Tolerating defiance?
Identification of treatment effects without monotonicity.*

Clément de Chaisemartin†

April 12, 2016

Abstract

Instrumental variable (IV) is a commonly used method to estimate the effect of a treatment. It estimates a causal effect if the instrument satisfies various assumptions, among which a monotonicity condition. This condition restricts the applicability of the IV method, as there are a large number of cases in which it is implausible. I show that IV estimates a causal effect under a weaker condition than monotonicity. I outline several criteria applied researchers can use to assess whether this condition is plausible in their studies, and I review examples where this weaker condition is applicable while monotonicity is not. The monotonicity condition has also been invoked to derive bounds for the average treatment effect. I show that these bounds are still valid under my weaker condition.

Keywords: monotonicity, defiers, instrumental variable, average treatment effect, partial identification

JEL Codes: C21, C26

*I am very grateful to Josh Angrist, Sascha Becker, Stéphane Bonhomme, Federico Bugni, Laurent Davezies, Xavier D’Haultfoeuille, Sara Geneletti, Walker Hanlon, Toru Kitagawa, Andrew Oswald, Azeem Shaikh, Roland Rathelot, Ed Vytlacil, 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, c.de-chaisemartin@warwick.ac.uk
1 Introduction

Applied economists 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). As is well-known, a valid instrument should be as good as randomly assigned and should not have a direct effect on the outcome. But even with an instrument satisfying these two conditions, the resulting 2SLS 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 estimates 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. Defers could be present in a large number of applications, and I will now give three 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), or receipt of disability insurance (see Maestas et al., 2013, French & Song, 2012, and Dahl et al., 2013). 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.

Second, defiers could be present in studies relying upon sibling-sex composition as an instrument for family size, because some parents are sex-biased. In the US, 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. Dahl & Moretti (2008) show that in the US, 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.

Third, defiers could be present in randomized controlled trials relying on an encouragement design. Dufo & 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 given a financial incentive upon attendance. Deci (1971) and Frey & Jegen (2001) provide evidence showing that financial incentives sometimes backfire because they crowd-out intrinsic motivation. Sometimes, the crowding-out effect even seems to dominate: Gneezy & Rustichini (2000) find that fining parents who pick their children late at day-care centers actually increased the number of late-coming parents. Accordingly, paying subjects to get treated in encouragement designs could lead some of them to forgo treatment.

In this paper, 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 on compliance types under which 2SLS estimates a LATE: if it is violated, 2SLS does not estimate a causal effect.

The CD condition is somewhat abstract, so I derive more interpretable sufficient conditions. I start by showing that CD holds if in each stratum of the population with the same value of their treatment effect there are more compliers than defiers. If that is the case, within each stratum one can form a subgroup of compliers with as many units as defiers. Pooling these subgroups across strata yields a subgroup of compliers accounting for the same percentage of the population as defiers and with the same LATE. I further show that with binary outcomes, CD holds if defiers’ LATE and the 2SLS coefficient have the same sign; or if defiers’ and compliers’ LATE have the same sign and the ratio of these two LATEs is lower than the ratio of the shares of compliers and defiers in the population; or if the difference between compliers’ and defiers’ LATEs is not larger than some upper bound which can be estimated from the data.

These results have practical applicability. Maestas et al. (2013) study the effect of disability insurance on labor market participation. Their 2SLS coefficient is negative. In standard labor supply models, disability insurance can only reduce labor market participation because it increases non-labor income. It is therefore plausible that defiers’ LATE is negative, and has the same sign as their 2SLS coefficient, thus implying that CD should hold in this study.
Therefore, even though their coefficient might not estimate the LATE of compliers, it follows from my results that it still estimates the LATE of a subgroup of compliers. Later in the paper, I argue that this restriction on the sign of defiers’ LATE is also plausible in French & Song (2012), Aizer & Doyle (2013), and Duflo & Saez (2003). The CD condition should hold in these studies as well. Angrist & Evans (1998) study the effect of having a third child on mothers labor market participation. I estimate the upper bound mentioned in the previous paragraph in their data, and find that it 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.

Overall, the 2SLS method is applicable in studies in which defiers could be present, provided one can reasonably assume that defiers’ LATE has the same sign as the 2SLS coefficient, or that compliers’ and defiers’ LATE do not differ too much. As I explain in more details later, my CD condition is also more likely to hold when the instrument has a large first stage.

2SLS is not the only statistical method requiring that there be no defiers. An important example are bounds for the average treatment effect (ATE) derived under the assumption that treatment effects have the same sign for all units in the population (see Bhattacharya et al., 2008, Chesher, 2010, Chiburis, 2010, Shaikh & Vytlačil, 2011, and Chen et al., 2012). All of these bounds rely on the assumption that there are no defiers in the population. Actually, I show that these bounds are still valid under my CD condition.

Other papers have studied relaxations of the “no-defiers” condition. Klein (2010) considers a model in which a disturbance uncorrelated with treatment effects 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 if in each stratum of the population with the same value of their two potential outcomes there are more compliers than defiers, a condition they refer to as “stochastic monotonicity”, then 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). Huber & Mellace (2012) consider a local monotonicity assumption which requires that there be only compliers or defiers conditional on each value of the outcome. The CD condition allows for both compliers and defiers conditional on the outcome. Finally, Fiorini et al. (2013) provide practitioners

\footnote{Actually, Chen et al. (2012) only require that the LATEs of compliers, never-takers, always-takers, and defiers all have the same sign.}
with recommendations as to how they should investigate the plausibility of the “no-defiers” condition in their applications.

The remainder of the paper is organized as follows. Section 2 concerns identification, Section 3 concerns inference, Section 4 concerns results of a simulation study, Section 5 concerns empirical applications, and Section 6 concludes. Most proofs are deferred to the appendix. For the sake of brevity, I consider some extensions in a paper gathering supplementary material. In this paper, I show that one can estimate quantile treatment effects among a subpopulation of compliers even if there are defiers, that one can test the CD condition, and that my results extend to multivariate treatment and instrument.

2 Identification

2.1 Identification of a LATE with defiers

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”. The results presented in this section extend to more general settings with multivariate instrument and treatment. These extensions are deferred to the supplementary material.

Imbens & Angrist (1994) 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 $Z$, $D \equiv D_Z$ and $Y \equiv Y_{DZ}$ are observed. Following Angrist et al. (1996), 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 $FS \neq 0$. I will further assume throughout the paper that $FS > 0$. 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. Under Assumption 1 (see below), this normalization implies that more subjects are compliers than defiers: $P(C) > P(F)$.

$^2$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.
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)*

\[P(F) = 0.\]

**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 Imbens & Angrist (1994) and Angrist et al. (1996).

**LATE Theorems** *(Imbens & Angrist, 1994 and Angrist et al., 1996)*

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

\[
FS = P(C) - P(F) \\
W = \frac{P(C)E(Y_1 - Y_0|C) - P(F)E(Y_1 - Y_0|F)}{P(C) - P(F)}. \tag{1}
\]

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

\[
FS = P(C) \tag{3} \\
W = E(Y_1 - Y_0|C). \tag{4}
\]

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

\[W = E(Y_1 - Y_0|C). \tag{5}\]
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 compliers 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 “comfers”. CD is weaker than Assumptions 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 comfers: 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

\[
P(C_V) = FS \tag{8}
\]

\[
E(Y_1 - Y_0 | C_V) = W. \tag{9}
\]

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

**Proof**

\[
FS = P(C) - P(F) = P(C_V) + P(C_F) - P(F) = P(C_V).
\]

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

\[ E(Y_1 - Y_0|C) = P(C_V|C)E(Y_1 - Y_0|C_V) + P(C_F|C)E(Y_1 - Y_0|C_F) \]

\[ = \frac{P(C) - P(F)}{P(C)}E(Y_1 - Y_0|C_V) + \frac{P(F)}{P(C)}E(Y_1 - Y_0|F), \]

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

\[ W = E(Y_1 - Y_0|C_V). \]

This proves that \( C_V \) satisfies (9).

\[ \Leftrightarrow \]

\[ P(C_F) = P(C) - P(C_V) = P(C) - FS = P(C) - (P(C) - P(F)) = P(F). \]

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

Then,

\[ E(Y_1 - Y_0|C) = \frac{FS}{P(C)}W + \frac{P(F)}{P(C)}E(Y_1 - Y_0|C_F), \]

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

\[ E(Y_1 - Y_0|F) = E(Y_1 - Y_0|C_F). \]

This proves that \( C_F \) satisfies (7).

QED.

This result is derived from Equations (1) and (2), after using the law of iterated expectations and invoking Assumption 5. The intuition underlying it goes as follows. Under CD, compliers and defiers cancel one another out, and the 2SLS coefficient is equal to the effect of the treatment for the remaining part of compliers. I hereafter refer to the \( C_V \) subpopulation as “compliers-survivors”, or “convivors”, as they are compliers who “out-survive” defiers.

