THE UNIVERSITY OF WARWICK

July 2014 No.197

# Tolerating defiance?
# Local average treatment effects without monotonicity

Clément de Chaisemartin

**WORKING PAPER SERIES**

Centre for Competitive Advantage in the Global Economy

Department of Economics

CAGE Competitive Advantage in the Global Economy

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

# Tolerating defiance?
# Local average treatment effects without monotonicity.*

Clément de Chaisemartin†

July 5, 2014

**Abstract**

We know that instrumental variable (IV) estimates a causal effect if the instrument satisfies a monotonicity condition. When this condition is not satisfied, we only know that IV estimates the difference between the effect of the treatment in two groups. This difference could be a very misleading measure of the treatment effect: it could be negative, even when the effect is positive in both groups. There are a large number of studies in which monotonicity is implausible. One might then question whether we should trust their estimates. I show that IV estimates a causal effect under a much weaker condition than monotonicity. I outline three criteria applied researchers can use to assess whether this condition is applicable in their studies. When this weaker condition is applicable, they can credibly interpret their estimates as causal effects. When it is not, they should interpret their results with caution.

Keywords: instrumental variable, two stage least squares, heterogeneous effects, monotonicity, defiers, internal validity
JEL Codes: C21, C26

*This paper is an extensively revised version of a paper circulated under the title "Defying the LATE? Identification of local treatment effects when the instrument violates monotonicity." I benefited from numerous discussions with Josh Angrist and Xavier D'Haultfœuille. I am also very grateful to Sascha Becker, Luc Behaghel, Stéphane Bonhomme, Federico Bugni, Laurent Davezies, Sara Geneletti, Marc Gurgand, Walker Hanlon, Toru Kitagawa, Andrew Oswald, Roland Rathelot, Pauline Rossi, Fabian Waldinger, Chris Woodruff, seminar participants at Crest, PSE, CORE, Brown, Harvard, MIT, Boston College, Columbia, Warwick, Bocconi, Université de Montréal, HEC Montréal, Duke, Chicago Booth, Cambridge, Bristol, Copenhagen Business School, Pompeu Fabra, St-Gallen, IFS, LSE, Oslo University, and Penn State for their helpful comments.

†Department of Economics, University of Warwick, clementdechaisemartin@gmail.com

# 1 Introduction

Standards of internal validity in applied work have increased dramatically over the past decades. Applied economists now seek out clean sources of variation to study difficult causal questions, including, for example, the effect of juvenile incarceration on educational attainment, or the effect of family size on mothers labor supply. On that purpose, they often use instruments that affect entry into the treatment being studied, and then estimate a two stage least squares regression (2SLS). But even with a randomly assigned instrument, the resulting estimate might not capture any causal effect.

People's treatment participation can be positively affected, unaffected, or negatively affected by the instrument. Those in the first group are called compliers, those in the second are called non-compliers, while those in the third are called defiers. Non-compliers reduce the instrument's statistical power as well as the external validity of the effect it estimates. But they do not threaten its internal validity. Indeed, Imbens & Angrist (1994) show that if the population only bears compliers and non-compliers, 2SLS estimates the average effect of the treatment among compliers, the so-called local average treatment effect (LATE). Defiers are a much more serious concern. If there are defiers in the population, we only know that 2SLS captures a weighted difference between the effect of the treatment among compliers and defiers (see Angrist et al., 1996). This difference could be a very misleading measure of the treatment effect: it could be negative, even when the effect of the treatment is positive in both groups. Defiers could be present in a large number of studies which have used 2SLS, and I will now give four examples which illustrate this situation.

First, a number of papers have used randomly assigned judges with different sentencing rates as an instrument for incarceration (see Aizer & Doyle, 2013 and Kling, 2006), receipt of disability insurance (see Maestas et al., 2013, French & Song, 2012, and Dahl et al., 2013), placement into foster care (see Doyle, 2007), or bankruptcy settlement (see Chang & Schoar, 2008). Imbens & Angrist (1994) argue that the "no-defiers" condition is likely to be violated in these types of studies. In this context, ruling out the presence of defiers would require that a judge with a high average of strictness always hands down a more severe sentence than that of a judge who is on average more lenient. Assume judge A only takes into account the severity of the offence in her decisions, while judge B is more lenient towards poor defendants, and more severe with well-off defendants. If the pool of defendants bears more poor than rich individuals, B will be on average more lenient than A, but she will be more severe with rich defendants. Whenever the number of cases per judge is large enough, the "no-defiers" condition has a strong testable implication: the ranking of judges in terms of their average severity should not vary over time or across subsamples. To my knowledge, this testable

implication has never been investigated.

Second, defiers could be present in randomized controlled trials relying on an encouragement design, because incentives might crowd-out subjects' intrinsic motivation for treatment. Duflo & Saez (2003) measure the effect of attending an information meeting on the take-up of a retirement plan. To encourage the treatment group to attend, subjects were told they would receive a financial incentive upon attendance. Frey & Jegen (2001) review a substantial body of empirical evidence showing that incentives sometimes reduce intrinsic motivation. In an encouragement design, paying subjects to get treated might negatively affect their perception of the benefits they should expect from treatment: if these benefits were high, there would be no need to pay them to get treated (see Benabou & Tirole, 2003). The incentive, therefore, could lead some of them to forgo treatment.

Third, defiers could be present in studies relying upon sibling-sex composition as an instrument for family size, because some parents are sex-biased. American parents are more likely to have a third child when their first two children are of the same sex. Angrist & Evans (1998) use this as an instrument to measure the effect of family size on mothers labor supply. However, some parents are biased towards one or the other sex. For example, Dahl & Moretti (2008) show that American fathers have a preference for boys. Because of sex-bias, some parents might want two sons, while others might want two daughters; such parents would be defiers. As I shall detail later, in Peru a non-negligible fraction of parents declare that the ideal composition of their family would be to have two sons and no daughters, or two daughters and no sons.

Fourth, defiers are present in studies that rely on quarter of birth as an instrument for school entry age, because some parents strategically delay entry to give children born late in the year more time to mature. In most countries, rules for school entry age should lead children born in the last quarter to enter at a younger age than those born in the first. Angrist & Krueger (1992) and Bedard & Dhuey (2006) use this as an instrument to measure the effect of entry age on later academic performance. Barua & Lang (2010) show that the distribution of school entry age of children born in the last quarter does not dominate that of children born in the first quarter, something we should observe if there were no defiers. This demonstrates that defiers are present in these studies. This is because parents are more prone to delaying school entry for children born late in the year, so-called redshirting. Children redshirted because they were born in the last quarter are defiers, as they would have entered school earlier had they been born in the first quarter.

These examples illustrate that defiers could be present in a number of studies, casting doubt on their results. This paper therefore addresses the following question: should we still trust results from a 2SLS study in which defiers are or could be present? I show that some of

these studies can still be trusted, while others should be interpreted with more caution. I also provide practitioners with a number of tools to assess in which category their study falls.

On the one hand, I show that 2SLS still estimates a LATE if the "no-defiers" condition is replaced by a weaker "compliers-defiers" condition. If a subgroup of compliers accounts for the same percentage of the population as defiers and has the same LATE, 2SLS estimates the LATE of the remaining part of compliers. "Compliers-defiers" is the weakest condition under which 2SLS estimates a LATE: if it is violated, 2SLS does not estimate a causal effect. I now outline three criteria applied researchers can use to assess whether "compliers-defiers" (CD) is likely to hold in their study.

CD is plausible in studies where there is little uncertainty on the sign of the treatment effect, as is the case in Maestas et al. (2013), for example. This study considers the effect of disability insurance on labor market participation. Its outcome is binary, and its 2SLS coefficient is negative. With a binary outcome, I show that CD automatically holds if the LATE of defiers has the same sign as the 2SLS coefficient. In Maestas et al., this will be satisfied if one is ready to assume that the effect of the treatment cannot be greater than zero for anyone. This appears to be a plausible restriction: by increasing non-labor income, disability insurance should unambiguously decrease labor market participation. Later in the paper, I argue that similar restrictions should also hold in French & Song (2012) and Aizer & Doyle (2013).

