It’s never too LATE: A new look at local average treatment effects with or without defiers

by

Christian M. Dahl, Martin Huber and Giovanni Mellace

Discussion Papers on Business and Economics
No. 2/2017

FURTHER INFORMATION
Department of Business and Economics
Faculty of Business and Social Sciences
University of Southern Denmark
Campusvej 55, DK-5230 Odense M
Denmark

E-mail: lho@sam.sdu.dk / http://www.sdu.dk/ivoe

It's 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; **University of Fribourg

February 14, 2017

Abstract: In heterogeneous treatment effect models with endogeneity, identification of the LATE typically relies on the availability of an exogenous instrument monotonically related to treatment participation. We demonstrate that a strictly weaker local monotonicity condition identifies the LATEs on compliers and on defiers. We propose simple estimators that are potentially more efficient than 2SLS, even under circumstances where 2SLS is consistent. Additionally, when easing local monotonicity to local stochastic monotonicity, our identification results still apply to subsets of compliers and defiers. Finally, we provide an empirical application, rejoining the endeavor of estimating returns to education using the quarter of birth instrument.

Keywords: instrumental variable, treatment effects, LATE, local monotonicity.

JEL classification: C14, C21, C26.

Previous versions of this paper have circulated under the title "Relaxing monotonicity in the identification of local average treatment effects". We have benefited from comments by Alberto Abadie, Joshua Angrist, Matias Cattaneo, Mario Fiorini, Markus Frölich, Stefan Hoderlein, Guido Imbens, Toru Kitagawa, Frank Kleibergen, Tobias Klein, Michael Lechner, Arthur Lewbel, Enno Mammen, Blaise Melly, Katrien Stevens, Tymon Słoczyński, 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. Addresses for correspondence: Christian M. Dahl (cmd@sam.sdu.dk), Giovanni Mellace (giome@sam.sdu.dk), and Martin Huber (martin.huber@unifr.ch).

1 Introduction

In heterogeneous treatment effect models with binary treatment, an instrument is conventionally required to satisfy two assumptions. Firstly, it must be independent of the joint distribution of potential treatment states and potential outcomes, which excludes direct effects on the latter and implies that the instrument is (as good as) randomly assigned. Secondly, the treatment state has to vary with the instrument in a weakly monotonic manner. For instance, an instrument based on the random assignment to some treatment state should weakly increase the actual treatment take-up of all individuals in the population (i.e., *globally*). This rules out the existence of defiers, who behave counter-intuitively to the instrument by participating in the treatment if not being assigned to treatment and by not participating in the treatment under treatment assignment.

Under these assumptions, Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) show that the local average treatment effect (LATE) on the subpopulation of compliers (i.e., subjects who with respect to treatment status react to the instrument in the intended way) is identified by the well known Wald ratio, which corresponds to the probability limit of 2SLS estimation. Imbens and Rubin (1997) demonstrate how to identify the potential outcome distributions (including the means) of the compliers under treatment and under non-treatment. Additionally, Imbens and Rubin (1997) show that by imposing data specific constraints in the form of monotonicity and independence, an estimator of the LATE that is more efficient than 2SLS can be obtained.

The first novel contribution of this paper is to show that LATEs are identified (and under particular assumptions $\sqrt{n}$-consistently estimated) conditionally on introducing a new assumption that can be characterized as strictly weaker than *global* monotonicity. We will refer to this condition as *local* monotonicity (LM). Crudely speaking, and in contrast to global monotonicity, LM allows for the existence of both compliers and defiers, but requires that they do not co-exist at any given point on the support of the potential outcomes for any given treatment state. That is, monotonicity is assumed to hold only locally in subregions of the marginal potential outcomes distributions, rather than over the entire support/region. More specifically, if we assume existence of a binary instrument, LM excludes the possibility that a subject is a defier if the difference in specific joint densities is positive, because this is a sufficient condition for the existence of compliers; see, e.g., 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. Conversely, in regions in which the differences in those joint densities are negative, defiers necessarily exist and LM rules out compliers. We show that LM is sufficient for the identification of the marginal potential outcome distributions of the compliers and the defiers in both treatment states.

Because defiers are no longer assumed away under LM we are not limited to only identifying (i) the LATE on the compliers, but can now also identify (ii) the LATE on the defiers as well as (iii) the LATE on the joint population of compliers and defiers. Furthermore, it becomes feasible to estimate the proportion of defiers (and any other subpopulation) in the sample which directly facilitates inference about the relevance of LM and of (ii) and (iii). It will also be shown that (i) and (iii) coincide with the standard LATE under monotonicity and equal the Wald ratio if defiers do not exist. If the proportion of defiers is larger than zero, (i), (ii), and (iii) generally differ, and the standard LATE approach is inconsistent unless the LATEs on compliers and defiers are homogeneous; see Angrist, Imbens, and Rubin (1996). However, even in the case of treatment effect homogeneity across subjects, the standard approach may not be desirable due to a weak instrument type problem that arises when the proportions of compliers and defiers are netting each other out in the first stage. Netting out does not occur in the methods suggested in this paper, implying that efficiency gains can be realized as demonstrated in the empirical application as well as in simulations presented in the online appendix.

Apart from the present work, other studies have considered deviations from monotonicity and their implications for the identification of LATEs. Small and Tan (2007) weaken (individual-level) monotonicity to stochastic monotonicity, requiring that the share of compliers weakly dominates the share of defiers. Small and Tan (2007) show that in this setting, albeit biased, 2SLS retains some desirable limiting properties, such as providing the correct sign of the LATE, yet they do not propose any method to fully identify the LATE. Klein (2010) develops methods to assess the sensitivity of the LATE to random departures from monotonicity and provide guidance on how to approximate the bias under various assumptions. In contrast, our framework admits full identification of the LATE under such non-random violations, given that LM is satisfied.

de Chaisemartin and D'Haultfoeuille (2012) characterize monotonicity by a latent index model, see Vytlacil (2002), in which the conventional rank invariance in the unobserved terms is relaxed to rank similarity, see Chernozhukov and Hansen (2005). Unobservable variables affecting the treatment outcomes may be a function of the instrument, hence admitting the existence of defiers. However, the distribution of these unobservable variables conditional on the potential outcomes must be unaffected by the instrument. In this situation the Wald ratio will identify a treatment effect on a specific mixture of subpopulations. de Chaisemartin (2016) suggests a new assumption which he terms compliers-defiers (CD). CD requires that if defiers are present, then there exists a subpopulation of compliers that has the same size and the same LATE as the defiers. Under CD, de Chaisemartin (2016) shows that the Wald ratio identifies the LATE on the remaining subpopulation of compliers, the so-called complier-survivors or "comvivors". de Chaisemartin (2016) discusses several conditions that imply CD. One sufficient condition enforcing CD is that compliers always outnumber defiers conditional on having the same treatment effect. A second sufficient condition is that the LATEs on defiers and compliers have the same sign and that the ratio of the LATEs is not "too" large. Importantly, CD and LM are not nested conditions and unlike CD, LM admits the identification of LATEs on the entire population of compliers and defiers.

In this paper, we will also reconsider a local stochastic monotonicity (LSM) assumption, which is weaker than LM and has been discussed in de Chaisemartin (2012). Differently from LM, LSM admits the existence of both compliers and defiers conditional on any potential outcome value, but requires that in regions where one of the two types outnumbers the other conditional on one potential outcome, this type would also outnumber the other conditional on both potential outcomes. Under LSM the parameters derived in this paper identify LATEs on subpopulations of compliers and defiers. Further, we show that CD and LSM are not nested. If both assumptions are satisfied, identification results based on LSM yield the LATE on a potentially larger complier subpopulation than those based on CD.

As the second main contribution of this paper, we propose estimators of the LATE that can be characterized asymptotically and are potentially efficient relative to 2SLS, similarly to the results of Imbens and Rubin (1997). Furthermore, the proposed estimators are simple and easily computed in two steps: In the first step, the support of the outcome variable is divided into two disjoint regions on a given treatment state; one where we assume no defiers and one with no compliers. If these regions are unknown and need to be estimated (as it is typically the case in empirical applications), this requires estimating differences in univariate densities, for which kernel methods are very well suited and readily available. Secondly, the LATEs of interest can be estimated based on the sample analogs of the two regions. We propose several estimation approaches in the main text and the online appendix, which all show encouraging finite sample behavior in simulation studies (see the online appendix). Interestingly, our estimators can be more efficient than 2SLS even when the latter consistently estimates a treatment effect. One such example is when global monotonicity holds, but that the aforementioned differences in densities are close to zero and possibly violated over a range of outcome values in the empirical distributions. This observation is in line with the findings of Imbens and Rubin (1997). We therefore argue that the estimators proposed in this paper might be preferred over 2SLS not only because they are more robust to deviations from global monotonicity, but also because their standard errors (and mean squared errors) can be smaller under the standard LATE assumptions.

The third and final contribution of the paper is an empirical application, where the proposed methods are used to estimate the returns to education for males born in 1940-49 (in the 1980 U.S. census data) by means of the quarter of birth as an instrument for education as in Angrist and Krueger (1991). Arguably, among children/students entering school in the same year, those who are born in an earlier quarter can drop out after fewer years of completed education at the age when compulsory schooling ends than those born in a later quarter (in particular after the end of the academic year). This suggests that education is monotonically increasing in the quarter of birth. However, the postponement of school entry due to redshirting or unobserved school policies as discussed in Aliprantis (2012), Barua and Lang (2009), and Klein (2010) may reverse the relation between education and quarter of birth for some individuals and thus violate monotonicity. Relaxing global monotonicity, we find statistically significant proportions of both compliers and defiers and positive returns to education of similar size in both subpopulations.

