanged, that is $\gamma \in \{\gamma_L, \gamma_H\}$ with $\gamma_L < 0$ and $\gamma_H > 0$, and the proportion of individuals with the two values of γ are given by p_{γ_L} and $p_{\gamma_H} = 1 - p_{\gamma_L}$. We also maintain that around the threshold value v_0 the assumptions of discontinuity in the probability of treatment (RD1), continuity in the conditional regression function (RD2) and independence (RD3) are satisfied.

[15] The latter can occur if a person different from the person undergoing a treatment is making the treatment decision.



(a) $\theta = 0.5$ (b) $\theta = 1$ (baseline) (c) $\theta = 2$

(d) Proportion of types (e) λ (f) ATE, LATEs and β^{IV}

Figure 5: Sensitivity to θ

3.4.1 Sharp RD

In a sharp design, treatment is known to depend in a deterministic way on Z. For instance, all individuals with $Z > v_0$ take treatment and viceversa. This is an example where all individuals are compliers. The model above would generate a sharp RD only if there is no sorting on gain and if γ is homogeneous: β not entering the selection equation rules out the existence of always-takers and never-takers while $p_{\gamma_L} = 0$ rules out the existence of defiers. Otherwise either β or γ introduce randomness in the treatment decision which results in a fuzzy RD.[16]

3.4.2 Fuzzy RD

The fuzzy design differs from the sharp design in that the treatment assignment is not a deterministic function of Z but there are additional variables unobserved by the econometrician that determine assignment to treatment. In the model these variables are β and γ. Thus, in the presence of sorting on gain or with $0 < p_{\gamma_L} < 1$ we have a fuzzy RD. Similarly to table 1, individuals can then be classified into types according to their individual return to treatment β and their response γ to crossing the threshold v_0:

Under independence, the distribution of types around the threshold v_0 would be the same as in figure 1. Thus, the probability of observing each of the four types, the ATE and all the LATEs can be computed as before. Importantly, the RD estimate of β can again be expressed as

[16] Alternatively, provided β does not enter the selection equation, γ could still be heterogeneous as long as its support is either strictly positive or strictly negative.

Table 3: Counterfactual Choices - RD

$\gamma = \gamma_L$			
	$\beta \leq 0$	$0 < \beta \leq -\gamma_L$	$\beta > -\gamma_L$
$[Z \leq v_0]$	$D = 0$	$D = 1$	$D = 1$
$[Z > v_0]$	$D = 0$	$D = 0$	$D = 1$
type	NT	DF	AT
$\gamma = \gamma_H$			
	$\beta \leq -\gamma_H$	$-\gamma_H < \beta \leq 0$	$\beta > 0$
$[Z \leq v_0]$	$D = 0$	$D = 0$	$D = 1$
$[Z > v_0]$	$D = 0$	$D = 1$	$D = 1$
type	NT	CM	AT

$$\beta^{RD} = \lambda \times LATE_{CM} + (1 - \lambda) \times LATE_{DF}$$

where

$$\lambda = \frac{p_{CM}}{p_{CM} - p_{DF}}$$

The only difference is that one relies on observations at the limit. The insights of sections 3.1-3.3 extend to the RD in a straightforward way.

3.5 Discussion

How informative is β^{IV} (or β^{RD}) when our economic world does not rule out essential heterogeneity and heterogeneous responses to the instrument (or forcing variable)? A extended Roy model suggests that the interpretation of the IV (fuzzy RD) estimate can be very sensitive to both essential heterogeneity and violations of the monotonicity condition. There are extreme cases when monotonicity is not an issue. These extreme cases are given by either the absence of heterogeneity in the treatment effect, the absence of sorting on gain, or the absence of heterogeneity in the response to the instrument. However, once we depart from these extreme cases, in the sense that *none* of them applies, then β^{IV} (or β^{RD}) can quickly become uninformative. The deviation from any LATE of interest depends on a variety of parameters, making it impossible to evaluate its size unless one is willing to make assumptions on all these parameters.[17]

4 Testing for Monotonicity - Stochastic Dominance

In this section we review a test of instrument validity. The strategies proposed in the literature are not tests of monotonicity alone, but tests of both independence and monotonicity. At the same time, the null hypothesis in any of these tests is a necessary but not sufficient condition for independence and monotonicity to hold. Thus, rejecting the null hypothesis is a clear indication that at least one of the two identification assumptions is unwarranted. Failing to reject the null hypothesis is a positive indication but not definite

[17]For the sake of brievity, we do not show results regarding sensitivity to a different $\overline{\beta}$ (ATE). Nevertheless, the same intuitions apply and results are available upon request.

evidence. This is an area of active research. Below we discuss a test for multivariate treatment settings which we later apply in section 5. We refer the readers to the work of Kitagawa (2015) and Huber and Mellace (2015) for tests when treatment is binary.

Angrist and Imbens (1995) point out that the independence and monotonicity assumption have a testable implication whenever the treatment takes more than two values: stochastic dominance (SD). Let treatment D be multivalued with support $\mathcal{D}$ and let F_j be the observed treatment CDF conditional on the j^{th} value of the instrument. Suppose $\{D(z)\}_{z\in\mathcal{Z}}$ is jointly independent of Z (as embedded in IV2) and for any two points of support $z, w \in \mathcal{Z}$, $D_i(z) \geq D_i(w)\ \forall i$ (monotonicity).

By *monotonicity* $P[D_i(z) \geq k] \geq P[D_i(w) \geq k]$ for every individual i and for every value k. Then $P[D(z) \geq k] \geq P[D(w) \geq k]\ \ \forall k$.

By *independence* $P[D(z) \geq k|Z = z] \geq P[D(w) \geq k|Z = w]\ \ \forall k$. Finally, by definition of $D(z)$ as the treatment value when $Z_i = z$, we have the SD result for the observed treatment outcomes:

$$P[D \geq k|Z = z] \geq P[D \geq k|Z = w]\ \ \forall k$$

or equivalently

$$F_w(d) \geq F_z(d)\ \forall d \in \mathcal{D}$$

Note that stochastic dominance does not exploit the full independence condition (IV2) but only require the weaker *type independence*:

IV5. $\{D(z)\}_{z\in\mathcal{Z}}$ is jointly independent of Z (type independence)

Stochastic dominance can be tested formally. Barrett and Donald (2003) show how to analytically compute the p-value when the two CDFs are derived from independent random samples. Let N_z and N_w be the number of observations under two alternative values of the instrument. Now let the null and alternative hypothesis be $H_0:\ F_w(d) \geq F_z(d)$ for all d and $H_1:\ F_w(d) < F_z(d)$ for some d. Thus we are testing the hypothesis that the CDF for the treatment values observed under $Z = z$ stochastically dominates the CDF for the treatment values observed under $Z = w$.[18] The test statistic for first-order stochastic dominance is given by

$$\widehat{S} = \left(\frac{N_z \times N_w}{N_z + N_w}\right)^{1/2} \sup_d \left(\widehat{F_z}(d) - \widehat{F_w}(d)\right)$$

Barrett and Donald (2003) show that one can compute a p-value by $\exp(-2(\widehat{S})^2)$.

5 Empirical Applications

In this section our goal is to maintain that monotonicity should be investigated like any other condition relying on a combination of economic insights and data analysis. We go

[18]The alternative hypothesis is simply the converse of the null and implies that there is at least some treatment value at which F_z is strictly larger than F_w. In other words stochastic dominance fails at some point. As formulated, one can in principle distinguish between the case where F_z and F_w coincide and the case where F_w dominates F_z by reversing the roles they play in the hypotheses and redoing the tests.

through two different studies in various settings that adopt either the IV or fuzzy RD estimator. For each study we try to make a case for or against monotonicity.

We do not scrutinize the independence condition since it is extensively discussed in these papers, and in applied work more generally. Similarly, we discuss but do not test essential heterogeneity. This would lengthen the discussion considerably and it is not our focus. Yet, at least on intuitive grounds we cannot see how essential heterogeneity could be ruled out a priori in any of these studies. Heckman, Urzua, and Vytlacil (2006) and Heckman, Schmierer, and Urzua (2010) provide a thorough discussion of essential heterogeneity and describe a way to test it.[19]

5.1 IV using School Entry Age regulation

Black, Devereux, and Salvanes (2011), henceforth BDS11, investigate the effect of school entry age ($D = EA$) on military IQ test scores and adult outcomes (Y) in Norway. To deal with endogeneity, they use Legal Entry Age as an instrument ($Z = LEA$). LEA is the age at which a child could start school given his/her birth date and given the country or state-specific school entry cutoff date. In Norway, school starts towards the end of August and children are expected to enter school in the calendar year they turn 7 (implying a January 1st cutoff date). BDS11 use two types of approaches: one including all months of birth, and one relying on a "discontinuity sample" which includes children born one month on either side of the cutoff date (December-January). Note that they only observe month of birth, so they cannot narrow the sample any further around the cutoff date, neither include a trend in date of birth. The approach using the discontinuity sample is implemented to account for either potential manipulation of the date of birth by parents and/or the seasonality of births.[20] In this section we focus on this discontinuity sample. Figure 6 plots the LEA by month of birth: the LEA is fully determined by the date of birth.[21] In this context, monotonicity clearly holds in the extreme case where all children start school on-time ($EA = LEA$): all December-born children would enter school 11 months older had they been born in January, and vice versa.

Estimating the causal effect of school entry age is problematic because of the endogeneity due to both sorting on levels and sorting on gain. For instance, children who are less intellectually and/or emotionally mature are more likely to face delayed school entry (a practice also known as "red-shirting"). Moreover, parents might make school entry decisions based on the (partial) knowledge of the gain from starting school later. For example, this gain could depend on the intellectual and/or emotional maturity of the child, or on the relative age of her classmates. This is plausibly a context with essential heterogeneity: (i) the gain from treatment is heterogeneous across the population and

[19] It is normal to imagine settings where at least some individuals take into account their expected return when deciding to get treatment, whether this is going to college, joining a union, buying health insurance, etc. Possible exceptions occur if the decision maker is different from the person being treated and this decision maker is not altruistic, or when the expected and actual return are uncorrelated.