CD is also plausible in studies where selection into being a complier or a defier is not directly based on gains from treatment, as is the case in Angrist & Evans (1998), for example. The outcome in this study is binary. With a binary outcome, I show that CD is also automatically satisfied if the difference between compliers' and defiers' LATEs is not larger than a quantity which can be estimated from the data. In Angrist & Evans (1998), this quantity is large. On the other hand, there is no reason to suspect that defiers and compliers have utterly different LATEs: selection into one or the other population is driven by parents preferences for one or the other sex, not by gains from treatment. Therefore, CD should also hold in this application.

Finally, CD is plausible in studies with large first-stage coefficients, as is the case in Barua & Lang (2010) or Duflo & Saez (2003), for example. In studies with large first stages, compliers largely outnumber defiers. Therefore, one will always be able to find a subgroup of compliers accounting for the same percentage of the population as defiers and with the same LATE, unless the effect of the treatment is utterly different in the two populations.

On the other hand, there are still instances in which even CD could fail, because none of the aforementioned criteria is satisfied. For instance, Doyle (2007) considers the effect of placement into foster care on teen motherhood and juvenile delinquency, and uses the placement rates of randomly assigned investigators as an instrument. Theory does not suggest that foster care

should impact these outcomes in one specific direction. Besides, in this application we do not know what drives selection into being a complier or a defier, because we do not observe the criteria investigators use when making their decisions. If some base their decision on a variable correlated to the effect of the treatment, defiers and compliers might have very different treatment effects. Finally, placement rates do not greatly vary across investigators, so the first stage is not very strong in this study.

Overall, estimates from 2SLS studies in which defiers could be present can still credibly be interpreted as causal effects, provided the CD condition I propose in this paper sounds plausible. Even this weaker condition could sometimes fail, so it should not be taken for granted. In the remainder of the paper, I outline various quantities applied researchers can estimate to assess whether CD is plausible in their studies.

Other papers have studied relaxations of the "no-defiers" condition. Klein (2010) considers a model in which a random disturbance uncorrelated with the effect of the treatment leads some subjects to defy. By contrast, under my CD condition the factors leading some subjects to defy can be correlated with treatment effects. Small & Tan (2007) show that under a stochastic monotonicity condition, 2SLS estimates a weighted average treatment effect. Nevertheless, some of their weights are greater than one, so their parameter does not capture the effect of the treatment for a well-defined subgroup, making it hard to interpret. Moreover, stochastic monotonicity is a stronger condition than CD. DiNardo & Lee (2011) derive a result similar to Small & Tan (2007). Finally, Hoderlein & Gautier (2012) consider a selection model where there can be both defiers and compliers. But they require a continuous instrument, while my results hold for a binary or multivariate instrument.

The remainder of the paper is organized as follows. In Section 2, I show that 2SLS still estimates a LATE under the CD condition. In Section 3, I derive easily interpretable conditions under which CD is satisfied, and I review various applications. Section 4 concludes. Proofs are deferred to Appendix A. For the sake of brevity, I consider a number of extensions in a web appendix (see de Chaisemartin, 2014). In this appendix, I show how to derive confidence intervals for the quantities I suggest to estimate.[1] I derive a testable implication of the CD condition, and I show how one can implement the test. I also show how one can estimate the mean of any characteristic (age, sex...) in the population whose LATE is estimated by 2SLS. Then, I show how to estimate quantile treatment effects in this population. Finally, I show how my results extend to multivariate treatment and instrument.

[1] A Stata program is available upon request.

## 2 LATE identification under the "compliers-defiers" assumption

In this section, I show that with a binary instrument at hand, one can identify the LATE of a binary treatment on some outcome under a weaker assumption than "no-defiers".

Angrist et al. (1996) study the causal interpretation of the coefficients of a 2SLS regression with binary instrument and treatment. Let $Z$ be a binary instrument. Let $D_z \in \{0;1\}$ denote a subject's potential treatment when $Z = z$. Let $Y_{dz}$ denote her potential outcomes as functions of the treatment and of the instrument. Only $Z$, $D = D_Z$ and $Y = Y_{DZ}$ are observed. Following Imbens & Angrist (1994), let never takers ($NT$) be subjects such that $D_0 = 0$ and $D_1 = 0$, let always takers ($AT$) be such that $D_0 = 1$ and $D_1 = 1$, let compliers ($C$) be such that $D_0 = 0$ and $D_1 = 1$, and let defiers ($F$)[2] be such that $D_0 = 1$ and $D_1 = 0$. Let $FS = P(D = 1|Z = 1) - P(D = 1|Z = 0)$ denote the probability limit of the coefficient of the first stage regression of $D$ on $Z$. Let $RF = E(Y|Z = 1) - E(Y|Z = 0)$ denote the probability limit of the coefficient of the reduced form regression of $Y$ on $Z$. Finally, let $W = \frac{RF}{FS}$ denote the probability limit of the coefficient of the second stage regression of $Y$ on $D$.

Angrist et al. (1996) make a number of assumptions. First, they assume that $FS > 0$. Under Assumption 1 (see below), this implies that more subjects are compliers than defiers: $P(C) > P(F)$. This is a mere normalization: if it appears from the data that $FS < 0$, one can switch the words "defiers" and "compliers" in what follows.

Second, they assume that the instrument is independent of potential treatments and outcomes.

**Assumption 1** *(Instrument independence)*

$$(Y_{00}, Y_{01}, Y_{10}, Y_{11}, D_0, D_1) \perp\!\!\!\perp Z.$$

Third, they assume that the instrument has no direct effect on the outcome.

**Assumption 2** *(Exclusion restriction)*
$\forall d \in \{0, 1\}$,

$$Y_{d0} = Y_{d1} = Y_d.$$

Last, they assume that there are no defiers in the population, or that defiers and compliers have the same average treatment effect.

**Assumption 3** *(No-defiers: ND)*

$$P(F) = 0.$$

[2] In most of the treatment effect literature, treatment is denoted by $D$. To avoid confusion, defiers are denoted by the letter $F$ throughout the paper.

**Assumption 4** *(Equal LATEs for defiers and compliers: ELATEs)*

$$E(Y_1 - Y_0|C) = E(Y_1 - Y_0|F).$$

The following proposition summarizes the three main results in Angrist et al. (1996).

**LATE Theorems (Angrist et al. (1996))**

1. *Suppose Assumptions 1 and 2 hold. Then,*

$$FS = P(C) - P(F) \tag{1}$$
$$W = \frac{P(C)E\left(Y_1 - Y_0|C\right) - P(F)E\left(Y_1 - Y_0|F\right)}{P(C) - P(F)}. \tag{2}$$

2. *Suppose Assumptions 1, 2, and 3 hold. Then,*

$$FS = P(C) \tag{3}$$
$$W = E\left(Y_1 - Y_0|C\right). \tag{4}$$

3. *Suppose Assumptions 1, 2, and 4 hold. Then,*

$$W = E\left(Y_1 - Y_0|C\right). \tag{5}$$

Under random instrument and exclusion restriction alone, $W$ cannot receive a causal interpretation, as it is equal to a weighted difference of the LATEs of compliers and defiers. If there are no defiers, (1) and (2) respectively simplify into (3) and (4). $W$ is then equal to the LATE of compliers, while $FS$ is equal to the percentage of the population they account for. Finally, when ND does not sound credible, $W$ can still capture the LATE of compliers provided one is ready to assume that defiers and compliers have the same LATE, as shown in (5).

In this paper, I substitute the following condition to Assumption 3 or 4.

