

It is never too LATE: a new look at local average treatment effects with or without defiers

CHRISTIAN M. DAHL[†], MARTIN HUBER[‡] AND GIOVANNI MELLACE[†]

[†]University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark.
E-mail: giome@sam.sdu.dk

[‡]University of Fribourg, Av. de l'Europe 20, 1700 Fribourg, Switzerland.

First version received: 22 February 2022; final version accepted: 8 December 2022.

Summary: In heterogeneous treatment effect models with endogeneity, identification of the local average treatment effect (LATE) typically relies on the availability of an exogenous instrument monotonically related to treatment participation. First, we demonstrate that a strictly weaker local monotonicity condition—invoked for specific potential outcome values rather than globally—identifies the LATEs on compliers and defiers. Second, we show that our identification results apply to subsets of compliers and defiers when imposing an even weaker local compliers-defiers assumption that allows for both types at any potential outcome value. We propose estimators that are potentially more efficient than two-stage least squares (2SLS) in finite samples, even in cases where 2SLS is consistent. Finally, we provide an empirical application to estimating returns to education using the quarter of birth instrument.

Keywords: *causal effects, IV, LATE, local CD, local monotonicity, principal stratification.*

JEL codes: C14, C21, C26.

1. INTRODUCTION

In heterogeneous treatment effect models with binary treatment, an instrument is conventionally required to satisfy two assumptions. First, it must be independent of potential treatments and potential outcomes, excluding direct effects on the latter and implying that the instrument is (as good as) randomly assigned. Second, the treatment must be weakly monotonic in the instrument. For instance, an instrument reflecting random treatment assignment must weakly increase actual treatment take-up of everyone (i.e., *globally*). This rules out defiers, who behave counter-intuitively by not taking the treatment when randomised in, but taking it when randomised out. Under these conditions, Imbens and Angrist (1994) and Angrist et al. (1996) demonstrate the identification of the local average treatment effect (LATE) on the subpopulation of compliers, whose treatment reacts to the instrument in the intended way, based on the so-called Wald ratio, which equals the probability limit of two-stage least squares (2SLS).

While the literature on instrumental variables (IV) typically devotes much attention to the IV independence, the monotonicity assumption is often less scrutinised, even though it appears questionable in a range of applications. Taking the estimation of the returns to education as example, Angrist and Krueger (1991) suggest using ‘quarter of birth’ as IV. Arguably, among students entering school in the same year, those born in an earlier quarter can drop out after

fewer years of completed education at the age of 16 (when compulsory schooling ends in the United States) than those born later, suggesting education to be monotonically increasing in the quarter of birth. However, strategic postponement of school entry, so-called redshirting, or school policies may reverse the relation of education and quarter of birth for some individuals, see Barua and Lang (2009), Klein (2010), and Aliprantis (2012) which might imply the existence of defiers and the violation of global monotonicity.

Our first contribution is to show that LATEs are identified under an assumption strictly weaker than *global* monotonicity, namely *local* monotonicity (LM). LM allows for both compliers and defiers, but requires that their marginal potential outcome distributions under either treatment do not overlap. That is, monotonicity is assumed to hold locally in subregions of the marginal potential outcome distributions, rather than over the entire outcome support. LM thus excludes the existence of defiers when conditional on IV independence, the joint distribution of the observed outcome, treatment, and (binary) instrument satisfies a sufficient condition for the existence of compliers; see for instance Balke and Pearl (1997) and Heckman and Vytlacil (2005). By ruling out defiers in such regions, the potential outcomes of the compliers are locally identified. A symmetric argument applies to regions in which defiers exist, where LM rules out compliers. We show that LM is sufficient for the identification of the marginal potential outcome distributions of compliers and defiers in either treatment state. Thus, we can identify the LATEs (a) on compliers, (b) defiers, and (c) the joint population of compliers and defiers.

If defiers do not exist, the identification results for (a) and (c) coincide with the standard LATE under monotonicity and equal the Wald ratio. If defiers exist, (a), (b), and (c) generally differ, and the standard LATE approach is inconsistent unless the LATEs on compliers and defiers are homogeneous; see Angrist et al. (1996). However, even in the case of treatment effect homogeneity across subjects, the standard approach may not be desirable due to a weak instrument problem that arises when the proportions of compliers and defiers net each other out in the first stage which is equivalent to their difference.

As a second contribution, we propose estimators that are potentially more efficient than 2SLS (even when the latter is consistent). Our estimators may be regarded as improved versions of the ‘non-negative’ estimator proposed in Imbens and Rubin (1997a) and are an alternative to the semi- and nonparametric empirical likelihood approaches of Cheng, Qin et al. (2009) and Cheng, Small et al. (2009). The efficiency gain comes from discarding observations for which the estimated complier density is negative. In 2SLS estimation, however, negative complier densities shift an (overall positive) first stage toward zero, which reduces efficiency. Discarding negative areas might induce some finite sample bias. However, in line with the previous literature (Imbens and Rubin, 1997a; Cheng, Qin et al., 2009; Cheng, Small et al., 2009), the bias is very small compared to the gain in precision in a range of simulation studies that we consider. In these simulations, our estimators have always a substantially smaller mean squared error than the 2SLS estimator (see the Online Appendix). Our estimators might therefore be preferred over 2SLS both in terms of robustness to deviations from global monotonicity and efficiency and consist of two steps. In the first step of our estimation procedure, we use kernel methods to estimate conditional outcome densities given the treatment and the instrument to partition the outcome support into complier and defier regions based on the differences in the densities within treatment states. Second, we estimate the LATEs separately within the complier and defier regions. We propose several estimation approaches, which all show an encouraging finite sample behaviour in simulation studies (see the Online Appendix) and confirm the efficiency gains of our estimators over 2SLS even when the latter is consistent.

Among other studies considering deviations from monotonicity is Klein (2010), who assesses the sensitivity of the LATE to random departures from monotonicity and approximates the bias under specific assumptions. In contrast, we consider point identification under nonrandom violations, given that LM holds. Small et al. (2017) weaken (individual-level) monotonicity to stochastic monotonicity, requiring that the share of compliers weakly dominates the share of defiers, and show that 2SLS, albeit biased in general, retains the correct sign of the LATE in the limit. Ding and Lu (2017) develop a general framework to address with post-randomisation complications, including noncompliance, and provide a sensitivity analysis under violations of monotonicity. De Chaisemartin (2017) suggests the ‘compliers-defiers’ (CD) assumption, which is weaker (and implied by) stochastic monotonicity and requires that a subpopulation of compliers is of the same size and has an identical LATE as the defiers. De Chaisemartin (2017) shows that under CD, the Wald ratio identifies the LATE on the remaining subpopulation of compliers, the so-called compliers-survivors or ‘comvivors’.¹ CD and LM are non-nested conditions. Unlike CD, LM allows for the identification of LATEs on the total of compliers and defiers, however, under the arguably strong condition that the marginal potential outcomes of these subpopulations do not overlap.

As third contribution, we suggest a relaxation of LM, which in the spirit of De Chaisemartin (2017) we call local compliers-defiers (LCD). In contrast to LM, LCD permits the coexistence of compliers and defiers at any potential outcome value under either treatment. However, for regions of outcome values in which compliers (defiers) outnumber defiers (compliers) in a specific treatment state, it requires that the ‘surviving’ compliers (defiers) under both treatment and nontreatment have identical distributions in terms of joint potential outcomes (i.e., with and without treatment). Under LCD, our identification results provide LATEs for subpopulations of compliers and defiers, rather than for the total respective group as under LM. Furthermore, we show that CD and LCD are not nested. If both assumptions hold, identification based on LCD provides the LATE for a potentially larger complier subset than that based on CD.

As empirical contribution, we apply our methods to estimate the returns to college for males born in 1940–1949 (in the 1980 U.S. census data) using the quarter of birth instrument of Angrist and Krueger (1991). Relaxing global monotonicity to LM or LCD, we estimate statistically significant proportions of both compliers and defiers and positive returns to college of similar size in both groups. We also find that defiers outnumber compliers in upper parts of the potential wage distribution. This can be rationalised by a structural model in which a family’s socio-economic status is positively correlated with a child’s wage as well as the probability to postpone school entry; see Bedard and Dhuey (2006) and Aliprantis (2012) for empirical evidence on this matter.

The remainder of this paper is organised as follows. Section 2 discusses the assumptions, the identification results for LM (along with implications in a structural model) and LCD, as well as connections and differences to alternative identifying assumptions. Section 3 proposes estimators of the parameters of interest based on kernel density estimation, while two further estimation approaches are discussed in the Online Appendix. Section 4 presents an empirical application to estimating the returns to education using the quarter of birth instrument. Section 5 concludes. A simulation study, technical proofs, and additional material are provided in the Online Appendix.

¹ Sufficient conditions for CD are for instances that the compliers outnumber defiers conditional on any treatment effect value or that the LATEs on the total of defiers and compliers have the same sign, while the ratio of their LATEs is not ‘too’ large.

Table 1. Subject types.

T	$D(1)$	$D(0)$	Subject type
a	1	1	Always-taker
c	1	0	Complier
$\bar{d}$	0	1	Defier
n	0	0	Never-taker

2. ASSUMPTIONS AND IDENTIFICATION

2.1. Notation

Suppose that we are interested in the causal effect of a binary treatment $D \in \{1, 0\}$ (e.g., graduating from high school) on an outcome Y (e.g., earnings). Under endogeneity, the effect of D on Y is confounded by unobserved factors. The treatment effect may nevertheless be identified if an instrument, denoted by Z , is available, which is correlated with the treatment, but does not have a direct effect on the outcome (i.e., any impact other than through D). In this section, we consider the case of a binary instrument ($Z \in \{0, 1\}$), such as a randomised treatment assignment, whereas the Online Appendix discusses the case of a bounded nonbinary instrument. Denote by $D(z)$ the potential treatment state that would occur when we set instrument $Z = z$, and denote by $Y(d)$ the potential outcome for treatment $D = d$ (see, e.g., Rubin, 1974, for a discussion of the potential outcome notation). In the data, only one potential outcome is observed for each subject because $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$.

As discussed in Angrist et al. (1996) and summarised in Table 1, the population can be categorised into four compliance types, denoted by $T \in \{a, c, \bar{d}, n\}$, depending on how the treatment changes with the instrument. The compliers (c) respond to the instrument by taking the treatment when $Z = 1$ and abstaining from it when $Z = 0$. The always-takers (a) are always treated irrespective of the instrument, the never-takers (n) are never treated, and the defiers ($\bar{d}$) only receive the treatment when $Z = 0$. One cannot directly infer anyone's type from the data because either $D(1)$ or $D(0)$ remains unknown, as the observed treatment status is given by $D = Z \cdot D(1) + (1 - Z) \cdot D(0)$. In fact, any observed combination of treatment and instrument status is consistent with two of the types listed in the first column of Table 1. For instance, $Z = 1, D = 1 \rightarrow D(1) = 1$ implies either $T = a$ or $T = c$, as can be seen from the first two rows of Table 1. Although the compliance types are not directly observable, we will demonstrate that the potential outcome distributions of compliers and defiers can be identified under conditions which are weaker than the conventional LATE assumptions of Imbens and Angrist (1994) and Angrist et al. (1996).

To formally characterise the identification problem, we follow Kitagawa (2009) and denote by $p_d(y)$ and $q_d(y)$ the joint densities of $Y = y$ and $D = d$ given $Z = 1$ and $Z = 0$, respectively:

$$\begin{aligned} p_1(y) &= f(y, D = 1|Z = 1), & p_0(y) &= f(y, D = 0|Z = 1), \\ q_1(y) &= f(y, D = 1|Z = 0), & q_0(y) &= f(y, D = 0|Z = 0). \end{aligned}$$

Furthermore, let $\mathcal{Y}$ denote the support of Y and $f(y(d))$ denote the marginal density of the potential outcome for $d \in \{0, 1\}$. The support of either potential outcome $Y(1)$ and $Y(0)$ might correspond to $\mathcal{Y}$ or only a subset thereof (and in the latter case, $Y(1)$ and $Y(0)$ do not have the

same support). We also define $f(y(d), T = t)$ as the joint density of potential outcome and type for $d \in \{0, 1\}$, $t \in \{a, c, \bar{d}, n\}$, and $y \in \mathcal{Y}$. Finally, we note that any of the observed joint densities, $p_d(y)$ and $q_d(y)$, are a function of the potential outcomes of two different types of subjects:

$$p_1(y) = f(y(1), T = c|Z = 1) + f(y(1), T = a|Z = 1), \quad (2.1)$$

$$q_1(y) = f(y(1), T = \bar{d}|Z = 0) + f(y(1), T = a|Z = 0), \quad (2.2)$$

$$p_0(y) = f(y(0), T = \bar{d}|Z = 1) + f(y(0), T = n|Z = 1), \quad (2.3)$$

$$q_0(y) = f(y(0), T = c|Z = 0) + f(y(0), T = n|Z = 0). \quad (2.4)$$

2.2. Assumptions and identification results

Our first assumption imposes independence between Z and the joint distribution of potential outcomes and treatments; see Imbens and Angrist (1994).