[20] For instance Buckles and Hungerman (2013) show that in the US season of birth is not random but is associated with maternal characteristics: winter births are disproportionally realized by teenagers and the unmarried. If date of birth is not random, instruments relying on it are likely to violate the independence assumption.

[21] In BDS11 LEA is defined as $7.7 - \frac{(\text{month of birth}-1)}{12}$. In constructing the figures we assume children are born on the first day of the month and that school starts September 1st.

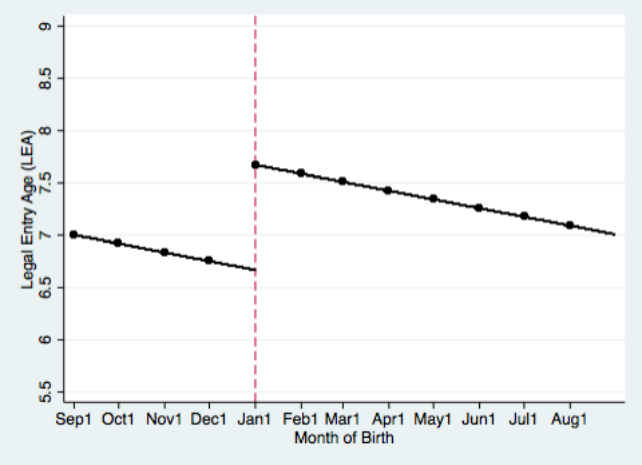


Figure 6: LEA by month of birth

(ii) there is some degree of sorting into treatment based on the gain from treatment.

A variety of studies use the same intuition to estimate the causal effect of school entry age for different countries and/or outcomes: Bedard and Dhuey (2006), Datar (2006), Puhani and Weber (2007), McEwan and Shapiro (2008), Elder and Lubotsky (2009), Muhlenweg and Puhani (2010), Muhlenweg, Blomeyer, Stichnoth, and Laucht (2012) and Fredriksson and Öckert (2014). None of these studies investigate the monotonicity assumption. The idea of using LEA as an instrument for school entry age is also very similar to the Angrist and Krueger (1991) idea of using quarter of birth as an instrument for schooling. In both cases the date of birth provides the variation in the instrument.

5.1.1 Discussing the Plausibility of the Monotonicity Assumption

Although BDS11 do not discuss monotonicity, they provide useful information. The table in figure 7a is taken from their paper and shows the proportion of children who enter school on-time, before and after the expected school entry age. Throughout the year, a very large fraction of children start school in the year they turn 7 (On Time). However, about 15% of December-borns are red-shirted (Late), while 10% of January-borns start school before the year they turn 7 (Early). Overall, the youngest children in an eligible school entry cohort (Oct-Dec borns) are the most likely to be redshirted, while the oldest ones (Jan-Feb borns) are the most likely to start school early.

This entry age behaviour is consistent with parents/educators making school-entry decisions based on either a child's absolute or relative age. Figure 7b replicates figure 6 but we now add the observed EA patterns as shown in the table. The size of the circles mirrors the proportions by month of birth. The largest circles are found along the LEA line, reflecting the high on-time entry rates. The smaller circles on the dotted line reflect early school entry ($EA < LEA$), which occurs mainly among children born in January-February. Instead, the smaller circles on the dashed lines reflect delayed school entry ($EA > LEA$), which is most common among October-December borns.

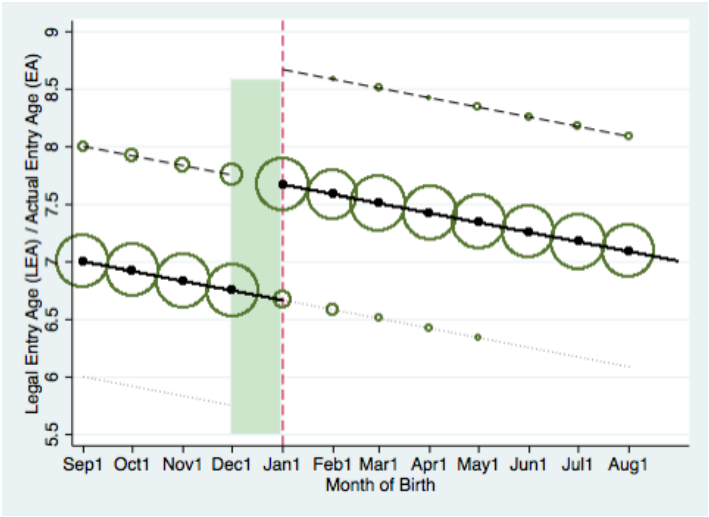
We now focus on December and January born children and consider counterfactuals to discuss monotonicity. Since all children are assumed to be born on a given day in either month, it is possible to distinguish 9 types based on actual and counterfactual EA

TABLE 1.—COMPLIANCE RATES BY MONTH OF BIRTH

	Early	On Time	Late
January	.10	.90	0.0
February	.04	.96	.01
March	.02	.97	.01
April	.01	.98	.01
May	.01	.98	.01
June	0.0	.98	.01
July	0.0	.98	.01
August	0.0	.98	.02
September	0.0	.97	.02
October	0.0	.96	.04
November	0.0	.93	.07
December	0.0	.85	.15

Each number in the "Early" column refers to the percentage of children in each birth month who started school before the year they turned 7 years old. Each number in the "On Time" column refers to the percentage of children in each birth month who started school the year they turned 7 years old. Each number in the "Late" column refers to the percentage of children in each birth month who started school after the year they turned 7 years old.

(a) Black, Devereux, and Salvanes (2011) page 458.



(b) LEA and observed school Entry Age

Figure 7: Monotonicity in Black, Devereux, and Salvanes (2011)

behaviour (see table 4). The sign in each cell indicates the change in EA if a child is born in January rather than December. For instance, type E represents December-born on-time school entrants who would also enter school on-time had they been born in January. This implies an increase in EA for members of type E: $EA_i(Dec) < EA_i(Jan),\ \forall\ type_i = t_E$. The (+) sign reflects the associated increase in EA. Assuming type independence, observed behaviour of January-born children can function as a counterfactual for actual December-borns, and vice versa. Figure 7b thus suggest that type E is the most prevalent type.

Table 4: Monotonicity

	January Born		
December born	Early	On Time	Late
Early	$t_A(+)$	$t_B(+)$	$t_C(+)$
On Time	$\mathbf{t_D}$(-)	$\mathbf{t_E}(+)$	$t_F(+)$
Late	t_G(-)	$\mathbf{t_H}$(-)	$t_I(+)$

The (+) term indicates that children enter school older when born in January. Viceversa, the (-) term indicates that children enter school younger when born in January.

Since late school entry among January-borns does not occur, we can ignore types C, F and I. Similarly, since early school entry among December-borns does not occur, we can ignore types A, B and C. The type independence assumption (IV5) ensures non-existence of any of these types. Finally, we find it unlikely that December-born late entrants would instead enter school early had they been born in January. Relying on this judgement regarding plausible behaviour, we ignore type G. Any other December-born on-time entrants would enter school early had they been born in January (type D). Crossing the cutoff date implies a drop in EA for this type: $EA_i(Dec) > EA_i(Jan),\ \forall\ type_i = t_D$. Since 10% of January-born children enter school early, existence of this type cannot be ruled out. Similarly, type H represent December-borns entering school late, but who would enter school on-time if they had been born in January. This again implies a drop in EA: $EA_i(Dec) > EA_i(Jan),\ \forall\ type_i = t_H$. Figures 7a and 7b again suggest that the existence of this type cannot be ruled out. Crucially, the existence of either of the $(-)$ types D or H alongside the most numerous type E (+) creates a violation of monotonicity.[22]

[22] We rule out the existence of type G individuals for simplicity, but including them in the discussion does not change any of the intuition provided. Nevertheless, ruling out the existence of type G individuals allows us to go one step further and fully determine the proportion of each type starting from the observed marginal distributions of EA for December and January borns. This is illustrated in the table below. Overall it would imply that for 25% of children entry age drops when born in January rather than December (types D and H), while for 75% entry age rises when crossing this threshold in date of birth (type E). Unfortunately, knowing these proportions is not sufficient to recover any LATE of interest.

	January Born			
December born	Early	On Time	Late	
Early	0	0	0	0
On Time	.10	.75	0	.85
Late	0	.15	0	.15
	.10	.90	0	1

To complement our discussion we apply the SD test. From table 7a and figure 7b we can derive the school entry age under each value of the instrument. Thus for children born December 1st $EA_i \in \{5.75, 6.75, 7.75\}$ depending on whether they enter early, on-time or late respectively. Similarly, for children born January 1st $EA_i \in \{6.67, 7.67, 8.67\}$. We can then use the proportions in table 7a to draw the CDFs. Note that none of the December born children enter school early ($EA = 5.75$) and none of the January born children enter late ($EA = 8.67$). Hence there are only four points of support in constructing the CDFs.