**Assumption 5** *(Compliers-defiers: CD)*
*There is a subpopulation of compliers $C_F$ which satisfies:*

$$P(C_F) = P(F) \tag{6}$$
$$E(Y_1 - Y_0|C_F) = E(Y_1 - Y_0|F). \tag{7}$$

CD is satisfied if a subgroup of compliers accounts for the same percentage of the population as defiers and has the same LATE. I call this subgroup "compliers-defiers", or "comfiers". The CD condition is somewhat abstract. In section 3, I derive a number of easily interpretable conditions under which it is satisfied. For now, let me just note that CD is weaker than Assumption 3 and 4. If there are no defiers, one can find a zero probability subset of compliers

with the same LATE as defiers. Similarly, if compliers and defiers have the same LATE, one can randomly choose $\frac{P(F)}{P(C)}\%$ of compliers and call them comfiers: this will yield a subgroup accounting for the same percentage of the population and with the same LATE as defiers.

I can now state the main result of this paper.

**Theorem 2.1** *Suppose Assumptions 1 and 2 hold.*
*If a subpopulation of compliers $C_F$ satisfies* (6) *and* (7)*, then $C_V = C \setminus C_F$ satisfies*

$$P(C_V) = FS \tag{8}$$

$$E\left(Y_1 - Y_0 | C_V\right) = W. \tag{9}$$

*Conversely, if a subpopulation of compliers $C_V$ satisfies* (8) *and* (9)*, then $C_F = C \setminus C_V$ satisfies* (6) *and* (7)*.*

**Proof**

$\Rightarrow$

$$FS = P(C) - P(F) = P(C_V) + P(C_F) - P(F) = P(C_V).$$

The first equality follows from (1), the last follows from (6). This proves that $C_V$ satisfies (8).

Then,

$$\begin{aligned} E\left(Y_1 - Y_0 | C\right) &= P(C_V|C)E\left(Y_1 - Y_0|C_V\right) + P(C_F|C)E\left(Y_1 - Y_0|C_F\right) \\ &= \frac{P(C) - P(F)}{P(C)} E\left(Y_1 - Y_0|C_V\right) + \frac{P(F)}{P(C)} E\left(Y_1 - Y_0|F\right), \end{aligned}$$

where the last equality follows from (6) and (7). Plugging this into (2) yields

$$W = E\left(Y_1 - Y_0|C_V\right).$$

This proves that $C_V$ satisfies (9).

$\Leftarrow$

$$P(C_F) = P(C) - P(C_V) = P(C) - FS = P(C) - (P(C) - P(F)) = P(F).$$

The second step follows from (8), the third follows from (1). This proves that $C_F$ satisfies (6).

Then,

$$\begin{aligned} E\left(Y_1 - Y_0 | C\right) &= P(C_V|C)E\left(Y_1 - Y_0|C_V\right) + P(C_F|C)E\left(Y_1 - Y_0|C_F\right) \\ &= \frac{FS}{P(C)} W + \frac{P(F)}{P(C)} E\left(Y_1 - Y_0|C_F\right), \end{aligned}$$

where the last equality follows from (8), (9), and (6). Plugging this Equation into (2) yields

$$E\left(Y_1 - Y_0|F\right) = E\left(Y_1 - Y_0|C_F\right).$$

This proves that $C_F$ satisfies (7).
**QED.**

The intuition for this result goes as follows. Under CD, comfiers and defiers cancel one another out, and the 2SLS coefficient captures the effect of the treatment for the remaining part of compliers. I hereafter refer to the $C_V$ subpopulation as "compliers-survivors", or "comvivors", as they are compliers who "out-survive" defiers.

Theorem 2.1 shows that 2SLS captures a LATE under a weaker assumption than ND. Notwithstanding, this LATE is not the same as the LATE of compliers, as it only applies to a subgroup of compliers. This raises the question of whether this LATE is an "interesting" parameter.

From most economists' perspective, a sufficient condition for a treatment effect parameter to be deemed interesting is its policy relevance. Some authors do not regard the LATE of compliers as policy relevant: to decide whether she should give some treatment to her population, a utilitarian social planner needs to know the average treatment effect (ATE), not the LATE (see e.g. Heckman & Urzúa (2010)). Proponents of the LATE of compliers generally put forward two reasons why this planner might still care about the LATE. I shall now summarize these two arguments, and argue that while they apply to the LATE of compliers in a world without defiers, they apply to the LATE of comvivors in a world with defiers.

Sometimes the policy the planner is contemplating is not whether she should give or not the treatment to her population, but whether she should marginally increase incentives for treatment. In judges papers, the relevant policy question is probably not whether the planner should send every defendant to jail, but whether defendants with marginal cases should go to jail, something the planner can manipulate by hiring marginally more severe or more lenient judges. In a world without defiers, compliers are the only subjects affected by marginal policy changes, so they are the relevant population the planner should consider when making this type of decision (see e.g. Imbens (2010)). In a world with defiers, compliers are no longer the only group affected by such policies. Marginal increases in incentives for treatment lead compliers to receive the treatment, and have the opposite effect on defiers. But under the CD assumption, comfiers and defiers cancel one another out, so the planner should not take them into account. The relevant population she should consider are comvivors, because they are affected by her policy, while their LATE is not netted out by that of another population.

Even when the planner contemplates whether she should give the treatment to her population, Imbens (2010) argues that knowing the LATE of compliers can be useful. As an example, he considers a randomized evaluation of the effect of a drug on survival rate with imperfect compliance. Because of imperfect compliance, the ATE can only be bounded (see Manski, 1990), and Imbens (2010) assumes that the bounds are $\left[-\frac{3}{16}, \frac{5}{16}\right]$. Despite imperfect compliance, one

can also credibly point-identify the LATE of compliers, provided there are no defiers. Imbens then argues that one should report the LATE of compliers along with the bounds on the ATE: "The bounds on the ATE can be consistent with a substantial negative average effect for compliers, lowering survival rates by $\frac{1}{4}$, or with a substantial positive average effect for compliers, raising survival rates by $\frac{1}{4}$. One would think that, in the first case, a decision maker would be considerably less likely to implement universal adoption of the treatment than in the second, and so reporting only the bounds might leave out relevant information." In a world with defiers, this argument is not valid anymore, as the LATE of compliers can no longer be credibly estimated. But under the CD assumption, one can credibly estimate the LATE of comvivors. It is this parameter which should be reported along with bounds on the ATE.

# 3 Sufficient conditions for "compliers-defiers" to hold

A great appeal of the ND condition is that it is simple to interpret. On the contrary, CD is an abstract condition. In this section, I try to clarify its meaning by deriving more interpretable conditions under which it is satisfied. All these conditions point towards the same interpretation of the CD condition. CD is more likely to hold if there are few defiers, or if defiers and compliers treatment effects are not too different. It is also more likely to hold when the instrument has a stronger first stage. Based on these results, I suggest various quantities practitioners can estimate to assess whether CD is likely to hold in their application.

## 3.1 Sufficient conditions with a binary outcome

Consider the two following assumptions.

**Assumption 6** *(Restriction on the sign of the LATE of defiers)*
$E(Y_1 - Y_0|F)$ *and* $W$ *have the same sign, or either of these two quantities is equal to* 0.

**Assumption 7** *(Restriction on the difference between compliers' and defiers' LATE)*

$$|E(Y_1 - Y_0|C) - E(Y_1 - Y_0|F)| \leq \Delta(P(F)) = \frac{|RF|}{FS + P(F)} = |W|\frac{FS}{FS + P(F)}.$$

Assumption 6 requires that the LATE of defiers has the same sign as the 2SLS coefficient. It is appealing in applications in which the sign of the treatment effect can be assumed to be known ex-ante, which is often called a monotone treatment response assumption (MTR, see Manski, 1997). MTR assumptions can be justified using economic theory or evidence from other sciences such as medicine. However, note that Assumption 6 is weaker than MTR as it only restricts the sign of the average effect for a subgroup.

Assumption 7 is a restriction on difference between the LATEs of defiers and compliers. The upper bound on LATEs difference, $\Delta(P(F))$, is decreasing in the share of defiers, and increasing in $|W|$ and $FS$. When $P(F)$ is low, compliers and defiers can have different LATEs under Assumption 7. When $P(F)$ is large, compliers and defiers LATEs should be fairly similar. Assumption 7 is also more credible when the instrument has large first and second stages.