ASSUMPTION 2.1. $Z \perp\!\!\!\perp (D(1), D(0), Y(1), Y(0))$, where $\perp\!\!\!\perp$ denotes independence (joint independence).

Assumption 2.1 implies the (quasi-)randomness of the instrument, such that it is unrelated to factors affecting the potential treatments and/or outcomes, and the absence of a direct effect on the potential outcomes. It follows that not only the potential outcomes, but also the types, which are defined by the potential treatments, are independent of the instrument. As discussed in Kitagawa (2009), equations (2.1) through (2.4) then simplify to

$$p_1(y) = f(y(1), T = c) + f(y(1), T = a), \quad (2.5)$$

$$q_1(y) = f(y(1), T = \bar{d}) + f(y(1), T = a), \quad (2.6)$$

$$p_0(y) = f(y(0), T = \bar{d}) + f(y(0), T = n), \quad (2.7)$$

$$q_0(y) = f(y(0), T = c) + f(y(0), T = n). \quad (2.8)$$

Imbens and Angrist (1994) and Angrist et al. (1996) show the identification of the local average treatment effect on the compliers, $E(Y(1) - Y(0)|T = c)$, when combining Assumption 2.1 with (global) monotonicity of D in Z , which rules out defiers:

ASSUMPTION 2.2. Define Z such that $\Pr(D = 1|Z = 1) \geq \Pr(D = 1|Z = 0)$. Then, $\Pr(D(1) \geq D(0)) = 1$ holds for all subjects in the population (global monotonicity).

Equations (2.5) through (2.8) then simplify to

$$p_1(y) = f(y(1), T = c) + f(y(1), T = a), \quad (2.9)$$

$$q_1(y) = f(y(1), T = a), \quad (2.10)$$

$$p_0(y) = f(y(0), T = n), \quad (2.11)$$

$$q_0(y) = f(y(0), T = c) + f(y(0), T = n). \quad (2.12)$$

To obtain the LATE based on these expressions, we note that subtracting (2.10) from (2.9) and (2.11) from (2.12), respectively, yields the joint densities of outcomes and compliance under treatment and nontreatment:

$$f(y(1), T = c) = p_1(y) - q_1(y), \quad (2.13)$$

$$f(y(0), T = c) = q_0(y) - p_0(y). \quad (2.14)$$

For this reason, integrating out y gives the share of compliers:

$$\Pr(T = c) = \int_{y \in \mathcal{Y}} (p_1(y) - q_1(y)) dy = E(D|Z = 1) - E(D|Z = 0), \quad (2.15)$$

$$\Pr(T = c) = \int_{y \in \mathcal{Y}} (q_0(y) - p_0(y)) dy = E(1 - D|Z = 0) - E(1 - D|Z = 1). \quad (2.16)$$

Furthermore, as $\int_{y \in \mathcal{Y}} y \cdot p_d(y) dy = E(y, D = d|Z = 1)$ and $\int_{y \in \mathcal{Y}} y \cdot q_d(y) dy = E(y, D = d|Z = 0)$, it follows

$$\begin{aligned} E(Y(1)|T = c) &= \int_{y \in \mathcal{Y}} y \cdot \frac{f(y(1), T = c)}{\Pr(T = c)} dy = \int_{y \in \mathcal{Y}} \frac{y \cdot (p_1(y) - q_1(y))}{E(D|Z = 1) - E(D|Z = 0)} dy \\ &= \frac{E(Y, D = 1|Z = 1) - E(Y, D = 1|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)}, \end{aligned} \quad (2.17)$$

where the second equality holds by Assumptions 2.1 and 2.2. In analogous manner, we obtain that

$$E(Y(0)|T = c) = \frac{E(Y, D = 0|Z = 0) - E(Y, D = 0|Z = 1)}{E(D|Z = 1) - E(D|Z = 0)}. \quad (2.18)$$

Given that $E(Y|Z = z) = E(Y, D = 0|Z = z) + E(Y, D = 1|Z = z)$, subtracting (2.18) from (2.17) yields the result of Imbens and Angrist (1994) that the LATE corresponds to the so-called Wald ratio (WR):

$$E(Y(1) - Y(0)|T = c) = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 0) - E(D|Z = 1)} = \text{WR}.$$

Our discussion illustrates that the WR assigns weights $\frac{p_1(y) - q_1(y)}{E(D|Z=1) - E(D|Z=0)}$ and $\frac{q_0(y) - p_0(y)}{E(D|Z=1) - E(D|Z=0)}$ to treated and nontreated observations, respectively. In addition, (2.13) and (2.14) provide testable implications of global monotonicity and Assumption 2.1, see Balke and Pearl (1997) and Heckman and Vytlacil (2005). As $f(y(1), T = c)$ and $f(y(0), T = c)$ cannot be negative for any $y \in \mathcal{Y}$, it must hold that

$$p_1(y) - q_1(y) \geq 0, \quad q_0(y) - p_0(y) \geq 0. \quad (2.19)$$

Huber and Mellace (2015), Kitagawa (2015), Mourifie and Wan (2017), and Farbmacher et al. (2022) provide formal tests of these or related constraints. Furthermore, Imbens and Rubin (1997a,b) propose estimators that impose (2.19) in the data to improve efficiency, see also the related semi- and nonparametric approaches by Cheng, Qin et al. (2009), Cheng, Small et al. (2009), and Lee et al. (2017).

However, when *global* monotonicity is violated (while maintaining Assumption 2.1), basing LATE evaluation on the WR appears unattractive both in terms of consistency and asymptotic efficiency. First, Angrist et al. (1996) show that the WR does in general not yield a causal effect

because the WR is equivalent to

$$\text{WR} = \frac{E(Y(1) - Y(0)|T = c) \cdot \Pr(T = c) - E(Y(1) - Y(0)|T = \bar{d}) \cdot \Pr(T = \bar{d})}{\Pr(T = c) - \Pr(T = \bar{d})}. \quad (2.20)$$

Therefore, the LATE on compliers is identified only if it is equal to the LATE on defiers. Second, even in this special case, the WR assigns negative weights to treated (nontreated) observations whenever $p_1(y) < q_1(y)$ ($q_0(y) < p_0(y)$). It is easy to see from (2.15) and (2.16) that negative weights shift the (positive) terms $E(D|Z = 1) - E(D|Z = 0)$ and $E(1 - D|Z = 0) - E(1 - D|Z = 1)$ toward zero. In applications, this could entail a poor finite sample performance of 2SLS due to weak instrument issues.

We therefore substitute *global* monotonicity by Assumption 2.3, denoted as *local* monotonicity (LM), which is weaker and allows for a violation of (2.19).

ASSUMPTION 2.3. For each $y(d) \in \mathcal{Y}$, either $\Pr(T = \bar{d}|Y(d) = y(d)) = 0$ or $\Pr(T = c|Y(d) = y(d)) = 0$ or $\Pr(T = \{c, \bar{d}\}|Y(d) = y(d)) = 0$, with $d \in \{0, 1\}$ (*local monotonicity*).

Assumption 2.3 (LM) is novel in the sense that it allows for both compliers and defiers in the population. LM, however, requires the potential outcome distributions of compliers and defiers in each treatment state to be nonoverlapping. Under LM, compliers and defiers inhabit disjoint regions of the support of $Y(1)$ and $Y(0)$, respectively.² We note that this does, e.g., not rule out that some compliers' potential outcomes under treatment are equal to some defiers' potential outcomes under nontreatment, because LM refers to the marginal potential outcome distributions only. More formally, note that under Assumption 2.1, equations (2.5) through (2.8) imply

$$\begin{aligned} p_1(y) - q_1(y) &= f(y(1), T = c) - f(y(1), T = \bar{d}), \\ q_0(y) - p_0(y) &= f(y(0), T = c) - f(y(0), T = \bar{d}), \end{aligned} \quad (2.21)$$

while adding Assumption 2.3 implies

$$p_1(y) > q_1(y) \Rightarrow f(y(1), T = c) > f(y(1), T = \bar{d}) \Rightarrow f(y(1), T = \bar{d}) = 0 \quad (\text{nodefiers}),$$

$$p_1(y) < q_1(y) \Rightarrow f(y(1), T = c) < f(y(1), T = \bar{d}) \Rightarrow f(y(1), T = c) = 0 \quad (\text{nocompliers}),$$

$$q_0(y) > p_0(y) \Rightarrow f(y(0), T = c) > f(y(0), T = \bar{d}) \Rightarrow f(y(0), T = \bar{d}) = 0 \quad (\text{nodefiers}),$$

$$q_0(y) < p_0(y) \Rightarrow f(y(0), T = c) < f(y(0), T = \bar{d}) \Rightarrow f(y(0), T = c) = 0 \quad (\text{nocompliers}).$$

Equation (2.21) shows that in regions of $\mathcal{Y}$ where (2.19) is satisfied and $f(y(d), T = c) > f(y(d), T = \bar{d})$, defiers are ruled out by Assumption 2.3. Similarly, a violation of (2.19) where $f(y(d), T = \bar{d}) > f(y(d), T = c)$, rules out compliers. Summarising these observations,

$$f(y(1), T = c) = (p_1(y) - q_1(y))I(p_1(y) > q_1(y)) = p_1(y) - \min(p_1(y), q_1(y)), \quad (2.22)$$

$$f(y(0), T = c) = (q_0(y) - p_0(y))I(q_0(y) > p_0(y)) = q_0(y) - \min(p_0(y), q_0(y)), \quad (2.23)$$

$$f(y(1), T = \bar{d}) = (q_1(y) - p_1(y))I(p_1(y) < q_1(y)) = q_1(y) - \min(p_1(y), q_1(y)), \quad (2.24)$$

$$f(y(0), T = \bar{d}) = (p_0(y) - q_0(y))I(q_0(y) < p_0(y)) = p_0(y) - \min(p_0(y), q_0(y)), \quad (2.25)$$

² We thank Joshua Angrist and Toru Kitagawa for a fruitful discussion regarding the interpretation of LM.

where $I(\cdot)$ denote the indicator function, which equals one if its argument holds true and is zero otherwise. Hence, the densities of potential outcomes under both treatment and nontreatment are identified for compliers as well as defiers. Also, their shares in the population are identified, i.e.,

$$\Pr(T = c) = \int_{y \in \mathcal{Y}} p_1(y) - \min(p_1(y), q_1(y)) dy = \Pr(D = 1|Z = 1) - \lambda_1, \quad (2.26)$$

$$\Pr(T = c) = \int_{y \in \mathcal{Y}} q_0(y) - \min(p_0(y), q_0(y)) dy = \Pr(D = 0|Z = 0) - \lambda_0, \quad (2.27)$$

$$\Pr(T = \bar{d}) = \int_{y \in \mathcal{Y}} q_1(y) - \min(p_1(y), q_1(y)) dy = \Pr(D = 1|Z = 0) - \lambda_1, \quad (2.28)$$

$$\Pr(T = \bar{d}) = \int_{y \in \mathcal{Y}} p_0(y) - \min(p_0(y), q_0(y)) dy = \Pr(D = 0|Z = 1) - \lambda_0, \quad (2.29)$$

where $\lambda_i = \int_{\mathcal{Y}} \min(p_i(y), q_i(y)) dy$ for $i = 0, 1$. This permits identifying the LATEs on the compliers, defiers, and the joint population of compliers and defiers. The identification results are stated in Proposition 2.1 and make use of the following expressions:

$$\mu_c = \frac{\int_{\mathcal{Y}} y \cdot (p_1(y) - \min(p_1(y), q_1(y))) dy}{\Pr(D = 1|Z = 1) - \lambda_1} - \frac{\int_{\mathcal{Y}} y \cdot (q_0(y) - \min(p_0(y), q_0(y))) dy}{\Pr(D = 0|Z = 0) - \lambda_0}, \quad (2.30)$$

$$\mu_{\bar{d}} = \frac{\int_{\mathcal{Y}} y \cdot (q_1(y) - \min(p_1(y), q_1(y))) dy}{\Pr(D = 1|Z = 0) - \lambda_1} - \frac{\int_{\mathcal{Y}} y \cdot (p_0(y) - \min(p_0(y), q_0(y))) dy}{\Pr(D = 0|Z = 1) - \lambda_0}, \quad (2.31)$$

$$\begin{aligned} \mu_{c,\bar{d}} &= \frac{\int_{\mathcal{Y}} y \cdot (\max(p_1(y), q_1(y)) - \min(p_1(y), q_1(y))) dy}{\Pr(D = 1|Z = 1) + \Pr(D = 1|Z = 0) - 2 \cdot \lambda_1} \\ &\quad - \frac{\int_{\mathcal{Y}} y \cdot (\max(p_0(y), q_0(y)) - \min(p_0(y), q_0(y))) dy}{\Pr(D = 0|Z = 0) + \Pr(D = 0|Z = 1) - 2 \cdot \lambda_0}. \end{aligned} \quad (2.32)$$

PROPOSITION 2.1. Assume Assumptions 2.1 and 2.3.