The LATE in Theorem 2.1 is harder to grasp than the LATE identified under the no-defiers assumption. It does not apply to all compliers, but only to a subset of them, the convivors subpopulation. Note that under the no-defiers assumption, compliers account for the same percentage of the population as convivors under the CD assumption. Therefore, the LATE in Theorem 2.1 does not apply to a smaller population than the LATE identified under the no-defiers assumption. Moreover, as I show in the next subsection, one can estimate the mean of any covariate (age, sex...) among convivors under a mild strengthening of the CD assumption. Thus, the analyst can assess whether convivors strongly differ from the entire
population. Still, convivors differ from compliers in that they are not fully characterized by their potential treatments. Knowing $D_0$ and $D_1$ is not sufficient to tell apart convivors from compliers. Actually, in most instances even knowing $Y_1 - Y_0$ is not sufficient to tell apart the two populations. If a convivor and a complier have the same value of $Y_1 - Y_0$, switching the convivor to the complier population, and the complier to the convivor population will not change the LATE and the size of the new convivor and complier populations. Thus, as soon as the supports of $Y_1 - Y_0$ in the two populations overlap, they are not uniquely defined.

This raises the question of whether this LATE is an interesting parameter. Some authors consider that treatment effect parameters are worth considering if they can inform treatment choice (see Manski, 2005). From that perspective, LATEs are not necessarily interesting: 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). However, other authors have argued that researchers should still report an estimate of the LATE of compliers, along with the bounds on the ATE (see Imbens, 2010). Their arguments can be summarized as follows: reporting only the bounds might leave out relevant information; the LATE of compliers can give researchers an idea of the magnitude of the treatment effect; under some assumptions this LATE can be extrapolated to other populations (see Angrist & Fernandez-Val, 2010). In a world with defiers, these arguments do not apply anymore. In such a world, the LATE of compliers is not even identified. Only the LATE of convivors can be identified. Accordingly, it is this parameter which should be reported along with bounds on the ATE.³

A great appeal of the ND condition is that it is simple to interpret. On the contrary, CD is an abstract condition. I try to clarify its meaning by deriving more interpretable conditions under which it is satisfied.

**A sufficient condition for CD to hold**

I start by considering a condition which is sufficient for CD to hold irrespective of the nature of the outcome. Let $R(P(F)) = 1 + \frac{FS}{R(F)}$. Notice that Equation (1) implies that $R(P(F)) = \frac{P(C)}{R(F)}$. Therefore, $R(P(F))$ is merely the ratio of the shares of compliers and defiers in the population.

**Assumption 6 (More compliers than defiers: MC)**

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

$$\frac{f_{Y_1 - Y_0}(\delta)}{f_{Y_1 - Y_0}(\delta)} \leq R(P(F)).$$

³The extrapolation strategy proposed in Angrist & Fernandez-Val (2010) under the no-defers assumption can also be used under the compliers-defiers assumption introduced in this paper.
I call this condition the more compliers than defiers condition. Indeed, as \( R(P(F)) = \frac{P(C)}{P(F)} \), Equation (10) is equivalent to

\[
P(F|Y_1 - Y_0) \leq P(C|Y_1 - Y_0).
\]  

(11) 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). For instance, their condition is satisfied if \( P(F|Y_0, Y_1) \leq P(C|Y_0, Y_1) \), i.e. if in each stratum of the population with the same value of their two potential outcomes there are more compliers than defiers.

As shown in Angrist et al. (1996), 2SLS estimates 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 (10) is then equal to 1, while its right-hand side is greater than 1.\(^4\) And MC also holds when there are no defiers, as the right hand side of (10) is then equal to +\( \infty \).

**Theorem 2.2** Assumption 6 \( \Rightarrow \) Assumption 5.

To convey the intuition of this Theorem, I consider the example displayed in Figure 1. \( Y_0 \) 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 1. MC holds as there are more compliers than defiers in each cell.

<table>
<thead>
<tr>
<th>( Y(1)-Y(0) )</th>
<th>Defiers</th>
<th>Compliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>f1 f2</td>
<td>c1 c2 c3</td>
</tr>
<tr>
<td>0</td>
<td>f3 f4 f5</td>
<td>c4 c5 c6 c7 c8</td>
</tr>
<tr>
<td>1</td>
<td>f6 f7</td>
<td>c9 c10 c11 c12 c13</td>
</tr>
</tbody>
</table>

Figure 1: A population in which MC is satisfied.

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 2. Comfiers account for the same percentage of the population as defiers and also have the same LATE.

<table>
<thead>
<tr>
<th>( Y(1)-Y(0) )</th>
<th>Defiers</th>
<th>Comfiers</th>
<th>Comvivors</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>f1 f2</td>
<td>c1 c2</td>
<td>c3</td>
</tr>
<tr>
<td>0</td>
<td>f3 f4 f5</td>
<td>c4 c5 c6</td>
<td>c7 c8</td>
</tr>
<tr>
<td>1</td>
<td>f6 f7</td>
<td>c9 c10</td>
<td>c11 c12 c13</td>
</tr>
</tbody>
</table>

Figure 2: This population also satisfies CD.

\(^4\)I have assumed, as a mere normalization, that \( FS > 0 \).
$R(P(F))$ is increasing in $FS$, and decreasing in $P(F)$, so Assumption 6 is more plausible in applications with a large first stage, and in applications where defiers are unlikely to account for a very large share of the population. Because $P(F)$ is not identified, neither is $R(P(F))$. To get a sense of the plausibility of Assumption 6, one can estimate $R(P(F))$ for plausible values of $P(F)$. If one does not want to make any assumption on $P(F)$, one can also derive a worst case lower bound for $R(P(F))$. Indeed,

$$P(F) \leq \min(P(D = 1|Z = 0), P(D = 0|Z = 1)) \equiv P(F). \quad (12)$$

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. $P(F) \leq P(F)$ implies the following worst-case lower bound for $R(P(F))$:

$$1 + \frac{FS}{P(F)} \leq R(P(F)). \quad (13)$$

**More sufficient conditions with a binary outcome**

While Assumption 6 is intuitive, there might be applications where it is hard to gauge its plausibility. I now derive conditions which are sufficient for CD to hold when the outcome is binary, and whose plausibility should be easy to assess in most applications.

Let $sgn[.]$ denote the sign function: for any real number $x$, $sgn[x] = 1\{x > 0\} - 1\{x < 0\}$. Let also $\Delta(P(F)) = \frac{|RF|}{FS+P(F)} = |W| \frac{FS}{FS+P(F)}$. Notice that Equation (1) implies that $FS+P(F) = P(CV)$. Therefore, $\Delta(P(F))$ is equal to the absolute value of the Wald ratio, weighted by the ratio of the shares of convivors and compliers in the population.

The three following conditions are sufficient for CD to hold when the outcome is binary.

**Assumption 7** (Restriction on the sign of the LATE of defiers)

$sgn[E(Y_1 - Y_0|F)] = sgn[W]$, or either $E(Y_1 - Y_0|F)$ or $W$ is equal to 0.

**Assumption 8** (Equal signs and bounded ratio of the LATE of defiers and compliers)

Either $sgn[E(Y_1 - Y_0|F)] = sgn[E(Y_1 - Y_0|C)] \neq 0$ and $\frac{E(Y_1 - Y_0|F)}{E(Y_1 - Y_0|C)} \leq R(P(F))$, or $E(Y_1 - Y_0|F) = 0$.

**Assumption 9** (Restriction on the difference between compliers’ and defiers’ LATE)

$$|E(Y_1 - Y_0|C) - E(Y_1 - Y_0|F)| \leq \Delta(P(F)).$$
Theorem 2.3 If \( Y_0 \) and \( Y_1 \) are binary and \( |W| \leq 1 \),\(^5\) Assumption 9 \( \Rightarrow \) Assumption 8 \( \Leftrightarrow \) Assumption 7 \( \Rightarrow \) Assumption 5.

The first implication and the equivalence follow after some algebra. The second implication states that if the LATE of defers 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

\[
RF = P(Y_1 - Y_0 = 1, C) - P(Y_1 - Y_0 = -1, C) - (P(Y_1 - Y_0 = 1, F) - P(Y_1 - Y_0 = -1, F)).
\]

To fix ideas, suppose that Assumption 7 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 defers’ positive LATE.