When the outcome is binary, Assumptions 6 and 7 are sufficient for CD to hold.

**Theorem 3.1** *If $Y_0$ and $Y_1$ are binary, Assumption 7 $\Rightarrow$ Assumption 6 $\Rightarrow$ Assumption 5.*

The first implication follows after some algebra. The second one states that if the LATE of defiers has the same sign as the 2SLS coefficient (or if either of those two quantities is equal to 0), CD is satisfied. The intuition for this result goes as follows. With binary potential outcomes, it follows from (2) that

$$\begin{aligned} RF &= P(Y_1 - Y_0 = 1, C) - P(Y_1 - Y_0 = -1, C) \\ &- (P(Y_1 - Y_0 = 1, F) - P(Y_1 - Y_0 = -1, F)) . \end{aligned}$$

To fix ideas, suppose that Assumption 6 is satisfied with $E(Y_1 - Y_0|F)$ and $W$ greater than 0. $W \geq 0$ implies $RF \geq 0$. $RF \geq 0$ combined with the previous equation implies that

$$P(Y_1 - Y_0 = 1, C) \geq P(Y_1 - Y_0 = 1, F) - P(Y_1 - Y_0 = -1, F).$$

Then, there are sufficiently many compliers with a strictly positive treatment effect to extract from them a subgroup that will compensate defiers' positive LATE.

**Applications of Theorem 3.1**

*Maestas et al. (2013)*

Maestas et al. (2013) study the effect of receiving DI on labor market participation. Their 2SLS coefficient is negative, so Assumption 6 will hold if $E(Y_1 - Y_0|F)$ is not greater than 0. A sufficient condition for this to be true is the following MTR condition: $Y_1 - Y_0 \leq 0$. Theory suggests this is a credible assumption. By increasing non-labor income, DI should unambiguously diminish labor supply.

*Aizer & Doyle (2013)*

Aizer & Doyle (2013) study the effect of juvenile incarceration on high school completion. Their 2SLS coefficient is negative, so Assumption 6 will hold if $E(Y_1 - Y_0|F) \leq 0$. This is a credible condition. Being incarcerated disrupts schooling and increases the chances a youth form relationships with non-academically oriented peers. This should increase the chances of drop-out. Prison education programs might have a positive effect on taste for schooling, but

it sounds implausible they can offset schooling disruption and negative peer effects. Moreover, it suffices that incarceration have on average a negative effect among defiers for CD to hold.

*Angrist & Evans (1998)*

Angrist & Evans (1998) study the effect of having a third child on mothers labor supply. Their 2SLS coefficient is negative, so Assumption 6 will hold if $E(Y_1 - Y_0|F)$ is not greater than 0. Here, theory is ambiguous on this restriction. Giving birth can have two countervailing effects on labor market participation (see e.g. Blau & Robins, 1988). The cost of daycare acts as a tax on mothers wage. This could reduce their participation. Other child costs (food...) act as a negative income shock. This could increase participation.
But if there are not too many defiers and defiers' and compliers' LATE do not differ too much, Assumption 7 will be satisfied. First, notice that

$$P(F) \quad \leq \quad \min(P(D=1|Z=0), P(D=0|Z=1)) = \overline{P}(F). \tag{10}$$

The share of defiers must be lower than the percentage of treated observations among those who do not receive the instrument, as this group includes always takers and defiers. It must also be lower than the percentage of untreated observations among those who receive the instrument, as this group includes never takers and defiers. In Angrist & Evans (1998), $\widehat{\overline{P}}(F) = 37.2\%$: there cannot be more than 37.2% of defiers.[3] The left axis of Figure 1 shows the sample counterpart of $\Delta(P(F))$ for all values of $P(F)$ included between 0 and 37.2%. The right axis shows the same quantity normalized by the standard deviation of the outcome. Assumption 7 is satisfied for values of $P(F)$ and $|E(Y_1 - Y_0|C) - E(Y_1 - Y_0|F)|$ below the green line. For instance, Assumption 7 holds if there are less than 5% of defiers and compliers and defiers LATEs differ by less than 7.2 percentage points, or 14.5% of a standard deviation.[4]

[3] In de Chaisemartin (2014), I derive a 95% confidence upper bound for $P(F)$ and find it is equal to 37.4%.

[4] The 95% confidence interval of $\Delta(0.05)$ is [0.044,0.100]. In Stata, one can use reg3 and nlcom to derive it.

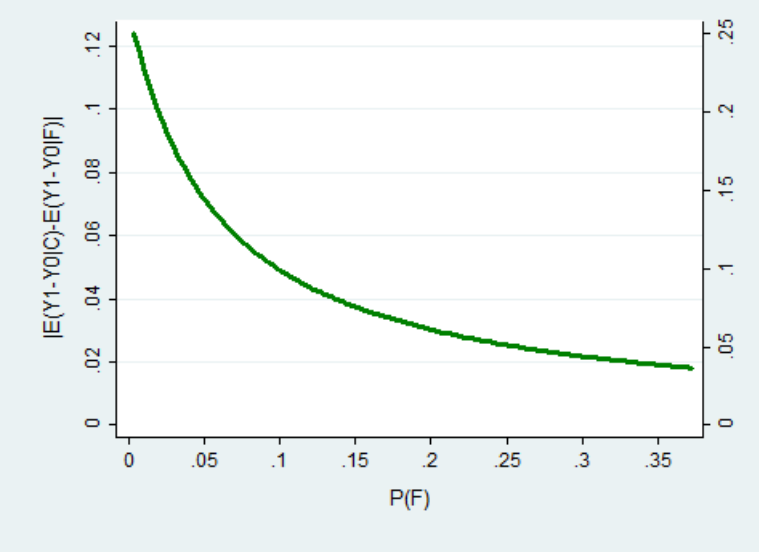


Figure 1: For all values of $P(F)$ and $|E(Y_1 - Y_0|C) - E(Y_1 - Y_0|F)|$ below the green line, CD is satisfied in Angrist & Evans (1998).

The very limited evidence available suggests that 5% is a reasonably conservative upper bound for the share of defiers in this application. In the 2012 Peruvian wave of the Demographic and Health Surveys, women were asked their ideal sex sibship composition. Among women whose first two kids is a boy and a girl, 1.8% had 3 children or more and retrospectively declare that their ideal sex sibship composition would have been two boys and no girl, or no boy and two girls. These women seem to have been induced to having a third child because their first two children were a boy and a girl. To the best of my knowledge, similar questions have never been asked in an American survey. There are many reasons why 1.8% could under or overestimate the share of defiers in the American population. But this figure is, as of now, the best piece of evidence available to make a guess on the percentage of defiers in Angrist & Evans (1998). 5% therefore sounds like a reasonably conservative upper bound.

15% of a standard deviation also sounds like a reasonably conservative upper bound for $|E(Y_1 - Y_0|C) - E(Y_1 - Y_0|F)|$ in this application. Compliers are couples with a preference for diversity, while defiers are sex-biased couples. Preference for diversity and sex bias are probably correlated with some of the variables entering into mothers labor market participation equation (mother's potential wage, preference for leisure...), but they are generally not directly included into this equation (see e.g. Blau & Robins, 1988). As a result, 15% of a standard deviation is probably a reasonably conservative upper bound for $|E(Y_1 - Y_0|C) - E(Y_1 - Y_0|F)|$, because selection into being a complier or a defier is not directly based on gains from treatment.

Estimates in Maestas et al. (2013), Aizer & Doyle (2013), and Angrist & Evans (1998) might not capture the LATE of compliers because of defiers, but they most probably capture the LATE of comvivors, because either Assumption 6 or 7 seem likely to hold in these applications.