- (1) The LATE on compliers is identified if $p_1(y) > q_1(y)$ for some y in $\mathcal{Y}$: $E(Y(1) - Y(0)|T = c) = \mu_c$.
- (2) The LATE on defiers is identified if $p_1(y) < q_1(y)$ for some y in $\mathcal{Y}$: $E(Y(1) - Y(0)|T = \bar{d}) = \mu_{\bar{d}}$.
- (3) The joint LATE on compliers and defiers is identified if $p_1(y) \neq q_1(y)$ for some y in $\mathcal{Y}$: $E(Y(1) - Y(0)|T = c, d) = \mu_{c,\bar{d}}$.
- (4) If $\Pr(T = \bar{d}) = 0$ and $\Pr(T = c) > 0$, then (2.32) is equivalent to $E(Y(1) - Y(0)|T = c) = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)}$, whereas $E(Y(1) - Y(0)|T = \bar{d})$ is not well defined.
- (5) If $\Pr(T = c) = 0$ and $\Pr(T = \bar{d}) > 0$, then (2.32) is equivalent to $E(Y(1) - Y(0)|T = \bar{d}) = \frac{E(Y|Z=0) - E(Y|Z=1)}{E(D|Z=0) - E(D|Z=1)}$, whereas $E(Y(1) - Y(0)|T = c)$ is not well defined.

Proof: Results 1, 2, and 3 of Proposition 2.1 follow from using (2.22) through (2.25) and (2.26) through (2.29) in $E(Y(d)|T = t) = \frac{E(Y(d), T=t)}{\Pr(T=t)} = \frac{\int_{\mathcal{Y}} y \cdot f(y(d), T=t) dy}{\Pr(T=t)}$ and taking the differences in mean potential outcomes under treatment and nontreatment. Result 4 follows from the fact that $\Pr(T = \bar{d}) = 0$ (global monotonicity) implies $p_1(y) \geq q_1(y)$ and $q_0(y) \geq p_0(y)$ for all $y \in \mathcal{Y}$,

see (2.28) and (2.29), such that (2.32) simplifies to the WR. Finally, Result 5 follows from the fact that $\Pr(T = c) = 0$ implies $p_1(y) \leq q_1(y)$ and $q_0(y) \leq p_0(y)$ for all $y \in \mathcal{Y}$; see (2.22) and (2.23), such that (2.32) simplifies accordingly. $\square$

Note that in contrast to the WR, the weights of the parameters defined in Proposition 2.1 cannot be negative. This is a potential advantage not only when the WR fails to identify the LATE on the compliers, but also in two additional scenarios. First, assume that Assumptions 2.1 and 2.3 hold, that *global* monotonicity fails, and that the LATEs on the compliers and the defiers are equal. If $\Pr(T = c) > \Pr(T = \bar{d})$, 2SLS consistently estimates the WR given by (2.20). The estimator, however, may suffer from weak instrument problems in finite samples in particular when the shares of compliers and defiers are not too different and thus largely netting out, making the denominator of (2.20) very small. In the limiting case, when $\Pr(T = c) = \Pr(T = \bar{d})$, the WR is infinity and 2SLS is inconsistent. In contrast, the LATEs given by Proposition 2.1 remain well defined even when $\Pr(T = c) = \Pr(T = \bar{d}) > 0$. Second, assume that *global* monotonicity is satisfied. In finite samples it may occur that the estimators of $p_d(y)$ and $q_d(y)$ will be close to or will actually be violating the constraints given by (2.19). In this scenario, the estimators based on the sample analogue of the LATEs in Proposition 2.1 may provide efficiency gains compared to 2SLS, as also noted by Imbens and Rubin (1997a). A simulation study presented in the Online Appendix provides support in favour of this statement.³

Figure 1 provides a graphical illustration of the identification results under the conditions of Assumptions 2.1 and 2.3. The compliers are located in the regions of the support, $\mathcal{Y}$, where $p_1(y) > q_1(y)$ and $q_0(y) > p_0(y)$, and in these regions the density of the potential outcome under treatment equals $\frac{p_1(y) - q_1(y)}{\Pr(T=c)}$ if $p_1(y) > q_1(y)$ and is zero otherwise. The share of compliers is given by the area between the two curves, $p_1(y)$ and $q_1(y)$, in areas of $\mathcal{Y}$ with $p_1(y) > q_1(y)$. Similarly, the density of the compliers' potential outcome under nontreatment equals $\frac{q_0(y) - p_0(y)}{\Pr(T=c)}$ if $q_0(y) > p_0(y)$ and is zero otherwise. Again, the area between the curves, $q_0(y)$ and $p_0(y)$, in areas of $\mathcal{Y}$ with $q_0(y) > p_0(y)$ yields the share of compliers. Symmetrically, the density of the defiers' potential outcomes under treatment and nontreatment are $\frac{q_1(y) - p_1(y)}{\Pr(T=d)}$ if $p_1(y) < q_1(y)$ and $\frac{p_0(y) - q_0(y)}{\Pr(T=d)}$ if $p_0(y) > q_0(y)$, respectively, and is zero otherwise. The share of defiers corresponds to the area between $q_1(y)$ and $p_1(y)$ for which $p_1(y) < q_1(y)$ as well as to the region between $p_0(y)$ and $q_0(y)$ for which $q_0(y) < p_0(y)$.

As pointed out by Kitagawa (2009), Assumption 2.1 implies the so-called scale constraint, which requires the share of any compliance type to be equal across treatment states:

$$\int_{\mathcal{Y}} f(y(1), T = t) dy = \int_{\mathcal{Y}} f(y(0), T = t) dy = \Pr(T = t) \quad \forall t = \{a, c, \bar{d}, n\}. \quad (2.33)$$

When imposing Assumption 2.3 in addition to Assumption 2.1, the scale constraint is testable because $f(y(1), T = t)$ and $f(y(0), T = t)$ are identified under these conditions. For the compliers, for example, the scale constraint implies that $\Pr(D = 1|Z = 1) - \lambda_1 = \Pr(D = 0|Z = 0) - \lambda_0$; see (2.26) and (2.27). Under specific regularity conditions discussed in Section 3, estimates of $\Pr(D = 1|Z = 1) - \lambda_1$ and $\Pr(D = 0|Z = 0) - \lambda_0$ are asymptotically normal. One may thus

³ It is worth noting that, even though Proposition 2.1 focusses on the identification of average effects among compliers and/or defiers, Assumptions 2.1 and 2.3 also permit identifying distributional treatment effects. Replacing, e.g., y in μ_c , $\mu_{\bar{d}}$, and $\mu_{c,\bar{d}}$ by the indicator function $I(y \leq y^*)$, with y^* being a value in $\mathcal{Y}$, yields the treatment effects on the cumulative outcome distribution functions of the respective compliance types at outcome value $Y = y^*$. Inverting the cumulative distribution functions of the potential outcomes under treatment and nontreatment allows identifying quantile treatment effects.

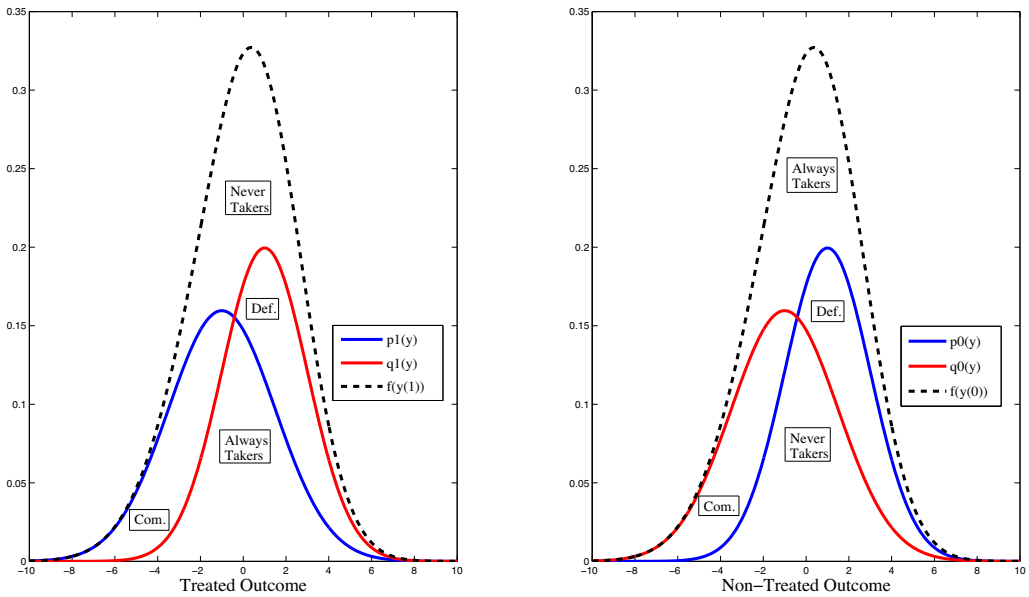


Figure 1. Identification under joint independence and *local* monotonicity. In the left graph, $q_1(y)$ corresponds to the joint density of the potential outcome under treatment and being an always-taker or defier, $p_1(y)$ to the respective density for both always-takers and compliers. By our assumptions, the difference $p_1(y) - q_1(y)$ corresponds to the outcome density among compliers whenever $p_1(y) > q_1(y)$, while $q_1(y) - p_1(y)$ corresponds to the outcome density among defiers whenever $p_1(y) > q_1(y)$. In the right graph, $p_0(y)$ corresponds to the joint density of the potential outcome under nontreatment and being a never-taker or defier. The difference $q_0(y) - p_0(y)$ corresponds to the outcome density among compliers whenever $q_0(y) > p_0(y)$, while $p_0(y) - q_0(y)$ corresponds to the outcome density among defiers whenever $q_0(y) < p_0(y)$.

test the equality of both parameters (i.e., the null hypothesis that the scale constraint holds) using two-sided hypothesis tests, e.g., based on the t -statistic or a bootstrap-based procedure, as considered in the empirical application of Section 4. In the Online Appendix we demonstrate that if the scale constraint holds for one type of subject, this provides a necessary and sufficient condition for (2.33) to hold for all types of subjects in the population. As an additional check of the plausibility of LM, one may plot the differences between $p_d(y)$ and $q_d(y)$ in order to see whether the location of compliers and defiers on $\mathcal{Y}$ is consistent with prior expectations based on theoretical or empirical grounds.