Assumption 7 requires that defers’ LATE have the same sign as \( W \). The sign of \( W \) is not known, but it can be inferred from the data using \( \hat{W} \) and an estimator of its standard deviation. When \( W < 0 \) is rejected and \( E(Y_1 - Y_0|F) \geq 0 \) is a plausible restriction in the application under consideration, one can invoke Theorem 2.3 to claim that \( \hat{W} \) consistently estimates the LATE of compliers. When \( W > 0 \) is rejected and \( E(Y_1 - Y_0|F) \leq 0 \) is a plausible restriction, one can also invoke Theorem 2.3. On the other hand, when one fails to reject \( W > 0 \) or \( W < 0 \), one cannot assess whether Assumption 7 is plausible because the data does not give sufficient guidance on the sign of \( W \).

Assumption 8 requires that defers’ and compliers’ LATE have the same sign, and that their ratio be lower than \( R(P(F)) \). Notice that \( R(P(F)) \) is greater than 1. Therefore, when it is plausible to assume that the two LATEs have the same sign, and that defers react less to the treatment thus implying that their LATE is closer to 0, one can invoke Theorem 2.3 to claim that \( \hat{W} \) consistently estimates the LATE of compliers.

Finally, Assumption 9 requires that the difference between defers’ and compliers’ LATEs be smaller in absolute value than \( \Delta(P(F)) \). \( \Delta(P(F)) \) is increasing in \( |W| \) and \( FS \), and decreasing in \( P(F) \). Therefore, Assumption 9 is more likely to be satisfied when the instrument has large

\(^5\) Assuming that \( |W| \leq 1 \) is without loss of generality. If \( |W| > 1 \), Assumption 5 cannot be true anyway as with a binary outcome there cannot be a subgroup of compliers with a LATE strictly greater or strictly lower than 1. In the supplementary material, I discuss testable implications of Assumption 5.
first and second stages, and when defiers are unlikely to account for a large fraction of the population. Here as well, one can estimate \( \Delta(P(F)) \) for plausible values of \( P(F) \). One can also estimate a worst case lower bound for \( \Delta(P(F)) \). Indeed, \( P(F) \leq \overline{P}(F) \) implies the following worst-case lower bound for \( \Delta(P(F)) \):

\[
|W| \frac{FS}{FS + \overline{P}(F)} \leq \Delta(P(F)).
\]  

(14)

2.2 Incorporating covariates into the analysis

Instruments are sometimes valid only after conditioning for some covariates. Theorem 2.4 below shows that identifying the LATE of compliers in such instances does not require a strengthening of the CD condition.

Let \( X \) denote a vector of covariates. Assume that instead of Assumption 1, the following assumption is satisfied.

**Assumption 10 (Instrument conditional independence)**

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

I prove the following result.

**Theorem 2.4** Suppose Assumptions 10, 2, and 5 hold. Then \( C_V = C \setminus C_F \) satisfies

\[
P(C_V) = E(E(D|Z = 1, X) - E(D|Z = 0, X))
\]

\[
E(Y_1 - Y_0|C_V) = \frac{E(E(Y|Z = 1, X) - E(Y|Z = 0, X))}{E(E(D|Z = 1, X) - E(D|Z = 0, X))}
\]

The estimand identifying the LATE in Theorem 2.4 is not the same as that in Theorem 2.1, but it is the same as the one considered in Frölich (2007). Frölich (2007) proposes an estimator and derives its asymptotic distribution.

Under the no-defiers condition, one can recover the mean of any covariate among compliers (this follows from Abadie (2003), for instance). This is a desirable property, as LATEs apply to subpopulations. Therefore, applied researchers often want to describe these subpopulations, so as to assess whether their LATEs are likely to extend to other populations. When the instrument is unconditionally independent of potential treatments and outcomes and when it is also independent of \( X \), one can recover the mean of \( X \) among compliers under a mild strengthening of Assumption 5.6

---

6When the instrument is not independent of \( X \), the mean of \( X \) among compliers is still identified if one is ready to assume that Equations (6) and (7) hold conditional on \( X \).
Assumption 11 (Conditional compliers-defiers)
There is a subpopulation of compliers $C_F$ which satisfies Equations (6) and (7), and

$$E(X|C_F) = E(X|F).$$  \hspace{1cm} (15)

Let $W_{XD} = \frac{E(XD|Z=1) - E(XD|Z=0)}{P(D=1|Z=1) - P(D=1|Z=0)}$.

Theorem 2.5 Suppose Assumptions 1, 2, and 11 hold, and $Z \perp \perp X$. Then $C_V = C \setminus C_F$ satisfies Equations (8), (9), and

$$E[X|C_V] = W_{XD}.\hspace{1cm} (16)$$

2.3 Partial identification of the ATE with defiers

Shaikh & Vytlacil (2011) consider a model with binary treatment and outcome, where the treatment and the outcome are both determined by threshold-crossing single-index equations. The sharp bounds for the ATE under their assumptions are tighter than those obtained under Assumptions 1 and 2 and studied in Manski (1990), Balk & Pearl (1997), or Kitagawa (2009).

In particular, the sign of the ATE is identified under their assumptions. Their single-index model for treatment implies that there cannot be defiers in the population. Similarly, their single-index model for the outcome implies that the sign of the treatment effect is the same for all units in the population. The next theorem shows that their result holds even if there are defiers in the population.

Assumption 12 (Sign restrictions on the LATEs of all subpopulations)
For every $(T_1, T_2) \in \{AT, NT, C, F\}^2$, $\text{sgn}[E(Y_1 - Y_0|T_1)] \times \text{sgn}[E(Y_1 - Y_0|T_2)] \geq 0$.

Theorem 2.6 Assume that $Y_0$ and $Y_1$ are binary, and that Assumptions 1, 2, 8, and 12 are satisfied.

1. If $RF > 0$,

$$RF \leq E(Y_1 - Y_0) \leq P(Y = 1, D = 1|Z = 1) - P(Y = 0, D = 0|Z = 0) + P(D = 0|Z = 1).$$

2. If $RF < 0$,

$$P(Y = 1, D = 1|Z = 1) - P(Y = 0, D = 0|Z = 0) - P(D = 1|Z = 0) \leq E(Y_1 - Y_0) \leq RF.$$

These bounds are sharp if for every $(y, d) \in \{0, 1\}^2$, $P(Y = y, D = d|Z = d) \geq P(Y = y, D = d|Z = 1 - d).$\hspace{1cm} \footnote{This condition is equivalent to the testable implication of the LATE assumptions studied by Kitagawa (2013) (Equation (1.1) in his paper).}
Assumption 12 requires that the LATEs of always takers, never takers, compliers, and defiers all have the same sign. This restriction is plausible in applications where selection into one or the other population is not directly based on gains from treatment, making it unlikely that LATEs switch sign across subpopulations. If one is further ready to assume that defiers are less affected by the treatment than compliers, thus implying that their LATE is closer to 0, one can use Theorem 2.6 to sign and bound the ATE, even if there are defiers in the population.

The bounds presented in this theorem are not new. They coincide with those in Bhattacharya et al. (2008), Chiburis (2010), and Chen et al. (2012), and with those in Chesher (2010) and Shaikh & Vytlacil (2011) with no covariates and a binary instrument. Assumption 12 has already been considered in Chen et al. (2012). The novelty is that here, I show that these bounds are valid even if there are defiers in the population provided Assumption 8 is satisfied.

The intuition for the lower bound goes as follows. Assume that $RF > 0$. If $\frac{E(Y_1 - Y_0 \mid C)}{E(Y_1 - Y_0 \mid C)} < \frac{P(C)}{P(F)}$ and $E(Y_1 - Y_0 \mid C)$ and $E(Y_1 - Y_0 \mid F)$ have the same sign, it is easy to see from Equation (2) that $E(Y_1 - Y_0 \mid C)$ and $E(Y_1 - Y_0 \mid F)$ must have the same sign as $RF$. $E(Y_1 - Y_0 \mid AT)$, $E(Y_1 - Y_0 \mid NT)$, $E(Y_1 - Y_0 \mid C)$, and $E(Y_1 - Y_0 \mid F)$ must therefore be positive. Moreover, it follows from Theorem 2.3 that CD is satisfied under the assumptions of Theorem 2.6. Therefore, there is a subgroup of units accounting for $FS\%$ of the population with a LATE equal to $W$. This combined with the fact that the remaining units must have a positive LATE yields $RF \leq E(Y_1 - Y_0)$.

These bounds are sharp when the standard LATE assumptions are not rejected. As noted in Balke & Pearl (1997) and Heckman & Vytlacil (2005), Assumptions 1, 2, and 3 have testable implications. Equation (1.1) in Kitagawa (2013) summarizes these testable implications. In many applications, Equation (1.1) is not rejected, so deriving sharp bounds under this restriction is without great loss of generality. Still, as I discuss in the supplementary material, Assumptions 1, 2, and the CD condition might hold while Kitagawa’s Equation (1.1) is violated. Deriving sharp bounds without this restriction is left for future work.