In other instances, Assumptions 6 and 7 might not sound credible. One might therefore want to investigate how the results of a study would be affected if defiers do not satisfy either of these two assumptions. To conclude this section, I conduct a worst-case analysis, and study the largest possible negative impact defiers can have on the external validity of 2SLS estimates. Let $C_V^1$ denote the largest subpopulation of compliers with a LATE equal to $W$. For any $(p,e) \in [0, \overline{P}(F)] \times [-1,1]$, let

$$\lambda(p,e) = \min\left(\max\left(0, 1 + \frac{p \times e}{RF}\right), 1\right).$$

**Theorem 3.2** *Assume $Y_0$ and $Y_1$ are binary. If $P(F) \leq p$, and either $RF > 0$ and $E(Y_1 - Y_0|F) \geq e$, or $RF < 0$ and $E(Y_1 - Y_0|F) \leq e$,*

$$\lambda(p,e) \times FS \leq P\left(C_V^1\right).$$

**Application of Theorem 3.2: Duflo & Saez (2003).**

I use Theorem 3.2 to show that results in Duflo & Saez (2003) are very robust to defiers. This paper studies the effect of attending an information meeting on the take-up of a retirement plan. The authors find $\widehat{FS} = 23\%$ and $\widehat{W} = 6.1\%$. If $E(Y_1 - Y_0|F) \geq 0$, it follows from Theorem 3.1 that $\widehat{W}$ consistently estimatates the effect of the treatment for a subgroup accounting for 23% of the population. But $E(Y_1 - Y_0|F) \geq 0$ might not sound credible. Here, defiers are people who do not like to do what they are being told, so the meeting might have a negative effect on their participation decision. Now, let $\overline{p} = \widehat{P}(D=1|Z=0) \times \widehat{P}(D=0|Z=1) = 3.5\%$. $\overline{p}$ is a worst case upper bound for the share of defiers under the assumption that the two potential treatments are positively correlated. It is only slightly smaller than $\widehat{\overline{P}}(F)$, the worst-case upper bound for the share of defiers, which is equal to 4.9%. The blue line on Figure 2 plots the sample counterpart of the lower bound of $P\left(C_V^1\right)$ derived in Theorem 3.2 for all possible values of $E(Y_1 - Y_0|F)$. The red line on this figure is at $E(Y_1 - Y_0|F) = -0.13$, which is equal to $-50\%$ of the standard deviation of the outcome in this study. The intersection between these two lines tells us that even if one is only ready to assume that $P(F) \leq 0.035$, and that $E(Y_1 - Y_0|F)$ is greater than -50% of the standard deviation of the outcome, one can still claim that $\widehat{W}$ consistently estimates the LATE of a population of compliers accounting for at least 15.5% of the population. Even for these extreme values of $P(F)$ and $E(Y_1 - Y_0|F)$, defiers might at most slightly reduce the external validity of the 2SLS coefficient.

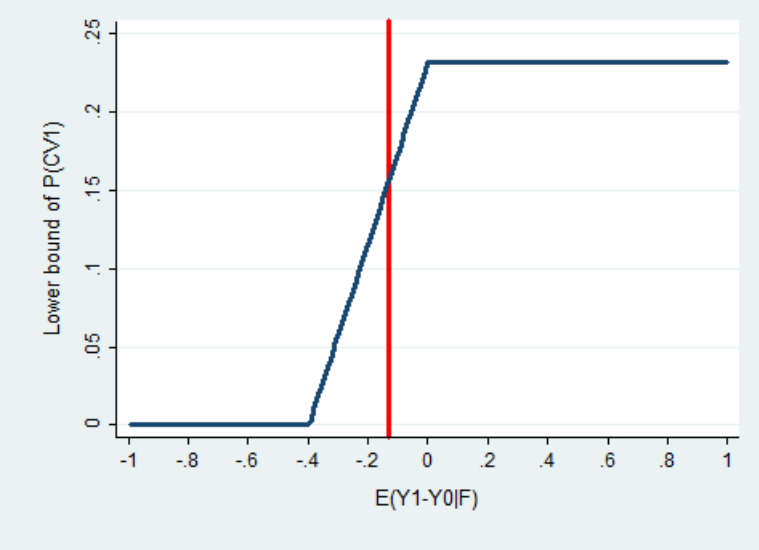


Figure 2: Lower bound for the size of the population to which results in Duflo & Saez (2003) apply, as a function of $E(Y_1 - Y_0|F)$.

In Duflo & Saez (2003), the reduced form is strong and two-sided imperfect compliance is close to being one sided. Applications meeting these two requirements are robust to almost any type of defiers. If the reduced form is large and, say, positive, $\lambda(p, e)$ is not decreasing too quickly when $e$ becomes negative. If two-sided imperfect compliance is close to being one-sided, meaning either that few subjects are treated when $Z = 0$ or that few remain untreated when $Z = 1$, there cannot be too many defiers. Duflo & Saez (2003) is not the only paper meeting these two requirements. In judges papers, one could redefine a binary instrument equal to 1 if a case was assigned to one of the most severe judges, say to one in the upper quartile of sentencing rates, and to 0 if it was assigned to one of the least severe judges. This instrument will have a large first stage, and is therefore likely to have a large reduced form. Imperfect compliance should be close to one-sided: the least severe judges probably have a low sentencing rate. If this robustness check yields similar results to that of the main specification, which most often in these papers directly uses sentencing rate as the instrument, this will be reassuring, as this modified instrument is robust to almost any type of defiers.

### 3.2 Sufficient condition with a general outcome

Consider the following condition:

**Assumption 8** *(More compliers than defiers: MC)*

*For every $\delta$ in the support of $Y_1 - Y_0$,*

$$\frac{f_{Y_1-Y_0|F}(\delta)}{f_{Y_1-Y_0|C}(\delta)} \leq R(P(F)) = 1 + \frac{FS}{P(F)}. \tag{11}$$

I call this condition the more compliers than defiers condition. Indeed, it follows from (1) that $R(P(F)) = \frac{P(C)}{P(F)}$. Therefore, (11) is equivalent to

$$P(F|Y_1 - Y_0) \leq P(C|Y_1 - Y_0). \tag{12}$$

(12) requires that each subgroup of the population with the same value of $Y_1 - Y_0$ comprise more compliers than defiers. This condition is weaker but closely related to the stochastic monotonicity assumption in Small & Tan (2007).

As shown in Angrist et al. (1996), 2SLS captures a LATE if there are no defiers, or if defiers and compliers have the same distribution of $Y_1 - Y_0$. These assumptions are "polar cases" of MC. MC holds when defiers and compliers have the same distribution of $Y_1 - Y_0$, as the left hand side of (11) is then equal to 1, while its right hand side is greater than 1.[5] And MC also holds when there are no defiers, as the right hand side of (11) is then equal to $+\infty$. In between those polar cases, MC holds in many intermediate cases. $R(P(F))$ is decreasing in $P(F)$. If there are many defiers, Assumption 8 will be satisfied if the distributions of $Y_1 - Y_0$ among compliers and defiers are not too different. Conversely, if these two distributions are very different, Assumption 8 can still be satisfied if there are few defiers. $R(P(F))$ is also increasing in $FS$: MC is more likely to hold when the instrument has a strong rather than a weak first stage.

The next theorem shows that MC is a sufficient condition for CD to hold.

**Theorem 3.3** *Assumption 8 $\Rightarrow$ Assumption 5.*

To convey the intuition of this Theorem, I consider the example displayed in Figure 3. $Y_0$ and $Y_1$ are binary. The population bears 20 subjects. 13 of them are compliers, while 7 are defiers. Those 20 subjects are scattered over the three $Y_1 - Y_0$ cells as shown in Figure 3. MC holds as there are more compliers than defiers in each cell.

| Y(1)-Y(0) | Defiers | Compliers |
|---|---|---|
| -1 | f1 f2 | c1 c2 c3 |
| 0 | f3 f4 f5 | c4 c5 c6 c7 c8 |
| 1 | f6 f7 | c9 c10 c11 c12 c13 |

Figure 3: A population in which MC is satisfied: there are more compliers than defiers in each subgroup with the same value of $Y_1 - Y_0$.

