Scrutinizing 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 3, 2021

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 must satisfy for their estimate to be interpretable as a LATE. However, applied economic work often omits a discussion of this important assumption. A possible explanation for this missing step is the lack of a clear framework to think about monotonicity in practice. In this paper, we use an extended Roy model to provide insights into the interpretation of IV and fuzzy RD estimates under various degrees of treatment effect heterogeneity, sorting on gain and violation of monotonicity. We then extend our analysis to two applied settings to illustrate how monotonicity can be investigated using a mix of economic insights, data patterns and formal tests. For both settings, we use a Roy model to interpret the estimate even in the absence of monotonicity. We conclude with a set of recommendations for the applied researcher.

JEL classification: C1, C21, C26, I1, I2, J01, J24

Keywords: essential heterogeneity, monotonicity assumption, LATE, average causal response, instrumental variable, regression discontinuity, education, health.

We thank Jerôme Adda, Colin Cameron, Clément de Chaisemartin, Ben Edwards, Margherita Fort, Martin Huber, Susumu Imai, Toru Kitagawa, Tobias Klein, Heather Royer, Kjell Salvanes, Peter Siminski, Olena Stavrunova, Matthew Taylor and seminar participants at the International Association for Applied Econometrics, Australasian Meetings of the Econometrics Society, Swedish Institute for Social Research (SOFI), 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. The LATE identification result applies in any context with *essential heterogeneity*: (i) the gain from treatment is heterogeneous across the population and (ii) there is sorting into treatment based on the gain from treatment. One of the LATE identifying assumptions is 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) point out that, under essential heterogeneity, the assumption of monotonicity is also needed for the identification of a LATE in a fuzzy regression discontinuity (RD) design. 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 often not investigated in applied studies. We found 22 articles published in the American Economic Review (AER) over the 2005-2019 period that explicitly identify a local average treatment effect using either an Instrumental Variable or fuzzy Regression Discontinuity approach. Only 11 of these articles (50%) include the word "monotonicity". Considering a broader set of articles published in the AER, by removing the term "local average treatment effect" from our search, only 21 articles out of 161 (13%) included the word monotonicity.[1] 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.

This paper makes two contributions. First, using numerical examples, we show how informative the IV and fuzzy RD estimates are under varying degrees of heterogeneity in treatment effects, varying degrees of sorting on gain and the degree of violation of

[1]The list of 22 papers is provided separately below the References. The papers that mention monotonicity are denoted by the § symbol. We used Google Scholar to search the AER (2005-2019) for articles that include the terms "local average treatment effect" combined with either "instrumental variable" or ("fuzzy" AND "regression discontinuity"). This returned a total of 24 unique articles, from which we manually removed two published in AER: Papers and Proceedings. We then added the word "monotonicity" which resulted in a subset of 11 articles. Removing the term "local average treatment effect" from the search resulted in the numbers reported. We restrict our search to the AER (excluding AER: Papers and Proceedings) because our focus is on applied studies published in a general audience journal. Note that we exclude "Papers and Proceedings" from any field in the broader search, which might remove some articles that only reference to an AER: Papers and Proceedings article.

monotonicity. The originality here is to do so in an extended Roy selection model for either binary or multivalued treatment (Heckman, Urzua, and Vytlacil (2006), Kline and Walters (2019)), and to consider structural instead of reduced form parameters as inputs into the sensitivity analysis. We find that the interpretation of the estimates can be very sensitive to both essential heterogeneity and violations of the monotonicity condition.

Second, since monotonicity is relevant for a wide range of applications, we investigate it in two existing applied studies that adopt either the IV or fuzzy RD estimator. The first study is Clark and Royer (2013) who use changes in compulsory schooling laws to investigate the effect of education on health in a fuzzy RD setting. The second study is 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. We show how the monotonicity assumption has sometimes been overlooked in the applied literature and how possible violations could have been detected and tested. In each case we also construct a Roy model to provide an interpretation of the estimate under a violation of monotonicity. We end with a set of recommended steps to frame a discussion of the monotonicity assumption and its implications.

Our paper speaks to the applied economist who is faced with estimating treatment effects under essential heterogeneity in an Instrumental Variable or fuzzy Regression Discontinuity context, as exemplified by the list of AER papers mentioned earlier. It is related to the literature that uses economic modelling in a potential outcomes framework to guide interpretation of reduced form estimates, such as Börklund and Moffitt (1987), Angrist, Graddy, and Imbens (2000), Vytlacil (2002), Aakvik, Heckman, and Vytlacil (2005), Heckman, Urzua, and Vytlacil (2006) and Mehta (2019). It also complements the literature that focuses on deriving testable implications of instrument validity (Huber and Mellace (2015), Kitagawa (2015), Mourifié and Wan (2017), Arai, Hsu, Kitagawa, Mourifié, and Wan (2018)), Kowalski (2019)) and the literature that looks at special cases of monotonicity failures that yield interpretable estimates (de Chaisemartin (2017), Dahl, Huber, and Mellace (2017), Klein (2010)).[2]

The remainder of the paper is organized as follows. Section 2 discusses the monotonicity condition in IV and RD settings. Section 3 illustrates the interpretability of the estimates in the context of the binary treatment Roy model, under different degrees of

[2]Throughout the paper our focus is on the identification of LATEs which rely on a single discrete instrument. Börklund and Moffitt (1987), Heckman and Vytlacil (2005), Heckman, Urzua, and Vytlacil (2006) and Cornelissen, Dustmann, Raute, and Schönberg (2016) discuss the identification of the Marginal Treatment Effect (MTE): the average treatment effect on the marginal individuals entering treatment. The MTE is a building block of the LATE, since the latter can be expressed as a weighted average of MTEs. However, identification of the MTE is generally more demanding, ideally requiring a continuous instrument, or a combination of multiple discrete instruments. Even though monotonicity is also an important condition to identify the MTE, our emphasis is on discrete instruments which are more commonly used in the empirical literature.

essential heterogeneity and violation of monotonicity. Section 4 generalizes the analysis to the multivalued treatment case. Section 5 provides a thorough discussion of the monotonicity condition in two existing studies. Section 6 provides recommendations for applied researchers. Section 7 concludes.

2 The Monotonicity Assumption

We adopt the Rubin (1974) potential outcomes framework in the context of a binary treatment. Let $Y_i(0)$ be the response without treatment for individual i. $Y_i(1)$ is the response with treatment for the same individual. D_i is an indicator of treatment. We observe D_i and $Y_i = Y_i(D_i) = D_iY_i(1) + (1 - D_i)Y_i(0)$. The individual's treatment effect $\beta_i = Y_i(1) - Y_i(0)$ is unobserved.

Essential heterogeneity arises when the treatment effect β is heterogeneous across individuals and when there is sorting on gain such that $E(\beta_i|D_i = 1) \neq E(\beta_i|D_i = 0)$. The model can be generalized to include covariates (X), in which case essential heterogeneity arises when $E(\beta_i|D_i = 1, X) \neq E(\beta_i|D_i = 0, X)$. In the rest of the paper we keep the conditioning on X implicit and, unless helpful, we drop the i subscript for notational simplicity.

2.1 Monotonicity in the IV design

Define $D(z)$ as the individual's treatment assignment when $Z = z$, for each $z \in \mathcal{Z}$. Imbens and Angrist (1994) show that β^{IV} identifies the average treatment effect for those individuals whose treatment assignment is affected by the instrument (LATE), provided a random variable Z is available that satisfies the following three conditions,.

IV1. $E[D|Z = z]$ 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 potential outcomes and potential treatment assignment. The independence between Z and the potential treatment assignment $\{D(z)\}_{z\in\mathcal{Z}}$ is sometimes referred to as type independence.[3] The monotonicity assumption IV3 is a condition on counterfactuals

[3]Individuals are classified into types depending on their counterfactual treatment assignment $\{D(z)\}_{z\in\mathcal{Z}}$. When both the treatment and the instrument are binary there are only four possible types: compliers, defiers, always-takers and never-takers. We define these types more formally in section 3.1.

that refers to an individual's treatment in two alternative states of the world, $Z = w$ and $Z = z$. It 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 a change in the value of the instrument, some individuals respond by getting the treatment ("switching in") while others stop getting it ("switching out"). The independence condition (IV2) together with monotonicity (IV3) are needed to identify a LATE whenever there is essential heterogeneity.

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

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

Following Angrist, Imbens, and Rubin (1996), and assuming both IV1 and IV2, we can interpret this estimand as,

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

where,

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

Equation (1) is very informative.

- The identification of $\beta^{IV}(z,w)$ does not rely on individuals who do not respond to changes in the value of the instrument: $D(z) - D(w) = 0$. This motivates the standard LATE interpretation.

- 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 (LATE-*in*: $E[Y(1) - Y(0)|D(z) - D(w) = 1]$) and those who switch out (LATE-*out*: $E[Y(1) - Y(0)|D(z) - D(w) = -1]$), and equal to the average treatment effect (ATE):

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

- If the return to treatment is heterogeneous and there is sorting on gain, then monotonicity matters. 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: $\beta^{IV}(z, w) =$ LATE-*in*. Alternatively, if monotonicity holds because $D_i(z) - D_i(w) \leq 0 \; \forall i$ then $\lambda = 0$, and IV estimates the effect for those individuals that stop getting the treatment because of the instrument: $\beta^{IV}(z, w) =$ LATE-*out*.

- If monotonicity does not hold, then either $\lambda < 0$ or $\lambda > 1$, and the IV estimate is neither a LATE-*in* or LATE-*out* nor a weighted average of the two LATEs. In this case Equation (1) 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. Even if the researcher has a very good guess on both proportions such that λ can be inferred, recovering one of the LATEs is still not possible. Hence, it is generally not possible to interpret $\beta^{IV}(z, w)$.[4]

Hahn, Todd, and Van der Klaauw (2001) point out that both the fuzzy Regression Discontinuity and the IV estimands can be expressed as a Wald estimand. A set of conditions similar to IV1-IV3 applies arbitrarily close to the RD threshold.[5] Therefore, the above discussion regarding the role of the monotonicity assumption in interpreting the IV estimand equally applies in the fuzzy RD design.[6]

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 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 limited sorting on gain? What if there is only a small violation of monotonicity? The answers to these questions cannot be derived from Equation (1) without an explicit selection model.

We set up an extended Roy model with an exogenous variable Z affecting the treatment decision, similar to Heckman, Urzua, and Vytlacil (2006), Kline and Walters (2019).

[4] de Chaisemartin (2017) discusses a special case where monotonicity does not hold and where $\beta^{IV}(z, w)$ can be interpreted as the $LATE$ for a subpopulation of switchers-in (see section 3.4). Dahl, Huber, and Mellace (2017) and Klein (2010) discuss special cases of monotonicity failures that nevertheless allow them to obtain a LATE for compliers.

[5] Dong (2018) shows that the local independence assumption can be replaced by a weaker local smoothness assumption.

[6] We refer the reader to Appendix A.1 for more detail on the RD case in a binary treatment setting.

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. Such a model encompasses a broad range of settings including selection into schooling and job market programs. The goal 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. At the same time, the model can also be used to gain insights into the kind of economic behaviour that would lead to interpretation concerns.

3.1 A Roy model with violation of monotonicity

Let,

$$Y_i(1) = \alpha + \overline{\beta} + U_i(1)$$
$$Y_i(0) = \alpha + U_i(0) \ .$$

Thus, the gain from treatment is heterogeneous and given by $\beta_i = Y_i(1) - Y_i(0) = \overline{\beta} + U_i(1) - U_i(0)$. Let the treatment be determined as follows:

$$D = \begin{cases} 1 & \text{if } Y_i(1) - Y_i(0) + \gamma_i Z_i > 0 \Leftrightarrow \beta_i > -\gamma_i Z_i \\ 0 & \text{if } Y_i(1) - Y_i(0) + \gamma_i Z_i \leq 0 \Leftrightarrow \beta_i \leq -\gamma_i Z_i \ . \end{cases}$$

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). Here γZ could be interpreted as a cost or taste for treatment. Dropping the i subscript for notational simplicity, let us consider the case of a binary instrument Z and a binary parameter γ: $z \in \{0, 1\}$, and $\gamma \in \{\gamma_L, \gamma_H\}$. Importantly, for monotonicity to be violated, we need to set $\gamma_L < 0$ and $\gamma_H > 0$. Thus, when the instrument changes from 0 to 1, it 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 a 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 four types based on their counterfactual treatment choices: always-takers (AT), never-takers (NT), compliers (CM) and defiers (DF).

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, $P(D =$

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

$1|Z = 1) = p_{AT} + p_{CM}$ 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 types are defined by the pair (β, γ), type independence requires that these parameters are jointly independent of Z. To simplify our discussion below, we set 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 normally distributed: $\beta \sim N(\overline{\beta}, \sigma_\beta)$.

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.[7]

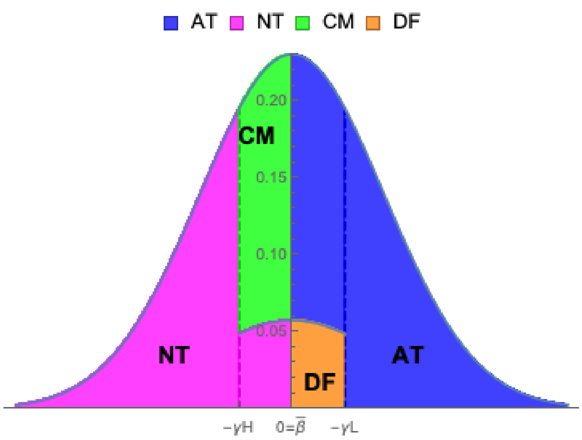


Figure 1: Types over the β distribution

- $p_{AT} = p_{\gamma_L} \times P[\beta > -\gamma_L] + p_{\gamma_H} \times P[\beta > 0]$. Always-takers take treatment irrespective of the instrument. Therefore, they must have a positive return to treatment. In particular, ATs with $\gamma_L < 0$ need a large positive β to compensate for the negative impact of the instrument and remain treated.

[7] Figure 1 is obtained using the parametrization described in Table 2a.