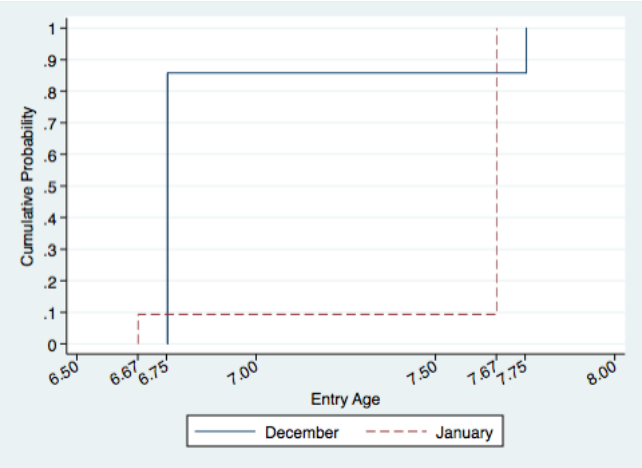


Figure 8: Stochastic Dominance under alternative values of the instrument

Figure 8 shows that the CDFs cross. We can test whether the crossing is statistically significant by applying the Barrett and Donald (2003) procedure. Let N_D and N_J be the number of the December and January born children respectively. Similarly let $F_D(x)$ and $F_J(x)$ be the CDFs of EA by month of birth (or equivalently, LEA). Finally let the null and alternative hypothesis be $H_0: \ F_D(x) \geq F_J(x)$ for all x and $H_1: \ F_D(x) < F_J(x)$ for some x. Thus we are testing the hypothesis that the CDF for the January born children stochastically dominates the CDF for the December born children. The test statistic for first-order stochastic dominance is given by

$$\widehat{S} = \left(\frac{N_D \times N_J}{N_D + N_J}\right)^{1/2} \sup_x \left(F_J(x) - F_D(x)\right)$$

BDS11 have a sample of N=104,023 children born in December and January. Setting the $\sup_x \left(F_J(x) - F_D(x)\right) = 0.15$ and assuming that $N_D = N_J = N/2$ leads to a test statistic of 24.189448 compared to a critical value of 1.5174 at a 1% level of significance. The p-value is zero. Thus we reject stochastic dominance.[23] We can also invert the null and alternative hypothesis to $H_0: \ F_J(x) \geq F_D(x)$ for all x and $H_1: \ F_J(x) < F_D(x)$ for some x but this obviously leads to an even larger test statistic of 120.94724. Thus the test indicates that either independence or monotonicity (or both) are violated. Together

[23]We do not know exactly how many children were born in December versus January. However, the conclusion is not sensitive to a (reasonable) imbalance in N_D and N_J. For example, if three quarters of children were born in December (N_D=78,017) and the rest in January (N_J=26,006), the test statistic is still much larger than the critical value ($\widehat{S}$=13.6066).

with the earlier discussion, this evidence raises concerns about the internal validity of the identification approach.

5.1.2 Interpreting β^{IV} without monotonicity

The data provide evidence that monotonicity does not hold. BDS11 write *"The high compliance rates are reassuring as they imply that our IV estimates can be interpreted as an approximation to the average treatment effect of school starting age rather than the usual local average treatment effect (LATE) interpretation."* This is because a change in the legal entry age instrument has an impact on the entry age of *every* child irrespective of whether they enter school On Time, Early or Late. In this section we investigate to what degree a violation of monotonicity invalidates this ATE interpretation. We do so by constructing a Roy model similar to the one used in section 3.1 but in the school entry age setting. Using the model we provide an interpretation of β^{IV}.

As a starting point, it is useful to spell out what the IV estimator measures. The treatment EA is a multivalued random variable measuring age at school entry (in monthly steps), with $EA \in \{6.67, 6.67 + \frac{1}{12}, 6.67 + \frac{2}{12}..., 7.75\}$. Given that we focus on the discontinuity sample, the instrument is binary, with two values of the legal entry age depending on whether a child is born in December or January. Therefore, we apply equation 7 from appendix B. Let $Y(k)$ be the military IQ test score of an individual who started school at age $EA = k$. $Y(k) - Y(k - \frac{1}{12})$ is then the gain in test score associated with starting school a month older. Hence,

$$\begin{aligned}\beta_1^{IV}(Jan, Dec) =& \frac{1}{\Omega} \times \sum_{s=1}^{11} \Big\{ \\ & E\left[Y\left(6.67 + \frac{s}{12}\right) - Y\left(6.67 + \frac{s-1}{12}\right) \Big| EA(Jan) \geq 6.67 + \frac{s}{12} > EA(Dec)\right] \\ & \times P\left[EA(Jan) \geq 6.67 + \frac{s}{12} > EA(Dec)\right] - \\ & E\left[Y\left(6.67 + \frac{s}{12}\right) - Y\left(6.67 + \frac{s-1}{12}\right) \Big| EA(Dec) \geq 6.67 + \frac{s}{12} > EA(Jan)\right] \\ & \times P\left[EA(Dec) \geq 6.67 + \frac{s}{12} > EA(Jan)\right] \Big\}\end{aligned} \tag{5}$$

where

$$\Omega = \sum_{s=1}^{11} \left(P\left[EA(Jan) \geq 6.67 + \frac{s}{12} > EA(Dec)\right] - P\left[EA(Dec) \geq 6.67 + \frac{s}{12} > EA(Jan)\right]\right)$$

To evaluate the effect of violating monotonicity we need to identify how children's entry age and outcomes are differently affected by respectively the instrument and treatment. This is generally impossible because counterfactuals are unobserved. However, in our earlier discussion we show that, under the very mild assumption that no December-born late entrants would instead enter school early had they been born in January, we can distinguish between three types of children:

$t_E(+)$ These children enter school On Time irrespective of their date of birth. Thus, going from a December to a January birth corresponds to a +11 months change in entry age: $EA(Dec) = 6.75 \rightarrow EA(Jan) = 7.67$. They form 75% of the discontinuity sample: $p_{t_E} = 0.75$.

$t_D(-)$ These children enter On Time if born in December but enter Early if born in January. Thus, going from a December to a January birth corresponds to a -1 month change in entry age: $EA(Dec) = 6.75 \rightarrow EA(Jan) = 6.67$. They form 10% of the discontinuity sample: $p_{t_D} = 0.1$.

$t_H(-)$ These children enter Late if born in December but enter On Time if born in January. Thus, going from a December to a January birth also corresponds to a -1 month change in entry age: $EA(Dec) = 7.75 \rightarrow EA(Jac) = 7.67$. They form 15% of the discontinuity sample: $p_{t_H} = 0.15$.

Using this information, we can rewrite (5) as

$$\begin{aligned}\beta_1^{IV}(Jan, Dec) =& \frac{1}{\Omega} \times \left\{ \sum_{s=1}^{11} \left\{ E\left[Y\left(6.75 + \frac{s}{12}\right) - Y\left(6.75 + \frac{s-1}{12}\right) \middle| t_e \right] \times p_{t_E} \right\} \right. \\ & - E\left[Y(6.75) - E(6.67) \middle| t_D \right] \times p_{t_D} \\ & \left. - E\left[Y(7.75) - E(7.67) \middle| t_H \right] \times p_{t_H} \right\}\end{aligned}$$

where

$$\Omega = (11 \times p_{t_E} - p_{t_D} - p_{t_H})$$

If we assume that the return to entering one month later is constant over age, at least within the observed points of support, then we can express (5) in terms of yearly rather than monthly return.[24] Let $\overline{\beta}_{t_i} = LATE_{t_i} = E[Y(EA = k+1\ year) - Y(EA = k)|t_i]$ for type $i = E, D, H$. Then

$$\beta_1^{IV}(Jan, Dec) = \frac{\frac{11}{12}\overline{\beta}_{t_E} p_{t_E} - \frac{1}{12}\overline{\beta}_{t_D} p_{t_D} - \frac{1}{12}\overline{\beta}_{t_H} p_{t_H}}{\frac{11}{12} p_{t_E} - \frac{1}{12} p_{t_D} - \frac{1}{12} p_{t_H}} \tag{6}$$

while $ATE = p_{t_E}\overline{\beta}_{t_E} + p_{t_D}\overline{\beta}_{t_D} + p_{t_H}\overline{\beta}_{t_H}$.

- If the monotonicity assumption was to hold, for instance because every child entered school On Time (t_E) irrespective of their month of birth then $p_{t_E} = 1$ while $p_{t_D} = p_{t_H} = 0$. Then $\beta^{IV} = \overline{\beta}_{t_E} = ATE$.
- If monotonicity does not hold but the β's are homogeneous or if there is no sorting on gain, such that the $E[\beta]$'s are homogeneous across types then $\beta^{IV} = ATE$, since

$$\beta^{IV}(Jan, Dec) = \frac{\overline{\beta} \times \left(\frac{11}{12} p_{t_E} - \frac{1}{12} p_{t_D} - \frac{1}{12} p_{t_H}\right)}{\frac{11}{12} p_{t_E} - \frac{1}{12} p_{t_D} - \frac{1}{12} p_{t_H}} = \overline{\beta}$$

[24] Constant monthly returns are also implied by the linear specification used in BDS11: $Y = b_0 + b_1 EA + e$ otherwise their IV approach breaks down. See Lochner and Moretti (2015) for a recent discussion of IV estimation with non-constant effects.

- If monotonicity does not hold, the β's are heterogeneous and there is sorting on gain then $\beta_1^{IV}(Jan, Dec)$ is different from the ATE.
- In BDS11 setting $\beta_1^{IV}(Jan, Dec)$ is close to $\overline{\beta}_{t_E}$ for two reasons:
  - Going from a December to a January birth corresponds to a +11 months change in entry age for type E children, as opposed to a -1 month change for type D and H children. Therefore, each type E child "counts" 11 times more than a type D or type H child.
  - Type E children are more numerous being 75% of the sample.

Without additional information about the different $E[\beta]$'s by type, it is impossible to be more precise about what β^{IV} measures, and how close it is to the ATE or to some LATE such as $\overline{\beta}_{t_E}$. Even if we constrain $f(\beta)$ to match the $\widehat{\beta}^{IV} = 0.167$ found in BDS11, equation (6) is still an equation in three unknowns with infinite combinations of $\overline{\beta}_{t_E}$, $\overline{\beta}_{t_H}$ and $\overline{\beta}_{t_D}$.[25] Nevertheless, we can use the entry age patterns observed in the data to construct a simple choice model and derive more restrictions on the distribution of the β's.

5.1.3 A simple choice model

Assume that:

EA1. β is normally distributed: $\beta \sim N(\mu_\beta, \sigma_\beta)$

EA2. There is sorting on gain. Children (and/or their parents) decide whether to enter school Early, On Time or Late based on their return.[26]

EA3. The return to entering school one month later is constant over the relevant age interval $[6.67, 7.75]$ (see footnote 24).

EA4. There is a cost from not entering school On Time. BDS11 explain that parents had to formally apply for an exception from the rule and the application had to be approved by health and school specialists as well as by the local government. This assumption is used to explain the high rate of On Time entrance. Since child care facilities were in short supply at the time, it is plausible that the overall cost of entering school Early is lower than the cost of entering Late. Thus we allow for a cost from entering early C_e that is lower than the cost of entering late C_ℓ: $C_e = \lambda C_\ell$ where we set $\lambda = 0.5$ This helps in matching the proportions of each type: p_{t_E}, p_{t_D} and p_{t_H}.