The remainder of this paper is organized as follows. Section 2 discusses identification. It presents the main assumptions and identification results, and illuminates and explains differences and links among global monotonicity, local monotonicity, local stochastic monotonicity, and the compliers-defiers assumption. Section 3 proposes estimators of the parameters of interest based on kernel density methods, while two further estimation approaches are discussed in the online appendix. Section 4 presents an empirical application, revisiting the challenging task of estimating 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.

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) evaluated at some point in time after the treatment. Under endogeneity, D and Y are confounded by unobserved factors. Treatment 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 the treatment variable D). In this section, we consider the case of a binary instrument ($Z \in \{0, 1\}$), such as a randomized treatment assignment, whereas the online appendix discusses the case of a bounded non-binary 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). Note that in the sample, only one potential outcome is observed for each subject because $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$.

Table 1: Subject types

T	D(1)	D(0)	Subject type
a	1	1	Always taker
c	1	0	Complier
d	0	1	Defier
n	0	0	Never taker

As discussed in Angrist, Imbens, and Rubin (1996) and summarized in Table 1, the population can be categorized into four types, denoted by $T \in \{a, c, d, n\}$, depending on how the treatment state changes with the instrument. The compliers respond to the instrument in the intended way by taking the treatment when $Z = 1$ and abstaining from it when $Z = 0$. For the remaining three types, $D(z) \neq z$ for either $Z = 1$ or $Z = 0$, or both: The always takers are always treated irrespective of the instrument state, the never takers are never treated, and the defiers only take up the treatment when $Z = 0$. Clearly, it is not possible to directly observe the subject type in the sample because $D(1)$ or $D(0)$ remains unknown, as the observed treatment status, D, is decided by $D = Z \cdot D(1) + (1 - Z) \cdot D(0)$. This implies that any subject with a particular combination of treatment and instrument status can belong to two of the types listed in the first column of Table 1. For instance, if the combination $Z = 1, D = 1 \rightarrow D(1) = 1$ is observed for a given subject, this is consistent with the subject belonging to either $T = a$ (the subject being an always taker) or $T = c$ (the subject being a complier) as can be seen from the first two rows of Table 1. Although the subject types are not directly observable, we will show how the potential outcome distributions of the compliers and the defiers can be identified under conditions that are weaker than the common LATE assumptions of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996).

In order to formally characterize the identification problem, we introduce a notation that borrows from Kitagawa (2009) and write the observed joint densities of outcome and treatment status conditional on the instrument as

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

Here, $p_d(y)$ and $q_d(y)$ represent the joint densities of $Y = y$ and $D = d$ given $Z = 1$ and

$Z = 0$, respectively. Furthermore, $\mathcal{Y}$ denotes the support of Y, and $f(y(d))$ denotes the marginal density of the potential outcome for $d \in \{0,1\}$. We define $f(y(d), T = t)$ as the joint density of potential outcome and type for $d \in \{0,1\}$, $t \in \{a,c,d,n\}$, and $y \in \mathcal{Y}$. Importantly, if we exploit that any of the observed joint densities, $p_d(y)$ and $q_d(y)$, depend on the potential outcomes of two different types of subjects, we can rewrite the joint densities as

$$\begin{aligned} p_1(y) &= f(y(1), T = c|Z = 1) + f(y(1), T = a|Z = 1), && (1) \\ q_1(y) &= f(y(1), T = d|Z = 0) + f(y(1), T = a|Z = 0), && (2) \\ p_0(y) &= f(y(0), T = d|Z = 1) + f(y(0), T = n|Z = 1), && (3) \\ q_0(y) &= f(y(0), T = c|Z = 0) + f(y(0), T = n|Z = 0). && (4) \end{aligned}$$

2.2 Assumptions and identification results

The first assumption we impose effectuates the independence between Z and the joint distribution of potential outcomes and treatment status, see Imbens and Angrist (1994).

Assumption 1 (joint independence): Let there exist a random variable Z such that $Z \perp (D(1), D(0), Y(1), Y(0))$, where $\perp$ denotes independence.

Assumption 1 is a commonly used condition in the literature on LATEs, which ensures the existence and randomness of the instrument and implies that the instrument cannot have a direct effect on the potential outcomes. The randomness of the instrument signifies that the instrument is unrelated to any factors potentially affecting the treatment states and/or potential outcomes. Noticably, it follows that not only the potential outcomes, but also the subject types, which are defined by the potential treatment states, are independent of the instrument. Therefore, as also discussed by Kitagawa (2009), equations (1) through

(4) simplify to

$$\begin{aligned} p_1(y) &= f(y(1), T = c) + f(y(1), T = a), && (5)\\ q_1(y) &= f(y(1), T = d) + f(y(1), T = a), && (6)\\ p_0(y) &= f(y(0), T = d) + f(y(0), T = n), && (7)\\ q_0(y) &= f(y(0), T = c) + f(y(0), T = n), && (8)\end{aligned}$$

While Assumption 1 alone does not admit identifying any treatment effects, Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) show that the local average treatment effect on the compliers given by $E(Y(1) - Y(0)|T = c)$ can be obtained by ruling out the defiers. In order to better understand our new identification results, we will provide a short illustrative derivation of the Wald ratio (WR) estimator under the assumption known as (global) monotonicity, where defiers do not exist. In short, this assumption writes:

Global monotonicity: Order 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 in addition to Assumption 1 implies that defiers cannot exist in the population and (5) through (8) simplify readily to

$$\begin{aligned} p_1(y) &= f(y(1), T = c) + f(y(1), T = a), && (9)\\ q_1(y) &= f(y(1), T = a), && (10)\\ p_0(y) &= f(y(0), T = n), && (11)\\ q_0(y) &= f(y(0), T = c) + f(y(0), T = n). && (12)\end{aligned}$$

The identification of the joint densities under treatment and non-treatment for the compliers can be verifyed by first subtracting (10) from (9) and (11) from (12):

$$\begin{aligned} f(y(1), T = c) &= p_1(y) - q_1(y), && (13) \\ f(y(0), T = c) &= q_0(y) - p_0(y). && (14) \end{aligned}$$

Then secondly, by integrating out y in both (13) and (14), the share of compliers in the population is given as:

$$\begin{aligned} \Pr(T = c) &= \int_{y\in\mathcal{Y}} (p_1(y) - q_1(y))\, dy = E(D|Z = 1) - E(D|Z = 0), && (15) \\ \Pr(T = c) &= \int_{y\in\mathcal{Y}} (q_0(y) - p_0(y))\, dy = E(1 - D|Z = 0) - E(1 - D|Z = 1). && (16) \end{aligned}$$

By further noticing that $\int_{y\in\mathcal{Y}} y \cdot p_d(y)dy = E(y, D = d|Z = 1)$ and $\int_{y\in\mathcal{Y}} y \cdot q_d(y)dy = E(y, D = d|Z = 0)$, we can write

$$\begin{aligned} E(Y(1)|T = c) &= \int_{y\in\mathcal{Y}} y \cdot f(y(1)|T = c)dy \\ &= \int_{y\in\mathcal{Y}} y \cdot \frac{f(y(1), T = c)}{\Pr(T = c)} dy \\ &= \int_{y\in\mathcal{Y}} \frac{y \cdot (p_1(y) - q_1(y))}{E(D|Z = 1) - E(D|Z = 0)} dy \\ &= \frac{E\left(y, D = 1|Z = 1\right) - E\left(y, D = 1|Z = 0\right)}{E(D|Z = 1) - E(D|Z = 0)}, && (17) \end{aligned}$$

where the second equality follows from basic probability theory and the third from the imposed assumptions. Similarly,

$$E(Y(0)|T = c) = \frac{E\left(y, D = 0|Z = 1\right) - E\left(y, D = 0|Z = 0\right)}{E(D|Z = 1) - E(D|Z = 0)}. \qquad (18)$$

Since $E\left(Y|Z = z\right) = E\left(Y, D = 0|Z = z\right) + E\left(Y, D = 1|Z = z\right)$, the result of Imbens and

Angrist (1994) showing that the LATE corresponds to the WR, follows imidiately from subtracting (18) from (17), that is

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

The derivation 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 non-treated observations, respectively. Furthermore, (13) and (14) provide necessary (albeit not sufficient) conditions for the satisfaction of global monotonicity and for Assumption 1.[1] Furthermore, as $f(y(1),T=c)$ and $f(y(0),T=c)$ cannot be negative for any $y\in\mathcal{Y}$, it follows directly from equations (13) and (14) that

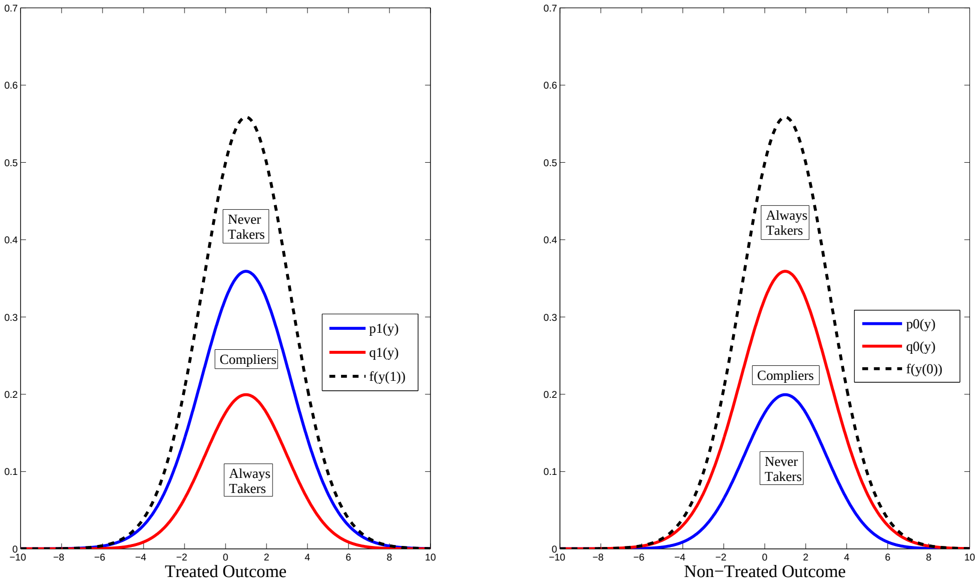
$$p_1(y)-q_1(y)\geq 0,\quad q_0(y)-p_0(y)\geq 0. \tag{19}$$

Imbens and Rubin (1997) propose an estimator that imposes (19) in an attempt to improve efficiency, while Kitagawa (2015), Huber and Mellace (2015), and Mourifie and Wan (2016) provide formal tests of these constraints. Figure 1 presents a graphical illustration of the identification under global monotonicity.[2] In Figure 1, Equation (19) is satisfied for all $y\in\mathcal{Y}$ implying that all the weights in the expression for the WR are non-negative.