As can be seen in points 1 and 2 of Theorem 2.6, the expression of the bounds depends on the sign of $RF$. This quantity is unknown but can be estimated. When $RF = 0$ is rejected and $\hat{RF} \geq 0$, one can use the sample counterpart of $RF$ and $P(Y = 1, D = 1\mid Z = 1) - P(Y = 0, D = 0\mid Z = 0) + P(D = 0\mid Z = 1)$ as lower and upper bounds of the ATE. When $RF = 0$ is rejected and $\hat{RF} \leq 0$, one can use the sample counterpart of $P(Y = 1, D = 1\mid Z = 1) - P(Y = 0, D = 0\mid Z = 0) - P(D = 1\mid Z = 0)$ and $RF$ as lower and upper bounds of the ATE. On the other hand, when $RF = 0$ is not rejected, the data does not give sufficient guidance on the sign of this quantity, so the ATE cannot be bounded and signed.

Finally, to draw inference on the ATE I refer the reader to Shaikh & Vytlacil (2005). In their Theorem 7.1, they develop a method to derive a confidence interval for the ATE based on the
bounds obtained in Theorem 2.6.

3 Inference

I briefly sketch how one can use results from Andrews & Soares (2010) to draw inference on $P(F)$ using the worst case upper bound derived in Equation (12). Following similar steps, one can also use their results to draw inference on $R(P(F))$ and $\Delta(P(F))$ using the worst case upper bounds derived in Equations (13) and (14).

It follows from Equation (12) that

$$P(F) \leq \min(P(D = 1|Z = 0), P(D = 0|Z = 1)).$$

This rewrites as

$$0 \leq E(D(1 - Z) - (1 - Z)P(F))
0 \leq E((1 - D)Z - ZP(F)).$$

This defines a moment inequality model. Because $D$ and $Z$ are binary, this model satisfies all the conditions necessary for Theorem 1 in Andrews & Soares (2010) to apply. One can therefore use their method to derive a uniformly valid confidence upper bound for $P(F)$.

The moment inequality model in the previous display also falls into the framework studied by Romano et al. (2014). Therefore, one could use their results to draw inference on $P(F)$. One advantage of their procedure relative to that of Andrews & Soares (2010) is that it does not rely on the choice of a tuning parameter. Instead, they propose a two-step procedure whereby they start selecting the binding moment conditions through a test of level $\beta$, and then derive a confidence interval for the target parameter using the $1 - \alpha + \beta$ critical values of the asymptotic distribution of the selected moment conditions. However, their procedure cannot accommodate for preliminary estimated parameters in the moment inequalities, contrary to that of Andrews & Soares (2010). The moment inequality models involving $R(P(F))$ and $\Delta(P(F))$ both have preliminary estimated parameters. For instance, one has

$$0 \leq E(D(1 - Z)E(Z)R(P(F)) - DZE(1 - Z))$$
$$0 \leq E((1 - D)ZE(1 - Z)R(P(F)) - (1 - D)(1 - Z)E(Z)).$$

This moment inequality model has one preliminary estimated parameter, $E(Z)$. Therefore, results from Romano et al. (2014) cannot be used to draw inference on $R(P(F))$ and $\Delta(P(F))$. 
4 A simulation study

In this section, I assess the validity of the CD condition in a trivariate normal selection model inspired from Heckman (1979). On that purpose, I consider a model in which potential treatments are determined through the following threshold-crossing selection equations: for every \( z \in \{0, 1\} \),

\[
D_z = 1\{V_z \geq v_z\}.
\]  

(17)

\( V_0 \) and \( V_1 \) are two random variables respectively representing one’s taste for treatment without and with the instrument. \( v_0 \) and \( v_1 \) are two real numbers. Without loss of generality, one can assume that \( V_0 \) and \( V_1 \) have the same marginal distributions, and that \( v_1 \leq v_0 \) to account for the fact that \( P(D_1 = 1) \geq P(D_0 = 1) \). Compliers satisfy \( \{V_0 < v_0, V_1 \geq v_1\} \): the instrument substantially diminishes their taste for treatment, which induces them not to get treated when they receive it.

Vytlačil (2002) shows that ND is equivalent to imposing \( V_0 = V_1 \). I will not make this assumption here to allow for defiers. On the other hand, I will assume that \((V_0, V_1, Y_1 - Y_0)\) is jointly normal:

\[
\begin{pmatrix}
V_0 \\
V_1 \\
Y_1 - Y_0
\end{pmatrix}
\sim
\mathcal{N}
\left(
\begin{pmatrix}
0 \\
0 \\
\mu
\end{pmatrix},
\begin{pmatrix}
1 & \rho_{V_0, V_1} & \sigma_{\Delta} \rho_{V_0, \Delta} \\
\rho_{V_0, V_1} & 1 & \sigma_{\Delta} \rho_{V_1, \Delta} \\
\sigma_{\Delta} \rho_{V_0, \Delta} & \sigma_{\Delta} \rho_{V_1, \Delta} & \sigma_{\Delta}^2
\end{pmatrix}
\right).
\]

Let \( \Sigma \) denote the variance of this vector. \( V_0 \) and \( V_1 \) are normalized to have mean 0 and variance 1. I further assume that \( \sigma_{Y_0}^2 = 1 \) and \( \sigma_{Y_1}^2 = \sigma_{Y_1}^2 \). The first assumption is a mere normalization, which corresponds to the common practice of standardizing the outcome by its standard deviation in empirical work. The second one is an homoscedasticity condition. Together, they imply that \( \sigma_{\Delta}^2 \leq 4 \). The data also imposes a number of restrictions on the parameters of this model. It reveals \( v_0 \) and \( v_1 \): \( v_z = \Phi^{-1}(P(D = 0|Z = z)) \), where \( \Phi(.) \) denotes the cdf of a standard normal variable. It also imposes that \( \rho_{V_1, \Delta} \) write as a function of \( \mu, \sigma_{\Delta}, \rho_{V_0, \Delta} \):

\[
\rho_{V_1, \Delta} = \frac{RF - \mu FS}{\sigma_{\Delta} \phi(v_1)} + \rho_{V_0, \Delta} \frac{\phi(v_0)}{\phi(v_1)},
\]

where \( \phi(.) \) is the pdf of a standard normal. Combining the last equation with \( 0 \leq \sigma_{\Delta} \leq \sqrt{4} \), \( -1 \leq \rho_{V_0, \Delta} \leq 1 \), and \( -1 \leq \rho_{V_1, \Delta} \leq 1 \), one can show that the data also bounds \( \mu \):

\[
\mu = \frac{RF - 2(\phi(v_0) + \phi(v_1))}{FS} \leq \mu \leq \bar{\mu} = \frac{RF + 2(\phi(v_0) + \phi(v_1))}{FS}.
\]
Overall, the parameters of the model are partially identified, and the identified set is defined by the following constraints:

$$\theta = (\mu, \sigma^2_\Delta, \rho_{V_0, V_1}, \rho_{V_0, \Delta}) \in \Theta = [\mu, \overline{\mu}] \times [0, 4] \times [-1, 1] \times [-1, 1]$$

$$\rho_{V_1, \Delta}(\theta) = \frac{RF - \mu FS}{\sigma_\Delta \phi(v_1)} + \rho_{V_0, \Delta} \frac{\phi(v_0)}{\phi(v_1)} \in [-1, 1]$$

$\Sigma$ is positive definite.

Finally, note that if $\rho_{V_0, \Delta} = \rho_{V_1, \Delta}$, CD is satisfied. Indeed, we then have $(V_0, V_1)|Y_1 - Y_0 \sim (V_1, V_0)|Y_1 - Y_0$, so $C_F = \{V_1 \geq v_0, V_0 < v_1\}$ satisfies Equations (6) and (7):

$$P(C_F) = P(V_1 \geq v_0, V_0 < v_1) = P(V_0 \geq v_0, V_1 < v_1) = P(F)$$

$$E(Y_1 - Y_0|C_F) = E(Y_1 - Y_0|V_1 \geq v_0, V_0 < v_1) = E(Y_1 - Y_0|V_0 \geq v_0, V_1 < v_1) = E(Y_1 - Y_0|F).$$

In my simulations, I consider a first numerical example in which $P(D = 1|Z = 1) = 0.4$, $P(D = 1|Z = 0) = 0.1$, and $W = 0.2$. This could for instance correspond to a randomized experiment with a first stage of 30%, and with a 2SLS coefficient equal to 20% of the standard deviation of the outcome. I also consider a second numerical example in which $P(D = 1|Z = 1) = 0.2$, $P(D = 1|Z = 0) = 0.1$, and $W = 0.2$. This could for instance correspond to a randomized experiment with a weaker first stage of 10%, and the same 2SLS coefficient. For each numerical example, I draw a sample of 4 000 vectors of parameters representative of the population of parameters compatible with the data. To do so, I draw values for $\theta$ from the uniform distribution on $\Theta$, and keep only those such that $\rho_{V_1, \Delta}(\theta) \in [-1, 1]$ and $\Sigma$ is positive definite. For each vector of parameters, I draw 100 000 realizations from the corresponding distribution of $(V_0, V_1, Y_1 - Y_0)$. This also gives me 100 000 realizations of $(D_0, D_1, Y_1 - Y_0)$. For each of these 4 000 empirical distributions of $(D_0, D_1, Y_1 - Y_0)$, I assess whether it satisfies the CD assumption using an algorithm presented in the appendix.