EA5. There is a strictly enforced social convention or law that no individual can enter school younger than six or older than eight. This assumption is used to explain the fact that no child is observed entering school outside of the 6-8 age interval.

[25] $\widehat{\beta}^{IV} = 0.167$ is derived from table 3, column (3) - 2SLS Discontinuity Sample - in BDS11, by summing the School starting age coefficient of -0.039 to the Age at test coefficient of 0.206.

[26] As pointed out earlier, without sorting on gain there is no interpretation issue.

Table 5: Entry age options and utility

Entry age		Supported by the data	Utility
December born			
Early	5.75	No	-
On Time	6.75	Yes	$Y(6.75)$
Late	7.75	Yes	$Y(7.75) - C_L$
Late+	8.75	No	-
January born			
Early+	5.67	No	-
Early	6.67	Yes	$Y(6.67) - C_E$
On Time	7.67	Yes	$Y(7.67)$
Late	8.67	No	-

EA6. No December-born late entrants would instead enter school early had they been born in January (Type G). This assumption is needed to identify the proportions of each type (see footnote 22).

EA7. There is no other factor driving the school entry decision other than those described in assumptions EA2, EA4 and EA5. Moreover, the utility is linear and separable in the test scores Y and the costs of the school entry decision C: $U(EA) = Y(EA) - C(EA)$ with β being the marginal benefit of waiting one more year before entering school. This assumption is used to keep the model simple but seems fairly realistic since the main benefits and costs are accounted for.[27]

The school entry age decision can be modelled as an optimal stopping problem: from the year a child is eligible, parents decide whether to exit the child care system (formal or informal) and enter the school system or whether to wait 1 more year (school intake occurs every 12 months). These children exit if the net utility of waiting one more year is (weakly) negative. In table 5 we describe the available choices and associated utilities.

Thus, December borns face a binary choice On Time vs Late. These children enter On Time if the return from entering school one year later is no larger than the cost from doing so:

$$Y(7.75) - C_\ell \leq Y(6.75) \Rightarrow \beta \leq C_\ell$$

January borns face a binary choice Early vs On Time. They enter Early if

$$Y(7.67) \leq Y(6.67) - C_e \Rightarrow \beta \leq -C_e$$

which implies that these children have a negative return from entering school one year later.[28] To identify the parameters of $f(\beta)$ we then exploit the following restrictions taken from the data:

[27] The alternative would be to use a full-fledged structural model, with heterogeneous costs and shocks of entering school at a given age, and including observable characteristics. However, since costs are generally unobserved, observed choices would not allow us to separately identify the individual β_i, C_i and shocks, only the net effect. To separate them we need to identify β_i in the first place.

[28] C_e could also be negative, implying a net benefit from having a child enter school ahead of time. Assumption EA6 rules out type G children. This assumption is violated if $C_\ell < \beta < -C_e$ for some child.

- $Pr(\beta \leq -C_e) = p_{t_D} = 0.1$
- $Pr(\beta > C_\ell) = p_{t_H} = 0.15$
- $\beta^{IV}(Jan, Dec) = \frac{\frac{11}{12}\overline{\beta}_{t_E} p_{t_E} - \frac{1}{12}\overline{\beta}_{t_D} p_{t_D} - \frac{1}{12}\overline{\beta}_{t_H} p_{t_H}}{\frac{11}{12} p_{t_E} - \frac{1}{12} p_{t_D} - \frac{1}{12} p_{t_H}} = \widehat{\beta}^{IV} = 0.167$

Based on these restrictions, we can identify C_ℓ, C_e, μ_β and σ_β. Intuitively, given σ_β, the cost parameters C_ℓ, C_e and μ_β are set to match the proportions of each type (the first two restrictions), with the LATEs determined accordingly. Then σ_β is set to ensure the $\beta^{IV}(Jan, Dec) = 0.167$ restriction is satisfied.[29]

The solution to this problem is shown in table 6 and in figure 9. Since there is a cost from not entering school On Time, children who do so must have either very negative returns (t_D) or very positive returns (t_H) from entering school one year later. The average return to postponing school entry with one year is 0.20 and the standard deviation is 0.395. Note that in the data used by BDS11, the military IQ test score is reported in stanine (Standard Nine) units, a method of standardizing raw scores into a nine point standard scale that has a discrete approximation to a normal distribution, a mean of 5, and a standard deviation of 2. The values of σ_β and μ_β reported in table 6a are thus very plausible in this setting. We can now establish how informative the β^{IV} estimate is. The difference $(\beta^{IV} - ATE)$ is about 8.6% of a standard deviation or about 17% of the ATE. The difference $(\beta^{IV} - \overline{\beta}_{t_E})$ is about 0.1% of a standard deviation, or 1.9% of the ATE. Hence β^{IV} is fairly close to a LATE of interest in spite of the monotonicity violation, albeit somewhat more distant from the ATE.

Table 6: Model coefficients and LATEs

(a) Model coefficients

C_e	C_ℓ	$\mu_\beta(ATE)$	σ_β	λ
-0.30548	0.61096	0.201195	0.395361	0.5

(b) LATEs

β^{IV}	$\overline{\beta}_{t_D} = \frac{\int_{-\infty}^{-C_e} bf(b)db}{p_{t_D}}$	$\overline{\beta}_{t_E} = \frac{\int_{-C_e}^{C_\ell} bf(b)db}{p_{t_E}}$	$\overline{\beta}_{t_H} = \frac{\int_{C_\ell}^{+\infty} bf(b)db}{p_{t_H}}$
0.167	-0.492656	0.170799	0.81574

(c) β^{IV} vs Parameters of Interest

$\beta^{IV} - ATE$	$\frac{\beta^{IV} - ATE}{\sigma_\beta}$	$\frac{\beta^{IV} - ATE}{\mu_\beta}$	$\beta^{IV} - \overline{\beta}_{t_E}$	$\frac{\beta^{IV} - \overline{\beta}_{t_E}}{\sigma_\beta}$	$\frac{\beta^{IV} - \overline{\beta}_{t_E}}{\mu_\beta}$
-0.034195	-0.0864907	-0.169959	-0.00379944	-0.00961007	0.0188844

A negative C_e would make this possible. However, for children to enter On Time irrespective of their month of birth (type E) we need $-C_e < \beta < C_\ell$ instead. Thus the existence of type G children rules out the existence of type E ones and viceversa. Finally, note that assumption EA4 only require the returns to be constant in the $[6.67, 7.75]$ interval.

[29] There is a unique solution to this problem. Satisfying the proportions results in the LATE of each type, ATE and the $\beta^{IV}(Jan, Dec)$ being linear and strictly increasing in σ_β. Hence satisfying $\beta^{IV}(Jan, Dec) = 0.167$ ensures there is a unique solution that pins down all the parameters.

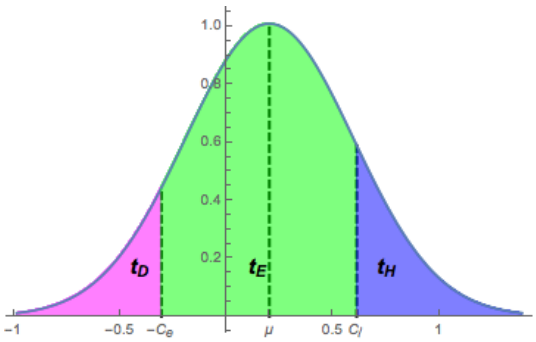


Figure 9: Types over the β distribution

IV vs fuzzy RD Suppose BDS11 had information on the date of birth. With enough observations we could create a stricter discontinuity sample with children born a day apart around the cutoff: 31 December - 1 January.[30] The resulting β^{IV} would be

$$\beta^{IV}(Jan\ 1st, Dec\ 31st) = \frac{\frac{364}{365}\overline{\beta}_{t_E}p_{t_E} - \frac{1}{365}\overline{\beta}_{t_D}p_{t_D} - \frac{1}{365}\overline{\beta}_{t_H}p_{t_H}}{\frac{364}{365}p_{t_E} - \frac{1}{365}p_{t_D} - \frac{1}{365}p_{t_H}} \approx \overline{\beta}_{t_E}$$

This is because each type E child now counts 364 times more than a type D or type H child. Even if the proportion of type E children becomes smaller as we approach the cutoff, it is unlikely to change enough to reverse $\beta^{IV} \approx \overline{\beta}_{t_E}$. Monotonicity is still violated because of the type D and H children, but the cost of violating monotonicity is small because these types do not carry much weight. An RD approach takes this to the extreme, by identifying β *exactly* at the threshold. In this setting, this is an important advantage of using a genuine RD approach as opposed to an IV approach. Alternatively, one could include a trend in date of birth which, if correctly specified, would allow to capture the effect right at the discontinuity while also using observations further from the cutoff date. In fact BDS11 estimate the effect of school entry age by also running a 2SLS procedure using all months of birth. This is numerically equivalent to a fuzzy RD:

$$\begin{aligned} Y =& b_0 + b_1 EA + b_2 X + e \\ EA =& a_0 + a_1 LEA + a_2 X + v \end{aligned}$$

where X includes month of birth (ranging between 1-12). However, we believe that even this alternative specification is problematic because of the linear trend. Since children are more likely to enter late (early) the closer they are born to the left (right) of the discontinuity the trend is not linear over the different months. Imposing a linear trend is a misspecification of the true process, and it will bias the estimate of the first stage. An optimal fuzzy RD design would have data relying on date rather than month of birth, use a small bandwidth and include a trend using local linear regression that is allowed to be different on each side of the threshold. That is possibly a robust solution in the school entry age setting.[31]

[30] Their discountinuity sample with children born December-January has 104,023 observations. Assuming births are equally likely on any given day, a sample 31 December - 1 January has 3,467 observations.

[31] Gelman and Imbens (2014) and Imbens and Kalyanaraman (2012) argue against the use of high-order polynomials in the Regression Discontinuity design.