[5] I have assumed, as a mere normalization, that $RF > 0$.

To construct $C_F$, one can merely pick up as many compliers as defiers in each of the three $Y_1 - Y_0$ strata. The resulting $C_F$ and $C_V$ populations are displayed in Figure 4. Comfiers account for the same percentage of the population as defiers and also have the same LATE.

| Y(1)-Y(0) | Defiers | Comfiers | Comvivors |
|---|---|---|---|
| -1 | f1 f2 | c1 c2 | c3 |
| 0 | f3 f4 f5 | c4 c5 c6 | c7 c8 |
| 1 | f6 f7 | c9 c10 | c11 c12 c13 |

Figure 4: This population satisfies CD: to create the subpopulation of comfiers, it suffices to pick as many compliers as defiers in each $Y_1 - Y_0$ subgroup.

**Application of Theorem 3.3: Barua & Lang (2010)**

Barua & Lang (2010) argue that using quarter of birth (QOB) as an instrument for school entry age might produce severely biased estimates of the effect of entry age on attainment because of defiers. First, they show that the cdf of entry age for children born in the fourth quarter of 1952 does not stochastically dominate that of those born in the first quarter. If there were no defiers, one should observe dominance (see Angrist & Imbens, 1995). Then, they argue that compliers and defiers probably have very different LATEs. Defiers are children redshirted by their parents; children who benefit the most from entering late are also the most likely to be redshirted. Finally, they use an example to illustrate the potential magnitude of the bias.

I now argue that QOB might still be a valid instrument despite violations of ND.

I first show that in Barua & Lang (2010), one can identify the share of compliers and defiers under a mild assumption. In their data, children born in Q4 enter school at either 3.75, 4.75, or 5.75 years old. Those born in Q1 enter at either 4.5, 5.5, or 6.5 years old. I assume QOB can affect entry age by one year at most: for instance, a child born in Q1 who entered when she was 4.5 years old would either have entered at 3.75 or 4.75 years old if she had been born in Q4. Under this restriction, one can use the distributions of school entry age of children born in Q4 and Q1 to recover the joint distribution of the two counterfactual entry ages of the same child if she had been born in Q4 or Q1. Let $A_0$ and $A_1$ denote these counterfactuals. Table 1 shows their joint distribution. The three groups on the diagonal are compliers: being born in Q1 induces them to enter three quarters later than if they had been born in Q4. The two groups below the diagonal are defiers: being born in Q1 induces them to enter one quarter sooner. The population therefore bears 65% of compliers and 35% of defiers.

| $A_0$ / $A_1$ | 4.5 | 5.5 | 6.5 |
|---|---|---|---|
| 3.75 | 6% | 0% | 0% |
| 4.75 | 3% | 46% | 0% |
| 5.75 | 0% | 32% | 13% |

Table 1: Joint distribution of school entry age if born in Q1 or Q4 in Barua & Lang (2010): there are 35% of defiers and 65% of compliers in the population.

Compliers outnumber defiers, and their entry age is more affected by the instrument than defiers'. As a result, 2SLS can still capture a LATE in this application under a mild "more compliers than defiers" assumption. To simplify the discussion, I assume that the effect of entering school at age $x+d$ quarters instead of $x$ depends on $d$ but not on $x$: $Y_{0.25(x+d)}-Y_{0.25x} = \Delta_d$.[6] Under this assumption, the reduced form regression of educational attainment on QOB captures a weighted difference of the effect of entering three quarters later for compliers, and one quarter later for defiers:

$$RF = E(\Delta_3|C)P(C) - E(\Delta_1|F)P(F) = 3E(\Delta_1|C)P(C) - E(\Delta_1|F)P(F). \quad (13)$$

If

$$\frac{f_{\Delta_1|F}(\delta)}{f_{\Delta_1|C}(\delta)} \leq 3 \times \frac{P(C)}{P(F)} = 5.57, \quad (14)$$

there is a subgroup of compliers denoted $C_F$ such that

$$E(\Delta_1|C_F) = E(\Delta_1|F)$$
$$P(C_F) = \frac{P(F)}{3}.$$

$E(\Delta_1|F)P(F)$ is netted out by $3E(\Delta_1|C_F)P(C_F)$ in (13), and the 2SLS coefficient finally captures the LATE of comvivors. This almost directly follows from Theorems 2.1 and 3.3. There is a slight difference though, arising from the multivariate nature of the treatment. (13) includes effects of entering school three quarters later for compliers, and only one quarter later for defiers. As a result, the right hand side of (14) is three times larger than that of (11).

I now consider a simple parametric model nested in the numerical example of Barua & Lang in which (14) is satisfied. Barua & Lang assume that entering school one quarter later always increases children's educational attainment, and that defiers and compliers LATEs are respectively equal to 1.5 and 0.5. I will further assume that entering school one quarter later cannot increase educational attainment by more than two years, and that treatment

[6] This assumption does not affect the substantive conclusions of the discussion.

effects for compliers and defiers follow truncated geometric distributions on $\{0,1,2\}$, with respective parameters $p_C$ and $p_F$. Under this geometric assumption, (14) holds if and only if $\left(\frac{1-p_F}{1-p_C}\right)^2 \leq 5.57$. Solving the two moments conditions imposed by Barua & Lang for $p_C$ and $p_F$ yields $p_C = 0.63$ and $p_F = 0.18$. For these values of $p_C$ and $p_F$, (14) is satisfied.

There are other parametric models nested in their numerical example in which MC fails. For instance, if all compliers have a treatment effect equal to 0.5 while all defiers have a treatment effect equal to 1.5, both MC and CD fail to hold. But the geometric example still shows that having many defiers with a very different LATE from that of compliers is not sufficient for the QOB instrument to fail. To regard this instrument as invalid, one should also have reasons to believe that models in which (14) fails are more credible than models in which it is satisfied.

In Barua & Lang (2010), the supports of the treatment variable when $Z = 0$ and when $Z = 1$ are disjoint. One can then identify the percentage of defiers under a mild assumption. In most applications, this disjoint support condition will fail. For instance, it will never be satisfied with a binary treatment. As a result, $P(F)$ and $R(P(F))$ are not identified. In such instances, one can estimate $R(P(F))$ for plausible values of $P(F)$ to assess the credibility of the MC condition. If one does not want to make any assumption on $P(F)$, one can also derive a worst case lower bound for $R(P(F))$. $P(F) \leq \overline{P}(F)$ indeed implies that

$$1 + \frac{FS}{\overline{P}(F)} \leq R(P(F)). \tag{15}$$

In the web appendix, I show how one can derive a confidence lower bound for $R(P(F))$ based on this inequality.

# 4 Conclusion

Until now, the causal interpretation of 2SLS coefficients has relied on a "no-defiers" assumption. This assumption is questionable in a large number of studies. I show that when it seems likely to fail, 2SLS estimates can still be credibly interpreted as causal effects provided the CD assumption I propose in this paper is satisfied. While CD sounds plausible in some applications, it is questionable in others. It should therefore not be taken for granted.

Here are the steps applied researchers should follow to assess the credibility of the CD condition. When their outcome is binary, CD will be satisfied if the LATE of defiers has the same sign as their 2SLS coefficient. If theory suggests this is a credible restriction, they can invoke my results to justify the validity of their estimates. With a binary outcome, CD will also be satisfied if the difference between compliers and defiers LATEs is not larger than the absolute value of their reduced form coefficient divided by the sum of their first stage and of

the percentage of defiers. They can estimate this quantity for reasonably conservative values of the percentage of defiers. If the resulting estimate is large while theory suggests compliers and defiers should not have utterly different LATEs, they can also invoke my results. Finally, if theory suggests defiers LATE might not have the desired sign and could be very different from that of compliers, they can perform a worst-case analysis to assess the maximum negative impact defiers can have on the external validity of their results.