- $p_{NT} = p_{\gamma_L} \times P[\beta \leq 0] + p_{\gamma_H} \times P[\beta \leq -\gamma_H]$. Never-takers do not take treatment irrespective of the instrument. Therefore, they must have a negative return to treatment. In particular, NTs with $\gamma_H > 0$ need a large negative β to compensate for the positive impact of the instrument and remain untreated.

- $p_{CM} = 0 + p_{\gamma_H} \times P[-\gamma_H < \beta \leq 0]$. Compliers 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 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 pull 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 B 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 subsection 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 $\sigma_{01} = -0.5$ in order to get a positive standard deviation of $\sigma_\beta = \sqrt{3}$.[8] We also set $p_{\gamma_L} = 0.25$ to ensure that the rank condition is satisfied: $p_{CM} > p_{DF}$. The resulting proportions of types are described in Table 2b. Most of the population are always-takers, followed by never-takers. However, compliers and defiers are also present. In this Roy model with a binary treatment and a binary instrument:

$$\beta^{IV} = \lambda \times LATE_{CM} + (1 - \lambda) \times LATE_{DF} \ , \tag{2}$$

[8] Note that as a result β and $Y(0)$ are negatively correlated since $Cov(\beta, U(0)) = \sigma_{01} - \sigma_0^2 = -1.5$, resulting in a negative selection bias $E[Y(0)|D = 1] - E[Y(0)|D = 0]$. This is similar to Heckman, Urzua, and Vytlacil (2006).

where,

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

Whenever both $p_{CM} > 0$ and $p_{DF} > 0$ 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).[9]

Table 2: Baseline

(a) Parametrization

γ_L	γ_H	p_{γ_L}	$\overline{\beta}$	μ_1	μ_0	σ_1	σ_0	σ_{01}
-1	1	0.25	0	0	0	1	1	-0.5

(b) Types

p_{AT}	p_{NT}	p_{CM}	p_{DF}	$p_{AT} + p_{NT} + p_{CM} + p_{DF}$	λ
0.445	0.336	0.164	0.055	1	**1.5**

All values are rounded to the third decimal digit.

(c) ATE, LATEs and IV estimate

ATE	$LATE_{AT}$	$LATE_{NT}$	$LATE_{CM}$	$LATE_{DF}$	β^{IV}
0	1.492	-1.818	-0.486	0.486	**-0.973**

All values are rounded to the third decimal digit.

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

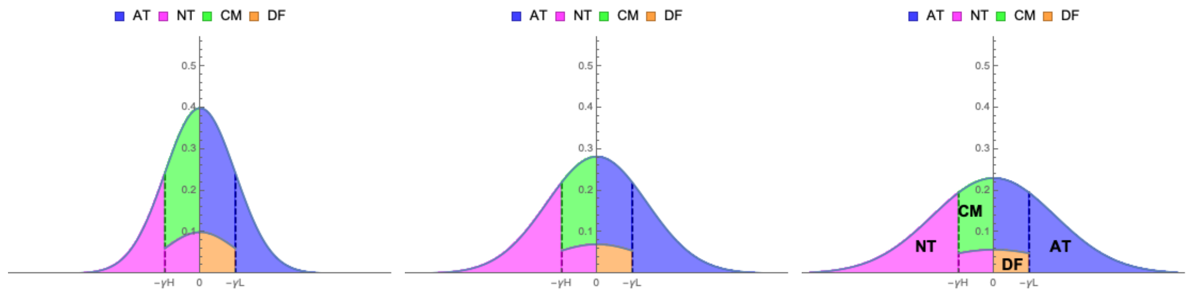
Here we alter the parameters in the baseline scenario and show how the decisions of individuals are affected. The goal is to illustrate whether β^{IV} becomes more or less interpretable as we change the degree of heterogeneity in the treatment effect, the degree of heterogeneity in the impact of the instrument and the extent of sorting on gain. We use simulations since the expressions defining the differences between β^{IV} and the LATEs are too complex to get clear and intuitive results.

[9] 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 directly related and even knowledge of λ would not allow us to recover any LATE.

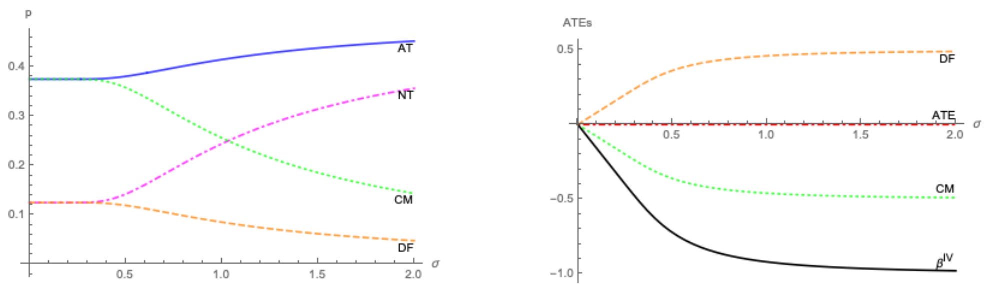
3.3.1 Heterogeneity in the treatment effect

In the baseline we set $\sigma_\beta = \sqrt{3} \approx 1.732$ 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.[10] As σ_β decreases the distribution narrows around the Average Treatment Effect ($\overline{\beta} = 0$). Figure 2d shows the proportion of each type: as σ_β goes towards zero all observations get concentrated within the $[-\gamma_H, -\gamma_L]$ interval. Thus, the proportions of always-takers and never-takers fall. 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 zero, λ is constant. Figure 2e shows the ATE, the LATEs for compliers and defiers, and β^{IV}. As σ_β tends to zero, β^{IV} and the LATEs converge to the ATE, suggesting that in the absence of treatment effect heterogeneity there is no monotonicity requirement. However, for positive values of σ_β, β^{IV} is always more extreme and far from the LATEs since $\lambda > 1$.

Figure 2: Sensitivity to σ_β



(a) $\sigma_\beta = \sqrt{1}$ (b) $\sigma_\beta = \sqrt{2}$ (c) $\sigma_\beta = \sqrt{3}$ (baseline)



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

3.3.2 Heterogeneity in the response to the instrument

Next, we investigate what happens when we vary the degree of heterogeneity in the response to the instrument. There are two ways to do so in the model. The first is to

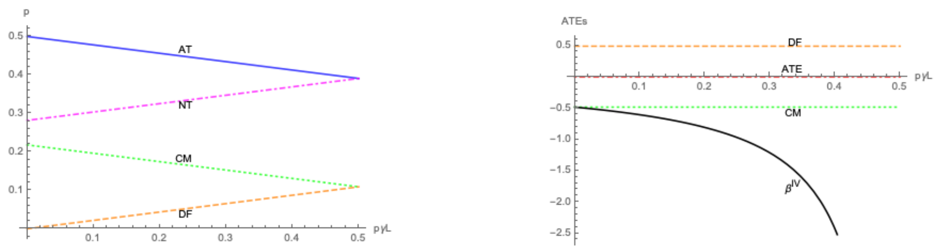
[10] We remain agnostic as to what drives the changes in σ_β since it is irrelevant for the key moments of the model in tables 2b and 2c.

alter p_{γ_L}: however, the only effect of varying this parameter is to change the proportion of compliers and defiers, while their LATEs are unchanged. Figure 3 shows the proportion of types and the treatment effects as p_{γ_L} varies between 0 and 0.5. Intuitively, the closer p_{γ_L} becomes to 0.5 the smaller the difference $p_{CM} - p_{DF}$, in turn leading to a larger λ in Equation (2) and β^{IV} becoming increasingly distant from the LATEs .

The second way to vary the degree of heterogeneity is to change the value of γ_H relative to γ_L. In our baseline $\gamma_H = -\gamma_L$, but it is possible that individuals are heterogeneous not only in the direction of their response but also in their magnitude. Thus, holding γ_H constant, shifting the value of γ_L leads to varying proportions of defiers while also changing the LATE of defiers.

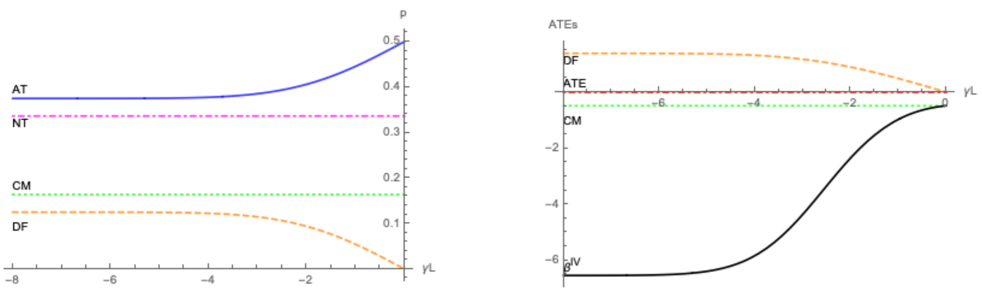
Figure 4a shows the proportion of types as a function of γ_L. For $\gamma_L = 0$ we have a limit case with heterogeneity in the response to the instrument but no defiers since no individual is pushed out of treatment by the instrument.[11] As γ_L becomes a larger negative number, the proportion of defiers increases. In turn, λ in Equation (2) grows larger than 1 while the LATE of defiers also increases. Figure 4b shows that as a result β^{IV} becomes more extreme and cannot be interpreted as a LATE of interest.

Figure 3: Sensitivity to p_{γ_L}



(a) Proportion of types (b) ATE, LATEs and β^{IV}

Figure 4: Sensitivity to γ_L



(a) Proportion of types (b) ATE, LATEs and β^{IV}

Note that under both scenarios the larger the first stage $p_{CM} - p_{DF}$, the closer β^{IV}

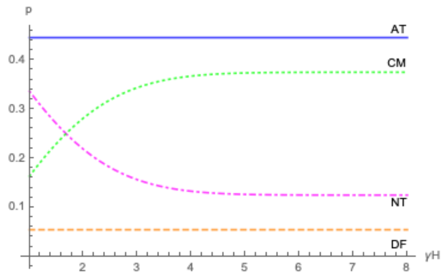
[11] This case is equivalent to a situation where γ is heterogeneous but only takes positive values.

is to the LATE of compliers. However, this is not always true. To see this, it is useful to rewrite (2) as

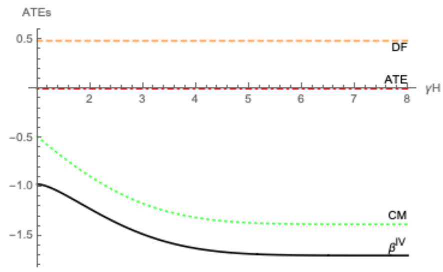
$$\beta^{IV} - LATE_{CM} = (1 - \lambda)(LATE_{DF} - LATE_{CM}) \tag{3}$$

Recall that in our model $\lambda > 1$, and $LATE_{DF} > LATE_{CM}$ due to sorting on gain. In Figure 3, the difference $\beta^{IV} - LATE_{CM}$ becomes smaller because λ converges to 1 while the LATEs are unchanged. In Figure 4, λ converges to 1 while the $LATE_{DF}$ becomes smaller and $LATE_{CM}$ stays unchanged. Now consider the case where we increase γ_H while holding constant the value of γ_L. Doing so increases the proportion of compliers, and makes their LATE more negative, while leaving the proportion of defiers and their LATE unaffected. This case is illustrated in figure 5. β^{IV} does not converge monotonically to $LATE_{CM}$ with a growing first stage. λ converges to 1 but $(LATE_{DF} - LATE_{CM})$ increases instead, and for certain values of γ_H the latter effect dominates resulting in β^{IV} diverging from $LATE_{CM}$. Hence, as long as there are defiers, a larger first stage is not sufficient to counteract a departure from the monotonicity assumption.

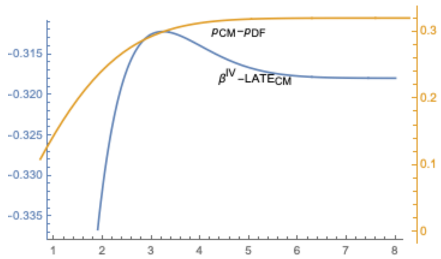
Figure 5: Sensitivity to γ_H



(a) Proportion of types



(b) ATE, LATEs and β^{IV}



(c) Bias vs First Stage

3.3.3 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. Monotonicity is not a problem in the limit case of no sorting on gain. However, to what extent is a departure from this limit case sufficient to render β^{IV} uninformative? To address this question, we adjust our model by altering the relative importance of the gain in the treatment decision. We do so by introducing an additional unobserved factor ε in the treatment decision process. For simplicity we assume that this factor is normally distributed and independent of all other parameters and variables of the model: $\varepsilon \sim N(0, \sigma_\varepsilon)$. Hence:

$$D = \begin{cases} 1 & \text{if } Y(1) - Y(0) + \gamma Z + \varepsilon > 0 \Leftrightarrow \beta + \varepsilon > -\gamma Z \\ 0 & \text{if } Y(1) - Y(0) + \gamma Z + \varepsilon \leq 0 \Leftrightarrow \beta + \varepsilon \leq -\gamma Z \ . \end{cases}$$

Intuitively, as σ_ε grows larger, ε becomes the main determinant of treatment. The introduction of ε in the decision process could be interpreted as additional cost or taste parameters. Alternatively it could be interpreted as individuals using $\beta + \varepsilon$ as their guess of the gain if they only have a noisy signal. While a larger σ_ε reduces the importance of sorting on gain, it also quickly reduces the importance of the instrument. To avoid this, we rescale the parameter associated with Z accordingly. In particular, we set $\tilde{\gamma}_L$ and $\tilde{\gamma}_H$ to be functions of σ_ε in a way that leaves the proportion of the different types unchanged.[12] The treatment D is then determined as follows:

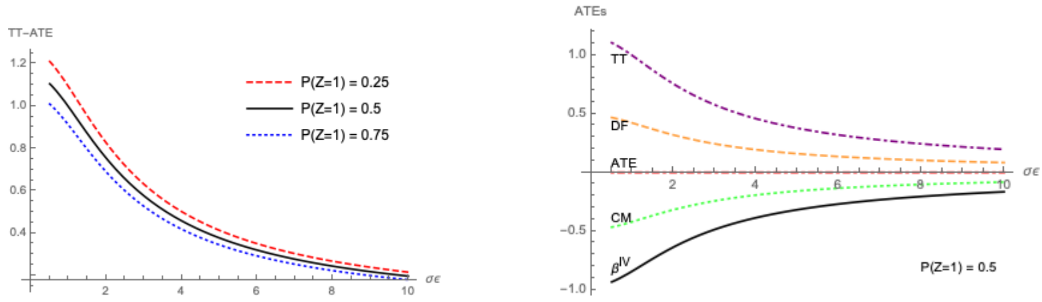
$$D = \begin{cases} 1 & \text{if } Y(1) - Y(0) + \tilde{\gamma} Z + \varepsilon > 0 \Leftrightarrow \beta + \varepsilon > -\tilde{\gamma} Z \\ 0 & \text{if } Y(1) - Y(0) + \tilde{\gamma} Z + \varepsilon \leq 0 \Leftrightarrow \beta + \varepsilon \leq -\tilde{\gamma} Z \ . \end{cases}$$

Define sorting on gain as the difference between the average treatment on the treated (TT) and ATE (see for instance Heckman and Vytlacil (2005)). Without sorting on gain we expect these averages to be the same, while with positive sorting we expect this difference to be larger than zero.