Barua and Lang (2009) We are not the first ones to discuss the monotonicity condition in the school entry age setting. argue that studies using legal entry age as an instrument *"may be severely biased because they violate the monotonicity assumption needed for LATE."* There are a number of differences with their work. First, based on the evidence that several US states have increased the minimum school entry age by shifting the entry cut-off, Barua and Lang (2009) aim to identify the effect of such policy change on children outcomes. To this extent they propose an alternative definition of treatment and instrument. Our focus is to understand whether using the legal entry age instrument so widely adopted in the literature is any helpful in identifying the ATE and/or any LATE. Second, they discuss monotonicity using US census date for cohorts born in the 1950s, for which they only observe quarter of birth. This data has not been used in any of the school entry age studies. We look at data actually used in a very recent study and with individuals born only a month before/after the cutoff. Third, they argue that monotonicity is violated by simply showing that stochastic dominance does not hold (while this could also be due to a violation of the independence assumption). We go further by emphasizing the role of sorting on gain on school entry decisions, how this leads to the different *types* of individuals and consequently to the violation of the monotonicity assumption. Overall, our findings are less negative than theirs since we find that β^{IV} is fairly close to a LATE of interest in spite of the monotonicity violation, albeit not close to the ATE.

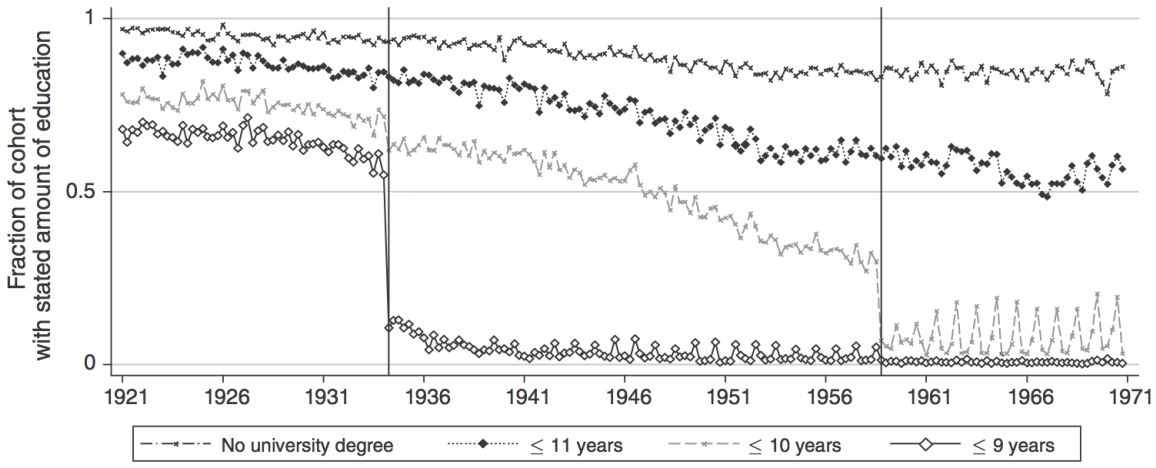
5.2 Fuzzy RD using changes in minimum school leaving age

Clark and Royer (2013), hereafter CR13, investigate the effect of years of schooling (D) on health (Y). To deal with endogeneity, they use two changes to British compulsory schooling laws that generated differences in educational attainment across cohorts (Z). The first reform raised the minimum school leaving age from 14 to 15: it was implemented on 1 April 1947 and therefore it affected individuals born from April 1933 onwards. The second reform raised the minimum school leaving age from 15 to 16: it was implemented on 1 September 1972 and it therefore affected individuals born from September 1957 onwards. Figure 10 is extracted from their paper and shows the impact of the schooling laws on different cohorts by quarter of birth. Both reforms had very large impacts. The first reform affected about 50% of the population while the second reform affected about 25% of the population.

To identify the treatment effect, CR13 use a fuzzy RD approach where the discontinuities are given by the 1 April 1933 and 1 September 1957 cutoffs in date of birth. In the estimation a local linear regression is adopted, selecting individuals born within a 43 to 105 months bandwidth depending on the reform and gender group. They also include trends in month of birth (see equation (1) in their paper). CR13 find no effect of education on health, a result that stands in sharp contrast to previous estimates of the effects of education on health.

The same British compulsory schooling reforms have been extensively used as an instrumental variable in various contexts such as earnings and labor activity (Harmon and Walker (1995), Oreopoulos (2006), Devereux and Hart (2010)); citizenship and political involvement (Milligan, Moretti, and Oreopoulos (2004)); health of offspring (Lindeboom, Llena-Nozal, and van der Klaauw (2009)); fertility and teenage childbearing (Silles

Figure 10: Clark and Royer (2013) page 2092



(2011)). None of these papers investigate monotonicity. Similar reforms have also been used in other countries to estimate the returns to education for a variety of outcomes.

5.2.1 Discussing the Plausibility of the Monotonicity Assumption

In this context, monotonicity holds in the case where the schooling reforms induced all individuals to get more years of schooling, or at least not less of it. CR13 do not discuss monotonicity but again we can use the information provided in the paper to scrutinise the assumption. The table in figure 11 is extracted from their paper and it shows the estimated effect of the compulsory schooling changes. The first column reports the effect on the years of schooling: the positive coefficients clearly show that both reforms increased the average years of schooling. The following columns report the effect by cumulate years of schooling: these columns show that the 1947 reform increased the proportion of individuals staying in education beyond 9 and 10 years of schooling but the same reform actually decreased the proportion of individuals staying in education beyond 11, 12 and 13 years of schooling. This latter result is mostly true for men as shown in the rectangular selection in the table.

We can use the results in their table to consider counterfactuals. There are several possible types of individuals based on actual and counterfactual behaviour which we summarize in table 7. The sign in each cell indicates the change in years of schooling if an individual is born before or after April 1933 (1947 reform). The reform certainly increased the average years of schooling, so for monotonicity to hold what is required is that no one belongs to a cell below the main diagonal (=), because that would imply a decrease in schooling due to the reform (-).

It is plausible that no individual who would be attaining 9 or fewer years of schooling before the 1947 reform would attain less education after a reform that made it illegal (even if the data show that not everyone obeyed the law). What we cannot exclude though, is that someone who would attain 11 or more years of education before the reform would attain fewer years after, as suggested by the significantly positive coefficients in figure

	Years of education	≤ 9 Years	≤ 10 Years	≤ 11 Years	≤ 12 Years	≤ 13 Years
Panel A. Impact of 1947 change						
	All (bandwidth = 46 months, N = 31,345)					
Estimate	0.450	−0.445	−0.040	0.009	0.011	0.015
	(0.035)	(0.009)	(0.009)	(0.008)	(0.008)	(0.007)
Outcome mean	15.11	0.58	0.70	0.83	0.88	0.90
	Men (bandwidth = 105 months, N = 33,337)					
	0.443	−0.478	−0.019	0.021	0.019	0.014
	(0.035)	(0.011)	(0.010)	(0.008)	(0.007)	(0.006)
	15.14	0.57	0.70	0.82	0.86	0.89
	Women (bandwidth = 69 months, N = 24,613)					
Estimate	0.524	−0.472	−0.064	0.004	0.003	0.005
	(0.036)	(0.010)	(0.010)	(0.009)	(0.009)	(0.007)
Outcome mean	15.07	0.59	0.71	0.84	0.89	0.91

Figure 11: Clark and Royer (2013) page 2103

Table 7: Monotonicity - Years of Schooling

	Born After April 1933					
Born before April 1933	9	10	11	12	13	14+
9	=	+	+	+	+	+
10	-	=	+	+	+	+
11	-	-	=	+	+	+
12	-	-	-	=	+	+
13	-	-	-	-	=	+
14+	-	-	-	-	-	=

The (+) term indicates that individuals would attain more schooling if born after the schooling reform. Viceversa, the (-) term indicates that individuals would attain less schooling.

11.[32] Note that we can exploit the information in figure 11 to explicitly derive the CDFs for men born before and after April 1933. This is illustrated in figure 12: the CDFs clearly cross.

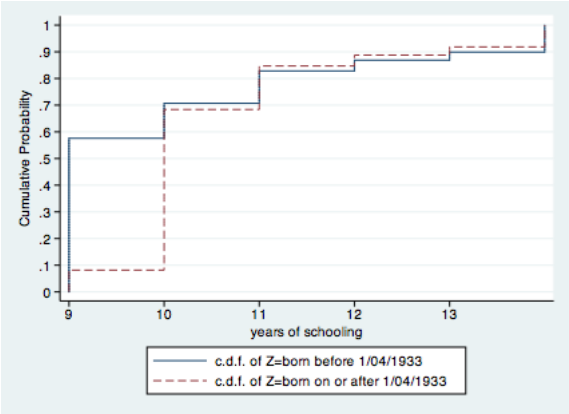


Figure 12: stochastic dominance

We can again test whether the crossing is statistically significant by applying Barrett and Donald (2003) procedure. Let N_B and N_A be the number of men born before and after April 1933. Similarly let $F_B(x)$ and $F_A(x)$ be the CDFs of years of schooling for the male cohorts before and after the reform. Finally let the null and alternative hypothesis be $H_0: \ F_B(x) \geq F_A(x)$ for all x and $H_1: \ F_B(x) < F_A(x)$ for some x. Thus we are testing the hypothesis that the CDF for the pre-reform cohorts stochastically dominates the CDF for the post-reform cohorts. The test statistic for first-order stochastic dominance is given by

$$\widehat{S} = \left(\frac{N_B \times N_A}{N_B + N_A} \right)^{1/2} \sup_x \left(F_A(x) - F_B(x) \right)$$

CR13 have a sample of N=33,337 men. Assuming that $N_B = N_A = N/2$ and setting $\sup_x \left(F_A(x) - F_B(x) \right) = 0.021$ leads to a test statistic of 1.91713 with a p-value of 0.00064. Thus we reject stochastic dominance.[33]

One might ask what kind of economic behaviour would explain this violation of the monotonicity condition. The fuzzy RD setting with trends implies that one needs to explain why a given individual born in March 1933 would obtain 12 years of schooling, while that same individual would reduce years of schooling if she was born in April 1933. Even if CR13 effectively use a 105 months bandwidth for men, the RD setting with trends still implies that all results have to be interpreted at the limit. That is the trend allows us to see the counterfactual choice for someone born arbitrarily far from the cut-off date had she been born in the March-April 1933 window. Importantly, the 1 April is not

[32] These coefficients reflect net flows: they could result from a large number of individuals taking less education after the reform that are not completely compensated by the flow of individuals taking more education after the reform.