[1] This feature has also been discussed by Balke and Pearl (1997) and Heckman and Vytlacil (2005).

[2] The illustration is similar to Figure 1 in Kitagawa (2015).

Figure 1: Graphical illustration of identification under global monotonicity



Under a violation of (19) and,therefore, also a violation of global monotonicity when Assumption 1 is maintained, basing the LATE estimation on the WR appears unattractive both in terms of consistency and efficiency. First, Angrist, Imbens, and Rubin (1996) show that the WR does not generally yield a treatment effect because the WR in this case is equivalent to

$$\text{WR} = \frac{E(Y(1)-Y(0)|T=c)\cdot \Pr(T=c) - E(Y(1)-Y(0)|T=d)\cdot \Pr(T=d)}{\Pr(T=c)-\Pr(T=d)}. \tag{20}$$

Therefore, the LATE on the compliers is identified only if it is equal to the LATE on the defiers. Second, even in this special case, the WR assigns negative weights to treated (non-treated) observations whenever $p_1(y) < q_1(y)$ ($q_0(y) < p_0(y)$). It is easy to see from (15) and (16) that negative weights decrease the terms $E(D|Z=1)-E(D|Z=0)$ and $E(1-D|Z=0)-E(1-D|Z=1)$, which reduces the efficiency of LATE estimation, even in large samples.

We will now proceed by replacing the assumption of global monotonicy by Assumption 2, which we will denote local monotonicity (LM). Importantly, LM is weaker than global monotonicity and admits a violation of (19).

Assumption 2 (local monotonicity, LM): For all subjects in the population, either $\Pr(D(1) \geq D(0)|Y(d) = y(d)) = 1$, or $\Pr(D(0) \geq D(1)|Y(d) = y(d)) = 1 \ \forall \ y(d) \in \mathcal{Y}$, where $d \in \{0, 1\}$.

Assumption 2 (LM) is novel in the sense that it allows the presence of both compliers and defiers in the population. LM, however, restricts their co-existence on a local scale. More precisely, LM requires the potential outcome distributions of compliers and defiers under each treatment state to be non-overlapping. Consequently, under LM, compliers and defiers inhabit disjoint regions of the support of $Y(1)$ and $Y(0)$, respectively.[3,4] More formally, note that under Assumption 1, Equations (5) through (8) imply

$$\begin{aligned} p_1(y) - q_1(y) &= f(y(1), T = c) - f(y(1), T = d), \\ q_0(y) - p_0(y) &= f(y(0), T = c) - f(y(0), T = d), \end{aligned} \tag{21}$$

while adding Assumption 2 implies

$$\begin{aligned} p_1(y) > q_1(y) &\Rightarrow f(y(1), T = c) > f(y(1), T = d) \Rightarrow f(y(1), T = d) = 0 \quad \textbf{(no defiers)}, \\ p_1(y) < q_1(y) &\Rightarrow f(y(1), T = c) < f(y(1), T = d) \Rightarrow f(y(1), T = c) = 0 \quad \textbf{(no compliers)}, \\ q_0(y) > p_0(y) &\Rightarrow f(y(0), T = c) > f(y(0), T = d) \Rightarrow f(y(0), T = d) = 0 \quad \textbf{(no defiers)}, \\ q_0(y) < p_0(y) &\Rightarrow f(y(0), T = c) < f(y(0), T = d) \Rightarrow f(y(0), T = c) = 0 \quad \textbf{(no compliers)}. \end{aligned}$$

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

[4] The online appendix presents two examples of structural models in which Assumptions 1 and 2 hold, while global monotonicity does not.

This signifies that in regions of $\mathcal{Y}$ where (19) is satisfied, implying $f(y(d), T = c) > f(y(d), T = d)$, defiers are ruled out by Assumption 2. Similarly, a violation of (19), implying $f(y(d), T = d) > f(y(d), T = c)$, rules out compliers. Summarizing these observation, we can conveniently write

$$\begin{aligned}
f(y(1), T = c) &= (p_1(y) - q_1(y))I(p_1(y) > q_1(y)) = p_1(y) - \min(p_1(y), q_1(y)), && (22)\\
f(y(0), T = c) &= (q_0(y) - p_0(y))I((q_0(y) > p_0(y))) = q_0(y) - \min(p_0(y), q_0(y)), && (23)\\
f(y(1), T = d) &= (q_1(y) - p_1(y))I(p_1(y) < q_1(y)) = q_1(y) - \min(p_1(y), q_1(y)), && (24)\\
f(y(0), T = d) &= (p_0(y) - q_0(y))I((q_0(y) < p_0(y))) = p_0(y) - \min(p_0(y), q_0(y)). && (25)
\end{aligned}$$

Hence, the densities of potential outcomes under both treatment and non-treatment are identified for compliers as well as defiers. Also their shares in the population are identified, i.e.,

$$\begin{aligned}
\Pr(T = c) &= \int_{y\in\mathcal{Y}} (p_1(y) - \min(p_1(y), q_1(y))dy = \Pr(D = 1|Z = 1) - \lambda_1, && (26)\\
\Pr(T = c) &= \int_{y\in\mathcal{Y}} (q_0(y) - \min(p_0(y), q_0(y))dy = \Pr(D = 0|Z = 0) - \lambda_0, && (27)\\
\Pr(T = d) &= \int_{y\in\mathcal{Y}} (q_1(y) - \min(p_1(y), q_1(y))dy = \Pr(D = 1|Z = 0) - \lambda_1, && (28)\\
\Pr(T = d) &= \int_{y\in\mathcal{Y}} (p_0(y) - \min(p_0(y), q_0(y))dy = \Pr(D = 0|Z = 1) - \lambda_0, && (29)
\end{aligned}$$

where $\lambda_i = \int_{\mathcal{Y}} \min(p_i(y), q_i(y))dy$ for $i = 0, 1$. These results admits identification of not only the LATE on the compliers, but also of the LATE on the defiers and on the joint population of compliers and defiers. These identification results are summarized accurately in the following Proposition 1:

Proposition 1 (identification of the LATEs): Let the conditions under Assumptions 1 and 2 hold. Then:

1. The LATE on compliers is given as

$$E(Y(1)-Y(0)|T=c) = \frac{\int_{\mathcal{Y}} y\cdot(p_1(y)-\min(p_1(y),q_1(y)))dy}{\Pr(D=1|Z=1)-\lambda_1} - \frac{\int_{\mathcal{Y}} y\cdot(q_0(y)-\min(p_0(y),q_0(y)))dy}{\Pr(D=0|Z=0)-\lambda_0}. \tag{30}$$

2. The LATE on defiers is given as

$$E(Y(1)-Y(0)|T=d) = \frac{\int_{\mathcal{Y}} y\cdot(q_1(y)-\min(p_1(y),q_1(y)))dy}{\Pr(D=1|Z=0)-\lambda_1} - \frac{\int_{\mathcal{Y}} y\cdot(p_0(y)-\min(p_0(y),q_0(y)))dy}{\Pr(D=0|Z=1)-\lambda_0}. \tag{31}$$

3. The joint LATE on compliers and defiers is given as

$$E(Y(1)-Y(0)|T=c,d) = \frac{\int_{\mathcal{Y}} y\cdot(\max(p_1(y),q_1(y))-\min(p_1(y),q_1(y)))dy}{\Pr(D=1|Z=1)+\Pr(D=1|Z=0)-2\cdot\lambda_1} - \frac{\int_{\mathcal{Y}} y\cdot(\max(p_0(y),q_0(y))-\min(p_0(y),q_0(y)))dy}{\Pr(D=0|Z=0)+\Pr(D=0|Z=1)-2\cdot\lambda_0}. \tag{32}$$

4. If $\Pr(T=d)=0$ and $\Pr(T=c)>0$, then (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 $E(Y(1)-Y(0)|T=d)$ is not identified.