Figure 6a shows that the degree of sorting on gain decreases with σ_ε irrespective of $P(Z = 1)$. There would be no sorting on gain only if $\sigma_\varepsilon = \infty$. For $P(Z = 1) = 0.5$, Figure 6b shows that β^{IV} converges to the ATE and the other LATEs as σ_ε increases but we would need to be in the limit case of $\sigma_\varepsilon = \infty$ for β^{IV} to be interpretable. This would occur only in a setting where individuals choose not to, or cannot take their gain

[12] For instance, let $V = \beta + \varepsilon$, with $V \sim N(0, \sqrt{\sigma_\beta^2 + \sigma_\varepsilon^2})$. The proportion of compliers is given by $p_{\gamma_H}\left(\Phi\left(\frac{0}{\sigma_V \sqrt{2}}\right) - \Phi\left(\frac{-\gamma_H}{\sigma_V \sqrt{2}}\right)\right)$ where Φ is the CDF. The first term remains constant but the second term becomes larger as σ_ϵ increases, thus reducing the proportion of compliers. We prevent this by setting $\tilde{\gamma}_H = -\gamma_H \frac{\sigma_V}{\sigma_\beta}$. Hence the impact of the instrument also increases in σ_ε, while the distribution of the gains is left unchanged. We set $\tilde{\gamma}_L$ in a similar manner.

Figure 6: Sensitivity to σ_ε



(a) $TT - ATE$ (b) ATE, LATEs and β^{IV}

into account.[13]

3.4 Special cases of monotonicity failures

3.4.1 Compliers-Defiers condition

In an attempt to assign a meaning to an otherwise uninterpretable IV parameter, de Chaisemartin (2017) discusses a special case where monotonicity is violated but the following two assumptions hold:

$$p_{CM_F} = p_{DF}$$
$$LATE_{CM_F} = LATE_{DF} \ ,$$

where CM_F is a subgroup of the population of compliers CM: $p_{CM} = p_{CM_V} + p_{CM_F}$. Under these assumptions de Chaisemartin (2017) shows that $\beta^{IV} = LATE_{CM_V}$, where CM_V is the remaining subgroup of compliers. The idea is that the LATE of the CM_F group cancels out the LATE of defiers, and β^{IV} then measures the LATE of the surviving compliers. In our Roy model where selection is driven only by the gain and the instrument (Figures 1-2), the support of β for compliers and defiers does not overlap, since each defier has a larger gain from treatment than any complier. Therefore, there is no scope for a subgroup of compliers to compensate for the LATE of defiers. Nevertheless, in a model where an additional factor drives treatment, such as in section 3.3.3, it might be possible. In this case it must be that the LATE of surviving compliers is more extreme than the LATE for the overall population of compliers. Thus, one can also use de Chaisemartin

[13]Not shown, the proportion of each types remains the same as in the baseline model due to our adjustment of the γ parameters. We also do not show the distribution of the types over the β distribution since they are no longer confined to certain intervals. Intuitively, relative to the baseline case of Figure 1, there are now some compliers with $\beta \leq -\gamma_H$ and some with $\beta > 0$. Then $LATE_{CM}$ converges to the ATE as σ_ε grows since $P[\beta \leq -\gamma_H] < P[\beta > 0]$. At the same time, there are now some defiers with $\beta \leq 0$ and some with $\beta > -\gamma_L$, and $LATE_{DF}$ also converges to the ATE as σ_ε grows since $P[\beta \leq 0] > P[\beta > -\gamma_L]$.

(2017) intuition to explain why β^{IV} is more extreme than $LATE_{CM}$ and $LATE_{DF}$ if monotonicity is violated.[14]

3.4.2 Local Monotonicity condition

Dahl, Huber, and Mellace (2017) show that the LATE for both compliers and defiers are identified when these types do not coexist at any point in the support of the marginal potential outcome distributions, $Y(0)$ and $Y(1)$. They label such assumption local monotonicity (LM). Formally:

> Either $P[D(1) \geq D(0)|Y(d) = y(d)] = 1$ OR $P[D(0) \geq D(1)|Y(d) = y(d)] = 1$, $\forall y(d)$ in the support of $Y(d)$ and $d \in \{0, 1\}$

Under LM, defiers can be ruled out in some regions of $Y(d)$. The potential outcomes and thus the LATE of compliers are then locally identified in these regions. The same reasoning applies to identifying the LATE for defiers in regions of Y (d) without compliers.

In our Roy model where selection is driven only by the gain and the instrument, compliers and defiers do not overlap in any region of the $\beta = Y(1) - Y(0)$ distribution. LM requires that such a lack of overlap applies to marginal outcome distributions instead. Such a situation occurs under a perfect negative correlation between U_0 and U_1 (i.e. $\sigma_{01} = -1$). The latter ensures $Var(\beta|Y(d)) = 0, \forall d = \{0, 1\}$, preventing compliers and defiers from coexisting.[15] When additional elements other than the gain and the instrument affect selection into treatment (section 3.3.3), $\sigma_{01} = -1$ is no longer sufficient to prevent compliers and defiers from coexisting at each potential outcome value.

3.5 Summary and Discussion

How informative is β^{IV} (or β^{RD}) when one cannot rule out essential heterogeneity and heterogeneous responses to the instrument (or forcing variable)? The analysis using our extended Roy model suggests that the interpretation can be very sensitive to both essential heterogeneity and violations of the monotonicity condition. There are extreme cases when a violation of 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.[16] However, once

[14]When the outcome is binary, de Chaisemartin (2017) provides sufficient conditions under which the two main assumptions hold and discusses them with some empirical applications. With a continuous outcome, a sufficient condition is that there are no fewer compliers than defiers for every value in the support of β. However, this is a difficult condition to test and would be unlikely to hold in our Roy model.

[15]$\sigma_{01} = 1$ yields a degenerate β-distribution and thus wipes out heterogeneity in the gain.

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

we depart from these extreme cases, then β^{IV} (or β^{RD}) quickly becomes uninformative. The deviation of β^{IV} from any LATE of interest depends on the underlying selection model. Evaluating its size requires making explicit assumptions on the model and its parameters.[17]

4 Monotonicity when the treatment is multivalued

In this section, we briefly discuss the monotonicity assumption and generalize the Roy model to the multivalued treatment case. This generalization is useful for our empirical examples later in the paper. 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 potential outcome of an individual under treatment value k, $\forall k \in \mathcal{D}$. Assume that IV1 (rank) and IV2 (independence) hold, we can express the IV estimand of β for any two points of support z, w in $\mathcal{Z}$ as follows

$$\begin{aligned}\beta^{IV}(z,w) = \frac{1}{\Omega} \times \sum_{k=1}^{K} \Big\{ & E[Y(k) - Y(k-1)|D(z) \geq k > D(w)] \times P[D(z) \geq k > D(w)] \\ & - E[Y(k) - Y(k-1)|D(w) \geq k > D(z)] \times P[D(w) \geq k > D(z)] \Big\}\end{aligned} \tag{4}$$

where

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

is the first stage. From equation (4) we can see that

- If assumption IV3 (monotonicity) holds such that no one decreases treatment intensity, $P[D(w) \geq k > D(z)] = 0 \ \ \forall k$. Equation (4) then simplifies to

$$\beta^{IV}(z,w) = \sum_{k=1}^{K} E[Y(k) - Y(k-1)|D(z) \geq k > D(w)] \times \frac{P[D(z) \geq k > D(w)]}{\sum_{k=1}^{K} P[D(z) \geq k > D(w)]} \tag{5}$$

  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. The ACR is the estimand of interest in the multivalued case. A similar discussion applies if instead no one increases treatment intensity such that $P[D(z) \geq k > D(w)] = 0 \ \forall k$.

[17] The Regression Discontinuity setting is very similar and is illustrated in Appendix C.

- If monotonicity does not hold, for a given change in the value of Z some individuals increase treatment, $D(z) \geq k > D(w)$, while others decrease it, $D(w) \geq j > D(z)$ for at least some $j, k \in \mathcal{D}$. Equation (4) shows that both the numerator and the denominator include contributions from individuals affected in either direction. While the denominator is always smaller than in (5), the impact on the numerator is ambiguous and depends on the treatment effects of those decreasing treatment. Without monotonicity, the IV estimate is thus not a weighted average of treatment effects and cannot be assigned a useful interpretation. It is an equation in many unknowns, which makes it impossible to back-out neither the ACR nor the LATE for any particular group.

- Monotonicity is not a concern only if there is no sorting on gain and average returns are constant across treatment levels. The former implies $E[Y(k) - Y(k-1)|D(z) \geq k > D(w)] = E[Y(k) - Y(k-1)] \;\; \forall k$, while the latter implies $E[Y(K) - Y(K-1)] = E[Y(K-1) - Y(K-2)] = \ldots = E[Y(1) - Y(0)]$. In that case (4) collapses to the ATE.

The close analogy between the fuzzy Regression Discontinuity design and the IV estimators extends to the case of multivalued treatment with a 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 estimand is still the same as that of the IV estimand.[18]

4.1 A Roy model when the treatment is multivalued

Let the treatment k be ordered, and let the potential outcome be

$$Y_i(k) = \alpha + U_i(k) \qquad \forall k \in \mathcal{D}$$

The choice of treatment level can be described by an optimal stopping problem. Let $\beta_{i,k\to k+1} \equiv Y_i(k+1) - Y_i(k)$ be the marginal benefit from increasing treatment, and let

[18] 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 of ACRs if the treatment is multivalued. However, if monotonicity is violated this LATE or ACR interpretation is lost. We refer the reader to Appendix A.2 for details of the multivalued instrument case.

$C(k+1)-\gamma_i(k+1)Z_i$ be the marginal cost. Treatment choices are determined as follows:

$$\begin{aligned} D=0 &\quad if \quad C(1)-\gamma_i(1)Z_i \geq \beta_{i,0\to 1} \\ D=k &\quad if \quad C(k)-\gamma_i(k)Z_i < \beta_{i,k-1\to k} \quad AND \\ &\qquad\quad C(k+1)-\gamma_i(k+1)Z_i \geq \beta_{i,k\to k+1} \quad \forall k=1,\ldots,K-1 \\ D=K &\quad if \quad C(K)-\gamma_i(K)Z_i < \beta_{i,K-1\to K} \end{aligned} \tag{6}$$

where $\gamma_i(k)Z_i$ is a taste or cost realization that is heterogeneous across individuals and $C(k)$ is a cost of treatment that is homogeneous across individuals. Both can differ across treatment levels.[19]

To keep the setting as a simple extension to the binary Roy model of section 3.1, assume that

- individuals can choose between three values of the treatment: $\mathcal{D}=\{0,1,2\}$.
- the return to increasing treatment by one unit is constant over treatment levels: $\beta_{i,0\to 1}=\beta_{i,1\to 2}=\beta_i$. This restriction implies that $U_i(2)=2U_i(1)-U_i(0)$. Therefore, the additional structural parameters $\{\mu_2,\sigma_2,\sigma_{02},\sigma_{12}\}$ are a deterministic function of the parameters in Table 2a.
- the cost of treatment is rising in k such that $C(2)>C(1)=0$
- the impact on the instrument $\gamma_i(k)$ is a binary parameter that can take values $\gamma_L(k)<0$ or $\gamma_H(k)>0 \quad \forall k=1,2$.
- the impact of the instrument is constant across treatment levels: $\gamma_i(k)=\gamma_i \quad \forall k=1,2$. Thus a proportion p_{γ_L} of individuals have $\gamma_L(1)=\gamma_L(2)<0$ in which case they might reduce treatment intensity, while the remaining $1-p_{\gamma_L}$ have $\gamma_H(1)=\gamma_H(2)>0$ which might raise treatment intensity. This allows for monotonicity to be violated at each treatment level.

Hence, for each individual the marginal benefit of treatment is constant whereas the marginal cost is rising. Given that $\mathcal{D}=\{0,1,2\}$ there are now nine possible types $(t_A,\ldots,t_I)$ based on treatment status under the alternative values of the binary instrument, as illustrated in Table 3.

Types t_B,t_C,t_F can be denoted as complier-types since they increase treatment in response to the instrument, with type t_C is a super-complier who experiences a multi-level increase in treatment. Similarly, types t_D,t_G,t_H are defier-types with type t_D being

[19] The expected reward from treatment level k is defined as $R_i(k)=Y_i(k)-\sum_{j=1}^{k}(C(j)-\gamma_i(j)Z_i)$. Only the k specific outcome $Y_i(k)$ is realized, whereas the $\gamma_i(k)Z_i$ and $C(k)$ realizations are experienced for all $j\leq k$.