Finally, we can investigate the robustness of LATE identification based on Proposition 2.1 to violations of LM by appropriately adapting the sensitivity analysis in Huber (2014) aimed at violations of global monotonicity. Considering the upper equation of expression (14) in Huber (2014) and only applying it among the subset of outcomes values y satisfying $p_1(y) > q_1(y)$ yields a sensitivity check for $E(Y(1)|T = c)$ to deviations from LM. Likewise, using the lower equation in that expression conditional on $q_0(y) > p_0(y)$ provides a check for $E(Y(0)|T = c)$. As in Huber (2014), the analysis depends on several sensitivity parameters to be chosen by the researcher: the assumed share of defiers in areas satisfying $p_1(y) > q_1(y)$ and $q_0(y) > p_0(y)$,

the level of (dis)similarity between $E(Y(1)|T = c)$ and $E(Y(1)|T = a)$, and the (dis)similarity between $E(Y(0)|T = c)$ and $E(Y(0)|T = n)$. In an analogous manner, one may conduct a robustness check for the LATE on defiers by assuming a nonzero share of compliers in subsets satisfying $p_1(y) < q_1(y)$ and $q_0(y) < p_0(y)$.

2.3. Structural models satisfying joint independence and local monotonicity

To understand the implications of Assumptions 2.1 and 2.3 in a structural model, consider the following two stage endogenous treatment selection model, with the first stage being characterised by a random coefficient model:⁴

$$\begin{aligned} Y_i &= \varphi(D_i, \epsilon_i), \\ D_i &= I(\gamma_0 + \gamma_i Z_i + v_i > 0), \end{aligned} \quad (2.34)$$

where i indexes a particular subject. φ is a general function, and ϵ_i, v_i denote the unobservables in the outcome and treatment equation and may be arbitrarily correlated. γ_0, γ_i denote the constant term and the random coefficient on the instrument, respectively. Our assumptions require that whenever $p_1(Y_i) \geq q_1(Y_i)$ or $q_0(Y_i) \geq p_0(Y_i)$, respectively, γ_i is large enough to satisfy $D_i(1) = I(\gamma_0 + \gamma_i + v_i > 0) \geq D_i(0) = I(\gamma_0 + v_i > 0)$, which rules out defiers locally. A sufficient condition for this is $\gamma_i \geq 0$. For $p_1(Y_i) \leq q_1(Y_i)$ or $q_0(Y_i) \leq p_0(Y_i)$, respectively, it must hold that γ_i is small enough to satisfy $D_i(0) = I(\gamma_0 + v_i > 0) \geq D_i(1) = I(\gamma_0 + \gamma_i + v_i > 0)$. A sufficient condition for this is $\gamma_i \leq 0$. Note that global monotonicity would restrict γ_i in either one or the other way of any i , while Assumption 2.3 restricts γ_i only locally.

To give an idea about possible setups in which Assumption 2.3 holds while global monotonicity does not, we provide two parametric examples that put further structure on the equations in (2.34). Firstly, assume that the outcome equation is characterised by the following model:

$$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 D_i \epsilon_i + \epsilon_i, \quad (2.35)$$

where α_0 is the constant, α_1, α_2 are the coefficients on the treatment and its interaction with ϵ_i (capturing individual effect heterogeneity), which is assumed to have finite first and second moments. In this case, $Y_i(0) = \alpha_0 + \epsilon_i$, $Y_i(1) = \alpha_0 + \alpha_1 + (1 + \alpha_2)\epsilon_i$ and the individual treatment effect is $\alpha_1 + \alpha_2 \epsilon_i$. Moreover, assume that the coefficient on Z in the first stage is deterministic in ϵ_i :

$$\gamma_i = \rho \epsilon_i, \quad (2.36)$$

where ρ is a coefficient on the unobserved term in the structural equation. For $\rho > 0$, it follows that $D_i(1) \geq D_i(0)$ for all $\epsilon_i \geq 0$ and $D_i(1) \leq D_i(0)$ for all $\epsilon_i \leq 0$, while the contrary holds for $\rho < 0$. As $Y_i(1)$ and $Y_i(0)$ are monotonic functions of ϵ_i (apart from the special case that α_2 is exactly -1 and $D_i = 1$), the potential outcomes of the compliers and defiers within a specific treatment state do not overlap so that LM is satisfied.

As an illustrative example for such a setup, consider a public information campaign aimed at encouraging vaccination against infectious diseases as an instrument for actually getting vaccinated (treatment), with an index for general life satisfaction as outcome. Furthermore, assume that ϵ_i reflects trust (positive values) or distrust (negative values) in the government or public institutions, with more positive values implying *ceteris paribus* a higher level of life satisfaction, as trust increases satisfaction with the government. For $\rho > 0$, individuals with a

⁴ We are indebted to Joshua Angrist for pointing us to such models.

positive level of trust might be encouraged to get vaccinated by the campaign, while under a negative level, i.e., distrust, the opposite may happen. Among some individuals willing to get vaccinated without campaign, the encouragement by a distrusted institution might raise doubts and suspicions and induce them to not get vaccinated, i.e., to defy.

As a further model, we consider a generalised Roy-type (1951) model as, for instance, considered in Heckman and Vytlacil (2005), which implies that the probability of treatment increases with the gains it creates. We maintain the previous outcome equation (2.35), but modify the first stage:

$$D_i = I(Y_i(1) - Y_i(0) + \gamma_i Z_i + v_i > 0) = I(\alpha_1 + (\alpha_2 + \rho Z_i)\epsilon_i + v_i > 0), \quad (2.37)$$

where the individual-level treatment effect, e.g., the returns to education or training, now influences the selection into treatment. In equation (2.37) v_i , if different from zero, may be interpreted as intrinsic utility of the treatment not reflected by the treatment effect. The instrument Z exogenously shifts participation, but the direction depends again on ϵ_i as specified in (2.36). The expression left of the equality follows from substituting $Y_i(1) - Y_i(0)$ by $\alpha_1 + \alpha_2 \epsilon_i$ and using (2.36). Again, this model implies a nonoverlapping support of the potential outcomes of compliers and defiers due to γ_i being a deterministic function of ϵ_i and Y_i being monotonic in ϵ_i .

2.4. Alternatives to local monotonicity

Albeit more general than *global* monotonicity, LM may also appear restrictive in applications.⁵ We therefore consider alternative relaxations of monotonicity, starting with the compliers-defiers (CD) assumption of De Chaisemartin (2017) for identifying the LATE on a subset of compliers:

ASSUMPTION 2.4. *There exists a subpopulation of compliers $c^{\bar{d}}$, such that $\Pr(T = c^{\bar{d}}) = \Pr(T = \bar{d})$ and $E(Y(1) - Y(0)|T = c^{\bar{d}}) = E(Y(1) - Y(0)|T = \bar{d})$ (compliers-defiers).*

The CD assumption states that a subset of compliers is of the same size and has an identical LATE as the defiers, known as compliers-defiers. The WR then identifies the LATE on the remaining subset of compliers, who do not necessarily resemble the defiers and are known as compliers-survivors or ‘comvivors’, denoted by c^s . By splitting the compliers into compliers-defiers and compliers-survivors in (2.20), we obtain

$$\begin{aligned} \text{WR} &= \frac{E(Y(1) - Y(0)|T = c^s) \cdot \Pr(T = c^s)}{\Pr(T = c^s) + \Pr(T = c^{\bar{d}}) - \Pr(T = \bar{d})} + \frac{E(Y(1) - Y(0)|T = c^{\bar{d}}) \cdot \Pr(T = c^{\bar{d}})}{\Pr(T = c^s) + \Pr(T = c^{\bar{d}}) - \Pr(T = \bar{d})} \\ &\quad - \frac{E(Y(1) - Y(0)|T = \bar{d}) \cdot \Pr(T = \bar{d})}{\Pr(T = c^s) + \Pr(T = c^{\bar{d}}) - \Pr(T = \bar{d})} = \frac{E(Y(1) - Y(0)|T = c^s) \cdot \Pr(T = c^s)}{\Pr(T = c^s)} \\ &= E(Y(1) - Y(0)|T = c^s). \end{aligned} \quad (2.38)$$

To show that LM and CD are not nested, assume that $\Pr(T = c) > \Pr(T = \bar{d})$ and split the outcome support into the following sets, depending on whether (2.19) is violated or not conditional on the treatment:

$$\begin{aligned} C_{q_1} &= \{y \in \mathcal{Y} : p_1(y) > q_1(y)\}, & C_{p_0} &= \{y \in \mathcal{Y} : q_0(y) > p_0(y)\} \\ C_{p_1} &= \{y \in \mathcal{Y} : q_1(y) > p_1(y)\}, & C_{q_0} &= \{y \in \mathcal{Y} : p_0(y) > q_0(y)\}. \end{aligned} \quad (2.39)$$

⁵ This applies in particular to outcomes with limited support. For binary outcomes Assumption 2.3 implies that the potential outcomes of compliers given a particular treatment state are either all zero or unity, while all defier outcomes take on the respective other value.

Let c^+ and $\bar{d}^+$ denote the compliers and the defiers, respectively, located in either C_{q_1} or C_{p_0} satisfying (2.19), and c^- and $\bar{d}^-$ those located in either C_{p_1} or C_{q_0} where the constraints are violated. Equation (2.20) then corresponds to

$$\begin{aligned} \text{WR} = & \frac{E(Y(1) - Y(0)|T = c^+) \cdot \Pr(T = c^+) - E(Y(1) - Y(0)|T = \bar{d}^+) \cdot \Pr(T = \bar{d}^+)}{(\Pr(T = c^+) - \Pr(T = \bar{d}^+)) - (\Pr(T = \bar{d}^-) - \Pr(T = c^-))} \\ & + \frac{E(Y(1) - Y(0)|T = c^-) \cdot \Pr(T = c^-) - E(Y(1) - Y(0)|T = \bar{d}^-) \cdot \Pr(T = \bar{d}^-)}{(\Pr(T = c^+) - \Pr(T = \bar{d}^+)) - (\Pr(T = \bar{d}^-) - \Pr(T = c^-))}. \end{aligned} \quad (2.40)$$

If CD holds, the share of complivers corresponds to the denominator in (2.40) and $\Pr(c^s) = (\Pr(T = c^+) - \Pr(T = \bar{d}^+)) - (\Pr(T = \bar{d}^-) - \Pr(T = c^-))$. Furthermore, by (2.38) and (2.40) it becomes evident that under CD, a weighted average of LATEs on subsets of c^+ and c^- , whose joint shares equal $\Pr(T = \bar{d}^+) + \Pr(T = \bar{d}^-)$ (i.e., those sets of c^+ and c^- not in c^s), corresponds to a weighted average of LATEs on $\bar{d}^+$ and $\bar{d}^-$. The weights depend on the shares of the various (subsets of) compliance types. One can therefore construct cases in which CD holds if (2.19) is violated. In contrast to CD, LM assumes $\Pr(\bar{d}^+) = \Pr(c^-) = 0$. As this is neither necessary nor sufficient for CD, the two assumptions are not nested. Even when Assumption 2.3 fails and $\Pr(\bar{d}^+) > 0$ and/or $\Pr(c^-) > 0$, expressions (2.30), (2.31), and (2.32) may (similarly to the WR under CD) identify effects on subsets of compliers and defiers. This is, e.g., the case when replacing LM by a weaker local *stochastic* monotonicity (LSM) assumption, see the results in De Chaisemartin (2012):

ASSUMPTION 2.5. *Let $y(d)$ be in $\mathcal{Y}$, for $d = 0, 1$. Then the condition $\Pr(T = c|Y(d) = y(d)) \geq \Pr(T = \bar{d}|Y(d) = y(d))$ implies that $\Pr(T = c|Y(1) = y(1), Y(0) = y(0)) \geq \Pr(T = \bar{d}|Y(1) = y(1), Y(0) = y(0))$. Similarly, the condition $\Pr(T = c|Y(d) = y(d)) \leq \Pr(T = \bar{d}|Y(d) = y(d))$ implies that $\Pr(T = c|Y(1) = y(1), Y(0) = y(0)) \leq \Pr(T = \bar{d}|Y(1) = y(1), Y(0) = y(0))$ (local stochastic monotonicity).*

LSM permits the existence of both compliers and defiers at any marginal potential outcome value. However, it requires that if the share of one type of subjects weakly dominates the share of the other type conditional on either $Y(1)$ or $Y(0)$, it must also dominate conditional on both potential outcomes jointly.⁶

Finally, we propose a local version of CD (LCD) that is even weaker than LSM. We show in the Online Appendix that LCD is the weakest possible assumption under which expressions (2.30) to (2.32) identify causal effects on subgroups of compliers and defiers.