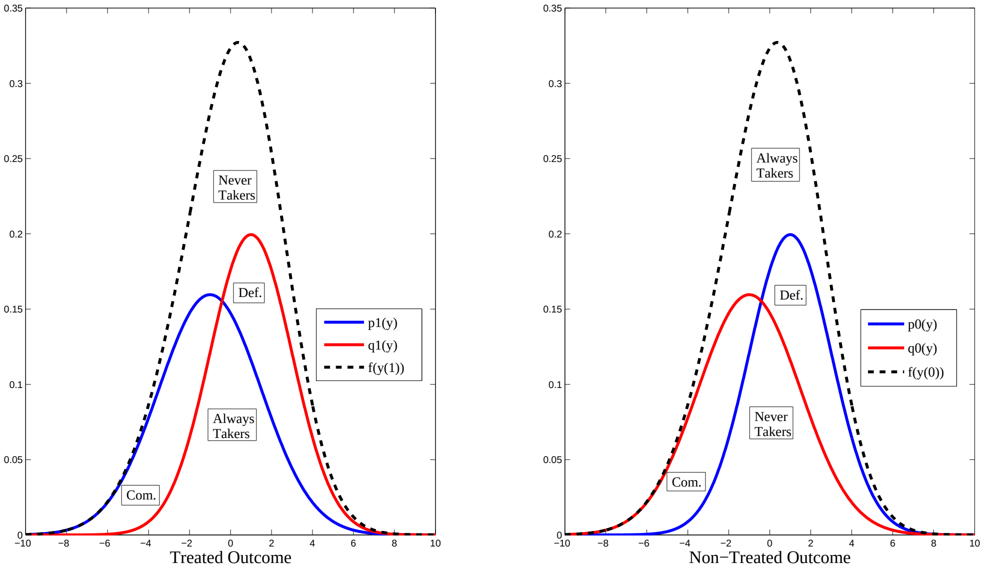
5. If $\Pr(T=c)=0$ and $\Pr(T=d)>0$, then (32) is equivalent to $E(Y(1)-Y(0)|T=d)=\frac{E(Y|Z=0)-E(Y|Z=1)}{E(D|Z=0)-E(D|Z=1)}$, whereas $E(Y(1)-Y(0)|T=c)$ is not identified.

Proof of Proposition 1: Results 1, 2, and 3 of Proposition 1 follow from using (22) through (25) and (26) through (29) in $E(Y(d)|T=t)=\frac{\int_{\mathcal{Y}} f(y(d),T=t)dy}{\Pr(T=t)}$ and taking the differences in mean potential outcomes under treatment and non-treatment. Result 4 follows from the fact that $\Pr(T = d) = 0$ (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 (28) and (29)), such that (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 (22) and (23)), such that (32) simplifies accordingly. □

Note that, different from the WR, the weights of the parameters defined in Proposition 1 cannot be negative. For example, consider Result 1, where the weights are given by $\frac{p_1(y)-\min(p_1(y),q_1(y))}{\int_{\mathcal{Y}} p_1(y)-\min(p_1(y),q_1(y))dy}$, and $\frac{q_0(y)-\min(p_0(y),q_0(y))}{\int_{\mathcal{Y}} q_0(y)-\min(p_0(y),q_0(y))dy}$, and are thus non-negative. This is a potential advantage not only when the WR fails to identify the LATE on the compliers, but also in at least two additional scenarios that we will briefly discuss. In the first scenario, assume that Assumptions 1 and 2 hold, that global monotonicity fails but that the LATEs on the compliers and the defiers are equal. If in this case $\Pr(T = c) > \Pr(T = d)$, then 2SLS consistently estimates the WR given by (20). The estimator, however, may suffer from severe weak instrument issues in finite samples particularly when the shares of compliers and defiers are not too different and are netting out (making the denominator of (20) very small). In the limiting case, when $\Pr(T = c) = \Pr(T = d)$, the WR does not exist and the consistency of 2SLS therefore no longer applies. In contrast, the LATEs given by Proposition 1 remains well defined even in the limiting case when $\Pr(T = c) = \Pr(T = d)$ facilitating the construction of more powerful estimators. In the second scenario, 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 (19). In this scenario, the estimators based on the sample analog of the LATEs in Proposition 1 provide substantial efficiency gains compared to 2SLS, as also noted by Imbens and Rubin (1997). A simulation study described in the online appendix provide supportive evidence in favor of this statement.

Figure 2 is a graphical illustration of the identification results under the conditions of Assumptions 1 and 2. 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 as the area between the two curves, $p_1(y)$ and $q_1(y)$, on the parts of $\mathcal{Y}$ where $p_1(y) > q_1(y)$. Similarly, the density of the compliers' potential outcome under non-treatment 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 $p_0(y)$, on the parts of $\mathcal{Y}$ where $q_0(y) > p_0(y)$ yields the share of compliers. Symmetrically, the density of the defiers' potential outcomes under treatment and non-treatment 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 $q_0(y) < p_0(y)$.

Figure 2: Graphical illustration of the identification of LATEs under the conditions of Assumptions 1 (instrument) and 2 (local monotonicity)



As pointed out by Kitagawa (2009), Assumptions 1 and 2 are actually testable by a measure referred to as the scale constraint. Accordingly, the share of any type of subjects in the population must be equal across treatment states. This condition writes

$$\int_{\mathcal{Y}} f(y(1), T = t)dy = \int_{\mathcal{Y}} f(y(0), T = t)dy = \Pr(T = t) \quad \forall t = \{a, c, d, n\}. \tag{33}$$

For the compliers, for example, the scale contraint implies that $\Pr(D = 1|Z = 1) - \lambda_1 = \Pr(D = 0|Z = 0) - \lambda_0$; see (26) and (27). 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 (33) to hold for all types of subjects in the population. As an additional check of the plausibility of LM, we suggest plotting 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. For example, one might wish to compare the distributions of observable covariates, denoted by X, across compliers and defiers to infer on their socio-economic differences. In fact, if Assumptions 1 and 2 are invoked in the presence of X, it is easy to show that for any x in the support of the covariate space,

$$\begin{aligned}
f(X = x, D = 1|Z = 1, p_1 > q_1) &- f(X = x, D = 1|Z = 0, p_1 > q_1) = f(X = x, T = c, D = 1),\\
f(X = x, D = 1|Z = 0, p_1 < q_1) &- f(X = x, D = 1|Z = 1, p_1 < q_1) = f(X = x, T = d, D = 1),\\
f(X = x, D = 0|Z = 0, p_0 < q_0) &- f(X = x, D = 0|Z = 1, p_0 < q_0) = f(X = x, T = c, D = 0),\\
f(X = x, D = 0|Z = 1, p_0 > q_0) &- f(X = x, D = 0|Z = 0, p_0 > q_0) = f(X = x, T = d, D = 0).
\end{aligned}$$

Furthermore, by $\Pr(T=t|D=d)=\Pr(T=t)$ under Assumption 1 it follows that

$$\begin{aligned}
f(X=x|T=c,D=1) &= \frac{f(X=x,T=c,D=1)}{\Pr(T=c,D=1)} = \frac{f(X=x,T=c,D=1)}{\Pr(T=c)\Pr(D=1)},\\
f(X=x|T=d,D=1) &= \frac{f(X=x,T=d,D=1)}{\Pr(T=d,D=1)} = \frac{f(X=x,T=d,D=1)}{\Pr(T=d)\Pr(D=1)},\\
f(X=x|T=c,D=0) &= \frac{f(X=x,T=c,D=0)}{\Pr(T=c,D=0)} = \frac{f(X=x,T=c,D=0)}{\Pr(T=c)\Pr(D=0)},\\
f(X=x|T=d,D=0) &= \frac{f(X=X,T=d,D=0)}{\Pr(T=d,D=0)} = \frac{f(X=x,T=d,D=0)}{\Pr(T=d)\Pr(D=0)}.
\end{aligned}$$

This permits us to contrast compliers and defiers in terms of observed characteristics and to verify the plausibility of the Assumptions 1 and 2 within each type of subjects. In fact, if X are covariates that cannot be affected by the treatment, implying that X itself rather than its potential values are independent of Z, then $f(X=x|T=t,D=1)=f(X=x|T=t,D=0)$, which for testing purposes is an easy operational hypothesis measure.

2.3 Alternatives to local monotonicity

Our discussion has shown that if Assumption 1 holds, the identification of LATE does not necessarily rely on global monotonicity. The LATEs introduced by Proposition 1 are equivalent to the WR if global monotonicity holds, but can also be identified under the weaker Assumption 2, which as shown is partially testable by the scale constraint. Moreover, if Assumption 2 does not hold, neither does global monotonicity, and in this case there appears to be no gains in assuming global monotonicity rather than local monotonicity. However, albeit more general, also LM may appear restrictive in some applications, in particular when outcomes have limited support. For binary outcomes, for example, the conditions of Assumption 2 imply that the potential outcomes of all compliers given a particular treatment state are either zero or unity, while all defier outcomes take on the exact opposite value. For this purpose it appears instructive to compare the identification under LM to an alternative relaxation of monotonicity offered in de Chaisemartin (2016), the so-called compliers-defiers (CD) assumption, which admits identification of the LATE on a subset of compliers:

Compliers-defiers (CD): There exists a subpopulation of compliers c^d, such that $\Pr(T = c^d) = \Pr(T = d)$ and $E(Y(1) - Y(0)|T = c^d) = E(Y(1) - Y(0)|T = d)$.