[33] We do not know exactly how many men are on either side of the April 1933 threshold. However, the conclusion is not sensitive to a (reasonable) imbalance in N_B and N_A. For example, if one third of men were born before the threshold (N_B=11,112) and the rest after (N_A=22,225), the test statistic is still larger than the critical value ($\widehat{S}$=1.8074916, p-value=0.00145).

the cut-off date of birth for starting school in a given year, implying that individuals born in March and April are probably sharing the same classrooms, school system (with the exception of the minimum school leaving age), peers and labour market conditions. Structural changes in the English education system are generally introduced on 1 April. It is the beginning of the financial year and also gives a few months to prepare for the start of the school year on 1 September.[34] Hence, we do not have a clear intuition as to what might be driving the monotonicity violation. For this reason, we do not attempt the Roy model analysis in this setting.

Finally, note that there would be no monotonicity violation if we were to redefine treatment as being binary, with $D = 0$ if the individual completed only 9 years of schooling and $D = 1$ if she completed 10 years or more. Obviously, that would also change the interpretation of the results.

6 Conclusion

Theoretical studies have emphasised that IV and fuzzy RD estimates can be interpreted as a LATE only if monotonicity is satisfied. This requirement applies in a context with essential heterogeneity, a situation with heterogeneous treatment effects and sorting into treatment based on the gain. Monotonicity is a restriction on the impact of the instrument (or discontinuity) on the treatment. It implies that a change in the value of the instrument (or crossing the discontinuity) affects the treatment of all individuals in the same direction. If monotonicity does not hold then the IV and fuzzy RD estimates are generally not interpretable. Surprisingly, very few of the applied studies that rely on IV and fuzzy RD designs discuss monotonicity. This is in stark contrast to the lengthy discussions dedicated to other conditions like the IV independence and rank conditions, and to the RD discontinuity (in the probability of treatment) and continuity (in the conditional regression function) conditions. Given the importance of the monotonicity condition, this is an important missing step in evaluating internal validity.

In this paper we first illustrate how informative the IV and fuzzy RD estimates are once monotonicity is violated, and do so under various degrees of treatment effect heterogeneity and sorting on gain. We find that, under essential heterogeneity, interpretability can be lost even for minor violations of monotonicity. This reinforces the importance of investigating the monotonicity assumption. We then investigate monotonicity in different applied studies using economic insights and data analysis. We show that monotonicity is debatable in some settings and the interpretation of the estimates needs to be adjusted accordingly. IV and fuzzy RD designs can be incredibly powerful in dealing with endo-

[34] We thank Sir Michael Barber for this insight. See also Barber (2000) for a discussion on the background, passage and effect of the 1944 Education act that raised the minimum school leaving age from 14 to 15, and laid the foundations for the subsequent increase to 16. Note that the Act also included other changes to the education system for secondary schools in England and Wales. Called the "Butler Act" after the Conservative politician R. A. Butler, it introduced the Tripartite System of secondary education and made all schooling, especially secondary education, free for all pupils. The new tripartite system consisted of three different types of secondary school: grammar schools, secondary technical schools and secondary modern schools. Age 11 was the decision point for sending children to higher levels. Still, it is no clear why any of these reforms would have a different impact on individuals being born a few days apart on either side of the 1 April 1933 cut-off.

geneity problems. A discussion of the monotonicity assumption is just an extra step that should be included to validate the results.

As the key role of the monotonicity condition gains recognition, recent work is trying to develop alternative tests. In addition, a number of recent papers have tried to establish alternative conditions under which it is still possible to interpret the IV and fuzzy RD estimates as a LATE even though monotonicity is violated (see Klein (2010), de Chaisemartin (2014) and Huber and Mellace (2012)). An extensive discussion of these literatures is beyond the scope of this paper but they are promising areas of current and future research.

Appendix A Interpreting the fuzzy RD estimate

Assume RD1-RD3 hold, and let $e > 0$ denote an arbitrary small number. Consider first the numerator of the Wald estimator in (3). The mean difference in outcomes for individuals above and below the discontinuity point is

$$
\begin{aligned}
&E[Y|Z = v_0 + e] - E[Y|Z = v_0 - e] = \\
&E[Y(0) + (Y(1) - Y(0))D|Z = v_0 + e] - E[Y(0) + (Y(1) - Y(0))D|Z = v_0 - e] = \\
&E[Y(0)|Z = v_0 + e] - E[Y(0)|Z = v_0 - e] + \\
&E[(Y(1) - Y(0))D|Z = v_0 + e] - E[(Y(1) - Y(0))D|Z = v_0 - e]
\end{aligned}
$$

By assumption RD2, the first two terms cancel out at the limits. By assumption RD3, the remaining terms can be written as

$$
\begin{aligned}
&E[(Y(1) - Y(0))D(v_0 + e)] - E[(Y(1) - Y(0))D(v_0 - e)] = \\
&E[(Y(1) - Y(0))(D(v_0 + e) - D(v_0 - e))] = \\
&1 \times E[Y(1) - Y(0)|D(v_0 + e) - D(v_0 - e) = 1] \times P[D(v_0 + e) - D(v_0 - e) = 1] - \\
&1 \times E[Y(1) - Y(0)|D(v_0 + e) - D(v_0 - e) = -1] \times P[D(v_0 + e) - D(v_0 - e) = -1]
\end{aligned}
$$

Then consider the denominator. The mean difference in treatment value for individuals above and below the discontinuity is

$$
\begin{aligned}
&E[D|Z = v_0 + e] - E[D|Z = v_0 - e] = \\
&E[D(v_0 + e)] - E[D(v_0 - e)] = \\
&E[D(v_0 + e) - D(v_0 - e)] = \\
&P[D(v_0 + e) - D(v_0 - e) = 1] - P[D(v_0 + e) - D(v_0 - e) = -1]
\end{aligned}
$$

The Wald estimator in (3) can then be expressed as follows:

$$
\begin{aligned}
&\frac{\lim_{e \downarrow 0} E[Y|v_0 + e] - \lim_{e \uparrow 0} E[Y|v_0 - e]}{\lim_{e \downarrow 0} E[D|v_0 + e] - \lim_{e \uparrow 0} E[D|v_0 - e]} = \\
&\lim_{e \to 0} \{\lambda \times E[Y(1) - Y(0)|D(v_0 + e) - D(v_0 - e) = 1] + \\
&(1 - \lambda) \times E[Y(1) - Y(0)|D(v_0 + e) - D(v_0 - e) = -1]\}
\end{aligned}
$$

where

$$
\lambda = \frac{P[D(v_0 + e) - D(v_0 - e) = 1]}{P[D(v_0 + e) - D(v_0 - e) = 1] - P[D(v_0 + e) - D(v_0 - e) = -1]}
$$

Appendix B Monotonicity when either treatment or instrument are multivalued

B.1 Monotonicity when the treatment is multi-valued

Angrist and Imbens (1995) discuss the interpretation of the IV estimate when the treatment D is a multivalued random variable with support $\mathcal{D} = \{0, 1, ..., K\}$ and $K > 1$. Let Y_k be the outcome of an individual under treatment value k. Assume that IV1 (rank) and IV2 (independence) hold. Extending the discussion by Angrist and Imbens (1995), we can express the IV estimate of β_1 for any two points of support z, w in $\mathcal{Z}$ as follows (proof in section B.1.1 below.)

$$\beta_1^{IV}(z,w) = \frac{1}{\Omega} \times \sum_{k=1}^{K} \Big\{ E[Y_k - Y_{k-1}|D(z) \geq k > D(w)] \times P[D(z) \geq k > D(w)] \\ - E[Y_k - Y_{k-1}|D(w) \geq k > D(z)] \times P[D(w) \geq k > D(z)] \Big\} \tag{7}$$

where

$$\Omega = \sum_{k=1}^{K} \left(P[D(z) \geq k > D(w)] - P[D(w) \geq k > D(z)] \right)$$

From equation (7) we can see that

- First, $\beta_1^{IV}(z,w)$ does not rely on individuals who are not affected by the instrument: $D(z) = D(w)$.
- Second, if assumption IV3 (monotonicity) holds, such that no one decreases treatment intensity: $P[D(w) \geq k > D(z)] = 0 \; \forall k$. Equation (7) then simplifies to

$$\beta_1^{IV}(z,w) = \sum_{k=1}^{K} E[Y_k - Y_{k-1}|D(z) \geq k > D(w)] \times \frac{P[D(z) \geq k > D(w)]}{\sum_{k=1}^{K} P[D(z) \geq k > D(w)]} \tag{8}$$

  Angrist and Imbens (1995) refer to this parameter as the average causal response (ACR). It is a weighted average of causal responses to a unit change in treatment, for those whose treatment status is affected by the instrument. A similar discussion applies if instead monotonicity holds such that $P[D(z) \geq k > D(w)] = 0 \; \forall k$.

- Third, if monotonicity does not hold, then for a given change in the value of Z some individuals increase treatment value $D(z) \geq k > D(w)$ while others decrease it $D(w) \geq k > D(z)$ for at least some k. Equation (7) shows that in this case both the numerator and the denominator include contributions from individuals affected in either direction. Without monotonicity, the IV estimate is thus not a weighted average of treatment effects and cannot be given a useful interpretation. It is an equation in many unknowns, which makes it impossible to back-out the LATE for any particular group.

The close analogy between the fuzzy Regression Discontinuity design and the IV estimators extends to the case of multivalued treatment with binary instrument in a straightforward way. Both can still be expressed as Wald estimators. Lee and Lemieux (2010) show that monotonicity is also required in the fuzzy RD setting with a multivalued treatment, in which case the interpretation of the RD estimator is still the same as that of the IV estimator.