ASSUMPTION 2.6. *Assume that the testable implication of LM hold such that $\lambda_1 = 1 - \delta_0$. (a) There exist two subpopulations of c^+ under treatment and nontreatment, $c_1^{\bar{d}*}$ and $c_0^{\bar{d}*}$, that satisfy $\Pr(T = c_1^{\bar{d}*}) = \Pr(T = c_0^{\bar{d}*}) = \Pr(T = \bar{d}^+)$, $E(Y(1)|T = c_1^{\bar{d}*}) - E(Y(0)|T = c_0^{\bar{d}*}) = E(Y(1) - Y(0)|T = \bar{d}^+)$. Furthermore, the remaining compliers under treatment and nontreatment, c_1^{s*} and c_0^{s*} , are of the same type in terms of joint potential outcomes: $c_1^{s*} = c_0^{s*} = c^{s*}$;*

⁶ It is worth noting that, together with Assumption 2.1, the satisfaction of LSM implies the scale constraint $\Pr(D = 1|Z = 1) - \lambda_1 = \Pr(D = 0|Z = 0) - \lambda_0$, just like the satisfaction of LM. Under LSM, however, $\Pr(D = 1|Z = 1) - \lambda_1$ and $\Pr(D = 0|Z = 0) - \lambda_0$ correspond to the shares of subsets of treated and nontreated compliers who outnumber defiers with the same joint potential outcomes, while under LM, the respective parameters correspond to the shares of all compliers.

(b) There exist two subpopulations of under treatment and nontreatment, $\bar{d}^-$, $\bar{d}_1^{c*}$, and $\bar{d}_0^{c*}$, that satisfy $\Pr(T = \bar{d}_1^{c*}) = \Pr(T = \bar{d}_0^{c*}) = \Pr(T = c^-)$, $E(Y(1)|T = \bar{d}_1^{c*}) - E(Y(0)|T = \bar{d}_0^{c*}) = E(Y(1) - Y(0)|T = c^-)$. Furthermore, the remaining defiers, $\bar{d}_1^{s*}$ and $\bar{d}_0^{s*}$, are of the same type in terms of joint potential outcomes: $\bar{d}_1^{c*} = \bar{d}_0^{c*} = \bar{d}^{s*}$ (local compliers-defiers).

LCD requires that any complivers in one treatment state, who are per definition situated in outcome regions satisfying (2.19), would inhabit outcome regions that also satisfy (2.19) in the other treatment state. Therefore, c_1^{s*} and c_0^{s*} have the same joint potential outcomes. Analogous results apply to the defiers that survive the compliers in regions where (2.19) is violated.

Proposition 2.2 states the identification results under Assumptions 2.1 and 2.6 based on expressions (2.30), (2.31), and (2.32).

PROPOSITION 2.2.

- (1) Under Assumptions 2.1 and 2.6(a), the LATE on compliers-survivors is identified if $p_1(y) > q_1(y)$ for some y in $\mathcal{Y}$: $E(Y(1) - Y(0)|T = c^{s*}) = \mu_c$.
- (2) Under Assumptions 2.1 and 2.6(b), the LATE on defiers-survivors is identified if $p_1(y) < q_1(y)$ for some y in $\mathcal{Y}$: $E(Y(1) - Y(0)|T = \bar{d}^{s*}) = \mu_{\bar{d}}$.
- (3) Under Assumptions 2.1, 2.6(a), and 2.6(b), the joint LATE on compliers-survivors and defiers-survivors is identified if $p_1(y) \neq q_1(y)$ for some y in $\mathcal{Y}$: $E(Y(1) - Y(0)|T = c^{s*}, \bar{d}^{s*}) = \mu_{c, \bar{d}}$.

Proof: See the Online Appendix. □

For illustration, Figure 2 provides hypothetical example in which LCD is satisfied, while LSM and CD are violated, thus showing that LCD and CD are not nested. There are four compliers (represented by a green square) and three defiers (represented by a blue circle). LSM does not hold because $\Pr(T = \bar{d}|Y(0) = 1) = 2/3 > \Pr(T = c|Y(0) = 1) = 1/3$ and $\Pr(T = \bar{d}|Y(1) = 1, Y(0) = 1) = 0 < \Pr(T = c|Y(1) = 1, Y(0) = 1) = 1$, as the only individual with $Y(1) = 1$ and $Y(0) = 1$ is complier 3. Furthermore, CD is violated, as it is not possible to create a subpopulation of compliers of equal number and with the same LATE as the defiers. The LATE on the defiers is $\frac{2}{3} \cdot 2 + \frac{1}{3} \cdot (-1) = 1$, while the two possible LATEs from using only three of the four compliers are either $(-1) \cdot \frac{3}{3} = -1$ (using compliers 1, 2, and 4) or $\frac{2}{3} \cdot (-1) + \frac{1}{3} \cdot 0 = -\frac{2}{3}$ (using compliers 1, 2, and 3) or $\frac{1}{3} \cdot (1) + \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot 0 = 0$ (using either compliers 1, 3, and 4 or 2, 3, and 4). LCD is satisfied and identifies the LATE on compliers 1 and 2 in this example, as we can use complier 3 under treatment and complier 4 under nontreatment to ‘neutralise’ defier 1, which is the only one in the region satisfying (2.19). An analogous result holds for the defiers.

Even in cases where both LCD and CD hold, our approach might be preferable because it refers to a possibly larger complier subpopulation. Note that $\Pr(T = c^{s*})$ weakly dominates the share of complivers under CD given by $\Pr(c^s) = (\Pr(T = c^+) - \Pr(T = \bar{d}^+)) - (\Pr(T = \bar{d}^-) - \Pr(T = c^-))$, since $\Pr(T = \bar{d}^-) - \Pr(T = c^-) \geq 0$. Thus, if both LCD and CD hold, LCD allows identifying the LATE on a potentially larger share of compliers than CD when $\Pr(T = \bar{d}^-) - \Pr(T = c^-) > 0$ (i.e., C_{p_1} and C_{q_0} are nonempty). Using the results of this paper may therefore lead to finite sample efficiency gains as well as higher external validity.

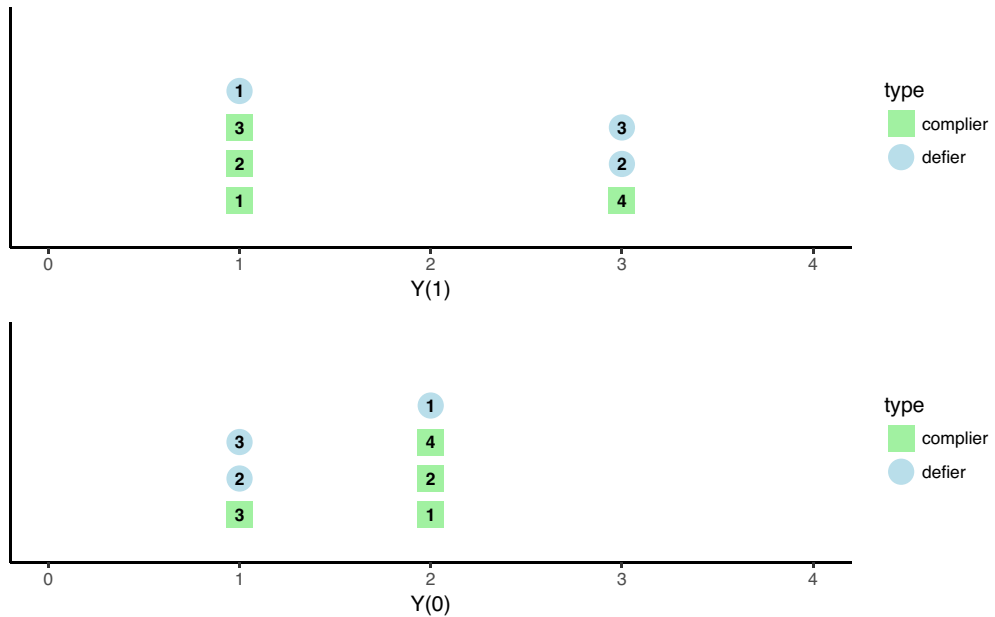


Figure 2. Graphical illustration of a data generating process where LCD is satisfied and CD and LSM are violated.

3. ESTIMATION

Estimation of the LATEs under LM or LCD is based on the sample analogues of expressions (2.30), (2.31), and (2.32), which can be rewritten as follows:

$$\begin{aligned}\mu_c &= \frac{\int_{y \in C_{q_1}} y \cdot (p_1(y) - q_1(y)) dy}{\int_{y \in C_{q_1}} p_1(y) - q_1(y) dy} - \frac{\int_{y \in C_{p_0}} y(q_0(y) - p_0(y)) dy}{\int_{y \in C_{p_0}} q_0(y) - p_0(y) dy}, \\ \mu_d &= \frac{\int_{y \in C_{p_1}} y \cdot (q_1(y) - p_1(y)) dy}{\int_{y \in C_{p_1}} q_1(y) - p_1(y) dy} - \frac{\int_{y \in C_{q_0}} y(p_0(y) - q_0(y)) dy}{\int_{y \in C_{q_0}} p_0(y) - q_0(y) dy}, \\ \mu_{c,d} &= \frac{\int_{y \in C_{q_1}} y \cdot (p_1(y) - q_1(y)) dy}{\int_{y \in C_{q_1}} p_1(y) - q_1(y) dy + \int_{y \in C_{p_1}} q_1(y) - p_1(y) dy} \\ &\quad - \frac{\int_{y \in C_{p_0}} y(q_0(y) - p_0(y)) dy}{\int_{y \in C_{p_0}} q_0(y) - p_0(y) dy + \int_{y \in C_{q_0}} p_0(y) - q_0(y) dy} \\ &\quad + \frac{\int_{y \in C_{p_1}} y(q_1(y) - p_1(y)) dy}{\int_{y \in C_{q_1}} p_1(y) - q_1(y) dy + \int_{y \in C_{p_1}} q_1(y) - p_1(y) dy} \\ &\quad - \frac{\int_{y \in C_{q_0}} y(p_0(y) - q_0(y)) dy}{\int_{y \in C_{p_0}} q_0(y) - p_0(y) dy + \int_{y \in C_{q_0}} p_0(y) - q_0(y) dy},\end{aligned}$$

where the level sets C_{q_1} , C_{p_0} , C_{p_1} , and C_{q_0} are defined in (2.39). Note that under Assumption 1 and LM, the numerator in the first right-hand expression of μ_c , $\int_{y \in C_{q_1}} y \cdot (p_1(y) - q_1(y)) dy$ corresponds to $E[Y(1), T = c]$, as $p_1(y) - q_1(y)$ gives the joint density of the potential outcome under treatment and being a complier for any $y \in C_{q_1}$. For the same reason, the denominator $\int_{y \in C_{q_1}} p_1(y) - q_1(y) dy$ corresponds to $\Pr(T = c)$, such that the ratio of the numerator and the denominator yields $E[Y(1)|T = c]$. Under Assumption 1 and LCD, the mean potential outcome under treatment is identified for the subset of compliers-survivors. Analogous arguments hold for the other parameters defined in terms of level sets. Their sample analogues, for which we use the $\hat{\curvearrowright}$ notation, correspond to

$$\hat{\mu}_c = \frac{\hat{\theta}_1}{\hat{P}_{1|1} - \hat{\lambda}_1} - \frac{\hat{\theta}_0}{\hat{P}_{0|0} - \hat{\lambda}_0}, \quad (3.1)$$

$$\hat{\mu}_{\bar{d}} = \frac{\hat{\theta}_2}{\hat{P}_{1|0} - \hat{\lambda}_1} - \frac{\hat{\theta}_3}{\hat{P}_{0|1} - \hat{\lambda}_0}, \quad (3.2)$$

$$\hat{\mu}_{c,\bar{d}} = \frac{\hat{\theta}_1 + \hat{\theta}_2}{\hat{P}_{1|1} + \hat{P}_{1|0} - 2\hat{\lambda}_1} - \frac{\hat{\theta}_0 + \hat{\theta}_3}{\hat{P}_{0|0} + \hat{P}_{0|1} - 2\hat{\lambda}_0}, \quad (3.3)$$

where

$$\begin{aligned} \hat{\theta}_0 &= \int_{C_{p_0}} y \cdot (\hat{q}_0(y) - \hat{p}_0(y)) dy, & \hat{\theta}_1 &= \int_{C_{q_1}} y \cdot (\hat{p}_1(y) - \hat{q}_1(y)) dy, \\ \hat{\theta}_2 &= \int_{C_{p_1}} y (\hat{q}_1(y) - \hat{p}_1(y)) dy, & \hat{\theta}_3 &= \int_{C_{q_0}} y \cdot (\hat{p}_0(y) - \hat{q}_0(y)) dy, \end{aligned}$$

and

$$\begin{aligned} \hat{\lambda}_d &= \int_y \min(\hat{p}_d(y), \hat{q}_d(y)) dy, \quad d = 0, 1, \\ \hat{P}_{d|z} &= \frac{1}{n} \sum_{i=1}^n \frac{I(D_i = d) I(Z_i = z)}{\frac{1}{n} \sum_{i=1}^n I(Z_i = z)}, \quad d, z = 0, 1. \end{aligned}$$