When they are interested in a non-binary outcome, CD will be satisfied if the ratio of the distributions of the treatment effect for defiers and compliers is not larger than the ratio of the percentages of compliers and defiers. They can estimate this second ratio for conservative values of the percentage of defiers. If the resulting estimate is large while theory suggests compliers and defiers should not have very different distributions of their treatment effects, CD should be satisfied.

When this is possible, they can also redefine their instrument in a way which maximizes its first stage. CD is indeed more likely to be satisfied when the instrument has a large rather than a weak first stage, so this modified instrument will be very robust to defiers.

When none of these exercises prove conclusive, they should be more careful when interpreting their results, as their estimate might fail to capture a causal effect.

## References


Aizer, A. & Doyle, J. J. (2013), Juvenile incarceration, human capital and future crime: Evidence from randomly-assigned judges, Technical report, NBER.

Angrist, J. D. & Evans, W. N. (1998), 'Children and their parents' labor supply: Evidence from exogenous variation in family size', *American Economic Review* **88**(3), 450–77.

Angrist, J. D. & Imbens, G. W. (1995), 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of the American Statistical Association* **90**(430), pp. 431–442.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996), 'Identification of causal effects using instrumental variables', *Journal of the American Statistical Association* **91**(434), pp. 444–455.

Angrist, J. D. & Krueger, A. B. (1992), 'The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples', *Journal of the American Statistical Association* **87**(418), 328–336.

Barua, R. & Lang, K. (2010), School entry, educational attainment and quarter of birth: A cautionary tale of late. Working Paper.

Bedard, K. & Dhuey, E. (2006), 'The persistence of early childhood maturity: International evidence of long-run age effects', *The Quarterly Journal of Economics* **121**(4), 1437–1472.

Benabou, R. & Tirole, J. (2003), 'Intrinsic and extrinsic motivation', *The Review of Economic Studies* **70**(3), 489–520.

Blau, D. M. & Robins, P. K. (1988), 'Child-care costs and family labor supply', *The Review of Economics and Statistics* pp. 374–381.

Chang, T. & Schoar, A. (2008), Judge specific differences in chapter 11 and firm outcomes, *in* 'American Law & Economics Association Annual Meetings', bepress, p. 86.

Dahl, G. B., Kostol, A. R. & Mogstad, M. (2013), Family welfare cultures, Technical report, NBER.

Dahl, G. B. & Moretti, E. (2008), 'The demand for sons', *The Review of Economic Studies* **75**(4), 1085–1120.

de Chaisemartin, C. (2014), 'Web appendix to: "tolerating defiance: local average treatment effects without monotonicity"'.

DiNardo, J. & Lee, D. S. (2011), *Program Evaluation and Research Designs*, Vol. 4 of *Handbook of Labor Economics*, Elsevier, chapter 5, pp. 463–536.

Doyle, J. J. (2007), 'Child protection and child outcomes: Measuring the effects of foster care', *The American Economic Review* pp. 1583–1610.

Duflo, E. & Saez, E. (2003), 'The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment', *The Quarterly Journal of Economics* **118**(3), 815–842.

French, E. & Song, J. (2012), The effect of disability insurance receipt on labor supply: a dynamic analysis, Technical report, Working Paper, Federal reserve Bank of Chicago.

Frey, B. S. & Jegen, R. (2001), 'Motivation crowding theory', *Journal of economic surveys* **15**(5), 589–611.

Heckman, J. J. & Urzúa, S. (2010), 'Comparing iv with structural models: What simple iv can and cannot identify', *Journal of Econometrics* **156**(1), 27 – 37.

Hoderlein, S. & Gautier, E. (2012), Estimating treatment effects with random coefficients in the selection equation, Technical report.

Imbens, G. W. (2010), 'Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009)', *Journal of Economic Literature* **48**, 399–423.

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–75.

Klein, T. J. (2010), 'Heterogeneous treatment effects: Instrumental variables without monotonicity?', *Journal of Econometrics* **155**(2).

Kling, J. R. (2006), 'Incarceration length, employment, and earnings', *American Economic Review* **96**(3), 863–876.

Maestas, N., Mullen, K. J. & Strand, A. (2013), 'Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt', *American Economic Review* **103**(5), 1797–1829.

Manski, C. F. (1990), 'Nonparametric bounds on treatment effects', *American Economic Review* **80**(2), 319–23.

Manski, C. F. (1997), 'Monotone treatment response', *Econometrica* **65**(6), 1311–1334.

Small, D. & Tan, Z. (2007), A stochastic monotonicity assumption for the instrumental variables method, Working paper, department of statistics, university of pennsylvania.

# A Proofs

For any random variable $X$, let $\mathcal{S}(X)$ denote the support of $X$. In the proofs, I assume the probability distributions of $Y_1 - Y_0$, $Y_1 - Y_0|C$ and $Y_1 - Y_0|F$ are all dominated by the same measure $\lambda$. Let $f_{Y_1-Y_0}$, $f_{Y_1-Y_0|C}$, and $f_{Y_1-Y_0|F}$ denote the corresponding densities. I also adopt the convention that $\frac{0}{0} \times 0 = 0$.

**Lemma A.1** *1. A subpopulation of compliers $C_F$ satisfies (6) and (7) if and only if there is a real-valued function $g$ defined on $\mathcal{S}(Y_1 - Y_0)$ such that*

$$0 \leq g(\delta) \leq f_{Y_1-Y_0|C}(\delta)P(C) \text{ for } \lambda\text{-almost every } \delta \in \mathcal{S}(Y_1 - Y_0) \tag{16}$$

$$\int_{\mathcal{S}(Y_1-Y_0)} g(\delta)d\lambda(\delta) = P(F) \tag{17}$$

$$\int_{\mathcal{S}(Y_1-Y_0)} \delta \frac{g(\delta)}{P(F)} d\lambda(\delta) = E(Y_1 - Y_0|F). \tag{18}$$

*2. A subpopulation of compliers $C_V$ satisfies (8) and (9) if and only if there is a real-valued function $h$ defined on $\mathcal{S}(Y_1 - Y_0)$ such that*

$$0 \leq h(\delta) \leq f_{Y_1-Y_0|C}(\delta)P(C) \text{ for } \lambda\text{-almost every } \delta \in \mathcal{S}(Y_1 - Y_0) \tag{19}$$

$$\int_{\mathcal{S}(Y_1-Y_0)} h(\delta)d\lambda(\delta) = FS \tag{20}$$

$$\int_{\mathcal{S}(Y_1-Y_0)} \delta \frac{h(\delta)}{FS} d\lambda(\delta) = W. \tag{21}$$

**Proof of Lemma A.1:**

In view of Theorem 2.1, the proof will be complete if I can show the if part of the first statement, the only if part of the second statement, and finally that if a function $h$ satisfies (19), (20), and (21), then a function $g$ satisfies (16), (17), and (18).

I start proving the if part of the first statement. Assume a function $g$ satisfies (16), (17), and (18). Densities being uniquely defined up to 0 probability sets, I can assume without loss of generality that those three equations hold everywhere. Let

$$p(\delta) = \frac{g(\delta)}{f_{Y_1-Y_0|C}(\delta)P(C)} 1\{f_{Y_1-Y_0|C}(\delta) > 0\}.$$

It follows from (16) that $p(\delta)$ is always included between 0 and 1. Then, let $B$ be a Bernoulli random variable such that $P(B = 1|C, Y_1 - Y_0 = \delta) = p(\delta)$. Finally, let $C_F = \{C, B = 1\}$.

$$\begin{aligned}
P(C_F) &= E(P(C_F|Y_1 - Y_0)) \\
&= E(P(C|Y_1 - Y_0)P(B = 1|C, Y_1 - Y_0)) \\
&= E\left(P(C|Y_1 - Y_0)\frac{g(Y_1 - Y_0)}{f_{Y_1-Y_0|C}(Y_1 - Y_0)P(C)}1\{f_{Y_1-Y_0|C}(Y_1 - Y_0) > 0\}\right) \\
&= E\left(\frac{g(Y_1 - Y_0)}{f_{Y_1-Y_0}(Y_1 - Y_0)}\right) \\
&= \int_{\mathcal{S}(Y_1-Y_0)} g(\delta)d\lambda(\delta) \\
&= P(F)
\end{aligned}$$