The CD assumption states that if defiers are present, a subset of compliers of the same relative size with identical LATE exists. In this case the WR identifies the LATE on the remaining subset of compliers, which are subjects that do not necessarily resemble subjects that are defiers. These compliers are the so-called compliers-survivors or "comvivors", denoted by c^s. By splitting the compliers into compliers-defiers and compliers-survivors in (20) we obtain

$$\begin{aligned}
\text{WR} &= \frac{E(Y(1) - Y(0)|T = c^s) \cdot \Pr(T = c^s)}{\Pr(T = c^s) + \Pr(T = c^d) - \Pr(T = d)} + \frac{E(Y(1) - Y(0)|T = c^d) \cdot \Pr(T = c^d)}{\Pr(T = c^s) + \Pr(T = c^d) - \Pr(T = d)} \\
&- \frac{E(Y(1) - Y(0)|T = d) \cdot \Pr(T = d)}{\Pr(T = c^s) + \Pr(T = c^d) - \Pr(T = d)} \\
&= \frac{E(Y(1) - Y(0)|T = c^s) \cdot \Pr(T = c^s)}{\Pr(T = c^s)} \\
&= E(Y(1) - Y(0)|T = c^s). \qquad (34)
\end{aligned}$$

We briefly discuss what CD and LM imply under violations of constraints (19) in order to see that the two assumptions indeed are not nested. To this end, assume that $\Pr(T = c) > \Pr(T = d)$ and separate the support of the outcome into the following level sets, depending on whether (19) is violated or not conditional on the treatment:

$$\begin{aligned}
C_{q_1} &= \{y \in \mathcal{Y} : p_1(y) > q_1(y)\}, \quad C_{p_0} = \{y \in \mathcal{Y} : q_0(y) > p_0(y)\} \\
C_{p_1} &= \{y \in \mathcal{Y} : q_1(y) > p_1(y)\}, \quad C_{q_0} = \{y \in \mathcal{Y} : p_0(y) > q_0(y)\}. \qquad (35)
\end{aligned}$$

Furthermore, let c^+ and d^+ denote the compliers and the defiers, respectively, located in either C_{q_1} or C_{p_0} (i.e., in areas satisfying the constraints), and let c^- and d^- denote those located in either C_{p_1} or C_{q_0}, where (19) is violated. It is easy to show that (20) now corresponds to

$$\begin{aligned}\text{WR} &= \frac{E(Y(1)-Y(0)|T=c^+)\cdot\Pr(T=c^+)-E(Y(1)-Y(0)|T=d^+)\cdot\Pr(T=d^+)}{(\Pr(T=c^+)-\Pr(T=d^+))-(\Pr(T=d^-)-\Pr(T=c^-))}\\ &+ \frac{E(Y(1)-Y(0)|T=c^-)\cdot\Pr(T=c^-)-E(Y(1)-Y(0)|T=d^-)\cdot\Pr(T=d^-)}{(\Pr(T=c^+)-\Pr(T=d^+))-(\Pr(T=d^-)-\Pr(T=c^-))} \qquad (36)\end{aligned}$$

Note that as defiers outnumber compliers in the violation areas, then $\Pr(T=d^-)-\Pr(T=c^-)>0$. If CD holds, the share of comvivors corresponds to the denominator in (36) and $\Pr(c^s)=(\Pr(T=c^+)-\Pr(T=d^+))-(\Pr(T=d^-)-\Pr(T=c^-))$. Furthermore, by inspection of (34) and (36) it becomes evident that under CD, a weighted average of LATEs on subsets of c^+ and c^-, whose joint shares equal $\Pr(T=d^+)+\Pr(T=d^-)$ (i.e., those sets of c^+ and c^- not pertaining to c^s), corresponds to a weighted average of LATEs on d^+ and d^-. The weights depend on the relative shares of the various (subsets of) subject types. One can therefore construct cases in which CD holds if (19) is violated. However, the plausibility of CD arguably decreases in the range of the support of C_{p_1} and C_{q_0} and in the share of d^-. In contrast to CD, LM assumes $\Pr(d^+)=\Pr(c^-)=0$. As this is neither necessary nor sufficient for CD, this shows that the two assumptions are not nested.

Even in the case where the conditions of Assumption 2 fail to hold, such that $\Pr(d^+)>0$ and/or $\Pr(c^-)>0$, the expressions of Proposition 1 may (similarly to the WR under CD) still identify treatment effects on subsets of compliers and defiers. de Chaisemartin (2012) shows that this is the case if LM is replaced by a weaker local *stochastic* monotonicity (LSM) assumption, which appears plausible in many empirical contexts:[5]

[5] We have also considered a local version of CD. However, this assumption turns out to be equivalent to LSM if C_{p_1} and C_{q_0} are non-empty and to CD if these sets are empty. More details ara available from the authors upon request.

Assumption 3 (local stochastic monotonicity, LSM): Let $y(d)$ be in $\mathcal{Y}$, for $d = 0,1$. Then the condition $\Pr(T = c|Y(d) = y(d)) \geq \Pr(T = d|Y(d) = y(d))$ implies that $\Pr(T = c|Y(1) = y(1), Y(0) = y(0)) \geq \Pr(T = d|Y(1) = y(1), Y(0) = y(0))$. Similarly, the condition $\Pr(T = c|Y(d) = y(d)) \leq \Pr(T = d|Y(d) = y(d))$ implies that $\Pr(T = c|Y(1) = y(1), Y(0) = y(0)) \leq \Pr(T = d|Y(1) = y(1), Y(0) = y(0))$.

Assumption 3 admits the existence of both compliers and defiers at any given value of the marginal potential outcome distribution. However, LSM requires that if the share of one type of subjects weakly dominates the share of the other subject conditional on either $Y(1)$ or $Y(0)$, it must also dominate conditional on both potential outcomes jointly. Under Assumption 1 alone, the data reveal such a dominance conditional on one of the two potential outcomes: $p_1(y) \geq q_1(y)$ implies that $\Pr(T = c|Y(1) = y) \geq \Pr(T = d|Y(1) = y)$, and similarly $p_1(y) \leq q_1(y)$ implies that $\Pr(T = c|Y(1) = y) \leq \Pr(T = d|Y(1) = y)$. Moreover, it follows from $q_0(y) \geq p_0(y)$ that $\Pr(T = c|Y(0) = y) \geq \Pr(T = d|Y(0) = y)$, and from $q_0(y) \leq p_0(y)$ that $\Pr(T = c|Y(0) = y) \leq \Pr(T = d|Y(0) = y)$. When enforcing Assumption 4, de Chaisemartin (2012) shows that the identification results of Proposition 1 apply to a subset of compliers outnumbering the defiers whenever $\Pr(T = c|Y(1) = y(1), Y(0) = y(0)) \geq \Pr(T = d|Y(1) = y(1), Y(0) = y(0))$, and similarly to a subset of defiers outnumbering the compliers whenever $\Pr(T = c|Y(1) = y(1), Y(0) = y(0)) \leq \Pr(T = d|Y(1) = y(1), Y(0) = y(0))$. Under Assumptions 1 and 3, Result 1 of Proposition 1 can be shown to correspond to

$$\begin{aligned}
& \frac{\int_{y \in C_{q_1}} y \cdot (p_1(y) - \min(p_1(y), q_1(y)))dy}{\Pr(D = 1|Z = 1) - \lambda_1} - \frac{\int_{y \in C_{p_0}} y \cdot (q_0(y) - \min(p_0(y), q_0(y)))dy}{\Pr(D = 0|Z = 0) - \lambda_0} \\
= \; & \frac{E(Y(1) - Y(0)|T = c^+) \cdot \Pr(T = c^+) - E(Y(1) - Y(0)|T = d^+) \cdot \Pr(T = d^+)}{\Pr(T = c^+) - \Pr(T = d^+)} \\
= \; & E(Y(1) - Y(0)|T = c^{s*}),
\end{aligned}$$

where

$$\Pr(D=1|Z=1)-\lambda_1=\Pr(D=0|Z=0)-\lambda_0=\Pr(T=c^{+})-\Pr(T=d^{+})=\Pr(T=c^{s*}).$$

Here c^{s*} denotes the "local" comvivors in C_{q_1} and C_{p_0}. Note that $\Pr(T=c^{s*})$ is greater than or equal to the share of comvivors under CD given by $\Pr(c^s)=\Pr(T=c^{+})-\Pr(T=d^{+})-(\Pr(T=d^{-})-\Pr(T=c^{-}))$. Since $\Pr(T=d^{-})-\Pr(T=c^{-})\geq 0$, if both LSM and CD hold, LSM admits identifying the LATE on a larger share of compliers than the latter when $\Pr(T=d^{-})-\Pr(T=c^{-})>0$ (i.e., C_{p_1} and C_{q_0} are non-empty). This may lead to important finite sample efficiency gains using estimators of the LATE given by Result 1 of Proposition 1 and potentially to higher external validity. Analogous results hold for Result 2 of Proposition 1 and thus also for the joint population of local comvivors and local defiers-survivors considered in de Chaisemartin (2012).

Finally, it is worth mentioning that Assumption 3 is a local version of stochastic monotonicity, i.e., $\Pr(T=c|Y(1),Y(0))\geq\Pr(T=d|Y(1),Y(0))$, see, e.g., Small and Tan (2007), which is stronger than and sufficient for CD, see the discussion in de Chaisemartin (2016). In contrast, Assumption 3 is neither sufficient nor necessary for CD. Recall that the latter holds if there exists some subset in c^{+} and c^{-} whose share equals $\Pr(T=d^{+})+\Pr(T=d^{-})$ and whose LATE equals the joint LATE on d^{+} and d^{-}. On the other hand, LSM implies that there exists a subset in c^{+} whose share equals $\Pr(T=d^{+})$ and whose LATE equals the LATE on d^{+} (and an analogous restriction for d^{-} and c^{-}, respectively). Similarly as for LM, a testable implication for Assumption 1 and LSM is that $\Pr(D=1|Z=1)-\lambda_1=\Pr(D=0|Z=0)-\lambda_0$.