The density estimates,

$$\hat{p}_d(y) = \frac{\hat{f}_{Y,D,Z}(y, D = d, Z = 1)}{\hat{f}_Z(Z = 1)}, \quad \hat{q}_d(y) = \frac{\hat{f}_{Y,D,Z}(y, D = d, Z = 0)}{\hat{f}_Z(Z = 0)},$$

are obtained by kernel density estimation:

$$\hat{f}_{Y,D,Z}(y, D = d, Z = z) = \frac{1}{n} \sum_{i=1}^n L_{(D_i, Z_i), (d, z)} W_{h, Y_i, y}, \quad \hat{f}_Z(Z = z) = \frac{1}{n} \sum_{i=1}^n L_{Z_i, z},$$

where L and W are product kernels, see Li and Racine (2007), which are defined as

$$L_{(D_i, Z_i), (d, z)} = I(D_i = d) I(Z_i = z), \quad W_{h, Y_i, y} = \frac{1}{h} w\left(\frac{y - Y_i}{h}\right), \quad L_{Z_i, z} = I(Z_i = z).$$

Where h denotes the bandwidth and w denotes the kernel function, corresponding to the Gaussian kernel in our simulations (presented in the Online Appendix) and empirical application of Section 4. Subscript $i \in \{1, 2, \dots, n\}$ denotes an observation in an independent and identically distributed (i.i.d.) sample of size n . The integrals can be computed numerically using any of

the many approximation methods available. For the benchmark estimators, we use the trapezoid rule and refer to them as ‘plug-in’ estimators. Moreover, in the Online Appendix we present estimators that are based on a computationally more convenient approximation of integrals of density functions, which imposes the restriction that a proper density must integrate to unity. We refer to these methods as ‘modified plug-in’ estimators. The Online Appendix also shows how our identification results and estimators simplify in the case of one-sided noncompliance.

To derive the asymptotic properties of the estimators for known level sets, we assume the following regularity conditions concerning the kernel estimators.

ASSUMPTION 3.1. (a) *The general non-negative bounded kernel function, $w(\cdot)$, satisfies (i) $\int w(v) dv = 1$, (ii) $w(v) = w(-v)$, and (iii) $\int v^2 w(v) dv = \kappa_2 > 0$; (b) $\sup_{y \in \mathcal{Y}} B(y, D_i = d, Z_i = z) = M_{(d,z),y} < \infty$ for all $d = 0, 1$ and $z = 0, 1$, where $B(y, D = d, Z = z) = \frac{1}{2} \kappa_2 \frac{\nabla^2(f_{Y,D|Z}(y, D=d|Z=z))}{(\nabla y)^2} \Big/ f_Z(Z = z)$; (c) for all $d = 0, 1$ (i) $\inf_{y \in \mathcal{Y}} p_d(y) = M_{p_i,y} > 0$, and (ii) $\inf_{y \in \mathcal{Y}} q_d(y) = M_{q_i,y} > 0$; (d) (Y_i, D_i, Z_i) for $i = 1, 2, \dots, n$ is an i.i.d. random vector with a joint mixed distribution given by $f_{Y,D,Z}(Y_i = y, D_i = d, Z_i = z)$ with support $\mathcal{Y} \times (0, 1) \times (0, 1)$, where $\mathcal{Y} \subseteq \mathbb{R}$. Furthermore, all absolute second order moments $E(|Y|^2)$, $E_{p_d}(|Y|^2)$ and $E_{q_d}(|Y|^2)$ exist for $d = 0, 1$ (kernel estimation).*

Assumption 3.1 is standard for kernel methods. Assumption 3.1(a) implies that the estimated density is well defined. Assumption 3.1(b) imposes twice continuous differentiability of p_d and q_d . Assumption 3.1(c) ensures that p and q are not truncated and rules out boundary effects. It simplifies the proof, but can be relaxed by using boundary kernels; see Li and Racine (2007). Assumption 3.1(d) is quite general and implies, for instance, the existence of the following moments: $E_{p_1}(Y_i|Y_i \in C_{q_1})$, $E_{q_1}(Y_i|Y_i \in C_{q_1})$, $E_{p_0}(Y_i|Y_i \in C_{p_0})$, $E_{q_0}(Y_i|Y_i \in C_{p_0})$, $E(|Y_i|)$, etc.

Theorem 3.1 establishes the asymptotic normality of the estimators of the LATEs identified by Propositions 2.1 and 2.2:

THEOREM 3.1. *Under Assumption 3.1 and knowledge of the level sets C_{p_0} , C_{q_0} , C_{p_1} , and C_{q_1} ,*

$$\sqrt{n}(\hat{\mu}_{c,d} - \mu_{c,d} - b_{c,d}) \xrightarrow{d} \mathcal{N}(0, \Omega_{c,d}),$$

$$\sqrt{n}(\hat{\mu}_c - \mu_c - b_c) \xrightarrow{d} \mathcal{N}(0, \Omega_c),$$

$$\sqrt{n}(\hat{\mu}_d - \mu_d - b_d) \xrightarrow{d} \mathcal{N}(0, \Omega_d),$$

for $n \rightarrow \infty$, $h \rightarrow 0$, and $\sqrt{nh^2} \rightarrow 0$, where $b_{c,d}$, b_c , and b_d denote finite sample biases that vanish asymptotically. Detailed expressions for $\Omega_{c,d}$, Ω_c , and Ω_d are provided in the Online Appendix.

Proof: See the Online Appendix. □

Theorem 3.1 implies that if the level sets C_{p_0} , C_{q_0} , C_{p_1} , and C_{q_1} are known, the LATE estimators defined in (3.1), (3.2), and (3.3) are $\sqrt{n}$ -consistent and asymptotically normal under relatively mild regularity conditions. To evaluate how well the asymptotic distributions of Theorem 1 approximate the finite sample distributions of the LATE estimators, the Online Appendix presents an extensive set of Monte Carlo simulations, in which properties such as one-sided coverage probabilities, bias, and efficiency are investigated. The simulation results suggest that the asymptotic

distributions of the LATE estimators satisfactorily approximate the finite sample behaviour. The Online Appendix also proposes further (and asymptotically equivalent) estimators, which do not require kernel smoothing and bandwidth selection for known level sets. This feature makes them particularly attractive from a practical and computational perspective. We refer to these methods as ‘bandwidth-free’ estimators.

In empirical applications, the level sets are typically unknown and need to be estimated. Anderson et al. (2012) and Mammen and Polonik (2013) suggest plug-in methods for estimating the level sets and show that they are $\sqrt{n}$ -consistent and asymptotically normal. Both approaches can be adapted to our setting, for example, by estimating C_{p_0} by $\widehat{C}_{p_0} = \{y \in C : \widehat{q}_0(y) - \widehat{p}_0(y) > c_n, \widehat{q}_0(y) > 0, \widehat{p}_0(y) > 0\}$, where c_n is a positive and data-dependent threshold parameter that approaches zero as the sample size goes to infinity.

Alternatively, one may choose the threshold parameter c_n such that the conditions given in Lemma 6 of Dong et al. (2023) are satisfied. In this case, Theorem 3.1 holds even when the level sets are unknown and have to be estimated. To discuss this approach more formally, let us define $\widehat{\Delta}_d(y) = \widehat{p}_d(y) - \widehat{q}_d(y)$. If there exists a threshold parameter c_n such that

$$c_n^{-1} \sup |\widehat{\Delta}_d(y) - \Delta_d(y)| = o_p(1) \quad (3.4)$$

and

$$c_n^2 (\sup |\widehat{\Delta}_d(y) - \Delta_d(y)|)^{-1} = o_p(1), \quad (3.5)$$

then the level set estimation error has no effect on the asymptotic distribution as the integrals over estimated and known level sets are asymptotically equivalent.

As in Dong et al. (2023), consider defining the threshold parameter as

$$c_n = 1.96 \cdot \sqrt{\text{Var}(\tilde{\Delta}_d(y))} = O_p\left((n\tilde{h})^{(-\frac{1}{2})}\right), \quad (3.6)$$

where $\tilde{\Delta}_d(y)$ is a preliminary estimator of $\Delta_d(y)$ based on a bandwidth $\tilde{h}$. The threshold parameter defined by (3.7) appears intuitive, as it trims estimates of $\Delta_d(y)$ which are insignificant at the 5% level of statistical significance. Given that $\widehat{p}_d(y)$ and $\widehat{q}_d(y)$ are regular kernel density estimators, we have that $\sup |\widehat{\Delta}_d(y) - \Delta_d(y)| = O_p\left((nh)^{(-\frac{1}{2})}\right)$ as in Dong et al. (2023). If we choose $\tilde{h}$ such that $\tilde{h}/h \rightarrow 0$ and $n\tilde{h}^2/h \rightarrow \infty$, then $\sqrt{\text{Var}(\tilde{\Delta}_d(y))} = O_p\left((n\tilde{h})^{(-\frac{1}{2})}\right) > O_p\left((nh)^{(-\frac{1}{2})}\right)$ and thus, the conditions given by (3.4) and (3.5) are satisfied.

In the simulations, we suggest a slight modification of (3.7), where we use entirely data driven bootstrap-based plug-in estimators as opposed to determining the appropriate bandwidth parameter satisfying (3.4)–(3.6). First, the estimator $\widehat{\Delta}_d(y)$, is obtained using a cross-validated bandwidth selection procedure. Let $\widehat{\Delta}_d^*(y)$ denote the bootstrap estimate of $\widehat{\Delta}_d(y)$. Then, the level sets are estimated using pointwise $(1 - \alpha)\%$ -confidence intervals, e.g.,

$$\widehat{C}_{p_d} = \left\{ y \in \mathfrak{R} : \text{Median}(\widehat{\Delta}_d^*(y)) + Z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\widehat{\Delta}_d^*(y))} < 0 \right\}, \quad (3.7)$$

$$\widehat{C}_{q_d} = \left\{ y \in \mathfrak{R} : \text{Median}(\widehat{\Delta}_d^*(y)) - Z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\widehat{\Delta}_d^*(y))} > 0 \right\}, \quad (3.8)$$

for $d = 0, 1$, where $Z_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$ -percentile of the standard normal distribution.⁷ This approach is numerically efficient and the simulations presented in the Online Appendix suggest

⁷ Here, the median may be preferred over the mean of $(\widehat{\Delta}_d^*(y))$ because of increased precision/robustness of the resulting bootstrap statistic.

that the properties of the estimators do not change substantially when replacing known level sets by estimated level sets based on (3.7) and (3.8). Moreover, in the simulations, this method outperforms the approaches of Anderson et al. (2012) and Mammen and Polonik (2013), the asymptotic behaviour of which is known.

We also consider a modified version of the bootstrap approach based on randomly splitting the data into two equally sized subsamples. We then use the first subsample for estimating the level sets $\widehat{C}_{p_d}$, $\widehat{C}_{q_d}$ and the second subsample for estimating the LATEs based on the level set estimates in the first subsample. Estimating level sets and treatment effects in different parts of the data aims at avoiding overfitting bias related to correlations of both estimation steps if conducted in the same data. Such sample splitting approaches for parameter estimation have, for instance, been suggested or applied in Bickel (1982), Schick (1986), Angrist and Krueger (1995), and Altonji and Segal (1996), among others, and are a key feature in the literature on causal machine learning, see, for instance, Chernozhukov et al. (2018). Apart for limited samples below 1,000 observations, the method performs quite decent in the simulations presented in the Online Appendix due to its competitively low finite sample bias. This, however, comes at the cost of a somewhat larger variance than under bootstrapping without sample splitting.