The first equality follows from the law of iterated expectations, the second from the definition of $C_F$ and Bayes, the third from the definition of $B$, the fourth from the fact that under (16), $f_{Y_1-Y_0|C}(\delta)P(C) = 0 \Rightarrow g(\delta) = 0$, and the last from (17). This proves that $C_F$ satisfies (6). Then,

$$\begin{aligned}
E(Y_1 - Y_0|C_F) &= \frac{E((Y_1 - Y_0)1\{C_F\})}{P(C_F)} \\
&= \frac{E((Y_1 - Y_0)P(C_F|Y_1 - Y_0))}{P(C_F)} \\
&= \frac{E\left((Y_1 - Y_0)\frac{g(Y_1-Y_0)}{f_{Y_1-Y_0}(Y_1-Y_0)}\right)}{P(C_F)} \\
&= \int_{\mathcal{S}(Y_1-Y_0)} \delta\frac{g(\delta)}{P(F)}d\lambda(\delta) \\
&= E(Y_1 - Y_0|F).
\end{aligned}$$

The first equality follows from the definition of a conditional expectation, the fourth from (6), and the fifth from (18). This proves that $C_F$ satisfies (7).

I now prove the only if part of the second statement. Assume a subpopulation of compliers $C_V$ satisfies (8) and (9). Then $h = f_{Y_1-Y_0|C_V}P(C_V)$ must satisfy (19). Otherwise $C_V$ would not be included in $C$. It must also satisfy (20) and (21). Otherwise $C_V$ would not satisfy (8) and (9).

I finally show the last point. Assume $h$ satisfies (19), (20), and (21). Then, it follows from (1) and (2) that $g = f_{Y_1-Y_0|C}P(C) - h$ satisfies (16), (17), and (18).

**QED.**

**Proof of Theorem 3.1:**

I assume that $0 \leq RF$. The proof is symmetric if $RF \leq 0$. I also assume that the data does not reject Assumption 5, i.e. that Equation (31) in de Chaisemartin (2014) is satisfied. This implies that $RF \leq FS$.

I start proving the first implication. Rearranging (2) using (1) yields

$$E(Y_1 - Y_0|C) - E(Y_1 - Y_0|F) = \frac{FS}{FS + P(F)} \left(W - E(Y_1 - Y_0|F)\right).$$

Assumption 7 is therefore equivalent to

$$|W - E(Y_1 - Y_0|F)| \leq W,$$

which implies that $E(Y_1 - Y_0|F) \geq 0$. This proves the first implication.

I now prove the second implication. On that purpose, I show that if Assumption 6 is satisfied, there is a function $h_1$ satisfying (19), (20), and (21). In view of Lemma A.1, this will prove the result.

As I have assumed $0 \leq RF$, Assumption 6 rewrites as $0 \leq E(Y_1 - Y_0|F)$. With binary outcomes this is equivalent to $0 \leq P(Y_1 - Y_0 = 1, F) - P(Y_1 - Y_0 = -1, F)$. With binary outcomes, (2) simplifies to

$$P(Y_1 - Y_0 = 1, C) - P(Y_1 - Y_0 = -1, C) = RF + P(Y_1 - Y_0 = 1, F) - P(Y_1 - Y_0 = -1, F). \quad (22)$$

Once combined with (22), Assumption 6 implies

$$RF \leq P(Y_1 - Y_0 = 1, C). \quad (23)$$

Then, notice that

$$\begin{aligned} & FS - RF - P(Y_1 - Y_0 = 0, C) \\ = \; & 2P(Y_1 - Y_0 = -1, C) - (2P(Y_1 - Y_0 = -1, F) + P(Y_1 - Y_0 = 0, F)) \quad (24) \\ & FS + RF - P(Y_1 - Y_0 = 0, C) \\ = \; & 2P(Y_1 - Y_0 = 1, C) - (2P(Y_1 - Y_0 = 1, F) + P(Y_1 - Y_0 = 0, F)). \quad (25) \end{aligned}$$

Now, consider the function $h_1$ defined on $\{-1, 0, 1\}$ and such that

$$\begin{aligned} h_1(-1) &= \max\left(0, \frac{FS - RF - P(Y_1 - Y_0 = 0, C)}{2}\right) \\ h_1(0) &= \min\left(P(Y_1 - Y_0 = 0, C), FS - RF\right) \\ h_1(1) &= \max\left(RF, \frac{FS + RF - P(Y_1 - Y_0 = 0, C)}{2}\right). \end{aligned}$$

If $FS - RF \leq P(Y_1 - Y_0 = 0, C)$,

$$\begin{aligned} h_1(-1) &= 0 \\ h_1(0) &= FS - RF \\ h_1(1) &= RF. \end{aligned}$$

$h_1(-1)$ is trivially included between 0 and $P(Y_1 - Y_0 = -1, C)$. $0 \leq h_1(0)$ follows from Equation (31) in de Chaisemartin (2014). By assumption, we also have $h_1(0) \leq P(Y_1 - Y_0 = 0, C)$ and $0 \leq h_1(1)$. $h_1(1) \leq P(Y_1 - Y_0 = 1, C)$ follows from (23). This proves that $h_1$ satisfies (19). It is easy to see that it also satisfies (20) and (21).
If $FS - RF > P(Y_1 - Y_0 = 0, C)$,

$$h_1(-1) = \frac{FS - RF - P(Y_1 - Y_0 = 0, C)}{2}$$
$$h_1(0) = P(Y_1 - Y_0 = 0, C)$$
$$h_1(1) = \frac{FS + RF - P(Y_1 - Y_0 = 0, C)}{2}.$$

$h_1(-1)$ is greater than 0 by assumption. $h_1(-1) \leq P(Y_1 - Y_0 = -1, C)$ follows from (24). $h_1(0)$ is trivially included between 0 and $P(Y_1 - Y_0 = 0, C)$. $h_1(1)$ is greater than 0 because it is greater than $h_1(-1)$. $h_1(1) \leq P(Y_1 - Y_0 = 1, C)$ follows from (25). This proves that $h_1$ satisfies (19). It is easy to see that it also satisfies (20) and (21).
**QED.**

**Proof of Theorem 3.2**

I only prove the result when $RF > 0$ (the proof is symmetric when $RF < 0$). If $e \geq 0$, the result directly follows from Theorem 3.1. If $pe \leq -RF$, the result is trivial. Now, let $(p, e)$ be such that $e < 0$ and $pe > -RF$. We then have $\lambda(p, e) = 1 + \frac{pe}{RF}$. To prove the result, I shall show that if $P(F) \leq p$ and $E(Y_1 - Y_0|F) \geq e$, there is a real-valued function $h$ defined on $\{-1, 0, 1\}$ satisfying (19) and

$$\int_{\mathcal{S}(Y_1 - Y_0)} h(\delta) d\lambda(\delta) = \lambda(p, e) FS \tag{26}$$
$$\int_{\mathcal{S}(Y_1 - Y_0)} \delta \frac{h(\delta)}{\lambda(p, e) FS} d\lambda(\delta) = W. \tag{27}$$

This will prove the result, following the logic of Lemma A.1.
$P(F) \leq p$ and $E(Y_1 - Y_0|F) \geq e$ implies $P(Y_1 - Y_0 = 1, F) - P(Y_1 - Y_0 = -1, F) \geq pe$. Combining this with (22) implies

$$\lambda(p, e) RF \leq P(Y_1 - Y_0 = 1, C). \tag{28}$$

Now, consider the function $h_5 = \lambda(p, e) h_1$. It follows from (28), (24), and (25) that $h_5$ satisfies (19). It is easy to see that it also satisfies (26) and (27).
**QED.**

**Proof of Theorem 3.3:**

Under Assumption 8, $g_1 = f_{Y_1 - Y_0|F} P(F)$ satisfies (16), (17), and (18).
**QED.**