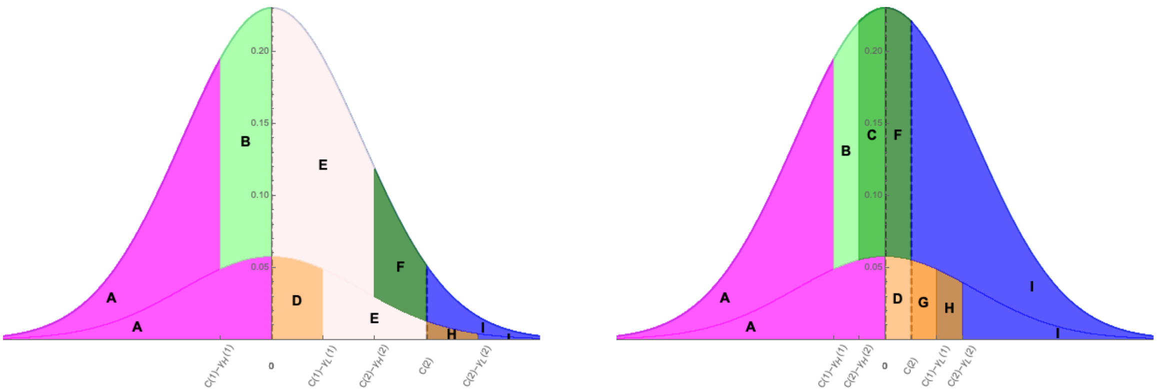
Table 3: Multivalued Roy model - Types

	$Z=1$		
$Z=0$	$D=0$	$D=1$	$D=2$
$D=0$	$t_A(=)$	$t_B(+)$	$t_C(+)$
$D=1$	t_D(-)	$t_E(=)$	$t_F(+)$
$D=2$	t_G(-)	t_H(-)	$t_I(=)$

The (+) indicates that individuals increase their level of treatment under $Z=1$ (vs. $Z=0$), and viceversa for the (-).

a super-defier. Figure 7(a) illustrates the location of each type over the β-distribution under the parametrization described in Table 4. Figure 7(b) illustrates this for a lower cost $C(2)$. In the latter case, both super-compliers t_C and super-defiers t_G exist.

Figure 7: Multivalued Roy model - Types



(a) High cost: $C(2)=3$ (b) Low cost: $C(2)=0.5$

The ACR is the estimand of interest and consists of a weighted average of LATEs for the complier-types:

$$ACR = \frac{p_B LATE_B + 2p_C LATE_C + p_F LATE_F}{p_B + 2p_C + p_F} \tag{7}$$

but the presence of defiers causes the IV estimand to be different from the ACR and makes it difficult to interpret[20]:

$$\beta^{IV} = \frac{p_B LATE_B - p_D LATE_D}{\Omega} + 2\frac{p_C LATE_C - p_G LATE_G}{\Omega} + \frac{p_F LATE_F - p_H LATE_H}{\Omega} \tag{8}$$

where $\Omega = (p_B + 2p_C + p_F) - (p_D + 2p_G + p_H)$. Note that super-complier and super-defier types, if they exist, contribute twice as much as the other complier and defier types to the IV and ACR estimands.

[20]There is no compact way to describe the difference between the IV estimand and the ACR.

Retaining the baseline parametrization from section 3.2, the only parameter still undefined is $C(2)$ which we initially set equal to $C(1) + \sigma^2_\beta = 3$. The full parametrization, types and Average Treatment Effects are described in Table 4. The first stage Ω is positive since there are more complier-types than defier-types, although both groups are present. There are no super-compliers or super-defiers.

Table 4: Multivalued Roy model - Baseline

(a) Parametrization

γ_L	γ_H	p_{γ_L}	$\overline{\beta}$	μ_0	μ_1	μ_2	
-1	1	0.25	0	0	0	0	
σ_0	σ_1	σ_2	σ_{01}	σ_{02}	σ_{12}	$C(1)$	$C(2)$
1	1	2.646	-0.5	-2	2.5	0	3

(b) Types and First Stage

p_A	p_B	p_C	p_D	p_E	p_F	p_G	p_H	p_I	$\sum$ p	Ω
0.336	0.164	0.	0.055	0.342	0.062	0.	0.008	0.034	1.	0.163

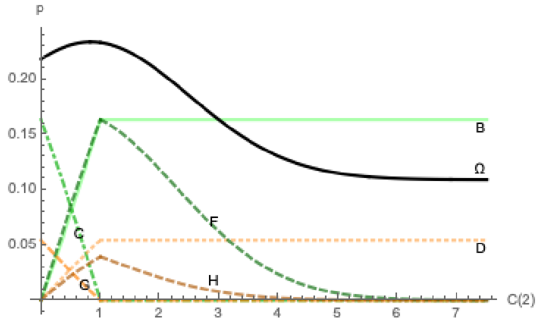
(c) ATE, LATEs, ACR and IV estimate

$LATE_A$	$LATE_B$	$LATE_C$	$LATE_D$	$LATE_E$	$LATE_F$	$LATE_G$	$LATE_H$	$LATE_I$
-1.818	-0.486	-	0.486	1.052	2.432	-	3.406	3.772
			ATE	β^{IV}	ACR			
			0.	0.109	0.314			

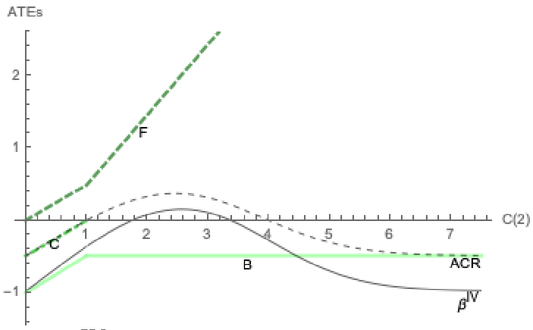
Due the monotonicity assumption not being satisfied, the IV estimand is different from the ACR. Figure 8 generalizes the discussion on the types and estimands for a range of values of the cost $C(2)$. For very large $C(2)$, no individual has $D = 2$, bringing us back to the binary treatment case of section 3.1. Similarly, for $C(2) = C(1) = 0$, individuals might take $D = 2$ but no one takes $D = 1$ resulting in a binary treatment where the first stage is now double the size. Importantly, the IV estimand is always lower than the ACR irrespective of the cost.[21] The IV estimand can however fall between the $LATEs$ of complier-types depending on the value of $C(2)$. Yet, in general, it is not a weighted average. Finally, as in the binary case, β^{IV} does not converge monotonically to the ACR with a growing first stage Ω, confirming that the latter is not necessarily sufficient to counteract a departure from the monotonicity assumption. This model can be generalized to unordered treatment choices as in Heckman, Urzua, and Vytlacil (2006) and Heckman, Urzua, and Vytlacil (2008), with the caveat that a different instrument is needed for every treatment level.

[21]Intuitively, in this Roy model, there is a $k+1 \to k$ defier-type for each $k \to k+1$ complier-type, with the defier-type located to the right-hand side of the corresponding complier-type in the β distribution such that $LATE_B < LATE_D, LATE_C < LATE_G, LATE_F < LATE_H$. The LATE of defier-types are also all positive.

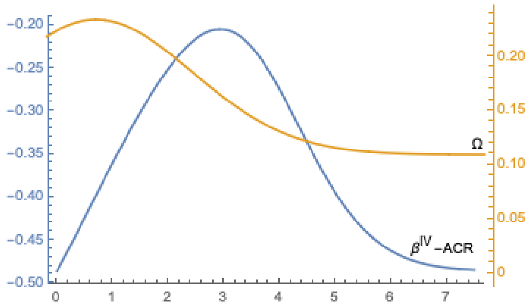
Figure 8: Multivalued Roy model - Sensitivity to $C(2)$



(a) Proportion of complier-types (B, C, F), defier-types (D, G, H), and first stage.



(b) β^{IV}, ACR and LATEs of complier-types (B, C, F)



(c) Bias vs First Stage

5 Empirical Applications

In this section we go through two different studies to illustrate how monotonicity can be investigated using economic insights and data analysis. We do not scrutinize the independence condition since it is extensively discussed in these papers and in applied work more generally. Most of our conclusions in this section are conditional on independence being satisfied. Similarly, we discuss but do not test essential heterogeneity.[22] On intuitive grounds we cannot see how essential heterogeneity could be ruled out a priori in any of these studies.[23] Since we find evidence that the monotonicity assumption does not hold, we impose some structure to assign an interpretation to the IV or fuzzy RD estimate.

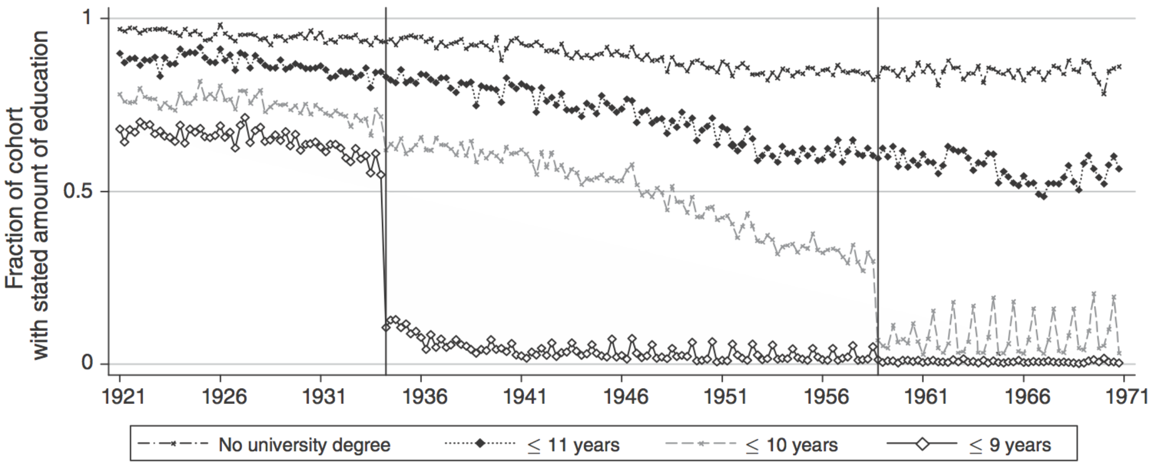
[22]This would lengthen the discussion considerably and for both studies the data is not publicly available. In section 6 we refer to a number of articles that provide a thorough discussion of essential heterogeneity and how to test for it.

[23]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.

5.1 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 birth 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 9 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.

Figure 9: Clark and Royer (2013) page 2092



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 used extensively as an instrumental variable for education in various contexts such as earnings and labor activity (Harmon and Walker (1995), Oreopoulos (2006), Devereux and Hart (2010), Grenet (2013)); citizenship and political involvement (Milligan, Moretti, and Oreopoulos (2004)); health of offspring (Lindeboom, Llena-Nozal, and van der Klaauw (2009)); fertility, teenage childbearing and marital outcomes (Silles (2011), Fort, Schneeweis, and Winter-Ebmer (2016), Geruso and Royer (2018)). 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.

5.1.1 Plausibility of the Monotonicity Assumption

It is worth discussing essential heterogeneity and the need for the monotonicity assumption in this context. CR13 use these compulsory schooling reforms to study the impact of education on health outcomes such as mortality, self-reported health status, self-reported health behaviours such as smoking and drinking and clinical health measures collected by a nurse. A heterogeneous return to education in terms of health seems plausible, as found in Conti and Heckman (2010) and Galama, Lleras-Muney, and van Kippersluis (2018). Sorting on gain is perhaps less intuitive in this setting. Would individuals take into account their health return to education when deciding how much education to attain? We might expect that the health return to education is at least a weaker driver of education decisions than the earnings return. However, the key issue is whether the health return to education (β) is correlated with educational attainment (D). For this to happen it is not necessary that individuals take into account β when making education decisions. Suppose that individuals do not know their health return but they do know and consider their earnings return ρ when making education decisions: $Cov(\rho, D) \neq 0$. Now suppose that the health return and earnings return to education are positively correlated: smarter and more motivated individuals might get larger returns to education for a variety of different outcomes such as earnings, health, networks, happiness. In this world it is likely that $Cov(\beta, D) \neq 0$ even though individuals do not directly take into account their health return when making education decisions. In other words, sorting on gain should be interpreted in the broad sense of a non-zero correlation between treatment and returns, resulting in $TT \neq ATE$. In the setting of CR13 this is not implausible.

In the CR13 context, monotonicity holds if the schooling reforms induce all individuals to get more years of schooling, or at least not less of it. The authors do not discuss monotonicity but we can use information provided in the paper to scrutinise the assumption. The table in Figure 10 is extracted from their paper and shows the estimated effect of the 1947 compulsory schooling reform. The first column reports the effect on years of schooling: the positive coefficients clearly show that both reforms increased the average years of schooling. The following columns report the effect by 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. In the remaining of this section we focus on this sub-population.

Discussion of types We can use the results in Figure 10 to consider counterfactuals. There are several possible types of individuals based on actual and counterfactual behaviour which we summarize in Table 5. The sign in each cell indicates the change in years of schooling if an individual is born before vs. after April 1933 (1947 reform). The reform certainly increased the average years of schooling. Monotonicity thus requires that no one belongs to a cell below the main diagonal (=), because that would imply a decrease in schooling due to the reform (-).

IMPACTS OF THE COMPULSORY SCHOOLING CHANGES ON EDUCATION

	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

Robust standard errors clustered by month-year of birth are presented in parentheses.

Figure 10: Clark and Royer (2013) page 2103

Table 5: 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, while the (-) term indicates that individuals would attain less schooling.

It is plausible that no individual who would attain 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. Nonetheless, we cannot exclude 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 10.[24]

Stochastic dominance test To complement our discussion we apply the stochastic dominance test as suggested by Angrist and Imbens (1995). The (type) independence and monotonicity assumptions combined have a testable implication whenever the treatment takes more than two values: stochastic dominance (SD) of observed treatment outcomes under different values of the instrument.[25] Combining the proportions changing schooling and the pre-reform outcome means in Figure 10 (panel 2) allows to derive the CDFs for men born before and after April 1933. This is illustrated in Figure 11: the CDFs clearly cross.