The main results from this exercise are as follows. First, CD is more likely to hold when the instrument has a large than a weak first stage. While in the first numerical example CD is satisfied for 67% of the 4000 DGPs considered, in the second example it is only satisfied for 43% of them. Second, CD is more likely to hold when the LATE of defiers has the same sign as the 2SLS coefficient. Most DGPs for which $E(Y_1 - Y_0|F) \geq 0$ satisfy CD. However, some DGPs for which $E(Y_1 - Y_0|F)$ is very large violate it. For instance, across the 4000 DGPs in the first numerical example, the DGP with the lowest positive value of $E(Y_1 - Y_0|F)$ for which CD is violated has $E(Y_1 - Y_0|F) = 0.86\sigma_{Y_0}$, a very large treatment effect. Third, the difference between $\rho_{V_1, \Delta}$ and $\rho_{V_0, \Delta}$ seems to be the main determinant of whether CD is satisfied or not in this model. A regression of a dummy for whether CD is satisfied on $|\rho_{V_1, \Delta} - \rho_{V_0, \Delta}|$ has an $R^2$ of 0.66. Adding $(\mu, \sigma^2_\Delta, \rho_{V_0, V_1}, \rho_{V_0, \Delta}, \rho_{V_1, \Delta})$ to this regression hardly adds any explanatory power.
These results might help applied researchers to assess whether CD is likely to hold when their outcome of interest is continuous. When their 2SLS coefficient is, say, positive, they can assess whether defiers are likely to have a negative or a very large positive treatment effect. If that sounds unlikely, CD is likely to hold. Similarly, when their first stage is large, they can be more confident that their results are robust to defiers than when it is weak.

To conclude this section, it is worth noting that results presented in this paper generalize to the local IV approach introduced in Heckman & Vytlacil (1999) and Heckman & Vytlacil (2005). These authors show that with a continuous instrument \( Z \) satisfying Assumptions 1 and 2, if Equation (17) is satisfied with i) \( V_z = V \) for every \( z \) in the support of \( Z \) and ii) \( v_z \) decreasing in \( z \), then under some regularity conditions \( \frac{\partial E(Y|P(D=1|Z=z)=p)}{\partial p} \) is equal to the average treatment effect of units at the \( 1-p^{th} \) quantile of the distribution of \( V \). This result can be extended to selection equations where \( V_z \) is allowed to vary across values of \( z \), under a generalization of the CD condition. For instance, if for every \( z_1 \) in the support of \( Z \) there is a \( z_0 < z_1 \) such that for every \( z \in [z_0, z_1] \) there is a subset of the \( \{V_{z_1} \geq v_{z_1}, V_z < v_z\} \) subpopulation accounting for the same percentage of the total population and with the same average treatment effect as the \( \{V_{z_1} < v_{z_1}, V_z \geq v_z\} \) subpopulation, then \( \frac{\partial E(Y|P(D=1|Z)=p)}{\partial p} \) is equal to the average treatment effect of units at the \( 1-p^{th} \) quantile of the distribution of \( V_{z_f} \), where \( z^p \) is the unique solution of \( P(D = 1|Z = z) = p \).

5 Applications

In this section, I show how one can use the previous results in various applications where it is likely that defiers are present.

Maestas et al. (2013) and French & Song (2012)

Maestas et al. (2013) study the effect of receiving disability insurance on labor market participation. They use average allowance rates of randomly assigned examiners as an instrument for receipt of DI. In this context, \( Y_1 \leq Y_0 \) is a plausible restriction.\(^8\) It is for instance satisfied in a static labor supply model under standard restrictions on agents’ utility functions. Assume agents’ utilities depend on consumption \( C \) and leisure \( L \). To simplify, assume agents can only work full-time or not work at all, which is denoted by a dummy \( Y \). To choose \( Y \), agents maximise \( U(C, L) \) subject to \( C = YW + I \) and \( L = T - HY \), where \( W, I, H, \) and \( T \) respectively denote agents’ wages, their non-labor income, the amount of time spent on a full-time job, and the total amount of time available. Let \( U_{CC}, U_{LL}, \) and \( U_{CL} \) respectively

\(^8\)Ex-ante restrictions on the sign of the treatment effect are usually called monotone treatment response assumptions and were first introduced by Manski (1997).
denote the second order and cross derivatives of $U$, and assume that $U_{CC} \leq 0$, $U_{LL} \leq 0$, and $U_{CL} \geq 0$, a property satisfied by most standard utility functions. Let $I_0 < I_1$ denote agents’ non-labor income without and with disability insurance, and let $Y_0$ and $Y_1$ denote their corresponding labor market participation decisions. As is well-known, $U_{CC} \leq 0$, $U_{LL} \leq 0$, and $U_{CL} \geq 0$ implies that $U(W + I, T - HY) - U(I, T)$ is increasing in $I$, which in turn implies that $Y_1 \leq Y_0$.

The 2SLS coefficient in this study is significantly negative. Following the discussion in the previous paragraph, Assumption 7 is plausible in this context: it will hold if $E(Y_1 - Y_0|F)$ is not strictly greater than 0, something which will be automatically satisfied if $Y_1 \leq Y_0$. Therefore, one can invoke Theorems 2.3 and 2.1 to claim that this coefficient consistently estimates the LATE of convivors, even though it might not be consistent for the LATE of compliers because of defiers. Moreover, $Y_1 \leq Y_0$ also implies that Assumption 12 is satisfied. One could then use Theorem 2.6 to estimate bounds for the ATE in this application.

Finally, French & Song (2012) also study the effect of disability insurance on labor supply and find a strictly negative 2SLS coefficient. Following the same line of argument as in the previous paragraph, the CD condition should also hold in this study.

**Aizer & Doyle (2013)**

Aizer & Doyle (2013) study the effect of juvenile incarceration on high school completion. They use average sentencing rates of randomly assigned judges as an instrument for incarceration. Here as well $Y_1 \leq Y_0$ sounds like a plausible restriction. 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. Their 2SLS coefficient is significantly negative, so Assumption 7 is also plausible in this context. Therefore, one can invoke Theorems 2.3 and 2.1 to claim that this coefficient consistently estimates the LATE of convivors.

**Angrist & Evans (1998)**

Angrist & Evans (1998) study the effect of having a third child on mothers labor supply. In their study, $\hat{P}(F) = 37.2\%$, and the 95% confidence upper bound for $P(F)$ constructed using Theorem 1 in Andrews & Soares (2010) is 37.4%. The left axis of Figure 3 shows the sample

---

9 The instrument used in Maestas et al. (2013) is multivariate. Theorem 2.3 can easily be extended to this type of setting, assuming that Assumption 7 holds within the sample of cases dealt with by each pair of judges. In the supplementary material of this paper, I cover in more details the case of multivariate instruments.

10 The DIODS data archives used in this paper contain personally identifiable information. It is only possible to access them at a secure location, after having signed an agreement with the US Social Security administration.
counterpart of $\Delta(P(F))$ for all values of $P(F)$ included between 0 and 37.4%. The right axis shows the same quantity normalized by the standard deviation of the outcome. Assumption 9 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, $\widehat{\Delta}(0.05) = 0.072$. Therefore, Assumption 9 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 of the outcome.

Figure 3: 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 limited evidence available suggests that 5% is a 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 my knowledge, similar questions have never been asked in a survey in the U.S.. 1.8% could under or overestimate the share of defiers in the U.S. population. But this figure is, as of now, the best piece of evidence available to assess the percentage of defiers in Angrist & Evans (1998). 5% therefore sounds like a reasonably conservative upper bound.

15% of a standard deviation is also 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’ decision to work (mothers’ potential wages,

\footnote{The 95% confidence interval of $\Delta(0.05)$ is [0.044,0.100]. It can be estimated using standard Stata commands. A code is available upon request.}
preferences for leisure...), but they are unlikely to enter directly into that decision. As a result, 15% of a standard deviation is arguably a 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.