B.1.1 Proof: Wald estimator under Multivalued Treatment

To write the IV estimate of β_1 in the case of a multivalued treatment and binary instrument, we take elements from the proof in the Appendix of Angrist and Imbens (1995) and generalize it by allowing violations of the monotonicity assumption. Let $\mathcal{D} = \{0, 1, \ldots, K\}$ be the treatment with $K > 1$, and let $\mathcal{Z} = \{w, z\}$ be the support of the (binary) instrument. Then let $\mathbb{1}(A)$ be an indicator function for the event A. Define the following indicators: $\psi_{Zk} = \mathbb{1}(D(Z) \geq k)$ for $Z = w, z$ and $k = 0, 1, \ldots, K+1$; $\chi = \mathbb{1}(Z = z)$ and thus $1 - \chi = \mathbb{1}(Z = w)$. Note that $\psi_{Z0} = 1$ and $\psi_{ZK+1} = 0$ for all Z. Let Y_{Dz} and Y_{Dw} be the counterfactual outcomes for an individual under treatment value $D(z)$ and $D(w)$ respectively. Then the outcome Y can be written as

$$\begin{aligned} Y &= \chi \times Y_{Dz} + (1 - \chi) \times Y_{Dw} \\ &= \left\{ \chi \times \sum_{k=0}^{K} Y_k(\psi_{zk} - \psi_{zk+1}) \right\} + \left\{ (1 - \chi) \times \sum_{k=0}^{K} Y_k(\psi_{wk} - \psi_{wk+1}) \right\} \end{aligned}$$

where Y_k is the outcome for an individual with $D = k$. Then, under the rank and independence conditions (IV1 and IV2), the numerator of the Wald estimator can be rewritten as

$$\begin{aligned} & E(Y|Z = z) - E(Y|Z = w) \\ &= E\left[\sum_{k=0}^{K} Y_k(\psi_{zk} - \psi_{zk+1} - \psi_{wk} + \psi_{wk+1}) \right] \\ &= E\left[\sum_{k=1}^{K} (Y_k - Y_{k-1})(\psi_{zk} - \psi_{wk}) \right] \\ &= \sum_{k=1}^{K} \{ 1 \times E\left[Y_k - Y_{k-1} | \psi_{zk} - \psi_{wk} = 1 \right] \times P[\psi_{zk} - \psi_{wk} = 1] \\ &\quad + 0 \times E\left[Y_k - Y_{k-1} | \psi_{zk} - \psi_{wk} = 0 \right] \times P[\psi_{zk} - \psi_{wk} = 0] \\ &\quad + (-1) \times E\left[Y_k - Y_{k-1} | \psi_{zk} - \psi_{wk} = -1 \right] \times P[\psi_{zk} - \psi_{wk} = -1] \} \\ &= \sum_{k=1}^{K} \{ E\left[Y_k - Y_{k-1} | D(z) \geq k > D(w) \right] \times P[D(z) \geq k > D(w)] \} \\ &- \sum_{k=1}^{K} \{ E\left[Y_k - Y_{k-1} | D(w) \geq k > D(z) \right] \times P[D(w) \geq k > D(z)] \} \end{aligned}$$

Similarly we can write the denominator of the Wald estimator as

$$
\begin{aligned}
&E(D|Z=z) - E(D|Z=w) \\
&= E\left[\sum_{k=0}^{K} k \times (\psi_{zk} - \psi_{zk+1} - \psi_{wk} + \psi_{wk+1})\right] \\
&= E\left[\sum_{k=1}^{K} (\psi_{zk} - \psi_{wk})\right] \\
&= \sum_{k=1}^{K} \{1 \times P[\psi_{zk} - \psi_{wk} = 1] + 0 \times P[\psi_{zk} - \psi_{wk} = 0] + (-1) \times P[\psi_{zk} - \psi_{wk} = -1]\} \\
&= \sum_{k=1}^{K} \{P[D(z) \geq k > D(w)] - P[D(w) \geq k > D(z)]\}
\end{aligned}
$$

Then, the Wald estimator is given by

$$
\begin{aligned}
\beta_1^{IV}(z,w) &= \frac{E(Y|Z=z) - E(Y|Z=w)}{E(D|Z=z) - E(D|Z=w)} \\
&= \frac{1}{\Omega} \sum_{k=1}^{K} \big\{ E\left[Y_k - Y_{k-1} | D(z) \geq k > D(w)\right] \times P[D(z) \geq k > D(w)] - \\
&\qquad E\left[Y_k - Y_{k-1} | D(w) \geq k > D(z)\right] \times P[D(w) \geq k > D(z)] \big\}
\end{aligned}
$$

where $\Omega = \sum_{k=1}^{K} \{P[D(z) \geq k > D(w)] - P[D(w) \geq k > D(z)]\}$

B.2 Monotonicity when the instrument is multi-valued

Let Z be a multivalued random variable with support $\mathcal{Z} = \{0, 1, ..., J\}$ and with $J > 1$. Assume that IV1 (rank) and IV2 (independence) hold. Define $g(Z)$ as a scalar function from the support of Z to the real space. Using $g(Z)$ as an instrument, the IV estimator is now given by[35]

$$\beta_1^{IV} \equiv \frac{Cov(Y, g(Z))}{Cov(D, g(Z))}$$

In order to interpret the IV estimator Imbens and Angrist (1994) and Angrist and Imbens (1995) supplement IV3 (monotonicity) with another condition

IV4. (i) either $\forall z, w \in \mathcal{Z}, E[D|Z=z] \leq E[D|Z=w]$ implies $g(z) \leq g(w)$; or, $\forall z, w \in \mathcal{Z}, E[D|Z=z] \leq E[D|Z=w]$ implies $g(z) \geq g(w)$ and
(ii) $Cov(D, g(Z)) \neq 0$

Note that while monotonicity is a condition across individuals, IV4 is a condition on the relation between $E[D|Z]$ and $g(Z)$: IV3 does not imply IV4 and viceversa.[36] Condition

[35] The case where $g(Z) = Z$ is the simplest case, but this notation generalizes the estimator to any functional form of $g(Z)$ and to the case where Z is a vector.

[36] Because IV3 implies that *every* individual has to respond to the instrument in the same direction, Heckman and Vytlacil (2005) and Heckman, Urzua, and Vytlacil (2006) rename IV3 with the term "uniformity".

IV4 is satisfied by construction when Z is binary, or when Z is a discrete random variable that enters $g(Z)$ in the form of mutually exclusive dummy variables. Otherwise IV4 is not guaranteed to hold.

Under IV1-IV4, let the points of support of Z be ordered such that $\ell < m$ implies $E[D|Z = \ell] < E[D|Z = m]$. Angrist and Imbens (1995) show that we can write the IV estimate of β_1 as a weighted average of Wald estimators:

$$\beta_1^{IV} = \sum_{j=1}^{J} \mu_j \beta_1^{IV}(j, j-1) \tag{9}$$

where

$$\beta_1^{IV}(j, j-1) = \frac{E[Y|Z = j] - E[Y|Z = j-1]}{E[D|Z = j] - E[D|Z = j-1]}$$

and

$$\mu_j = (E[D|Z = j] - E[D|Z = j-1]) \frac{\sum_{\ell=j}^{J} \pi_\ell (E[D|Z = \ell] - E[D])}{\sum_{\ell=0}^{J} \pi_\ell E[D|Z = \ell](E[D|Z = \ell] - E[D])}$$

with $\pi_\ell = P[Z = \ell]$. Given the ordering in Z we also have that $0 \leq \mu_j \leq 1$ and $\sum_{j=1}^{J} \mu_j = 1$. Equation (9) indicates that, under a binary treatment,

- when monotonicity IV3 and IV4 are satisfied then β_1^{IV} is a weighted average of LATEs $E[Y(1) - Y(0)|D(j) - D(j-1) = 1]$. This is because each Wald estimator has a LATE interpretation (given IV3) and the μ_j weights are non-negative (given IV4).

- when monotonicity IV3 is satisfied but IV4 is not then β_1^{IV} is not a weighted average of LATEs. While each Wald estimator has a LATE interpretation, some μ_j weights can be negative.

- when monotonicity IV3 is violated but IV4 is satisfied then β_1^{IV} is again not a weighted average of LATEs. Now some Wald estimators no longer have a LATE interpretation, even though all the μ_j weights are non-negative.

When the treatment is multi-valued the same conclusions apply with the only difference that each Wald estimator has an ACR interpretation (under monotonicity).

The analogy between the fuzzy Regression Discontinuity design and the IV estimators is less direct now. Let Z be the running variable in an RD setting, with $v \in \mathcal{Z}$. Consider now the case where there are multiple discontinuities or cutoffs (v_f) for $f = 1, \ldots, F$ and $F > 1$, such that the expected treatment value jumps at each threshold

$$\lim_{v \downarrow v_f} E[D|Z = v] \neq \lim_{v \uparrow v_f} E[D|Z = v] \qquad \forall f$$

One way to approach this problem is to split the sample and run a separate RD regression at each cutoff, doing so for both the treatment and the outcome equation. This approach yields a vector of F treatment effects, each interpretable as a LATE (or ACR) if monotonicity is satisfied. Note that this is not exactly how discrete multivalued

instruments are used. If Z has support $\mathcal{Z} = \{z_0, z_1, \ldots, z_J\}$ then the instrument often enters the $g(Z)$ function in the form of mutually exclusive dummy variables (similarly to RD) but the outcome equation is estimated over the whole sample (contrary to separate RD regressions). The result is a weighted average of LATEs (or ACRs) as described earlier. Of course one could take the same approach for the fuzzy RD by pooling all the observations together while using dummy variables to identify threshold-specific effects.

For instance, using a linear spline specification, the RD design can now be described by:

$$\begin{aligned} D =& \alpha_0 + \delta_0^D \times (Z - v_1) + \sum_{f=1}^{F} \Big\{ \alpha_{1f} \times \mathbb{1}[Z > v_f] + \delta_f^D \times \mathbb{1}[Z > v_f] \times (Z - v_f) \Big\} \\ Y =& \beta_0 + \beta_1 D + \delta_0^Y \times (Z - v_1) + \sum_{f=1}^{F} \Big\{ \delta_f^Y \times \mathbb{1}[Z > v_f] \times (Z - v_f) \Big\} \end{aligned}$$

See for instance Brollo, Nannicini, Perotti, and Tabellini (2013), Clark and Royer (2013) and Dobbie and Skiba (2013) for settings with multiple discontinuities. The case of a continuous instrument is instead unlikely in an RD setting since that would imply a continuum of cutoffs.