3 Estimation

Estimation of the LATEs presented in Proposition 1 will be based on the sample analogy principle. For that purpose, rewrite Results 1-3 of Proposition 1 as

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

where the level sets $C_{q_1}, C_{p_0}, C_{p_1}$, and C_{q_0} are defined by (35). By sample analogy, the estimators of interest can then be obtained as

$$
\begin{aligned}
\widehat{\mu}_c &= \frac{\widehat{\theta}_1}{\widehat{P}_{1|1}-\widehat{\lambda}_1} - \frac{\widehat{\theta}_0}{\widehat{P}_{0|0}-\widehat{\lambda}_0}, && (37) \\
\widehat{\mu}_d &= \frac{\widehat{\theta}_2}{\widehat{P}_{1|0}-\widehat{\lambda}_1} - \frac{\widehat{\theta}_3}{\widehat{P}_{0|1}-\widehat{\lambda}_0}, && (38) \\
\widehat{\mu}_{c,d} &= \frac{\widehat{\theta}_1+\widehat{\theta}_2}{\widehat{P}_{1|1}+\widehat{P}_{1|0}-2\widehat{\lambda}_1} - \frac{\widehat{\theta}_0+\widehat{\theta}_3}{\widehat{P}_{0|0}+\widehat{P}_{0|1}-2\widehat{\lambda}_0}, && (39)
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{\theta}_0 &= \int_{C_{p_0}} y\left(\widehat{q}_0(y)-\widehat{p}_0\left(y\right)\right)dy, \quad \widehat{\theta}_1 = \int_{C_{q_1}} y\left(\widehat{p}_1(y)-\widehat{q}_1\left(y\right)\right)dy, \\
\widehat{\theta}_2 &= \int_{C_{p_1}} y\left(\widehat{q}_1(y)-\widehat{p}_1\left(y\right)\right)dy, \quad \widehat{\theta}_3 = \int_{C_{q_0}} y\left(\widehat{p}_0(y)-\widehat{q}_0\left(y\right)\right)dy, \qquad (40)
\end{aligned}
$$

and

$$\widehat{\lambda}_d = \int_{\mathcal{Y}} \min\left(\widehat{p}_d\left(y\right), \widehat{q}_d\left(y\right)\right) dy, \quad d = 0, 1, \quad \widehat{P}_{d|z} = \frac{1}{n}\sum_{i=1}^{n} \frac{I\left(D_i = d\right) I\left(Z_i = z\right)}{\frac{1}{n}\sum_{i=1}^{n} I\left(Z_i = z\right)}, \quad d, z = 0, 1.$$

Let $I\left(\cdot\right)$ denote the indicator function, which equals one if its argument holds true and is zero otherwise. Standard kernel based methods are used to obtain estimators of the relevant densities, i.e.,

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

for

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

Here, L and W are product kernels, see Li and Racine (2007), pp. 164-165, defined as

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

where h denotes the bandwidth. The Gaussian kernel is used throughout the simulations (presented in the online appendix) and in the empirical application (presented in Section 4). Integrals can be computed numerically using any of the many approximation methods available. For the benchmark estimators, we use the trapezoid rule. We refer to these estimators as “plug-in” estimators. Moreover, in the online appendix we present a set of estimators that are based on a computationally more convenient approximation of integrals of density functions. This approximation imposes the restriction that a proper density must integrate to unity. We will refer to estimators based on this type of approximation as “modified plug-in” estimators.

To derive the asymptotic properties of the estimators for known level sets we introduce the following assumptions on the kernel estimators.

Assumption 4 (kernel estimation): (**a**) The general nonnegative 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 $B(y, D = d, Z = z) = \frac{1}{2}\kappa_2 \frac{\nabla^2\left(f_{Y,D|Z}(y,D=d|Z=z)\right)}{(\nabla y)^2}\Big/ f_Z(Z = z)$; (**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 $i = 1,2,...,n$ is an *i.i.d.* distributed 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 $E\left(|Y|^2\right)$, $E_{p_d}\left(|Y|^2\right)$ and $E_{q_d}\left(|Y|^2\right)$ exist for $d = 0,1$.

Assumption 4 is standard for kernel-based estimation methods. Assumption 4(a) implies that the estimated density is well defined. Assumption 4(b) imposes twice continuous differentiability of p_d and q_d. It is worth noticing that Assumption 4(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 4(d) is written in general terms 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.

We can now establish the following asymptotic properties of the estimators of the LATEs given by Proposition 1:

Theorem 1 (asymptotics): Let the conditions of Assumption 4 hold and let the level sets $C_{p_0}, C_{q_0}, C_{p_1}$, and C_{q_1} be known. Then

$$\begin{aligned}
\sqrt{n}\left(\widehat{\mu}_{c,d} - \mu_{c,d} - b_{c,d}\right) &\xrightarrow{d} \mathcal{N}\left(0, \Omega_{c,d}\right), \\
\sqrt{n}\left(\widehat{\mu}_{c} - \mu_{c} - b_{c}\right) &\xrightarrow{d} \mathcal{N}\left(0, \Omega_{c}\right), \\
\sqrt{n}\left(\widehat{\mu}_{d} - \mu_{d} - b_{d}\right) &\xrightarrow{d} \mathcal{N}\left(0, \Omega_{d}\right),
\end{aligned}$$

for $n \to \infty$, $h \to 0$, and $\sqrt{n}h^2 \to 0$, where $b_{c,d}, b_c$, and b_d denote finite sample bias terms that vanish asymptotically. Detailed expressions for $\Omega_{c,d}, \Omega_c$, and Ω_d are provided in the online appendix.

Proof of Theorem 1 See the online appendix. □

Theorem 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 (37),(38), and (39) 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, we have run an extensive set of Monte Carlo simulations in which we investigate properties such as one-sided coverage probabilities, bias, and efficiency. The simulation results are very encouraging and suggest that the asymptotic distribution of the LATE estimators satisfactorily approximates the finite sample behavior.

In the online appendix we propose a set of asymptotically equivalent estimators. For known level sets, these estimators do not require kernel smoothing and selection of bandwidth parameters. This feature makes them particularly attractive from a practical and computational perspective. Throughout the discussion we will refer to these estimators as "bandwidth-free" estimators.

A caveat of our discussion so far is that in empirical applications the level sets are typically unknown and need to be estimated. Anderson, Linton, and Whang (2012) and Mammen and Polonik (2013) suggest plug-in methods for estimating the level sets. One possible candidate could be $\widehat{C}_{p_0} = \{y \in C : \widehat{q}_0(y) - \widehat{p}_0(y) > c_n, \widehat{q}_0(y) > 0, \widehat{p}_0(y) > 0\}$, where c_n is a positive and data-dependent threshold parameter that approaches zero as the sample size goes to infinity. We recommend an alternatively and novel bootstrap-based plug-in estimator. Define $\widehat{\Delta}_d(y) = \widehat{p}_d(y) - \widehat{q}_d(y)$, which is estimated based on cross-validated bandwidth selection, and denote by $\widehat{\Delta}_d^*(y)$ the bootstrap estimate of $\widehat{\Delta}_d(y)$. The level sets can then be estimated using the following pointwise $(1-\alpha)\%$-confidence intervals, e.g.,

$$\widehat{C}_{p_d} = \left\{ y \in \Re : \text{Median}\left(\widehat{\Delta}_d^*(y)\right) + Z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}\left(\widehat{\Delta}_d^*(y)\right)} < 0 \right\}, \tag{41}$$

$$\widehat{C}_{q_d} = \left\{ y \in \Re : \text{Median}\left(\widehat{\Delta}_d^*(y)\right) - Z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}\left(\widehat{\Delta}_d^*(y)\right)} > 0 \right\}, \tag{42}$$

for $d = 0, 1$, where $Z_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$-percentile of the standard normal distribution.[6]

Deriving the asymptotic properties of our estimators when using estimated level sets is outside the scope of this paper. However, the simulations presented in the online appendix strongly suggest that the properties of the estimators do not change substantially when we replac known level sets with the estimated level sets given by (41) and (42).

In the simulation study, we also compare the LATE estimators given by (37), (38), and (39) to the commonly used 2SLS. We find that in cases where there are no defiers, these LATE estimators can perform better in terms of smaller variance and smaller mean squared error relatively to 2SLS, which is evidence in support of Imbens and Rubin (1997). This implies that even in cases where monotonicity is a reasonable assumption it is recommendable to use the LATE estimators (37), (38), and (39), as they are efficient

[6] As it is commonly argued, the median is prefered over the mean of $\left(\widehat{\Delta}_d^*(y)\right)$ because of increased precision/robustness of the resulting bootstrap statistic

relative to 2SLS. Our simulation results suggest that efficiency gains can be substantial in small samples with a relatively weak instrument.

4 Empirical application

This section provides an application to the 1980 U.S. census data analyzed by Angrist and Krueger (1991), which (among other cohorts) contain 486,926 males born in 1940-49. Angrist and Krueger (1991) assess the effect of education on wages by using the quarter of birth as an instrument to control for potential endogeneity (for example, due to unobserved ability) between the treatment and the outcome. The idea is that the quarter of birth instrument affects education through age-related schooling regulations. As documented in Angrist and Krueger (1992), 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 Appendix 2 in Angrist and Krueger (1991), 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 decide to 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 they turn six will 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, Jaeger, and Baker (1995) challenge the validity of the exclusion restriction and present empirical results that point to systematic patterns in the seasonality of birth (for instance w.r.t. 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 realized by teenagers and unmarried women) based on U.S. birth certificate data and census data. For this reason, we will 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 treatment D is a binary indicator that is equal to zero for individuals receiving high school education or less (i.e., up to 12 years of education) and unity if obtaining at least some higher education (i.e., 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.

Secondly, 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 behavior 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 enroll 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. As discussed in Klein (2010), 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 will just have started the 10th grade when turning 16, some of them may decide to drop out directly, with only nine years of completed education. In contrast, pupils born earlier will be 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 individuals, compulsory schooling decreases in the quarter of birth and therefore violates monotonicity.