The results of Theorem 3.1 are also useful for comparing our estimators and 2SLS under global monotonicity. Theorem 3.1 implies that the two estimators are both consistent and asymptotically normal. As $p_1(y) - q_1(y) > 0$ and $q_0(y) - p_0(y) > 0$ over the entire outcome support under global monotonicity, the level sets are known. Because these differences might nevertheless be negative in small samples due to sampling noise, the 2SLS first stage can be pushed towards 0, while this has no impact on our estimator. However, imposing non-negativity might introduce finite sample bias and therefore, we investigate two different simulation designs without defiers such that global monotonicity holds. We compare a subset of our LATE estimators to 2SLS and the semi-parametric estimator of Cheng, Qin et al. (2009). The idea behind these simulations is to check whether our estimators are competitive alternatives to estimators which are known to be more efficient than 2SLS in finite samples, but impose non-negativity at the cost of introducing additional structural assumptions. The first data generating process (DGP) considered satisfies the structural assumptions underlying the empirical likelihood approach of Cheng, Qin et al. (2009). As documented in the Online Appendix, we find that our estimators clearly outperform 2SLS in terms of root mean squared error, but do slightly worse than Cheng, Qin et al. (2009). In the second DGP the assumptions of Cheng, Qin et al. (2009) are violated and we find that our estimators perform substantially better than both 2SLS and Cheng, Qin et al. (2009) if the sample size exceeds 1,000. We obtain similar results when first estimating the level sets. This implies that even in cases where global monotonicity is a reasonable assumption, our estimators may be attractive alternatives to conventional approaches.

We note that in many applications where it does not appear credible that Assumption 2.1 holds unconditionally, it is common to control on observable pre-instrument characteristics, denoted by X . An example is Card (1995), who evaluates the returns to education using geographic proximity to college as instrument for the potentially endogenous education decision. The intuition is that proximity should induce some individuals to attend college who would otherwise not, e.g., due to costs associated with not living at home. However, the instrument might be correlated with factors like local labour market conditions or family background, which could be related to the earnings outcome, thus implying a violation of Assumption 2.1. For this reason, Card (1995) includes control variables such as parental education, ethnicity, urbanity, and geographic region in his IV estimations.

Controlling for covariates might also be required for the satisfaction of LM, implying that Assumption 2.3 does not hold unconditionally, but only conditional on X . This may be the case if the effect of Z is heterogeneous in X in particular ways. Consider, for instance, the scenario that for specific values of the potential outcome $Y(d)$ (with $d \in \{0, 1\}$), both compliers and defiers exist, but are fully distinct in terms of their values in X . This implies that Assumption 2.3 is conditionally valid when controlling for X . A further example is a scenario in which Z causally affects X and X causally affects D in particular ways. Consider again the case that both compliers and defiers exist for specific values of $Y(d)$. Furthermore, assume that defiance is exclusively caused by Z positively affecting X and X negatively affecting Z , while compliance is unrelated to X . In this case, controlling for (the mediating effect of) X will eliminate defiance such that only compliers remain and LM holds. In analogous ways, controlling for X can also have implications for the conditional or unconditional validity of other monotonicity assumptions, like Assumption 2.5. When conditioning on X , it follows that any testable implication of the identifying assumptions (like the scale constraint) need to hold conditional on X .

When the covariates to be conditioned on are discrete and their number is limited, one may control for them by estimating the LATEs within cells of values of X and averaging over cells to obtain the effect of interest. For example, assume that X is a binary variable and let $\hat{\mu}_c^1$ and $\hat{\mu}_c^0$ denote the LATEs estimated in the subsamples with $X = 1$ and $X = 0$, respectively. Then, the unconditional LATE is obtained as $\hat{\mu}_c = [\frac{1}{n} \sum_{i=1}^n (X_i)] \cdot \hat{\mu}_c^1 + [\frac{1}{n} \sum_{i=1}^n (1 - X_i)] \cdot \hat{\mu}_c^0$. However, if X is continuous and high dimensional, the nonparametric estimation of $f(y, D = d | Z = z, X = x)$ appears challenging in finite samples. One might therefore prefer controlling for the instrument propensity score $\Pr(Z = 1 | X)$ instead, see, for instance, Frölich (2007). This is asymptotically equivalent to controlling for X , as in Rosenbaum and Rubin (1983), but reduces the dimensionality of the conditioning set to one. Therefore, a practically feasible approach consists of estimating $P(Z = 1 | X)$ parametrically (e.g., by logit), discretising the propensity score distribution into a finite number of cells (binning), estimating the LATEs in each cell, and taking a weighted average across cells. See Lee (2009) for the application of such a procedure in a different estimation problem.

4. EMPIRICAL APPLICATION

This section provides an application to the 1980 U.S. census data analysed by Angrist and Krueger (1991), which (among other cohorts) contain 486,926 males born in 1940–1949. Angrist and Krueger (1991) assess the effect of education on wages by using the calendar quarter of birth as an instrument to control for the potential endogeneity (for example, due to unobserved ability) of education. The idea is that the quarter of birth instrument affects education through age-related schooling regulations. State-specific rules require that a child must have attained the first grade admission age, which is six years in most cases, on a particular date during the year. Because schooling is compulsory until the age of 16 in most states, see Angrist and Krueger (1991, app. 2), pupils who are born early in the year are in the 10th grade when turning 16. As the school year usually starts in September and ends in July, these pupils have nine years of completed education if they quit education as soon as possible. In contrast, pupils born after the end of the academic year, but still entering school in the same year that they turn six have ten years of completed education at age 16. This suggests education to be monotonically increasing in the quarter of birth.

However, the quarter of birth instrument is far from being undisputed. For instance, Bound et al. (1995) challenge the validity of the exclusion restriction and present empirical results that point to systematic patterns in the seasonality of birth (for instance, with respect to performance in school, health, and family income), which may cause a direct association with the outcome. In line with these arguments, Buckles and Hungerman (2013) document large differences in maternal characteristics for births throughout the year (with winter births being more often realised by teenagers and unmarried women) based on U.S. birth certificate data and census data. For this reason, we only consider quarters two and three in our analysis (i.e., the warmer seasons of the year). We acknowledge that this may not completely dissipate concerns about seasonality, but nevertheless assume that Assumption 1 is satisfied for the subsample born in the second or third quarters of the year (244,512 observations). The instrument Z is equal to zero for individuals born in the second quarter and equal to unity if born in the third quarter. The binary treatment D is equal to zero for individuals receiving high school education or less (up to 12 years of education) and one if obtaining at least some higher education (13 years or more). That is, we are interested in the returns to having at least some college education. According to our definition, roughly 48% (52%) of our sample receive lower (higher) education. The outcome variable Y is the log weekly wage.

Second, a crucial question related to standard IV estimation is whether positive monotonicity holds for all individuals. This appears unlikely in the light of strategic school entry behaviour as documented by Barua and Lang (2009), which may entail deviations from the schooling regulations. The authors present empirical evidence of redshirting based on 1960 U.S. census data implying that many parents did not enrol their children at the earliest permissible entry age, but postponed school entry. This occurred particularly often for children born late in the year. Aliprantis (2012) provides further empirical support for redshirting based on the Early Childhood Longitudinal Study. Moreover, Klein (2010) acknowledges that redshirting may also be induced by schools, which are generally not obliged to admit all children who turn six before the state-wide cutoff date. Both redshirting and school policies may reverse the relation between education and the instrument for some individuals. Because children with postponement are close to seven when entering school and just have started the 10th grade when turning 16, some of them might drop out directly, with only nine years of completed education. In contrast, pupils born earlier are at an advanced stage of the 10th grade when turning 16 and may therefore decide to complete the grade, thus having at least 10 years of completed education. For these defying individuals, compulsory schooling decreases in the quarter of birth and therefore violates *global* monotonicity.⁸

The implausibility of *global* monotonicity may motivate the use of the weaker local monotonicity, while maintaining Assumption 2.1. LM is satisfied if wage is a monotonic function of socio-economic status and if postponement is deterministic in socio-economic status, see the structural model in Section 2.3 (with ϵ_i reflecting socio-economic status). Parents with a high socio-economic status could be more inclined to delay their children's school entry, for example, because they can more easily afford child care costs for an extra year and/or behave more strategically in terms of schooling decisions compared to other groups. However, assuming that potential

⁸ We point out that there might also exist individuals for whom redshirting actually increases rather than reduces the level of education. This is the case if the higher age of almost seven when entering school entails a better academic performance to an extent that it induces those individuals to attend college, while they would have dropped out prior to college under a lower school entry age due to doing worse in school. Those individuals are compliers in spite of redshirting, and therefore, they do not belong to the group of defiers, for whom redshirting induces dropping out of school rather than going to college.

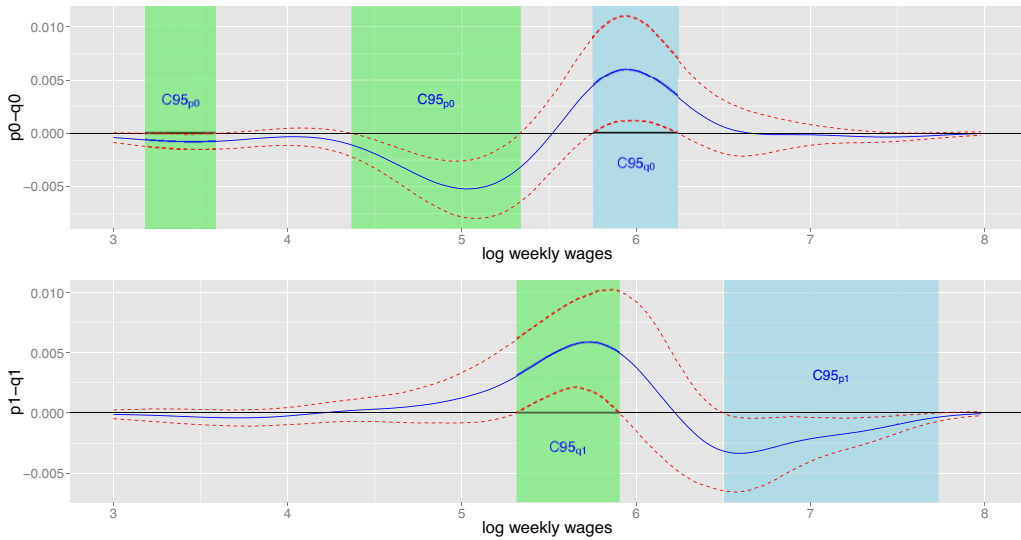


Figure 3. Estimated differences in densities of log weekly wages, $\hat{p}_d(w) - \hat{q}_d(w)$ (blue solid line) under treatment ($d = 1$, lower panel) and nontreatment ($d = 0$, upper panel). The red dashed lines indicate 95% confidence bands. Estimations of the curves are based on 6,000 bootstrap replications using a Gaussian kernel and cross-validated bandwidth selection. In green regions, compliers statistically significantly dominate defiers, while in blue regions, defiers statistically significantly dominate compliers.

wages of compliers and defiers do not overlap at all appears questionable as postponement is most likely not fully deterministic in socio-economic status, but also influenced by other factors (like parenting style). The weaker local compliers-defiers (LCD) therefore appears more realistic than LM and permits estimating the effects for subsets of compliers and defiers. Under LCD, we still suspect the defiers to outnumber the compliers at the upper part of the conditional wage distribution given education, but do no longer rule out overlaps in the wages. Empirical evidence pointing in this direction is provided by Bedard and Dhuey (2006), who report that children with a higher socio-economic status are over-represented among those who redshirt, and Aliprantis (2012), who finds that children whose enrolment is delayed are disproportionately wealthy with better-educated parents and more books at home.

The plausibility of our conjecture that defiers dominate among higher wages can be checked graphically by plotting the estimated differences between p_d and q_d for $d = 1, 0$. We do this and estimate the level sets with 95% confidence intervals based on the bootstrap procedure outlined in Section 3, see (3.8) and (3.9). In each of the 6,000 bootstraps, the densities $p_d(y)$ and $q_d(y)$ are estimated by kernel methods and cross-validated bandwidth on an equidistant grid of 1,000 values on the support of the empirical weekly log wage distributions. The results are plotted in Figure 3. In the blue regions, the defiers outnumber the compliers with 95% confidence, which indeed consist of the upper parts of the log wage distributions in either treatment state. In the green regions, the compliers outnumber the defiers, which correspond to lower parts of the wage distributions. The green regions thus contain the outcomes for the compliers-survivors under LCD (or the total of compliers under the less realistic LM), while the blue regions contain the outcomes for the defiers-survivors under LCD (or the total of defiers

under LM). Results are similar when 90% confidence bands are considered, as shown in the Online Appendix.