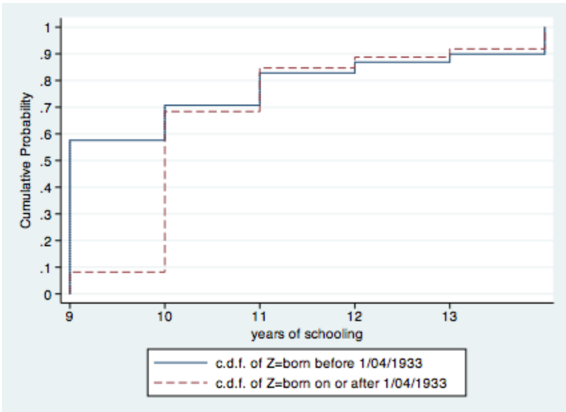


Figure 11: Stochastic dominance

We can formally test whether the crossing is statistically significant by applying the 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 post-reform cohorts stochastically dominates the CDF for the pre-reform cohorts. The test statistic for first-order stochastic dominance is given by:

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

[24]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 a large number of individuals taking more education after the reform.

[25]We refer the reader to section 6 for testing when the treatment is binary.

Barrett and Donald (2003) show that one can compute a p-value by $\exp(-2(\widehat{S})^2)$.[26] CR13 have a sample of N = 33,337 men. Assuming $N_B = N_A = N/2$ and given $\sup_x (F_A(x) - F_B(x)) = 0.021$, we obtain a test statistic of 1.91713 with a p-value of 0.00064. Thus we reject stochastic dominance.[27] The test indicates that either independence or monotonicity are violated. In the discussion that follows we assume independence holds but monotonicity is violated.[28]

5.1.2 Interpreting β^{IV} under a violation of monotonicity

CR13 interpret their RD estimates as the causal impact of education on health, driven by a reform that affected a large share of the relevant cohorts. *"Therefore, our estimates should be closer to the population-average effects of education on health [. . .] than to the effects for smaller subpopulations that maybe be of limited interest"* (p. 2088). In this section we investigate to what extent a violation of monotonicity might alter this interpretation.

As a starting point, it is useful to spell out what the RD estimand measures. The treatment S is a multivalued random variable measuring years of schooling, with support $\{9, 10, \ldots, 14\}$.[29] The CR13 fuzzy RD design compares individuals born before the 1 April 1933 cutoff with individuals born on or after that date who face a different compulsory schooling age. Even though the authors use a bandwidth of 105 months for males, they include a linear trend in month-year of birth which allows them to interpret any measured impact at the limit (v_0), i.e., at the cut-off date 1 April 1933. Formally, $S(v_0 - e)$ and $S(v_0 + e)$ are the counterfactual schooling outcomes for an individual born on either side of the cut-off. Let $Y(k)$ be the health outcome of an individual who obtains k years of schooling. In what follows we assume one-way flows between consecutive years of schooling to simplify the discussion. This restricts individuals to move only in one direction within a pair of consecutive schooling levels, but it allows the

[26] See also Donald, Hsu, and Barrett (2012) for on overview of different methods for testing stochastic dominance, including testing conditional on covariates.

[27] 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).

[28] CR13 provide arguments in support of independence: (i) by including a linear trend in month-year of birth, they are effectively comparing individuals born only one month apart; (ii) the April threshold relevant to the 1947 reform is not a threshold that matters for school entry age, (iii) the estimates are not sensitive to the chosen bandwidth and (iv) they do not find discontinuities in the predetermined characteristics at birth.

[29] We assume for simplicity that no one obtains less than 9 years and no one with more than 14 years of schooling changes schooling in response to the reform.

direction to differ across pairs.[30] The RD estimand then simplifies to:

$$\beta^{RD} = \lim_{e \to 0} \Big\{ \Big(E[Y(10) - Y(9)|S(v_0 + e) \geq 10 > S(v_0 - e)] \times 0.478 \tag{9}$$
$$+E[Y(11) - Y(10)|S(v_0 + e) \geq 11 > S(v_0 - e)] \times 0.019$$
$$-E[Y(12) - Y(11)|S(v_0 - e) \geq 12 > S(v_0 + e)] \times 0.021$$
$$-E[Y(13) - Y(12)|S(v_0 - e) \geq 13 > S(v_0 + e)] \times 0.019$$
$$-E[Y(14) - Y(13)|S(v_0 - e) \geq 14 > S(v_0 + e)] \times 0.014 \Big) \times \frac{1}{\Omega} \Big\} ,$$

where

$$\Omega = \Big(P[S(v_0 + e) \geq 10 > S(v_0 - e)] + P[S(v_0 + e) \geq 11 > S(v_0 - e)]$$
$$-P[S(v_0 - e) \geq 12 > S(v_0 + e)] - P[S(v_0 - e) \geq 13 > S(v_0 + e)]$$
$$-P[S(v_0 - e) \geq 14 > S(v_0 + e)] \Big) = 0.443 .$$

Instead, the ACR is a weighted average of LATEs for the first two groups in (9), i.e. for those who increase schooling. Both in the ACR and β^{RD} the treatment effect for individuals moving from $S = 9$ to $S = 10$ receives a much higher weight. Yet $|\beta^{RD} - ACR|$ could be non-negligible if the returns for those decreasing treatment are large. As explained in section 4, monotonicity is not a concern only if there is no sorting on gain and average returns are constant across schooling levels.

Finally, it is worth noting that there would be no monotonicity violation if one were to redefine treatment as binary, with $D = 0$ for $S = 9$ and $D = 1$ for $S \geq 10$. This would exploit a large part of the variation induced by the reform, but it would answer a different policy question.

5.1.3 What can we learn from the Roy model?

To be more specific about the effect of violating monotonicity we need to assign exact values to each of the treatment effects in (9). Even though we are unsure about the exact mechanism that drives defier behaviour[31], we can use and calibrate the multivalued Roy model from section 4.1 to gain insights into the LATEs of the different types and interpret the RD estimate.

[30] This is what drives the monotonicity failure. We refer the reader to Appendix D for further discussion.

[31] 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. One possibility is that individuals suddenly constrained to stay in school by such arbitrary criteria had a behavioural response, inducing them to drop out as soon as otherwise possible. An alternative explanation is that the reform had different effects on the pre and post-reform cohorts that are distant from the cut-off and that a misspecified trend is failing to fully control for this. In this case, we can no longer maintain the interpretation at the threshold and instead we might be comparing individuals that did not share the same classroom, school system, school quality, etc. A doubling of the number of pupils enrolled in school at age 15 might hurt the quality of education at this age, through for example class size, peer motivation and teacher quality. This in turn could reduce some pupils' motivation and skills needed to obtain higher levels of post-compulsory (11+ years of) education.

We simplify the CR13 setting by focusing on 3 treatment levels: $\mathcal{D} = \{0, 1, 2\}$, with $D = 0$ if $S = 9$, $D = 1$ if $S = 10, 11$, and $D = 2$ if $S \geq 12$. This simplified Roy model is sufficient to provide key insights into the RD estimate. We focus on the CR13 result regarding mortality rates, where $\beta > 0$ captures a reduction in health outcomes. To be consistent with the Roy model from section 4.1, we reverse the sign of the outcome such that $\beta > 0$ is a gain in health instead.

Discussion of types There are 9 possible types under the alternative values of the binary instrument $(t_A, t_B, \ldots, t_I)$. Under our one-way flow assumption, the second panel in Figure 10 suggests that the reform pushes a large group of men to increase D from 0 to 1, while a small group of more educated men reduces D from 2 to 1.[32] Table 6 highlights the existing types in bold and indicates proportions in square brackets: 47.8% are type t_B (complier-type) and 5.4% are type t_H (defier-type) who respond to the reform. There are no other complier or defier types. Types t_A, t_E and t_I exist but their schooling is unaffected by the reform.

Table 6: Multivalued Roy with $D = 0, 1, 2$ - Types in CR13

	Born after April 1933			
Born before April 1933	0	1	2	$\sum$
0	$\mathbf{t_A}(=)$ 0.092	$\mathbf{t_B}(+)$ 0.478	$t_C(+)$ 0	0.570
1	$t_D(-)$ 0	$\mathbf{t_E}(=)$ 0.196	$t_F(+)$ 0	0.196
2	$t_G(-)$ 0	$\mathbf{t_H}(-)$ 0.054	$\mathbf{t_I}(=)$ 0.18	0.234
$\sum$	0.092	0.728	0.180	1

The values in the table reflect the proportions of each type.

Define $\overline{\beta}_{t_j} \equiv E[Y(k+1) - Y(k) \mid t_j]$. Then:

$$\beta^{RD} = \left(\overline{\beta}_{t_B} \times p_{t_B} - \overline{\beta}_{t_H} \times p_{t_H} \right) \times \frac{1}{\Omega} \tag{10}$$
$$with\ \Omega = p_{t_B} - p_{t_H} = 0.478 - 0.054 = 0.424$$

while the ACR is the LATE for type B ($\overline{\beta}_{t_B}$) since they are the only complier-type.

Recall the multivalued Roy model

$$\begin{aligned}
D = 0 \quad & if \quad C(1) - \gamma_i(1) Z_i \geq \beta_{i,0\to1} \\
D = 1 \quad & if \quad C(1) - \gamma_i(1) Z_i < \beta_{i,0\to1} \quad AND \quad C(2) - \gamma_i(2) Z_i \geq \beta_{i,1\to2} \\
D = 2 \quad & if \quad C(2) - \gamma_i(2) Z_i < \beta_{i,1\to2}
\end{aligned}$$

We now make the following assumptions to help calibrate the model:

[32] To keep the model tractable, we ignore the small fraction of compliers increasing S from 10 to 11 (they remain at treatment level $D = 1$ in our Roy model) and bunch the various groups of defiers at $S = 12, 13, 14$ into the group decreasing D from 2 to 1. As will become clear later on, this grouping will tend to understate the impact of defiers on β_{RD}.

EH1. β is normally distributed: $\beta \sim N(\mu_\beta, \sigma_\beta)$

EH2. The individual health return is constant over D: $\beta_i = \beta_{i,0\to1} = \beta_{i,1\to2}$

EH3. The cost of schooling is increasing in D: $C(1) < C(2)$

EH4. The impact on the instrument $\gamma_i(k)$ is a binary parameter: $\gamma_i(k) \in \{\gamma_L, \gamma_H\}$ with $\gamma_L(k) < 0$ and $\gamma_H(k) > 0, \quad \forall k = 1, 2$.

EH5. There are one-way flows between consecutive years of schooling. For all individuals, the reform reduces the cost of obtaining $D = 1$ while it increases the cost of obtaining $D = 2$: $p_{\gamma_L}(1) = 0$, and $p_{\gamma_L}(2) = 1$.

Assumption EH5 implies that the impact of the instrument is homogeneous across individuals at any level of the treatment: $\gamma_i = \{\gamma_H(1), \gamma_L(2)\} \quad \forall i$. Monotonicity is violated because the impact changes across treatment levels, affecting individuals differently depending on their gain from treatment β. We now calibrate the parameters of this Roy model to fit the proportion of each type and the RD estimate:[33]

- $p_{t_A} = P[\beta \leq C(1) - \gamma_H(1)] = 0.092$
- $p_{t_B} = P[C(1) - \gamma_H(1) < \beta \leq C(1)] = 0.478$
- $p_{t_E} = P[C(1) < \beta \leq C(2)] = 0.196$
- $p_{t_H} = P[C(2) < \beta \leq C(2) - \gamma_L(2)] = 0.054$
- $p_{t_I} = 1 - \sum_{j=A,B,E,H} p_{t_j} = P[C(2) - \gamma_L(2) < \beta] = 0.180$
- $\beta^{RD} = \dfrac{\overline{\beta}_{t_B} \times p_{t_B} - \overline{\beta}_{t_H} \times p_{t_H}}{p_{t_B} - p_{t_H}} = -0.009$

We can calibrate the parameters of the model up to a normalization. Here we normalize σ_β to 1. Intuitively, for any given β-distribution $(\mu_\beta, \sigma_\beta)$, the parameters $C(1)$, $C(2)$, $\gamma_H(1)$ and $\gamma_L(2)$ are set to match the proportions of each type, with the LATEs determined accordingly. Fitting the β^{RD} estimate then pins down μ_β by assigning a suitable LATE to types B and H. Table 7 shows the solution to this calibration exercise and Figure 12 shows the types over the β distribution.

The ATE is positive and $\gamma_H(1) > |\gamma_L(2)|$ to match the larger proportion of complier-type B. These are the low educated men induced to take more schooling by the reform. Their LATE is below the ATE but positive overall, although some of them have a negative return: $0 < \overline{\beta}_{t_B} < ATE$. In addition, sorting on gain implies that the defier-type H is instead located higher up in the β-distribution, with $\overline{\beta}_{t_H} > ATE$, pulling down the numerator in equation

[33] The β^{RD} moment is the impact on the mortality rate for men, taken from CR13 (Table A.2 in the Online Appendix: mortality rate, by sex).

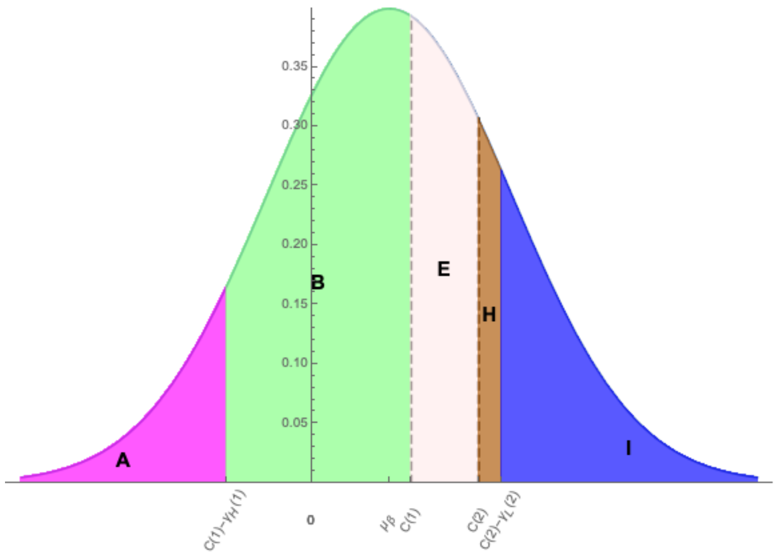


Figure 12: Types over the β distribution - CR13

Table 7: Model coefficients and LATEs

(a) Model coefficients

$C(1)$	$C(2)$	$\gamma_H(1)$	$\gamma_L(2)$	μ_β (ATE)	σ_β
0.809	1.358	1.505	-0.190	0.632	1

(b) LATEs

β^{RD}	$\overline{\beta}_{t_A}$	$\overline{\beta}_{t_B}$ (ACR)	$\overline{\beta}_{t_E}$	$\overline{\beta}_{t_H}$	$\overline{\beta}_{t_I}$
-0.009	-1.162	0.156	1.072	1.450	2.090

(c) β^{RD} vs Estimands of Interest

$\beta^{RD} - ACR$	$\frac{\beta^{RD}-ACR}{\mu_\beta}$	$\beta^{RD} - ATE$	$\frac{\beta^{RD}-ATE}{\mu_\beta}$
-0.156	-0.261	-0.641	-1.014

(10). As a result, β^{RD} is 0.165 of a standard deviation smaller than the ACR even though the proportion of the defier-type is small. Hence, our model provides two potential reasons for the small and insignificant health return to schooling found by CR13: the ATE is larger than the ACR, and the monotonicity violation is causing the ACR to be larger than β^{RD}.[34]

5.2 IV using School Entry Age regulation

Black, Devereux, and Salvanes (2011), henceforth BDS11, investigate the effect of school entry age ($D \equiv EA$) on military IQ test scores and adult outcomes (Y) in Norway. To deal with endogeneity, the authors use Legal Entry Age as an instrument ($Z \equiv LEA$). LEA is the age at which a child can 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 different 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, i.e. December-January. Note that only month of birth is observed, so the authors cannot narrow the sample any further around the cutoff date, nor include a trend in the exact date of birth. The approach using the discontinuity sample is implemented to account for potential manipulation of the date of birth by parents and the seasonality of births.[35] In this section we focus on the discontinuity sample, a setting with binary instrument and multivalued treatment. Figure 13 plots the LEA by month of birth: the LEA is fully determined by the date of birth.[36] 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.