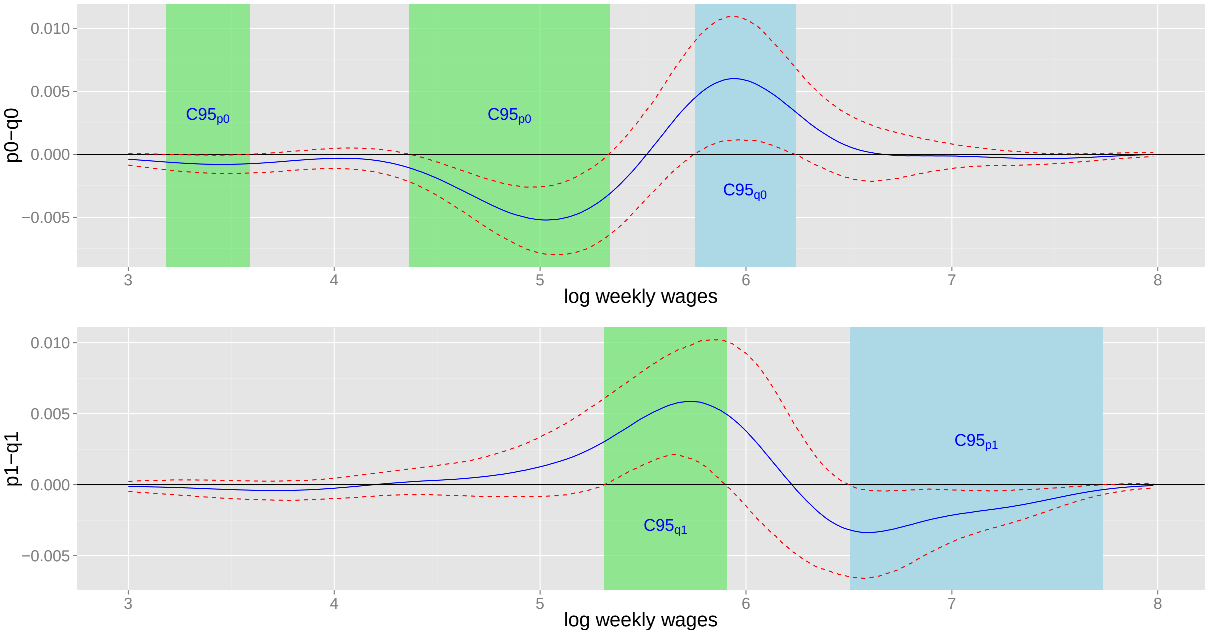
The implausibility of monotonicity motivates the use of the weaker Assumption 2 imposing local monotonicity only, while Assumption 1 (i.e., the exclusion restriction) will be maintained. LM is satisfied if wage is a positive function of socio-economic status and if socio-economic status also determines postponement. This is the case if parents with a high socio-economic status are 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. Empirical evidence pointing in this direction is provided by Bedard and Dhuey (2006), who report that children from the top quarter of the socio-economic distribution are over-represented among those who redshirt, and Aliprantis (2012), who finds that children whose enrollment is delayed are disproportionately wealthy with better-educated parents and more books at home. To summarize, it would be anticipated to find defiers to be situated in regions in the upper part of the wage distribution conditional on a particular level of education and compliers to come from regions in the lower part.

The plausibility of this hypothesis can be checked graphically by plotting the estimated differences between p_d and q_d for $d = 1, 0$. For this purpose, the bootstrap procedure outlined in Section 3 is applied and the respective 95% confidence bands are computed accordingly. In each of the 6000 bootstrap replications the densities $p_d(y)$ and $q_d(y)$ are estimated by kernel methods and cross-validated bandwidth on an equidistant grid of 1000 values on the support of the empirical weekly log wage distributions. The results are plotted in Figure 3. Under LM, the compliers are located in the green areas with 95% confidence, which, in agreement with our hypothesis, are in the lower part of the wage distribution both under treatment and non-treatment. The defiers are located in the blue areas, which, as conjectured, are in the upper part of the log wage distributions in both treatment states. Results are similar when 90% confidence bands are used, as shown in the online appendix. Note that if Assumption 2 fails, potential wage distributions of compliers and defiers overlap. This is a possibility here, as redshirting probably is not solely a function of socio-economic status but also of other factors, and wage is therefore not the only determinant of subject type. However, if the weaker Assumption 3 holds instead, the green and blue areas provide the areas for the compliers-survivors and defiers-survivors, respectively; see the discussion in Section 2.3.

After having obtained estimators of all the level sets based on the bootstrap according to (41) and (42), we can estimate the LATEs for the compliers, the defiers, and their joint distribution by the bandwidth-free estimators $\widehat{\mu}_c^a$, $\widehat{\mu}_d^a$, and $\widehat{\mu}_{c,d}^a$, the plug-in estimators $\widehat{\mu}_c$, $\widehat{\mu}_d$, and $\widehat{\mu}_{c,d}$, as well as the modified plug-in estimators denoted by $\widehat{\mu}_c^b$, $\widehat{\mu}_d^b$, and $\widehat{\mu}_{c,d}^b$ (see Section 3). 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.[7] Likewise, 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 of interest 6000 times to approximate their distributions, and the associated p-values are computed by assessing the rank of the estimates in their respective re-centered bootstrap distributions. We use two-sided hypothesis tests when computing p-values of the scale constraint and the LATE estimators and compute p-values associated with one-sided tests for complier and

[7] This particular choice of bandwidth (i.e., $h = b^{(3/2)}$) is motivated by Theorem 1 and is sufficient for the condition $\sqrt{n}h^2 \to 0$ to be satisfied.

Figure 3: Estimated differences in densities of log weekly wages (w), i.e., $\hat{p}_d(w) - \hat{q}_d(w)$ (blue solid line) under treatment ($d = 1$, lower panel) and non-treatment ($d = 0$, upper panel). The red dashed lines indicate 95% confidence bands. Estimations of the curves are based on 6000 bootstrap replications using a Gaussian Kernel and cross-validated bandwidth selection. Compliers are classified/estimated to exists in the green areas, whereas defiers exist in the blue areas.



defier shares (as the theoretical lower bound of these magnitudes is zero).

Table 2 presents the estimation results based on estimated level sets (using 95% (C95) and 90% (C90) confidence bands) in addition to 2SLS (which is numerically equivalent to the Wald estimator). The first six columns contain the estimates of the complier and defier shares using the bandwidth-free estimators and the two plug-in type estimators, respectively. Either share is, albeit very small (0.3% to 0.4%), significant at the 5% level irrespectively of the estimation method. This indicates a violation of global monotonicity, rendering 2SLS generally 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 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 not only come with potential gains in robustness, but also in efficiency. The next three columns give the estimates of $\Pr(D = 1|Z = 1) - \lambda_1 - (\Pr(D = 0|Z = 0) - \lambda_0)$ for testing the scale constraint, which must not be statistically different from zero under the null (see Section 2). For each of our estimation approaches and inference methods, the estimates are close to zero and not statistically significant at any conventional level, so that the identifying assumptions cannot be rejected.

The lower part of Table 2 presents the LATE estimators on the joint population of defiers and compliers (LATEcd), on the compliers only (LATEc), and on the defiers only (LATEd) as well as the 2SLS estimator. The precision of the 2SLS estimator is relatively low, implying that zero returns to higher education cannot be rejected. In contrast, the LATEs on compliers, defiers, and the joint population are highly significant, and the results suggest that education increases weekly wages by roughly 80 to 100%. In conclusion, our results suggest that the wage effects of higher education are substantial and quite homogeneous across compliers and defiers.[8] We wish again to point out that if Assumption 2 is violated but Assumption 3 holds, then the estimated shares and treatment effects still apply but in this case only to subpopulations of compliers and defiers. Alternatively, invoking CD instead and thereby focussing on 2SLS, would not permit us to draw conclusions about the returns to education due to the apparent weak instrument problem. Moreover, under the CD assumption, the estimated share of (global) compliers-survivors, which is equivalent to the first stage of a 2SLS regression, would not be significantly different from zero, while the estimated share of local compliers-survivors is small, but statistically significant according to our methods.

[8] We would like to point out that our estimation results are robust to educated modifications of the level sets. The level set robustness results are available from the authors upon request.

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**			
	$P(T=c)^a$	$P(T=c)$	$P(T=c)^b$	$P(T=d)^a$	$P(T=d)$	$P(T=d)^b$	$Test^a$	$Test$	$Test^b$	
C90										
estimator	0.004	0.004	0.004	0.003	0.004	0.004	-0.001	-0.001	-0.001	
p-value	0.011	0.013	0.013	0.021	0.014	0.014	0.675	0.704	0.707	
C95										
estimator	0.003	0.003	0.003	0.003	0.003	0.003	-0.001	-0.001	-0.001	
p-value	0.012	0.011	0.011	0.026	0.026	0.027	0.724	0.699	0.700	
Level set	**LATEcd**			**LATEc**			**LATEd**			**WALD**
	$\widehat{\mu}^a_{c,d}$	$\widehat{\mu}_{c,d}$	$\widehat{\mu}^b_{c,d}$	$\widehat{\mu}^a_c$	$\widehat{\mu}_c$	$\widehat{\mu}^b_c$	$\widehat{\mu}^a_d$	$\widehat{\mu}_d$	$\widehat{\mu}^b_d$	2SLS
C90										
estimator	0.907	0.906	0.906	0.801	0.816	0.816	1.023	0.995	0.995	1.379
p-value	0.003	0.003	0.003	0.010	0.010	0.010	0.023	0.015	0.017	0.213
C95										
estimator	0.928	0.921	0.921	0.898	0.880	0.880	0.963	0.969	0.969	1.379
p-value	0.001	0.001	0.001	0.007	0.006	0.006	0.018	0.019	0.018	0.213

Note: $\widehat{\mu}^a$ 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, $b^{\frac{3}{2}}$, where b is Silverman's rule-of-thumb bandwidth. Bootstrap p-values are based on 6000 resamples.