**Application to Duflo & Saez (2003)**

Duflo & Saez (2003) conduct a randomized experiment with an encouragement design to study the effect of an information meeting on the take-up of a retirement plan. To encourage the treatment group to attend, subjects were given a financial incentive upon attendance. In this context, $Y_1 \geq Y_0$ sounds like a plausible restriction. Unless it is poorly designed, the meeting should not reduce take-up. The authors’ 2SLS coefficient is significantly positive, so Assumption 7 is also plausible in this context. Therefore, one can invoke Theorems 2.3 and 2.1 to claim that this coefficient consistently estimates the LATE of compliers.

**6 Conclusion**

Applied economists often use instruments affecting the take-up of a treatment to estimate its effect. When doing so, the methods they use rely on a monotonicity assumption. In many instances, this assumption is not applicable. In this paper, I show that these methods are still valid under a weaker condition than monotonicity. Doing so, I extend the applicability of these methods. Specifically, I show that researchers can confidently use them in applications where one can reasonably assume that defiers’ LATE has the same sign as the reduced form effect of the instrument on the outcome, or that compliers’ and defiers’ LATE do not differ too much. My weaker condition is also more likely to hold when the instrument has a strong first stage. I put forward examples where my weaker condition is likely to hold, while monotonicity is likely to fail.
References

*Journal of Econometrics* **113**(2), 231-263.


Balk, A. & Pearl, J. (1997), ‘Bounds on treatment effects from studies with imperfect compliance’, 


Kitagawa, T. (2009), Identification region of the potential outcome distributions under instrument independence, CeMMAP working papers CWP30/09, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.


A The CD algorithm

In this section, I present the CD algorithm used in Section 4 to assess whether a joint distribution of $(D_0, D_1, Y_1 - Y_0)$ satisfies Assumption 5.

**Theorem A.1** Assume that $Y_1 - Y_0 | C$ is dominated by the Lebesgue measure on $\mathbb{R}$, and that its density relative to this measure is strictly positive on the support of $Y_1 - Y_0 | C$.\(^{12}\)

If $RF \geq 0$, one can use the following algorithm to assess whether Assumption 5 is satisfied:

1. If $E((Y_1 - Y_0)1\{Y_1 - Y_0 \geq 0\}1\{C\}) < RF$, Assumption 5 is violated.

2. Else, let $\delta_0 \geq 0$ solve $E((Y_1 - Y_0)1\{Y_1 - Y_0 \geq \delta\}1\{C\}) = RF$.
   
   If $P(Y_1 - Y_0 \geq \delta_0, C) > FS$, Assumption 5 is violated.

3. Else, if $E((Y_1 - Y_0)1\{Y_1 - Y_0 \leq \delta_0\}1\{C\}) \leq 0$, let $\delta_1$ solve $E((Y_1 - Y_0)1\{Y_1 - Y_0 \in [\delta, \delta_0]\}1\{C\}) = 0$.
   
   (a) If $P(Y_1 - Y_0 \geq \delta_1, C) \geq FS$, Assumption 5 is satisfied.
   
   (b) Else, Assumption 5 is violated.

4. Else, if $E((Y_1 - Y_0)1\{Y_1 - Y_0 \leq \delta_0\}1\{C\}) > 0$, let $\delta_2$ solve $E((Y_1 - Y_0)1\{Y_1 - Y_0 \leq \delta\}1\{C\}) = RF$.
   
   (a) If $P(Y_1 - Y_0 \leq \delta_2, C) \geq FS$, Assumption 5 is satisfied.
   
   (b) Else, Assumption 5 is violated.

If $RF < 0$, one can substitute $-(Y_1 - Y_0)$ to $Y_1 - Y_0$ in the previous algorithm.

The intuition for this theorem goes as follows. Assume $RF \geq 0$. If CD holds, there must be a subpopulation of compliers such that $P(C_V) = FS$ and $E((Y_1 - Y_0)1\{C_V\}) = RF$. If $E((Y_1 - Y_0)1\{Y_1 - Y_0 \geq 0\}1\{C\}) < RF$, CD must be violated, because for any subpopulation of compliers, $E((Y_1 - Y_0)1\{C_V\}) \leq E((Y_1 - Y_0)1\{Y_1 - Y_0 \geq 0\}1\{C\})$. Even summing the treatment effects for all compliers who gain from treatment is not enough to reach the numerator of the 2SLS coefficient. Similarly, if $P(Y_1 - Y_0 \geq \delta_0, C) > FS$, CD must be violated: even the smallest subpopulation of compliers such that $E((Y_1 - Y_0)1\{C_V\}) = RF$ is already too large. The following steps of the algorithm follow from similar arguments.

\(^{12}\)This ensures that the numbers $\delta_0$, $\delta_1$, and $\delta_2$ introduced hereafter are uniquely defined.
B  Proofs

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 B.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 \( S(Y_1 - Y_0) \) such that

\[
0 \leq g(\delta) \leq f_{Y_1-Y_0|C}(\delta) P(C) \text{ for } \lambda\text{-almost } \delta \in S(Y_1 - Y_0) \tag{19}
\]

\[
\int_{S(Y_1-Y_0)} g(\delta) d\lambda(\delta) = P(F) \tag{20}
\]

\[
\int_{S(Y_1-Y_0)} \delta \frac{g(\delta)}{P(F)} d\lambda(\delta) = E(Y_1 - Y_0|F). \tag{21}
\]

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

\[
0 \leq h(\delta) \leq f_{Y_1-Y_0|C}(\delta) P(C) \text{ for } \lambda\text{-almost } \delta \in S(Y_1 - Y_0) \tag{22}
\]

\[
\int_{S(Y_1-Y_0)} h(\delta) d\lambda(\delta) = FS \tag{23}
\]

\[
\int_{S(Y_1-Y_0)} \delta \frac{h(\delta)}{FS} d\lambda(\delta) = W. \tag{24}
\]

Proof of Lemma B.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 (22), (23), and (24), then a function \( g \) satisfies (19), (20), and (21).

I start proving the if part of the first statement. Assume a function \( g \) satisfies (19), (20), and (21). 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 (19) 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 $C_F = \{C, B = 1\}$.

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

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 (19), $f_{Y_1 - Y_0|C}(\delta)P(C) = 0 \Rightarrow g(\delta) = 0$, and the last from (20). This proves that $C_F$ satisfies (6).

Then,

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

The fourth equality follows from (6) and the fifth from (21). This proves that $C_F$ satisfies (7).

I now prove the only if part of the second statement. Assume a subset of $C$ denoted $C_V$ satisfies (8) and (9). Then $h = f_{Y_1 - Y_0|C_V} P(C_V)$ must satisfy (22), otherwise we would not have $C_V \subseteq C$. It must also satisfy (23) and (24), otherwise $C_V$ would not satisfy (8) and (9).

I finally show the last point. Assume $h$ satisfies (22), (23), and (24). Then, it follows from (1) and (2) that $g = f_{Y_1 - Y_0|C} P(C) - h$ satisfies (19), (20), and (21).

QED.

Proof of Theorem 2.2:

Under Assumption 6, $g_1 = f_{Y_1 - Y_0|F} P(F)$ satisfies (19), (20), and (21).

QED.
Proof of Theorem 2.3:

I only prove the result when $RF > 0$. The proof follows from a symmetric reasoning when $RF < 0$. When $RF = 0$, proving the equivalence and the first implication becomes trivial. To prove the second implication, if $E(Y_1 - Y_0|F) \geq 0$ one can use the same reasoning as that used for $RF > 0$, while if $E(Y_1 - Y_0|F) \leq 0$ one can use the same reasoning as that used for $RF < 0$.

I first prove that Assumption 9 ⇒ Assumption 7. As I have assumed $0 < RF$, Assumption 7 implies that $0 \leq E(Y_1 - Y_0|F)$. Rearranging Equation (2) yields

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

Assumption 9 is therefore equivalent to

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

which implies that $0 \leq E(Y_1 - Y_0|F)$. This proves the result.

Then, I prove that Assumption 7 ⇐ Assumption 8. Let Assumption 7 be satisfied with $0 \neq E(Y_1 - Y_0|F)$. As I have assumed $0 < RF$, Assumption 7 implies that $0 < E(Y_1 - Y_0|F)$. Then, it follows from Equation (2) that $E(Y_1 - Y_0|C)$ must also be strictly positive. Finally, rearranging Equation (2) yields

$$ E(Y_1 - Y_0|C) \leq \frac{P(C)}{P(F)} E(Y_1 - Y_0|F). $$

This proves that Assumption 8 is satisfied. If Assumption 7 is satisfied with $E(Y_1 - Y_0|F) = 0$, Assumption 8 is also trivially satisfied. Conversely, if Assumption 8 is satisfied with $E(Y_1 - Y_0|F) \neq 0$, one has $1 \leq \frac{P(C)E(Y_1 - Y_0|C)}{P(F)E(Y_1 - Y_0|F)}$. This in turn implies that $0 \leq \frac{RF}{P(F)E(Y_1 - Y_0|F)}$, thus proving that either $RF = 0$ or $E(Y_1 - Y_0|F)$ has the same sign as $RF$. This proves that Assumption 7 is satisfied. If Assumption 8 is satisfied with $E(Y_1 - Y_0|F) = 0$, Assumption 7 is also trivially satisfied. This proves the result.