Parents might make school entry decisions based on some knowledge of the gain from starting school later. This gain could depend on the intellectual and emotional maturity of the child, or on the relative age of her classmates. The practice of delayed school entry is also known as

[34] A few remarks about our model. First, grouping defiers in the simplified $D = 0, 1, 2$ model actually understates the impact of the defier-type on the RD estimate: defiers at $S > 12$ will have a higher LATE than those currently modelled. For instance, with $\mathcal{D} \in \{0, 1, \ldots, 4\}$, the calibration would lead to $C(4) > C(3) > C(2)$ and defiers spread across these three margins. Since $C(2)$ remains unchanged in that case, it would result in an overall larger LATE for the defier-types. In addition, ignoring the complier-type moving from 10 to 11 years of schooling (see Table 10) reduces the LATE for the complier-type. Hence we somewhat understate the LATEs of both complier and defier-types. Second, changing the value of σ_β rescales the parameters but does not alter the qualitative conclusions of this section. Third, a reduced degree of sorting on the health gain would bring $\overline{\beta}_{t_H}$ closer to $\overline{\beta}_{t_B}$ but again would not alter the qualitative conclusions.

[35] 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 unmarried parents. If date of birth is not random, instruments relying on it are likely to violate the independence assumption. BDS11 include family characteristics in the regression and show that the resulting estimates are very close to estimates without these controls. In addition, they are also able to include family fixed effects.

[36] 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.

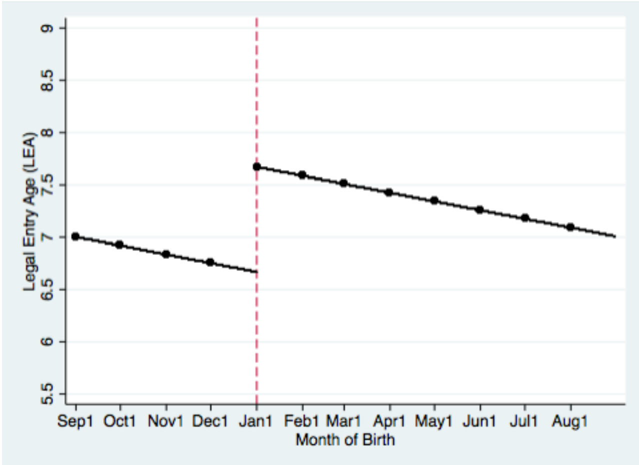


Figure 13: LEA by month of birth

red-shirting. The BDS11 setting plausibly a context with essential heterogeneity: (i) the gain from treatment is heterogeneous across the population and (ii) there is some degree of sorting into treatment based on the gain.

A variety of studies use the same intuition to estimate the causal effect of school entry age for different countries and 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), Fredriksson and Öckert (2014) and Dee and Sievertsen (2018). 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.[37]

5.2.1 Plausibility of the Monotonicity Assumption

Although BDS11 do not discuss monotonicity, the paper contains useful information. The table in Figure 14a 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.

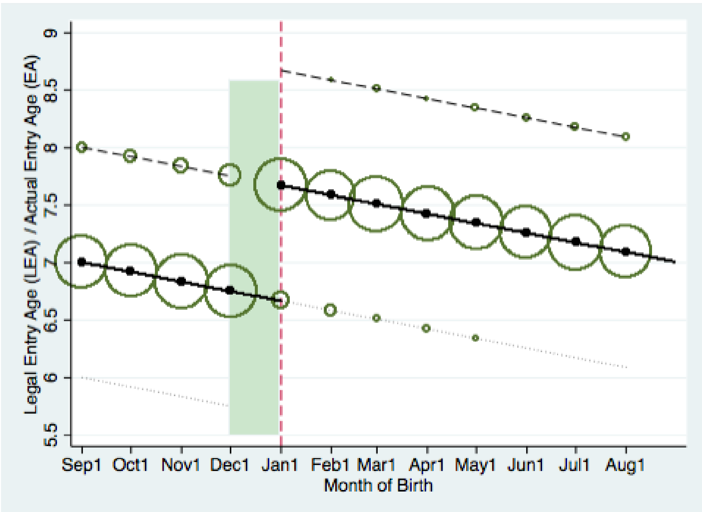
This entry age behaviour is consistent with parents/educators making school-entry decisions based on either a child's absolute or relative age. Figure 14b replicates Figure 13 but we now add the observed EA patterns as shown in the table. The size of the circles mirrors the proportions

[37]Whether monotonicity fails is, nevertheless, highly context-dependent. The discussion below provides insights into monotonicity and implications of a monotonicity failure in the context specific to BDS11.

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

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



(b) LEA and observed school Entry Age

Figure 14: Monotonicity in Black, Devereux, and Salvanes (2011)

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. This figure also clearly illustrates that all children are affected by the instrument in this setting: EA necessarily changes with LEA.

Discussion of types We now focus on December and January born children and consider counterfactuals to assess 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 behaviour. Table 8 shows these types denoted by t_j for $j = A, \ldots, I$. The sign in each cell indicates the change in EA if a child is born in January rather than December. Thus, 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 t_E: $EA_i(Jan) > EA_i(Dec)$.[38] Assuming type independence, observed behaviour of January-born children can function as a counterfactual for actual December-borns, and vice versa. Figure 14b thus suggests that type E is the most prevalent type. Since late school entry among January-borns does not occur, types C, F and I are absent in this setting. Similarly, since early school entry among December-borns does not occur, types A and B are also absent. Finally, it is highly improbable that December-born late entrants would instead enter school early had they been born in January. Relying on this judgement regarding plausible behaviour, we rule out type G. Thus 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: $EA_i(Jan) < EA_i(Dec)$. 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. Figures 14a and 14b again suggest that the existence of this type cannot be ruled out. Crucially, the existence of either defier-type D ($-$) or H ($-$) alongside the more numerous complier-type E (+) creates a violation of monotonicity.

Stochastic dominance test From Figures 14a and 14b we can derive the school entry age under each value of the instrument. Thus for children born December 1st the support of EA is $\{5.75, 6.75, 7.75\}$ depending on whether they enter early, on-time or late respectively. Similarly, for children born January 1st the support of EA is $\{6.67, 7.67, 8.67\}$. We can then use the proportions in Figure 14a 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 15 shows that the CDFs cross. Let N_D and N_J be the number of the December

[38] In contrast to Table 3 in the multivalued Roy model in section 4.1, here everyone is affected by the instrument and therefore the types on the main diagonal face an increase in treatment.

Table 8: 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.

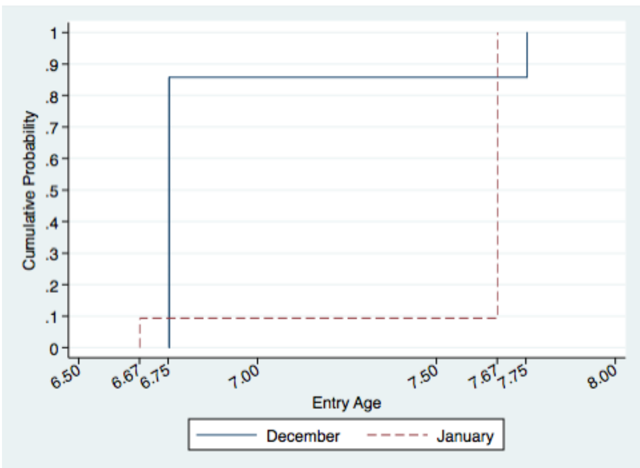


Figure 15: Stochastic Dominance under alternative values of the instrument

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 Barrett and Donald (2003) 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. From the paper we know that $\sup_x (F_J(x) - F_D(x)) = 0.15$ and assuming that $N_D = N_J = N/2$ leads to a test statistic of 24.189 compared to a critical value of 1.517 at a 1% level of significance. The p-value is zero, leading us to reject stochastic dominance.[39] Together with the earlier discussion, this evidence raises concerns about the internal validity of the identification approach.[40]

5.2.2 Interpreting β^{IV} under a violation of monotonicity

The data provide evidence that monotonicity does not hold. BDS11 (p. 458) state *"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.

In an empirical setting it is generally impossible to identify how children's entry age and outcomes are differently affected by the instrument and treatment respectively. This is because counterfactuals are unobserved. However, 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 and their relative proportions. This is illustrated in the table below.

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

[39] 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).

[40] Alternatively, one can test for stochastic dominance using a Wald test. In this setting, the CDF for the January born children stochastically dominates the CDF for the December born children only if $F_J(6.67) = 0$ and $F_D(6.75) = 1$. The Wald test also rejects stochastic dominance with a p-value of 0.

	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

$t_D(-)$ 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: $(EA(Jan) = 6.67, EA(Dec) = 6.75)$. 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: $(EA(Jan) = 7.67, EA(Dec) = 7.75)$. They form 15% of the discontinuity sample: $p_{t_H} = 0.15$.

Using the estimand for multivalued treatment and assuming that the return to entering one month later is constant over age for the observed points of support, we can express β^{IV} in terms of yearly rather than monthly return.[41] Define $\overline{\beta}_{t_j} \equiv LATE_{t_j} = E[Y(EA = k+1\ year) - Y(EA = k) \mid t_j]$. Then[42]

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

while $ATE = p_{t_E}\overline{\beta}_{t_E} + p_{t_D}\overline{\beta}_{t_D} + p_{t_H}\overline{\beta}_{t_H}$ and $ACR = \overline{\beta}_{t_E}$ since only one complier-type exists.

If monotonicity does not hold but β is homogeneous or there is no sorting on gain, such that $\overline{\beta}_{t_j}$ is identical across types, then $\beta^{IV} = ATE$ since,

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

If monotonicity does not hold, β is heterogeneous and there is sorting on gain, then β^{IV} is different from the ATE. In BDS11, β^{IV} is likely 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.

[41] 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 discussion of IV estimation with non-constant effects.

[42] Proof in Appendix D.2.

Without additional information about the different LATE'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 $\overline{\beta}_{t_E}$.

5.2.3 What can we learn from the Roy model?

In this section we present a simple Roy model adapted to the BDS11 setting, where the treatment is multivalued but individuals essentially make a binary decision (On Time school entry vs. not). To help calibrate the model, we assume that:

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

EA2. The return to entering school one month later is constant over the relevant age interval $[6.67, 7.75]$.

EA3. There is a strictly enforced social norm or law that no individual can enter school younger than six or older than eight. This captures the observation that no child is observed entering school outside of the 6-8 age interval.

EA4. There is a cost from not entering school On Time: C_e is the cost of entering early, while C_ℓ is the cost of entering late. These could be psychic costs of deviating from social norms, time or monetary costs. For instance, 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 entry.

EA5. No December-born late entrants would instead enter school early had they been born in January (type G in section 5.2.1). This assumption is needed to identify the proportions of each type as discussed above.

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 formal or informal child care system and enter the school system or whether to wait 1 more year. These children exit if the marginal benefit of waiting one more year (β) does not exceed the marginal cost (C). In Table 9 we describe the potential and observed entry age choices.

Table 9: Potential and observed entry age values

Entry age		Observed	Entry age		Observed
December born			January born		
Early	5.75	No	Early+	5.67	No
On Time	6.75	Yes	Early	6.67	Yes
Late	7.75	Yes	On Time	7.67	Yes
Late+	8.75	No	Late	8.67	No

Hence December borns face a binary choice On Time vs Late start: they choose On Time if

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

while January borns face a binary choice Early vs On Time school start: they choose Early if

$$Y(7.67) \le Y(6.67) - C_e \Rightarrow \beta \le -C_e \ ,$$

which implies that these children have a negative return from entering school one year later.[43] To identify the parameters of $f(\beta)$ we then exploit the following restrictions taken from BDS11:

- $p_{t_D} = P[\beta \le -C_e] = 0.1$
- $p_{t_H} = P[\beta > C_\ell] = 0.15$
- $p_{t_E} = 1 - \sum_{j=D,H} p_{t_j} = P[-C_e < \beta \le C_\ell] = 0.75$
- $\beta^{IV} = \frac{\frac{11}{12}\overline{\beta}_{t_E} p_{t_E} - \frac{1}{12}\overline{\beta}_{t_D} p_{t_D} - \frac{1}{12}\overline{\beta}_{t_H} p_{t_H}}{\frac{11}{12}p_{t_E} - \frac{1}{12}p_{t_D} - \frac{1}{12}p_{t_H}} = \widehat{\beta}^{IV} = 0.167$.[44]

As in the CR13 setting, we can calibrate the parameters of the model up to a normalization. Again, we normalize σ_β to 1. For any given β-distribution $(\mu_\beta, \sigma_\beta)$, the cost parameters C_ℓ, C_e are set to match the proportions of each type. Then μ_β is set to ensure the $\beta^{IV} = 0.167$ restriction is satisfied.[45]

The solution to this problem is shown in Table 10 and in Figure 16. Since there is a cost from not starting school On Time, children who do so must have either very negative returns $(\overline{\beta}_{t_D})$ or very positive returns $(\overline{\beta}_{t_H})$ from starting school one year later. The average return to postponing school entry with one year is 0.25. 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 34% of the ATE. The difference $(\beta^{IV} - \overline{\beta}_{t_E})$ is about 0.1% of a standard deviation, or 4% of the ATE. Hence β^{IV} is fairly close to the ACR in spite of the monotonicity violation, albeit somewhat more distant from the ATE. In addition to the two reasons mentioned earlier, this result is also driven by the complier-type being located in between the two defier-types. Since the LATEs of the defier-types have opposite signs, their contribution to β^{IV} in (11) partly cancels out. Yet, $0 < \beta^{RD} < ACR$ since the monotonicity failure reduces the numerator more than the denominator.