Appendix C Average Treatment Effects in the Roy model

Using figure 1, we can write down the average treatment effects for the different types.

- Average Treatment Effect (ATE):

$$E[\beta] = \overline{\beta} = \int_{-\infty}^{+\infty} \beta \; f(\beta) \; \mathrm{d}\beta$$

- ATE on compliers ($LATE_{CM}$):

$$E[\beta|CM] = \frac{1}{P[-\gamma_H < \beta \leq 0]} \int_{-\gamma_H}^{0} \beta \, f(\beta) \; \mathrm{d}\beta$$

  The ATE on compliers is the expected β over the interval where compliers are located. Each β is weighted by the probability density $f(\beta)$, adjusted for the fraction of individuals in that interval ($P[-\gamma_H < \beta \leq 0]$).[37]

- ATE on defiers ($LATE_{DF}$):

$$E[\beta|DF] = \frac{1}{P[0 < \beta \leq -\gamma_L]} \int_{0}^{-\gamma_L} \beta \; f(\beta) \; \mathrm{d}\beta$$

  The ATE on defiers is obtained in a similar way as that for compliers, but over the interval of β where defiers are located: $[0, -\gamma_L]$.

- ATE on always-takers ($LATE_{AT}$):

$$\begin{aligned} E[\beta|AT] = w_{AT} \; & \frac{1}{P[0 < \beta \leq -\gamma_L]} \int_{0}^{-\gamma_L} \beta \; f(\beta) \; \mathrm{d}\beta \\ & + (1 - w_{AT}) \frac{1}{P[\beta > -\gamma_L]} \int_{-\gamma_L}^{+\infty} \beta \; f(\beta) \; \mathrm{d}\beta \end{aligned}$$

  The always-takers are spread over two different intervals of β. Over each interval there is an expected β. Each of these are then assigned a weight equal to the fraction of always-takers that are located in that interval (w_{AT} and $1 - w_{AT}$, see below).

- ATE on never-takers ($LATE_{NT}$):

$$\begin{aligned} E[\beta|NT] = w_{NT} \; & \frac{1}{P[-\gamma_H < \beta \leq 0]} \int_{-\gamma_H}^{0} \beta \; f(\beta) \; \mathrm{d}\beta \\ & + (1 - w_{NT}) \frac{1}{P[\beta \leq -\gamma_H]} \int_{-\infty}^{-\gamma_H} \beta \; f(\beta) \; \mathrm{d}\beta \end{aligned}$$

[37] Note that the sum of weights $\frac{1}{P[-\gamma_H < \beta \leq 0]} \int_{-\gamma_H}^{0} f(\beta) \; \mathrm{d}\beta = \frac{1}{F(0) - F(-\gamma_H)} \int_{-\gamma_H}^{0} f(\beta) \; \mathrm{d}\beta = 1$.

Also never-takers are spread over 2 intervals of β. Over each interval there is an expected β. Each of these are then assigned a weight equal to the fraction of all never-takers that are located in that interval (respectively w_{NT} and $1 - w_{NT}$, see below).

The weights w_{AT} and w_{NT} are as follows:

$$
\begin{aligned}
w_{AT} &= \frac{p_{\gamma_H}\ P[0 < \beta < -\gamma_L]}{p_{AT}} \\
1 - w_{AT} &= 1 - \frac{p_{\gamma_H}\ P[0 < \beta < -\gamma_L]}{p_{AT}} = \frac{P[\beta > -\gamma_L]}{p_{AT}} \\
w_{NT} &= \frac{p_{\gamma_L}\ P[-\gamma_H < \beta < 0]}{p_{NT}} \\
1 - w_{NT} &= 1 - \frac{p_{\gamma_L}\ P[-\gamma_H < \beta < 0]}{p_{NT}} = \frac{P[\beta < -\gamma_H]}{p_{NT}}
\end{aligned}
$$

References

Angrist, Joshua D. and Guido W. Imbens. 1995. "Two-stage least squares estimation of average causal effects in models with variable treatment intensity." *Journal of the American statistical Association* 90 (430):431–442.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434):444–455.

Angrist, Joshua D. and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106 (4):979–1014.

Barber, Michael. 2000. *Making of the 1944 Education Act*. Bloomsbury Publishing.

Barrett, Garry F. and Stephen G. Donald. 2003. "Consistent tests for stochastic dominance." *Econometrica* 71 (1):71–104.

Barua, Rashmi and Kevin Lang. 2009. "School Entry, Educational Attainment and Quarter of Birth: A Cautionary Tale of LATE." Working Paper 15236, National Bureau of Economic Research.

Bedard, Kelly and Elizabeth Dhuey. 2006. "The Persistence Of Early Childhood Maturity: International Evidence Of Long-Run Age Effects." *The Quarterly Journal of Economics* 121 (4):1437–1472.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. 2011. "Too Young to Leave the Nest: The Effects of School Starting Age." *The Review of Economics and Statistics* 93 (2):455–467.

Brollo, Fernanda, Tommaso Nannicini, Roberto Perotti, and Guido Tabellini. 2013. "The Political Resource Curse." *The American Economic Review* 103 (5):1759–96.

Buckles, Kasey and Daniel M. Hungerman. 2013. "Season of Birth and Later Outcomes: Old Questions, New Answers." *The Review of Economics and Statistics* 95 (5):711–724.

Clark, Damon and Heather Royer. 2013. "The Effect of Education on Adult Mortality and Health: Evidence from Britain." *American Economic Review* 103 (6):2087–2120.

Datar, Ashlesha. 2006. "Does delaying kindergarten entrance give children a head start?" *Economics of Education Review* 25 (1):43–62.

de Chaisemartin, Clément. 2014. "Tolerating defiance? Local average treatment effects without monotonicity." *Warwick Economics Research Paper Series* (1020).

Devereux, Paul J. and Robert A. Hart. 2010. "Forced to be Rich? Returns to Compulsory Schooling in Britain." *The Economic Journal* 120 (549):1345–1364.

Dobbie, Will and Paige Marta Skiba. 2013. "Information Asymmetries in Consumer Credit Markets: Evidence from Payday Lending." *American Economic Journal: Applied Economics* 5 (4):256–82.

Elder, Todd E. and Darren H. Lubotsky. 2009. "Kindergarten Entrance Age and Children's Achievement: Impacts of State Policies, Family Background, and Peers." *The Journal of Human Resources* 44 (3):641–683.

Fredriksson, Peter and Björn Öckert. 2014. "The effect of school starting age on school and labor market performance." *The Economic Journal* 124 (579):977–1004.

Gelman, Andrew and Guido Imbens. 2014. "Why High-order Polynomials Should not be Used in Regression Discontinuity Designs." Working Paper 20405, National Bureau of Economic Research.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69 (1):201–09.

Harmon, C. and I. Walker. 1995. "Estimates of the Economic Return to Schooling for the United Kingdom." *The American Economic Review* 85 (5):1278–86.

Heckman, James J., Daniel Schmierer, and Sergio Urzua. 2010. "Testing the Correlated Random Coefficient Model." *Journal of Econometrics* 158 (2):177–203.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables In Models With Essential Heterogeneity." *The Review of Economics and Statistics* 88 (3):389–432.

Heckman, James J. and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73 (3):669–738.

Huber, Martin and Giovanni Mellace. 2012. "Relaxing monotonicity in the identification of local average treatment effects." Economics Working Paper Series 1212, University of St. Gallen, School of Economics and Political Science.

———. 2015. "Testing instrument validity for LATE identification based on inequality moment constraints." *The Review of Economics and Statistics* 97 (2):398–411.

Imbens, Guido and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies* 79 (3):933–959.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2):467–75.

Kitagawa, Toru. 2015. "A test for instrument validity." *Econometrica* 83 (5):2043–2063.

Klein, Tobias J. 2010. "Heterogeneous treatment effects: Instrumental variables without monotonicity?" *Journal of Econometrics* 155 (2):99–116.

Lee, David S. and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2):281–355.

Lindeboom, Maarten, Ana Llena-Nozal, and Bas van der Klaauw. 2009. "Parental education and child health: Evidence from a schooling reform." *Journal of Health Economics* 28 (1):109–131.

Lochner, Lance and Enrico Moretti. 2015. "Estimating and Testing Models with Many Treatment Levels and Limited Instruments." *The Review of Economics and Statistics* 2 (97):387–397.

McEwan, Patrick J. and Joseph S. Shapiro. 2008. "The Benefits of Delayed Primary School Enrollment: Discontinuity Estimates Using Exact Birth Dates." *The Journal of Human Resources* 43 (1):1–29.

Milligan, Kevin, Enrico Moretti, and Philip Oreopoulos. 2004. "Does education improve citizenship? Evidence from the United States and the United Kingdom." *Journal of Public Economics* 88 (910):1667–1695.

Mourifie, Ismael and Yuanyuan Wan. 2014. "Testing Local Average Treatment Effect Assumptions." Working Papers tecipa-514, University of Toronto, Department of Economics.

Muhlenweg, Andrea, Dorothea Blomeyer, Holger Stichnoth, and Manfred Laucht. 2012. "Effects of age at school entry (ASE) on the development of non-cognitive skills: Evidence from psychometric data." *Economics of Education Review* 31 (3):68–76.

Muhlenweg, Andrea M. and Patrick A. Puhani. 2010. "The Evolution of the School-Entry Age Effect in a School Tracking System." *The Journal of Human Resources* 45 (2):407–438.

Oreopoulos, Philip. 2006. "Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter." *American Economic Review* 96 (1):152–175.

Puhani, Patrick and Andrea Weber. 2007. "Does the early bird catch the worm?" *Empirical Economics* 32 (2-3):359–386.

Silles, Mary A. 2011. "The Effect of Schooling on Teenage Childbearing: Evidence Using Changes in Compulsory Education Laws." *Journal of Population Economics* 24 (2):761–777.