Finally, I prove that Assumption 7 ⇒ Assumption 5. To do so, I show that if Assumption 7 is satisfied, there is a function $h_1$ satisfying (22), (23), and (24). In view of Lemma B.1, this will prove the result.

As I have assumed $0 < RF$, Assumption 7 implies that $0 \leq E(Y_1 - Y_0|F)$. With binary potential outcomes this is equivalent to $0 \leq P(Y_1 - Y_0 = 1, F) - P(Y_1 - Y_0 = -1, F)$. With binary potential outcomes, (2) simplifies to

$$ P(Y_1 - Y_0 = 1, C) - P(Y_1 - Y_0 = -1, C) = RF + P(Y_1 - Y_0 = 1, F) - P(Y_1 - Y_0 = -1, F). \quad (25) $$

Once combined with (25), Assumption 7 implies

$$ RF \leq P(Y_1 - Y_0 = 1, C). \quad (26) $$
Then, notice that
\[
FS - RF - P(Y_1 - Y_0 = 0, C)
= 2P(Y_1 - Y_0 = -1, C) - (2P(Y_1 - Y_0 = -1, F) + P(Y_1 - Y_0 = 0, F))
\tag{27}
\]
\[
FS + RF - P(Y_1 - Y_0 = 0, C)
= 2P(Y_1 - Y_0 = 1, C) - (2P(Y_1 - Y_0 = 1, F) + P(Y_1 - Y_0 = 0, F)).
\tag{28}
\]

Now, consider the function \( h_1 \) defined on \( \{-1, 0, 1\} \) and such that
\[
\begin{align*}
    h_1(-1) &= \max \left( 0, \frac{FS - RF - P(Y_1 - Y_0 = 0, C)}{2} \right) \\
    h_1(0) &= \min (P(Y_1 - Y_0 = 0, C), FS - RF) \\
    h_1(1) &= \max \left( RF, \frac{FS + RF - P(Y_1 - Y_0 = 0, C)}{2} \right).
\end{align*}
\]

If \( FS - RF \leq P(Y_1 - Y_0 = 0, C) \),
\[
\begin{align*}
    h_1(-1) &= 0 \\
    h_1(0) &= FS - RF \\
    h_1(1) &= RF.
\end{align*}
\]

\( h_1(-1) \) is trivially included between 0 and \( P(Y_1 - Y_0 = -1, C) \). 0 \( \leq h_1(0) \) follows from the fact that by assumption \(|W| \leq 1\). 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 (26). This proves that \( h_1 \) satisfies (22). It is easy to see that it also satisfies (23) and (24).

If \( FS - RF > P(Y_1 - Y_0 = 0, C) \),
\[
\begin{align*}
    h_1(-1) &= \frac{FS - RF - P(Y_1 - Y_0 = 0, C)}{2} \\
    h_1(0) &= P(Y_1 - Y_0 = 0, C) \\
    h_1(1) &= \frac{FS + RF - P(Y_1 - Y_0 = 0, C)}{2}.
\end{align*}
\]

\( h_1(-1) \) is greater than 0 by assumption. \( h_1(-1) \leq P(Y_1 - Y_0 = -1, C) \) follows from (27). \( h_1(0) \) 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 (28). This proves that \( h_1 \) satisfies (22). It is easy to see that it also satisfies (23) and (24).

QED.

Proof of Theorem 2.4:
Following the same steps as those used by Angrist et al. (1996) to prove Equations (1) and (2), one can show that under Assumptions 10 and 2, for every \( x \) in the support of \( X \),

\[
E(D|Z = 1, X = x) - E(D|Z = 0, X = x) = P(C|X = x) - P(F|X = x)
\]

\[
E(Y|Z = 1, X = x) - E(Y|Z = 0, X = x) = E(Y_1 - Y_0|C, X = x) P(C|X = x) - E(Y_1 - Y_0|F, X = x) P(F|X = x).
\]

Therefore,

\[
E(E(D|Z = 1, X) - E(D|Z = 0, X)) = P(C) - P(F)
\]

\[
E(E(Y|Z = 1, X) - E(Y|Z = 0, X)) = E(Y_1 - Y_0|C) P(C) - E(Y_1 - Y_0|F) P(F).
\]

Under Assumption 5, one can apply to the right hand side of the previous display the same steps as in the proof of Theorem 2.1. One finally obtains

\[
E(E(D|Z = 1, X) - E(D|Z = 0, X)) = P(C_V)
\]

\[
E(E(Y|Z = 1, X) - E(Y|Z = 0, X)) = E(Y_1 - Y_0|C_V) P(C_V).
\]

This proves the result.

QED.

**Proof of Theorem 2.5:**

In view of Theorem 2.1, it is sufficient to show that if a subpopulation of compliers \( C_F \) satisfies Equations (6), (7), and (15), then \( C_V = C \setminus C_F \) satisfies (16). Using the same steps as those used in Angrist et al. (1996) to prove Equation (2), one can show that

\[
W_{XD} = \frac{P(C)E[X|C] - P(F)E[X|F]}{P(C) - P(F)}.
\]

Then, it follows from Equations (6) and (15) that

\[
E[X|C] = \frac{P(C) - P(F)}{P(C)} E[X|C_V] + \frac{P(F)}{P(C)} E[X|F].
\]

Plugging this equation into the previous one yields the result.

QED.

**Proof of Theorem 2.6:**

I only prove the result when \( RF > 0 \) and for the lower bound. The proof is symmetric when \( RF < 0 \), and follows from similar arguments for the upper bound.

I first prove that the lower bound is valid. If Assumption 8 is satisfied, Equation (2) implies that \( E(Y_1 - Y_0|C) \) must have the same sign as \( RF \). Assumption 12 then implies that \( E(Y_1 - Y_0|AT), E(Y_1 - Y_0|NT), E(Y_1 - Y_0|C), \) and \( E(Y_1 - Y_0|F) \) must all be weakly greater than 0.
Moreover, it follows from Theorem 2.3 that Assumption 5 is satisfied under the assumptions of the theorem. Therefore, it follows from Theorem 2.1 that compliers can be partitioned into subpopulations $C_F$ and $C_V$ respectively satisfying Equations (6) and (7), and (8) and (9). Thus,

\[
E(Y_1 - Y_0) = P(C_V)E(Y_1 - Y_0|C_V) + P(C_F)E(Y_1 - Y_0|C_F)
+ P(AT)E(Y_1 - Y_0|AT) + P(NT)E(Y_1 - Y_0|NT) + P(F)E(Y_1 - Y_0|F)
= RF + P(AT)E(Y_1 - Y_0|AT) + P(NT)E(Y_1 - Y_0|NT) + 2P(F)E(Y_1 - Y_0|F)
\geq RF.
\]

This proves that the bound is valid.

Let

\[
P^*(Y_0 = 0, Y_1 = 0, D_0 = 1, D_1 = 1) = P(Y = 0, D = 1|Z = 0)
\]
\[
P^*(Y_0 = 1, Y_1 = 1, D_0 = 1, D_1 = 1) = P(Y = 1, D = 1|Z = 0)
\]
\[
P^*(Y_0 = 0, Y_1 = 0, D_0 = 0, D_1 = 0) = P(Y = 0, D = 0|Z = 1)
\]
\[
P^*(Y_0 = 1, Y_1 = 1, D_0 = 0, D_1 = 0) = P(Y = 1, D = 0|Z = 1)
\]
\[
P^*(Y_0 = 0, Y_1 = 1, D_0 = 0, D_1 = 1) = RF
\]
\[
P^*(Y_0 = 0, Y_1 = 0, D_0 = 0, D_1 = 1) = P(Y = 0, D = 1|Z = 1) - P(Y = 0, D = 1|Z = 0)
\]
\[
P^*(Y_0 = 1, Y_1 = 1, D_0 = 0, D_1 = 1) = P(Y = 1, D = 0|Z = 0) - P(Y = 1, D = 0|Z = 1),
\]

and let $P^*(Y_0 = y_0, Y_1 = y_1, D_0 = d_0, D_1 = d_1) = 0$ for all other possible values of $(y_0, y_1, d_0, d_1) \in \{0, 1\}^4$. Equation (1.1) in Kitagawa (2013) ensures that $P^*$ is a probability measure. It is easy to see that it is compatible with the data and with the assumptions of the theorem, and that it attains the lower bound. This proves that the lower bound is sharp.