After having obtained estimates for the level sets based on bootstrapping, we estimate the LATEs on the compliers, the defiers, and both groups jointly based on three approaches, see Section 3: the bandwidth-free estimators $\hat{\mu}_c^a$, $\hat{\mu}_d^a$, $\hat{\mu}_{c,d}^a$; the plug-in estimators $\hat{\mu}_c$, $\hat{\mu}_d$, $\hat{\mu}_{c,d}$; and the modified plug-in estimators denoted by $\hat{\mu}_c^b$, $\hat{\mu}_d^b$, and $\hat{\mu}_{c,d}^b$. For the plug-in and modified plug-in estimators we use Silverman's rule-of-thumb bandwidth, denoted b , raised to the power of $3/2$.⁹ Furthermore, we test the scale constraint $\Pr(D = 1|Z = 1) - \lambda_1 - (\Pr(D = 0|Z = 0) - \lambda_0) = 0$ based on all three estimation methods.

Concerning inference, we bootstrap the estimators 6,000 times to approximate their distributions and obtain p -values based on the ranks of the estimates in their respective re-centred bootstrap distributions. We use two-sided hypothesis tests for computing the p -values of the scale constraint and the LATE estimators and one-sided tests for the complier and defier shares (as the theoretical lower bound of the shares is zero).

Table 2 presents the results based on the estimated level sets using 95% (C95) and 90% (C90) confidence bands, as well as the 2SLS estimates. The first six columns in the upper panel present the estimated complier and defier shares (e.g., $\hat{\Pr}(T = c)^a$ and $\hat{\Pr}(T = d)^a$) based on expressions (2.27) to (2.29), using the bandwidth-free and plug-in type estimators, respectively. Either share is, albeit very small (0.3% to 0.4%), statistically significant at the 5% level irrespective of the estimation method. The statistical significance of the estimated defier shares point to a violation of global monotonicity (Assumption 2.2), implying that 2SLS may be inconsistent. Furthermore, the difference between complier and defier shares is very small, entailing a weak instrument problem for 2SLS for reasons discussed in Section 2. Indeed, a first stage ordinary least squares (OLS) regression of D on a constant and Z yields a t -value of only 0.778 for the coefficient on Z , so that one would incorrectly conclude that the share of compliers is not statistically different from zero when incorrectly assuming *global* monotonicity. For this reason, our proposed methods do not only come with potential gains in terms of robustness, but also in terms of efficiency. The next three columns provide the estimates of $\Pr(D = 1|Z = 1) - \lambda_1 - (\Pr(D = 0|Z = 0) - \lambda_0)$ for testing the scale constraint, which must necessarily be zero if the identifying assumptions hold (see Section 2). For any estimation approach, the estimates are close to zero and not statistically significant at any conventional level, so that the identifying assumptions for LM or LCD cannot be rejected.

The lower part of Table 2 presents the LATE estimates on the joint groups of defiers and compliers (LATEcd), on the compliers only (LATEc), and on the defiers only (LATEd) as well as the 2SLS estimate. The precision of the last is relatively low, such that zero returns to higher education cannot be rejected. In contrast, the LATEs on compliers, defiers, and the joint groups are highly significant, suggesting that attending college increases weekly wages by roughly 80% to 100%. In conclusion, our results suggest that under the satisfaction of IV independence and LCD, the returns to college are substantial and quite homogeneous across compliers-survivors and defiers-survivors. This suggests that the effects might be rather similar across the ensemble of compliers and defiers.

⁹ This particular choice of bandwidth (i.e., $h = b^{(3/2)}$) is motivated by Theorem 2.1 and is sufficient for the condition $\sqrt{nh^2} \rightarrow 0$ to be satisfied.

Table 2. LATEs of having completed more than 12 years of education on log weekly wage among cohorts born in the second and third quarters of the year in the 1940s ($n = 244,512$).

Level set	Compliers		Defiers		Test	
	$\widehat{\Pr}(T = c)^a$	$\widehat{\Pr}(T = c)^b$	$\widehat{\Pr}(T = \bar{d})^a$	$\widehat{\Pr}(T = \bar{d})^b$	T_{est}^a	T_{est}^b
C90						
Estimator	0.004	0.004	0.004	0.004	−0.001	−0.001
p -value	0.011	0.013	0.014	0.014	0.675	0.704
C95						
Estimator	0.003	0.003	0.003	0.003	−0.001	−0.001
p -value	0.012	0.011	0.026	0.027	0.724	0.699
Level set	LATEcd		LATEc		LATEd	
	$\widehat{\mu}_{c,\bar{d}}^a$	$\widehat{\mu}_{c,\bar{d}}^b$	$\widehat{\mu}_c$	$\widehat{\mu}_c^b$	$\widehat{\mu}_d^a$	$\widehat{\mu}_d^b$
C90						
Estimator	0.907	0.906	0.816	0.816	1.023	0.995
p -value	0.003	0.003	0.010	0.010	0.023	0.015
C95						
Estimator	0.928	0.921	0.880	0.880	0.963	0.969
p -value	0.001	0.001	0.006	0.006	0.018	0.019

Notes: $\widehat{\mu}^a$ denotes the bandwidth-free estimator, $\widehat{\mu}$ denotes the plug-in estimator, while $\widehat{\mu}^b$ is the modified plug-in estimator. A similar notation is used for all the other parameters. C90 and C95 denote the level sets based on 90% and 95% confidence bands, respectively. For the plug-in and the modified plug-in estimators we use a bandwidth, $b^{\frac{2}{3}}$, where b is Silverman's rule-of-thumb bandwidth. Bootstrap p -values are based on 6,000 resamples.

5. CONCLUSION

We demonstrated that local average treatment effects (LATEs) are identified under strictly weaker conditions than the standard assumptions invoked in the literature. Under the assumption of joint independence of the instrument and the potential treatment states/outcomes, the (global) monotonicity of the treatment in the instrument may be weakened to local monotonicity (LM). The latter implies that defiers may exist, but must have marginal potential outcome distributions that do not overlap with those of the compliers. In this case, the LATEs on the compliers, defiers, and the joint population of defiers and compliers are identified. Even if global monotonicity is satisfied, using the proposed LATE estimators might result in efficiency gains compared to the 2SLS estimator. We also showed that LM can be further weakened to permit overlapping potential outcome distributions of compliers and defiers. Under the proposed local compliers-defiers assumption (LCD), compliers that outnumber defiers conditional on specific outcome values in one treatment state are required to have the same joint potential outcomes and size as such compliers in the other treatment state. An analogous argument applies to defiers outnumbering compliers for specific outcome values. Under this empirically more attractive restriction, the proposed approach yields treatment effects on subsets of compliers and defiers.

We applied our methods to U.S. census data to estimate the returns to college for males born in 1940–1949 based on the quarter of birth instrument. We provided evidence for the existence of both compliers and defiers and illustrated that standard LATE estimation is not robust to ignoring defiers. In particular, 2SLS was very imprecisely estimated because of compliers and defiers netting out in the first stage, thereby creating a weak instrument problem. In contrast, the suggested LATE estimators were much more precise and predicted large returns to college for both compliers and defiers.

ACKNOWLEDGEMENTS

We have benefited from comments by Alberto Abadie, Joshua Angrist, Matias Cattaneo, Clément de Chaisemartin, Mario Fiorini, Markus Frölich, Stefan Hoderlein, Guido Imbens, Toru Kitagawa, Frank Kleibergen, Tobias Klein, Michael Lechner, Arthur Lewbel, Enno Mammen, Blaise Melly, Tymon Słoczyński, Katrien Stevens, and participants at several seminars, workshops, and conferences (see the Online Appendix for a complete list). Martin Huber gratefully acknowledges financial support from the Swiss National Science Foundation grant PBSGP1_138770.

REFERENCES

- Aliprantis, D. (2012). Redshirting, compulsory schooling laws, and educational attainment. *Journal of Educational and Behavioral Statistics* 37, 316–38.
- Altonji, J. G. and L. M. Segal (1996). Small-sample bias in GMM estimation of covariance structures. *Journal of Business and Economic Statistics* 14(3), 353–66.
- Anderson, G., O. Linton and Y.-J. Whang (2012). Nonparametric estimation and inference about the overlap of two distributions. *Journal of Econometrics* 171, 1–23.
- Angrist, J. D., G. Imbens and D. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of American Statistical Association* 91, 444–72.

- Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106, 979–1014.
- Angrist, J. D. and A. B. Krueger (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business and Economic Statistics* 13(2), 225–35.
- Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92, 1171–6.
- Barua, R. and K. Lang (2009). School entry, educational attainment, and quarter of birth: A cautionary tale of late. Working paper 15236, NBER, Cambridge, MA.
- Bedard, K. and E. Dhuey (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *Quarterly Journal of Economics* 121, 1437–72.
- Bickel, P. J. (1982). On adaptive estimation. *Annals of Statistics* 10(3), 647–71.
- Bound, J., D. A. Jaeger and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–50.
- Buckles, K. S. and D. M. Hungerman (2013). Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics* 95, 711–24.
- Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In L. Christofides, E. Grant and R. Swidinsky (Eds.), *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, 201–22. Toronto: University of Toronto Press.
- Cheng, J., J. Qin and B. Zhang (2009). Semiparametric estimation and inference for distributional and general treatment effects. *Journal of the Royal Statistical Society, Series B* 71(4), 881–904.
- Cheng, J., D. S. Small, Z. Tan and T. R. Ten Have (2009). Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika* 96(1), 19–36.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* 21(1), C1–C68.
- De Chaisemartin, C. (2012). All you need is late. Ms., Paris School of Economics.
- De Chaisemartin, C. (2017). Tolerating defiance? Identification of treatment effects without monotonicity. *Quantitative Economics* 8, 367–96.
- Ding, P. and J. Lu (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79(3), 757–77.
- Dong, Y., Y.-Y. Lee and M. Gou (2023). Regression discontinuity designs with a continuous treatment. *Journal of the American Statistical Association* 118, 208–21.
- Farbmacher, H., R. Guber and S. Klaassen (2022). Instrument validity tests with causal forests. *Journal of Business and Economic Statistics* 40, 605–14.
- Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics* 139(1), 35–75.
- Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73, 669–738.
- Huber, M. (2014). Sensitivity checks for the local average treatment effect. *Economics Letters* 123, 220–3.
- Huber, M. and G. Mellace (2015). Testing instrument validity for late identification based on inequality moment constraints. *Review of Economics and Statistics* 97, 398–411.
- Imbens, G. W. and J. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica* 62, 467–75.
- Imbens, G. W. and D. B. Rubin (1997a). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies* 64, 555–74.

- Imbens, G. W. and D. B. Rubin (1997b). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* 25(1), 305–27.
- Kitagawa, T. (2009). Identification region of the potential outcome distribution under instrument independence. Working paper 30/09, CeMMAP, London, UK.
- Kitagawa, T. (2015). A test for instrument validity. *Econometrica* 83(5), 2043–63.
- Klein, T. J. (2010). Heterogeneous treatment effects: Instrumental variables without monotonicity? *Journal of Econometrics* 155, 99–116.
- Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies* 76(3), 1071–102.
- Lee, K., B. B. Bhattacharya, J. Qin and D. S. Small (2017). A nonparametric likelihood approach for inference in instrumental variable models. *arXiv: Statistics, Methodology* 1605.03868.
- Li, Q. and J. S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*, Princeton, NJ: Princeton University Press.
- Mammen, E. and W. Polonik (2013). Confidence regions for level sets. *Journal of Multivariate Analysis* 122, 202–14.
- Mourifie, I. and Y. Wan (2017). Testing local average treatment effect assumptions. *Review of Economics and Statistics* 99, 305–13.
- Rosenbaum, P. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.
- Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, 135–46.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 688–701.
- Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Annals of Statistics* 14(3), 1139–51.
- Small, D. S., Z. Tan, R. R. Ramsahai, S. A. Lorch and M. A. Brookhart (2017). Instrumental variable estimation with a stochastic monotonicity assumption. *Statistical Science* 32, 561–79.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

Co-editor Petra Todd handled this manuscript.