[43] C_e could also be negative, implying a net benefit from having a child enter school ahead of time. Assumption EA5 rules out type G children. This assumption is violated if $C_\ell < \beta < -C_e$ for some child. 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 and viceversa.

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

[45] The data used in BDS11 is not publicly available. Therefore, there is no additional information that can be used to identify the parameters of the model or to relax the assumptions made earlier.

Table 10: Model coefficients and LATEs

(a) Model coefficients

C_e	C_ℓ	μ_β (ATE)	σ_β
1.028	1.29	0.253	1

(b) LATEs

β^{IV}	$\overline{\beta}_{t_D}$	$\overline{\beta}_{t_E}$ (ACR)	$\overline{\beta}_{t_H}$
0.167	-1.501	0.177	1.808

(c) β^{IV} vs Estimands of Interest

$\beta^{IV} - ATE$	$\frac{\beta^{IV} - ATE}{\mu_\beta}$	$\beta^{IV} - ACR$	$\frac{\beta^{IV} - ACR}{\mu_\beta}$
-0.086	-0.341	-0.01	-0.038

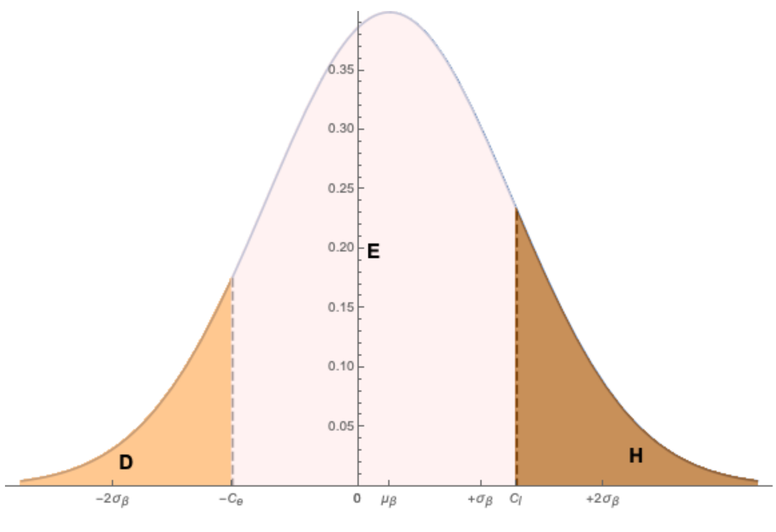


Figure 16: Types over the β distribution - BDS11

Link to Roy Model in Section 3.1 We can also express the multivalued treatment level as a function of the binary choice, linking to the Roy model in section 3.1.

$$
\begin{aligned}
EA &= 6.75 + 1 \times \mathbf{I}[\beta > C_\ell] + \left(\frac{11}{12} - 1 \times \mathbf{I}[\beta \leq C_e] - 1 \times \mathbf{I}[\beta > C_\ell] \right) Z \\
&= 6.75 + 1 \times \mathbf{I}[\beta > C_\ell] + \gamma Z; \qquad \gamma \in \left\{ \frac{11}{12}, \frac{-1}{12} \right\}
\end{aligned}
$$

where $\mathbf{I}[.]$ is an indicator function. This formulation of the model makes explicit that i) entry age changes value with the instrument even if the individual always enter school On Time and ii) the impact of the instrument (γ) is a deterministic function of the gain (β).

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.[46] 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 in the first stage. Since children are more likely to start late (early) the closer they are born to the left (right) of the discontinuity, the trend in EA 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

[46] Their discontinuity 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.

ideal 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.[47]

Barua and Lang (2016) We are not the first ones to discuss the monotonicity condition in the school entry age setting. Barua and Lang (2016) 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 (2016) 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 data 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 more 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. We go further by emphasizing the role of sorting on gain in school entry decisions, how this leads to the different types of individuals and consequently to the violation of the monotonicity assumption. While they conclude that, in the absence of monotonicity, empirical studies do not provide consistent estimates of the LATE, we find that β^{IV} is fairly close to a LATE of interest albeit not close to the ATE.

6 Recommendations for applied researchers

This section describes a thought process for the applied researcher seeking to scrutinise the monotonicity assumption and its implications. Figure 17 summarises the main ideas.

Q1. **Can Essential Heterogeneity ($Cov(\beta, D) \neq 0$) be ruled out?**
Can either heterogeneous treatment effects or selection into treatment based on the gain be ruled out? If so, monotonicity is not a required assumption: $\beta^{IV} = ATE$.[48] Two tests of essential heterogeneity have been suggested in the literature. Both apply to the binary treatment case. Heckman, Urzua, and Vytlacil (2006) present a propensity score-based test that relies on either a multivalued or continuous instrument, or a vector of instruments, while Mourifié, Henry, and Méango (2020)'s test uses selection shifters that

[47] Gelman and Imbens (2018) and Imbens and Kalyanaraman (2012) argue against the use of high-order polynomials in the Regression Discontinuity design.

[48] If the treatment is multivalued, an ATE interpretation also requires average returns to be constant across treatment levels.

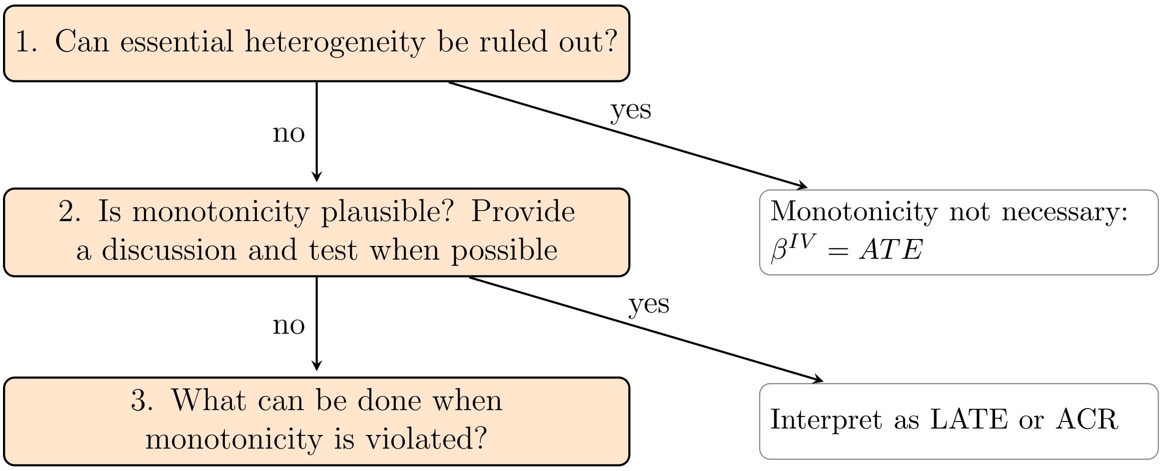


Figure 17: Recommendations for applied researchers

might be distinct from the instrument used for identification.[49]

Q2. **Is monotonicity plausible? Provide a discussion in your context and test when possible.**
What does monotonicity imply in your context? Can you think of a reason why it might be violated?
Whenever possible, one should also test for the monotonicity assumption. The tests available in the literature are joint tests of instrument independence and monotonicity. These two conditions are refutable but non-verifiable. As a result, we recommend a combination of economic insights and formal testing to assess the monotonicity condition whenever possible.

- Multivalued treatment case: stochastic dominance as suggested by Imbens and Angrist (1994) and fornally tested by Barrett and Donald (2003).
- Binary treatment case: see Kowalski (2019) when the instrument is binary, Kitagawa (2015) when the instrument is discrete, Huber and Mellace (2015) and Mourifié and Wan (2017) when the instrument is either binary or discrete.

If monotonicity seems plausible and is not refuted by the test, the estimate can be interpreted as a LATE or ACR .

Q3. **What can be done when monotonicity is violated?**
In this case IV or RD strategies do not generally identify a causal treatment effect. Now what can be done? Here we make a few alternative suggestions:

- Spell out the IV or RD estimand and quantify the weights attached to the different types of individuals. This step can guide interpretation.

[49]In the empirical examples discussed, we do not implement the suggested tests since the treatment is not binary and we do not have access to the data used in the original analyses.

- Give more structure. For instance, define and calibrate a Roy model to interpret the IV or RD estimate as illustrated in section 5. Even in the absence of a full calibration, this step can help locate the average treatment effects of the complier and defier-types.
- Redefine the treatment in a way that avoids a monotonicity failure. Example are provided in both empirical applications discussed in section 5. A different treatment definition might, however, answer a very different question.
- Consider whether your setting classifies as a special case of monotonicity failure that still allows to uncover a LATE of interest: compliers-defiers condition (de Chaisemartin (2017)) or local monotonicity (Dahl, Huber, and Mellace (2017)), both are discussed in section 3.4; or violations of monotonicity at random, which allows to approximate the bias caused by monotonicity failure (Klein (2010)).[50]
- If none of the above yields sufficient insights, an alternative estimation strategy might be needed. See e.g. Kline and Walters (2019) for an overview.

7 Conclusion

Identification of a LATE (or ACR) in an IV or fuzzy RD design with essential heterogeneity relies on the monotonicity assumption. In this paper, we provide a framework to scrutinize monotonicity and assess the impact of a monotonicity failure on the interpretation of the estimate. We focus on the case with a binary instrument, and either binary or multivalued treatment. First, using a Roy model for treatment selection, we find that interpretation of the estimates can be lost even under limited essential heterogeneity and for minor violations of monotonicity. We also show that a larger first stage impact does not necessarily counteract a departure from monotonicity. Second, we investigate monotonicity in two applied studies using economic insights and data analysis. Clark and Royer (2013) exploit changes in compulsory schooling laws to investigate the effect of education on health in a fuzzy RD setting. Since the reform affected a large share of the relevant cohorts, the authors suggest that their estimate should be closer to the ATE than a LATE. Black, Devereux, and Salvanes (2011) use school entry age cutoffs as an instrument to investigate the effect of entering school older on IQ test scores and adult outcomes. Given the high compliance rates, they also argue that their IV estimates can be interpreted as an approximation to the ATE rather than a LATE. We find that the data patterns reported in these papers are indicative of a monotonicity failure and confirm this using formal testing. By constructing and calibrating a Roy model, we conjecture that the authors' suggested interpretation of their estimates might be incorrect, and propose an alternative one. We conclude our paper with a set of recommendations for the empirical researcher seeking a clear framework to scrutinize the monotonicity assumption.

[50] We do not discuss Klein (2010) further because his setting does not translate easily into our Roy model.

Appendix A Generalizations of the binary treatment and instrument IV setting

A.1 Monotonicity in the Fuzzy RD design

Consider the potential outcomes framework with a binary treatment (as in section 2), but with Z taking on a continuum of values ($v \in \mathcal{Z}$). In an RD design we exploit the discontinuity in the treatment probability at a threshold $Z = v_0$:

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

with the difference being smaller than 1.[51] Hahn, Todd, and Van der Klaauw (2001) point out that both the fuzzy Regression Discontinuity and the IV estimands can be expressed as a Wald estimand:

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

which measures a similar LATE to $\beta^{IV}(z, w)$ if $z = v_0 + \epsilon$, $w = v_0 - \epsilon$ and ϵ is arbitrarily small. Under essential heterogeneity, regression discontinuity estimation identifies β for those individuals with $Z = v_0$ who are affected by the threshold (LATE at v_0) under the following conditions

RD1. $\lim_{v \downarrow v_0} E[D|Z = v] \neq \lim_{v \uparrow v_0} E[D|Z = v]$ (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. 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) and keeping the conditioning on the running variable implicit, we can interpret the RD estimate of β as

$$\begin{aligned} \beta^{RD} = \lim_{e \to 0} \Big\{ & \lambda \times E[Y(1) - Y(0)|D(v_0 + e) - D(v_0 - e) = 1] + \\ & (1 - \lambda) \times E[Y(1) - Y(0)|D(v_0 + e) - D(v_0 - e) = -1] \Big\} \end{aligned} \tag{12}$$

[51] 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.

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 (12) is the equivalent of equation (1) 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-*in* or LATE-*out* nor a weighted average of the two LATEs.

A.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[52]

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

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.[53] Condition 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 as a weighted average of Wald estimators:

$$\beta^{IV} = \sum_{j=1}^{J} \mu_j \beta^{IV}(j, j-1) \tag{13}$$

where

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

[52]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.

[53]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".

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 (13) indicates that, under a binary treatment,

- when monotonicity IV3 and IV4 are satisfied then β^{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^{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^{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 B 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 of 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 of 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 ($P[-\gamma_H < \beta \leq 0]$).[54]

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

- ATE of 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).