QED.

**Proof of Theorem A.1**

I only prove the result when $RF \geq 0$ (the proof is symmetric when $RF < 0$).

Assume $E((Y_1 - Y_0)1\{Y_1 - Y_0 \geq 0\}1\{C\}) < RF$. If CD is satisfied, it follows from Equation (8) and (9) that there is a subpopulation of compliers $C_V$ such that

\[
RF = E((Y_1 - Y_0)1\{C_V\}) \leq E((Y_1 - Y_0)1\{Y_1 - Y_0 \geq 0\}1\{C\}) < RF,
\]

a contradiction. CD must therefore be violated. This proves the first point.
Then, assume \( P(Y_1 - Y_0 \geq \delta_0, C) > FS \). Assume first that \( \delta_0 > 0 \). If CD is satisfied,
\[
0 = RF - RF
\]
\[
= E((Y_1 - Y_0)1\{Y_1 - Y_0 \geq \delta_0\}1\{C\}) - E((Y_1 - Y_0)1\{C\})
\]
\[
= E((Y_1 - Y_0)1\{Y_1 - Y_0 \geq \delta_0\}(1\{C\} - 1\{C\})) - E((Y_1 - Y_0)1\{Y_1 - Y_0 < \delta_0\}1\{C\})
\]
\[
\geq \delta_0 (P(Y_1 - Y_0 \geq \delta_0, C) - P(Y_1 - Y_0 \geq \delta_0, C_V) - P(Y_1 - Y_0 < \delta_0, C_V))
\]
\[
\geq \delta_0 (P(Y_1 - Y_0 \geq \delta_0, C) - FS)
\]
\[
> 0,
\]
a contradiction. CD must therefore be violated. Now, assume \( \delta_0 = 0 \). If CD is satisfied,
\[
0 \geq E((Y_1 - Y_0)1\{Y_1 - Y_0 < 0\}1\{C\})
\]
\[
= RF - E((Y_1 - Y_0)1\{Y_1 - Y_0 \geq 0\}1\{C\})
\]
\[
\geq RF - E((Y_1 - Y_0)1\{Y_1 - Y_0 \geq 0\}1\{C\})
\]
\[
= 0.
\]
Therefore, \( P(Y_1 - Y_0 < 0, C_V) = 0 \), which in turn implies that \( 1\{Y_1 - Y_0 \geq 0\}1\{C\} = 1\{Y_1 - Y_0 \geq 0\}1\{C_V\} \) almost everywhere, a contradiction. This proves the second point.

Then, assume \( P(Y_1 - Y_0 \geq \delta_0, C) \leq FS \) and \( P(Y_1 - Y_0 \geq \delta_1, C) \geq FS \). Let
\[
h_2(\delta) = \begin{cases} P(C)f_{Y_1-Y_0}(\delta) & \text{if } \delta \geq \delta_0; \\ \frac{FS - P(Y_1 - Y_0 \geq \delta_0, C)}{P(Y_1 - Y_0 \geq \delta_1, C)} P(C)f_{Y_1-Y_0}(\delta) & \text{if } \delta \in [\delta_1, \delta_0); \\ 0 & \text{otherwise}. \end{cases}
\]
h_2 satisfies (22), (23), and (24). This proves point 3.a), following Lemma B.1.

Then, assume \( P(Y_1 - Y_0 \geq \delta_1, C) < FS \). Assume first that \( \delta_1 < 0 \). If CD is satisfied,
\[
0 = E((Y_1 - Y_0)1\{Y_1 - Y_0 \geq \delta_1\}1\{C\}) - E((Y_1 - Y_0)1\{C\})
\]
\[
\geq \delta_1 (P(Y_1 - Y_0 \geq \delta_1, C) - FS)
\]
\[
> 0,
\]
a contradiction. CD must therefore be violated. Now, assume \( \delta_1 = 0 \). Then, we must also have \( \delta_0 = 0 \), so we can use the same reasoning as in the proof of the second point to show that CD must be violated.

Then, assume \( P(Y_1 - Y_0 \leq \delta_2, C) \geq FS \). Let \( \delta_3 \) solve \( E((Y_1 - Y_0)1\{Y_1 - Y_0 \leq \delta\}1\{C\}) = 0 \). First assume that \( P(Y_1 - Y_0 \in [\delta_3, \delta_2), C) \leq FS \). Let
\[
h_3(\delta) = \begin{cases} 0 & \text{if } \delta \geq \delta_2; \\ P(C)f_{Y_1-Y_0}(\delta) & \text{if } \delta \in [\delta_3, \delta_2); \\ \frac{FS - P(Y_1 - Y_0 \in [\delta_3, \delta_2), C)}{P(Y_1 - Y_0 \leq \delta_2, C) - P(Y_1 - Y_0 \in [\delta_3, \delta_2), C)} P(C)f_{Y_1-Y_0}(\delta) & \text{otherwise}. \end{cases}
\]
\( h_3 \) satisfies (22), (23), and (24).

Now, assume that \( P(Y_1 - Y_0 \in [\delta_3, \delta_2), C) > FS \). For any \( \delta \in [\delta_3, \delta_0] \), let \( \eta(\delta) \) solve \( E((Y_1 - Y_0)1\{Y_1 - Y_0 \in [\delta, \eta(\delta))\}1\{C\}) = RF \). \( \eta(\delta_3) = \delta_2 \), and \( \eta(\delta_0) = \overline{\gamma} \), the sup of the support of \( Y_1 - Y_0 | C \). It is easy to see that \( \eta(\delta) \) is increasing in \( \delta \). I show now that \( P(Y_1 - Y_0 \in [\delta, \eta(\delta)), C) \) is decreasing in \( \delta \). Consider \( \delta^a \leq \delta^b \) in \( [\delta_3, \delta_0] \). Assume first that \( \delta^b \leq \eta(\delta^a) \).

\[
0 = E((Y_1 - Y_0)1\{Y_1 - Y_0 \in [\delta^b, \eta(\delta^b))\}1\{C\}) - E((Y_1 - Y_0)1\{Y_1 - Y_0 \in [\delta^a, \eta(\delta^a))\}1\{C\}) \\
= E((Y_1 - Y_0)1\{Y_1 - Y_0 \in [\eta(\delta^a), \eta(\delta^b))\}1\{C\}) - E((Y_1 - Y_0)1\{Y_1 - Y_0 \in [\delta^a, \delta^b)\}1\{C\}) \\
\geq \eta(\delta^a)P(Y_1 - Y_0 \in [\eta(\delta^a), \eta(\delta^b)), C) - \delta^bP(Y_1 - Y_0 \in [\delta^a, \delta^b), C) \\
\geq \delta^b \left( P(Y_1 - Y_0 \in [\delta^b, \eta(\delta^b)), C) - P(Y_1 - Y_0 \in [\delta^a, \eta(\delta^a)), C) \right).
\]

This proves the result because \( \delta^b \geq 0 \). If \( \delta^b > \eta(\delta^a) \), the proof follows from a similar but simpler argument. Now, as \( P(Y_1 - Y_0 \in [\delta_3, \eta(\delta_3)), C) > FS \) and \( P(Y_1 - Y_0 \in [\delta_0, \eta(\delta_0)), C) \leq FS \), let \( \delta^* \) solve \( P(Y_1 - Y_0 \in [\delta, \eta(\delta)), C) = FS \), and let

\[
h_4(\delta) = \begin{cases} 
P(C)f_{Y_1 - Y_0|C}(\delta) & \text{if } \delta \in [\delta^*, \eta(\delta^*)) \\
0 & \text{otherwise}. 
\end{cases}
\]

\( h_4 \) satisfies (22), (23), and (24). This completes the proof of point 4.a), following Lemma B.1.

Finally, assume \( P(Y_1 - Y_0 \leq \delta_2, C) < FS \). Assume first that \( \delta_2 > 0 \). If CD is satisfied,

\[
0 = E((Y_1 - Y_0)1\{Y_1 - Y_0 \leq \delta_2\}1\{C\}) - E((Y_1 - Y_0)1\{C_Y\}) \\
\leq \delta_2 (P(Y_1 - Y_0 \geq \delta_2, C) - FS) \\
< 0,
\]
a contradiction. CD must therefore be violated. Now, assume \( \delta_2 = 0 \). One must then have \( \delta_3 = RF = 0 \). \( \delta_3 = 0 \) implies \( 1\{Y_1 - Y_0 \leq 0\}1\{C\} = 0 \). Combined with \( RF = 0 \), this implies \( 1\{C_Y\} = 0 \), so CD must be violated. This proves point 4.b).

QED.