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

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

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

  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:

$$w_{AT} = \frac{p_{\gamma_H} \ P[0 < \beta < -\gamma_L]}{p_{AT}}$$

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

$$w_{NT} = \frac{p_{\gamma_L} \ P[-\gamma_H < \beta < 0]}{p_{NT}}$$

$$1 - w_{NT} = 1 - \frac{p_{\gamma_L} \ P[-\gamma_H < \beta < 0]}{p_{NT}} = \frac{P[\beta < -\gamma_H]}{p_{NT}}$$

Appendix C RD in the Roy model

The same Roy model can be used to explain the importance of monotonicity in a fuzzy Regression Discontinuity Design. Let Z be a continuous variable. At the threshold v_0, 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.[55]

C.1 Sharp RD

In a sharp design, treatment is known to depend in a deterministic way on Z. For instance, near the threshold v_0 all individuals with $Z > v_0$ take treatment while those with $Z \leq v_0$ do not take treatment. This is an example where all individuals are compliers. The model above would generate such a sharp RD when β is a negative random variable, γ is a positive one, and $|\min\{\beta\}| < \min\{\gamma\}$ so that the instrument's impact is strong enough to push everyone into treatment.[56]

C.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 the IV case, 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, at the threshold v_0 the RD estimate of β can again be expressed as:

[55] Here we focus on the threshold but the selection equation could be more complex elsewhere, for instance including a function of the running variable Z.

[56] Alternatively, a sharp RD can originate if there is no sorting on gain and if $\gamma > 0$ and homogeneous. Intuitively, sharp RD is likely to emerge in settings where there are strict rules based on Z. Thus, individuals have no discretion in selecting into treatment resulting in no sorting on gain.

Table 11: Counterfactual Choices - RD

$\gamma = \gamma_L$			
	$\beta \leq 0$	$0 < \beta \leq -\gamma_L$	$\beta > -\gamma_L$
$\lim_{e \to 0} Z = v_0 + e$	$D = 0$	$D = 1$	$D = 1$
$\lim_{e \to 0} Z = v_0 - e$	$D = 0$	$D = 0$	$D = 1$
type	NT	DF	AT
$\gamma = \gamma_H$			
	$\beta \leq -\gamma_H$	$-\gamma_H < \beta \leq 0$	$\beta > 0$
$\lim_{e \to 0} Z = v_0 + e$	$D = 0$	$D = 0$	$D = 1$
$\lim_{e \to 0} Z = v_0 - e$	$D = 0$	$D = 1$	$D = 1$
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 with the IV setting is that one relies on observations at the limit. The insights of sections 3.1-3.3 extend to the RD setting in a straightforward way.

Appendix D Applications

D.1 Clark and Royer (2013)

The treatment S is a multivalued random variable measuring years of schooling, with support $\mathcal{S} = \{9, 10, 11, 12, 13, 14\}$. Applying equation (4) from Section 4 to the regression discontinuity setting:

$$\beta^{RD} = \lim_{e \to 0} \sum_{k=10}^{14} \Big\{ E[Y(k) - Y(k-1)|S(v_0+e) \geq k > S(v_0-e)] \times \frac{P[S(v_0+e) \geq k > S(v_0-e)]}{\Omega} \\ -E[Y(k) - Y(k-1)|S(v_0-e) \geq k > S(v_0+e)] \times \frac{P[S(v_0-e) \geq k > S(v_0+e)]}{\Omega} \Big\} \tag{14}$$

where

$$\Omega = \sum_{k=10}^{14} \big(P[S(v_0+e) \geq k > S(v_0-e)] - P[S(v_0-e) \geq k > S(v_0+e)] \big)$$

Information from Clark and Royer (2013) can be used to assign values to some of the moments in equation (14). The table in figure 10, however, only shows us the net flows between $k-1$ and k years of schooling. We can simplify the setting by assuming that these net flows are the actual gross flows that occurred.

RD5. for all treatment levels $k \in \mathcal{S}$, either $S_i(v_0+e) \geq k \geq S_i(v_0-e)\ \forall i$, Or $S_i(v_0+e) \leq k \leq S_i(v_0-e)\ \forall i$ (one-way flows)

Under RD5, the date inform us about which of the two-way flows can be discarded at each k, so that equation (14) simplifies to equation (9).

D.2 Black, Devereux, and Salvanes (2011)

The treatment EA is a multivalued random variable measuring age at school entry in monthly steps, with support $\{6.67, 6.67+\frac{1}{12}, 6.67+\frac{2}{12}, \ldots, 7.75\}$. Given that we focus on the discontinuity sample, the instrument is binary with the two values of the legal entry age depending on whether a child is born in December or January. Therefore, we apply equation (4). 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,

$$\beta^{IV}(Jan, Dec) = \frac{1}{\Omega} \times \sum_{s=1}^{11} \Bigg\{$$
$$E\left[Y\left(6.67+\frac{s}{12}\right) - Y\left(6.67+\frac{s-1}{12}\right) \Big| EA(Jan) \geq 6.67+\frac{s}{12} > EA(Dec)\right]$$
$$\times P\left[EA(Jan) \geq 6.67+\frac{s}{12} > EA(Dec)\right] -$$
$$E\left[Y\left(6.67+\frac{s}{12}\right) - Y\left(6.67+\frac{s-1}{12}\right) \Big| EA(Dec) \geq 6.67+\frac{s}{12} > EA(Jan)\right]$$
$$\times P\left[EA(Dec) \geq 6.67+\frac{s}{12} > EA(Jan)\right] \Bigg\}$$

where

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

Since we know there are only three types based on counterfactual school start behaviour (t_D, t_E, t_H, with t_E facing a jump of 11 months in EA in response to the instrument, while both t_D and t_H face a drop of 1 month in EA), we can rewrite the equation above as

$$\beta^{IV}(Jan, Dec) = \frac{1}{\Omega} \times \Bigg\{ \sum_{s=1}^{11} \left\{ E\left[Y\left(6.75+\frac{s}{12}\right) - Y\left(6.75+\frac{s-1}{12}\right) \Big| t_E\right] \times p_{t_E} \right\}$$
$$- E\left[Y(6.75) - E(6.67) \Big| t_D\right] \times p_{t_D}$$
$$- E\left[Y(7.75) - E(7.67) \Big| t_H\right] \times p_{t_H} \Bigg\}$$

where

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

Finally, using the assumption that an individual's return to entering one month later is constant with age over the observed points of support

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

□

References

Aakvik, Arild, James J Heckman, and Edward J Vytlacil. 2005. "Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs." *Journal of Econometrics* 125 (1-2):15–51.

Angrist, Joshua D, Kathryn Graddy, and Guido W Imbens. 2000. "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish." *The Review of Economic Studies* 67 (3):499–527.

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.

Arai, Yoichi, Yu-Chin Hsu, Toru Kitagawa, Ismael Mourifié, and Yuanyuan Wan. 2018. "Testing identifying assumptions in fuzzy regression discontinuity designs." Tech. rep., cemmap working paper.

Barrett, Garry F. and Stephen G. Donald. 2003. "Consistent Tests For Stochastic Dominance." *Econometrica* 71 (1):71–104.

Barua, Rashmi and Kevin Lang. 2016. "School Entry, Educational Attainment, and Quarter of Birth: A Cautionary Tale of a Local Average Treatment Effect." *Journal of Human Capital* 10 (3):347–376.

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.

Börklund, Anders and Robert Moffitt. 1987. "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models." *The Review of Economics and Statistics* 69 (1):42–49.

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.

Conti, Gabriella and James J. Heckman. 2010. "Understanding the early origins of the education–health gradient: A framework that can also be applied to analyze gene–environment interactions." *Perspectives on Psychological Science* 5 (5):585–605.

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg. 2016. "From LATE to MTE: Alternative methods for the evaluation of policy interventions." *Labour Economics* 41:47–60.

Dahl, Christian M., Martin Huber, and Giovanni Mellace. 2017. "It's never too LATE: A new look at local average treatment effects with or without defiers." Discussion Papers of Business and Economics 2/2017, University of Southern Denmark, Department of Business and Economics.

Datar, Ashlesha. 2006. "Does Delaying Kindergarten Entrance Give Children A Head Start?" *Economics of Education Review* 25 (1):43–62.

de Chaisemartin, C. 2017. "Tolerating Defiance? Identification Of Treatment Effects Without Monotonicity." *Quantitative Economics* 8 (2).

Dee, Thomas S. and Hans Henrik Sievertsen. 2018. "The gift of time? School starting age and mental health." *Health economics* 27 (5):781–802.

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.

Donald, Stephen G, Yu-Chin Hsu, and Garry F Barrett. 2012. "Incorporating covariates in the measurement of welfare and inequality: methods and applications." *The Econometrics Journal* 15 (1):C1–C30.

Dong, Yingying. 2018. "Alternative assumptions to identify LATE in fuzzy regression discontinuity designs." *Oxford Bulletin of Economics and Statistics* 80 (5):1020–1027.

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.

Fort, Margherita, Nicole Schneeweis, and Rudolf Winter-Ebmer. 2016. "Is Education Always Reducing Fertility? Evidence from Compulsory Schooling Reforms." *Economic Journal* 126 (595):1823 – 1855.

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.

Galama, Titus, Adriana Lleras-Muney, and Hans van Kippersluis. 2018. "The Effect of Education on Health and Mortality: A Review of Experimental and Quasi-Experimental Evidence."

Gelman, Andrew and Guido Imbens. 2018. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics* 0 (0):1–10.

Geruso, Michael and Heather Royer. 2018. "The Impact of Education on Family Formation: Quasi-Experimental Evidence from the UK." Working Paper 24332, National Bureau of Economic Research.

Grenet, Julien. 2013. "Is Extending Compulsory Schooling Alone Enough To Raise Earnings? Evidence From French And British Compulsory Schooling Laws." *The Scandinavian Journal of Economics* 115 (1):176–210.

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, Sergio Urzua, and Edward Vytlacil. 2006. "Understanding instrumental variables in models with essential heterogeneity." *The Review of Economics and Statistics* 88 (3):389–432.

———. 2008. "Instrumental variables in models with multiple outcomes: The general unordered case." *Annales d'Economie et de Statistique* :151–174.

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. 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.

Kline, Patrick and Christopher R. Walters. 2019. "On Heckits, LATE, and Numerical Equivalence." *Econometrica* 87:677–696.

Kowalski, Amanda E. 2019. "Counting Defiers." Working Paper 25671, National Bureau of Economic Research.

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.

Mehta, Nirav. 2019. "An Economic Approach to Generalizing Findings from Regression-Discontinuity Designs." *Journal of Human Resources* 54 (4):953–985.

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 (9-10):1667–1695.

Mourifié, Ismaël, Marc Henry, and Romuald Méango. 2020. "Sharp Bounds and Testability of a Roy Model of STEM Major Choices." *Journal of Political Economy* 128 (8):3220–3283.

Mourifié, Ismael and Yuanyuan Wan. 2017. "Testing Local Average Treatment Effect Assumptions." *The Review of Economics and Statistics* 99 (2):305–313.

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.

Rubin, Donald. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66.

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.

Vytlacil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica* 70 (1):331–341.

List of AER Papers from Google Scholar Search

The papers that mention monotonicity are denoted by the § symbol.

Andrews, Rodney J. 2016. "Coordinated admissions program." *American Economic Review* 106 (5):343–47. §.

Buonanno, Paolo and Steven Raphael. 2013. "Incarceration and incapacitation: Evidence from the 2006 Italian collective pardon." *American Economic Review* 103 (6):2437–65.

Carneiro, Pedro, James J Heckman, and Edward J Vytlacil. 2011. "Estimating marginal returns to education." *American Economic Review* 101 (6):2754–81. §.

Crost, Benjamin, Joseph Felter, and Patrick Johnston. 2014. "Aid under fire: Development projects and civil conflict." *American Economic Review* 104 (6):1833–56.

Deming, David J, Justine S Hastings, Thomas J Kane, and Douglas O Staiger. 2014. "School choice, school quality, and postsecondary attainment." *American Economic Review* 104 (3):991–1013.

Dinkelman, Taryn. 2011. "The effects of rural electrification on employment: New evidence from South Africa." *American Economic Review* 101 (7):3078–3108.

Dobbie, Will, Jacob Goldin, and Crystal S Yang. 2018. "The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges." *American Economic Review* 108 (2):201–40. §.

Dobbie, Will and Jae Song. 2015. "Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection." *American Economic Review* 105 (3):1272–1311. §.

Doyle Jr, Joseph J. 2007. "Child protection and child outcomes: Measuring the effects of foster care." *American Economic Review* 97 (5):1583–1610. §.

Huang, Zhangkai, Lixing Li, Guangrong Ma, and Lixin Colin Xu. 2017. "Hayek, local information, and commanding heights: Decentralizing state-owned enterprises in China." *American Economic Review* 107 (8):2455–78.

Maestas, Nicole, Kathleen J Mullen, and Alexander Strand. 2013. "Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt." *American economic review* 103 (5):1797–1829. §.

Markevich, Andrei and Ekaterina Zhuravskaya. 2018. "The economic effects of the abolition of serfdom: Evidence from the Russian Empire." *American Economic Review* 108 (4-5):1074–1117.

Martin, Gregory J and Ali Yurukoglu. 2017. "Bias in cable news: Persuasion and polarization." *American Economic Review* 107 (9):2565–99.

McCrary, Justin and Heather Royer. 2011. "The effect of female education on fertility and infant health: Evidence from school entry policies using exact date of birth." *American Economic Review* 101 (1):158–95. §.

Moser, Petra, Alessandra Voena, and Fabian Waldinger. 2014. "German Jewish émigrés and US invention." *American Economic Review* 104 (10):3222–55.

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. §.

Pinotti, Paolo. 2017. "Clicking on heaven's door: The effect of immigrant legalization on crime." *American Economic Review* 107 (1):138–68. §.

Pons, Vincent. 2018. "Will A Five-minute Discussion Change Your Mind? A Countrywide Experiment On Voter Choice In France." *American Economic Review* 108 (6):1322–63. §.

Schmieder, Johannes F, Till von Wachter, and Stefan Bender. 2016. "The effect of unemployment benefits and nonemployment durations on wages." *American Economic Review* 106 (3):739–77.

Simcoe, Timothy. 2012. "Standard setting committees: Consensus governance for shared technology platforms." *American Economic Review* 102 (1):305–36.

Stephens Jr, Melvin and Dou-Yan Yang. 2014. "Compulsory education and the benefits of schooling." *American Economic Review* 104 (6):1777–92. §.

Von Wachter, Till and Stefan Bender. 2006. "In the right place at the wrong time: The role of firms and luck in young workers' careers." *American Economic Review* 96 (5):1679–1705.