5 Conclusion

We have 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). This implies that defiers no longer need to be assumed away, so that also the LATEs on the defiers as well as on the joint population of defiers and compliers are identified. Furthermore, even if monotonicity is satisfied, using the novel LATE estimators might result in substantial efficiency gains compared to the 2SLS estimator.

Even when relaxing monotonicity, LM might still be considered restrictive in some applications. In this case, however, the proposed approach still identifies treatment effects on subsets of compliers and defiers if the conditions of the weaker LSM, which might be more plausible in some empirical applications, are satisfied.

We have applied our new methods to U.S. census data previously analyzed by Angrist and Krueger (1991) in an attempt to estimate the returns to higher education for males born in 1940-49 by using the birth quarter as an instrument for education. We have provided evidence in support of the existence of both compliers and defiers in this population and illustrated that traditional LATE estimation is not robust to ignoring defiers. In particular, the 2SLS estimator is very imprecisely estimated because compliers and defiers net out, thereby creating a weak instrument problem. In contrast, the new LATE estimators introduced are much more efficient and predict large returns to higher education for compliers and defiers that are very similar in magnitude. Finally, we have illustrated how a visual inspection of the estimated complier and defier distributions can help in assessing the plausibility of the global and local monotonicity assumptions.

References

ALIPRANTIS, D. (2012): "Redshirting, Compulsory Schooling Laws, and Educational Attainment," *Journal of Educational and Behavioral Statistics*, 37, 316–338.

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., G. IMBENS, AND D. RUBIN (1996): "Identification of Causal Effects using Instrumental Variables," *Journal of American Statistical Association*, 91, 444–472 (with discussion).

ANGRIST, J., AND A. KRUEGER (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, 106, 979–1014.

——— (1992): "The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples," *Journal of the American Statistical Association*, 87, 328–336.

BALKE, A., AND J. PEARL (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

BARUA, R., AND K. LANG (2009): "School Entry, Educational Attainment, and Quarter of Birth: A Cautionary Tale of LATE," *NBER Working Paper 15236.*

BEDARD, K., AND E. DHUEY (2006): "The persistence of early childhood maturity: International evidence of long-run age effects," *Quarterly Journal of Economics*, 121, 1437–1472.

BOUND, J., D. A. JAEGER, AND R. M. BAKER (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

BUCKLES, K. S., AND D. M. HUNGERMAN (2013): "Season of Birth and Later Outcomes: Old Questions, New Answers," *Review of Economics and Statistics*, 95, 711–724.

CHERNOZHUKOV, V., AND C. HANSEN (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

DE CHAISEMARTIN, C. (2012): "All you need is LATE," *mimeo, Paris School of Economics.*

——— (2016): "Tolerating Defiance? Identification of treatment effects without monotonicity," *mimeo, University of Warwick.*

DE CHAISEMARTIN, C., AND X. D'HAULTFOEUILLE (2012): "LATE again, with defiers," *mimeo, CREST.*

Frölich, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.

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

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, 467–475.

Imbens, G. W., and D. Rubin (1997): "Estimating outcome distributions for compliers in instrumental variables models," *Review of Economic Studies*, 64, 555–574.

Kitagawa, T. (2009): "Identification Region of the Potential Outcome Distribution under Instrument Independence," *CeMMAP working paper 30/09.*

Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83(5), 2043–2063.

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

Li, Q., and J. S. Racine (2007): *Nonparametric econometrics: theory and practice.* Princeton University Press, Princeton and Oxford.

Mammen, E., and W. Polonik (2013): "Confidence regions for level sets," *Journal of Multivariate Analysis*, 122, 202–214.

Mourifie, I., and Y. Wan (2016): "Testing Local Average Treatment Effect Assumptions," *forthcoming in the Review of Economics and Statistics.*

Roy, A. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3, 135–146.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Silverman, B. (1986): *Density estimation for statistics and data analysis.* Chapman and Hall, London.

Small, D. S., and Z. Tan (2007): "A Stochastic Monotonicity Assumption for the Instrumental Variables Method," *Technical report, Department of Statistics, Wharton School, University of Pennsylvania.*

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

Online Appendix to
It's 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; **University of Fribourg

February 14, 2017

Abstract: This online appendix contains the proofs of the main theoretical results. Results of an extensive simulation study are also reported. Finally, it is shown how to generalize the framework to the case of multivalued instruments and a simple structural model is derived in which local monotonicity is satisfied.

Keywords: instrumental variable, treatment effects, LATE, local monotonicity.

JEL classification: C14, C21, C26.

We have benefited from comments by Alberto Abadie, Joshua Angrist, Matias Cattaneo, Mario Fiorini, Markus Frölich, Stefan Hoderlein, Guido Imbens, Toru Kitagawa, Frank Kleibergen, Tobias Klein, Michael Lechner, Arthur Lewbel, Enno Mammen, Blaise Melly, Katrien Stevens, Tymon Słoczyński and seminar/conference participants at the University of St. Gallen (Oct 2011), Brown University (Nov 2011), the Austrian-Swiss labor meeting in Stuben (Jan 2012), Harvard (Feb 2012), MIT (Mar 2012), the Spring Meeting of Young Economists in Mannheim (Apr 2012), Boston College (Apr 2012), the North American Summer Meeting of the Econometric Society in Evanston (June-July 2012), the European Meeting of the Econometric Society in Malaga (Aug 2012), the European Association of Labour Economists meeting in Bonn (Sept 2012), the Italian Congress of Econometrics and Empirical Economics in Genoa (Jan 2013), Erasmus University Rotterdam (Jun 2013), the International Association for Applied Econometrics in London (Jun 2014), IHS Vienna (Oct 2015), University of Duisburg-Essen (Feb 2016), the Ce^2 workshop in Warsaw (Jun 2016), the COMPIE in Milan (Oct 2016), and the University of Mannheim (Nov 2016). Martin Huber gratefully acknowledges financial support from the Swiss National Science Foundation grant PBSGP1_138770. Addresses for correspondence: Christian M. Dahl (cmd@sam.sdu.dk), and Giovanni Mellace (giome@sam.sdu.dk), Martin Huber (martin.huber@unifr.ch).

A Proof of the equivalence of the four scale constraints

We first show that under Assumptions 1 and 2, the defiers' scale constraint $\Pr(D=1|Z=0)-\lambda_1=\Pr(D=0|Z=1)-\lambda_0$ (with $\lambda_d=\int_{\mathcal{Y}}\min(p_d(y),q_d(y))dy$) is equivalent to the constraints faced by the compliers, i.e.,

$$\begin{aligned}
\Pr(D=1|Z=0)-\lambda_1 &= \Pr(D=0|Z=1)-\lambda_0\\
1-\Pr(D=0|Z=0)-\lambda_1 &= 1-\Pr(D=1|Z=1)-\lambda_0\\
\Pr(D=1|Z=1)-\lambda_1 &= \Pr(D=0|Z=0)-\lambda_0.
\end{aligned}$$

where the last line is the scale constraint for the compliers. The scale constraints for the always takers and the never takers, respectively, are given by

$$\lambda_1 = 1-\delta_0, \qquad \text{and} \qquad 1-\delta_1=\lambda_0,$$

where $\delta_d=\int_{\mathcal{Y}}\max(p_d(y),q_d(y))dy$. Considering $\lambda_1=1-\delta_0$,

$$\begin{aligned}
\lambda_1 &= 1-\delta_0\\
\Pr(D=1|Z=1)-\lambda_1 &= \Pr(D=1|Z=1)-(1-\delta_0)\\
\Pr(D=1|Z=1)-\lambda_1 &= \delta_0-\Pr(D=0|Z=1)\\
\Pr(D=1|Z=1)-\lambda_1 &= \int_{\mathcal{Y}}\max(p_0(y),q_0(y))dy-\int_{\mathcal{Y}}p_0(y)dy\\
\Pr(D=1|Z=1)-\lambda_1 &= \int_{\mathcal{Y}}p_0(y)dy+\int_{\mathcal{Y}}q_0(y)dy-\int_{\mathcal{Y}}\min(p_0(y),q_0(y))dy-\int_{\mathcal{Y}}p_0(y)dy\\
\Pr(D=1|Z=1)-\lambda_1 &= \Pr(D=0|Z=0)-\lambda_0.
\end{aligned}$$

This completes the proof.□

B Structural models satisfying Assumptions 1 and 2

To judge the implications of Assumptions 1 and 2 in a structural model, consider the following two stage endogenous treatment selection model, with the first stage being characterized by a random coefficient model:[9]

$$\begin{aligned} Y_i &= \varphi(D_i, \epsilon_i), \\ D_i &= I\left(\gamma_0 + \gamma_i Z_i + \nu_i > 0\right), \end{aligned} \tag{B.1}$$

where i indexes a particular subject. $I(\cdot)$ is the indicator function which is equal to one if its argument holds true and zero otherwise. φ is a general function, and ϵ_i, ν_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 $D_i(1) = I(\gamma_0 + \gamma_i + \nu_i > 0) \geq D_i(0) = I(\gamma_0 + \nu_i > 0)$, which locally rules out defiers. 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 $D_i(0) = I(\gamma_0 + \nu_i > 0) \geq D_i(1) = I(\gamma_0 + \gamma_i + \nu_i > 0)$. 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 restricts γ_i only locally.

To give an idea about possible setups in which Assumption 2 holds while monotonicity does not, we provide two parametric examples that put further structure on the equations in (B.1). Firstly, assume that the outcome equation is characterized by the following model:

$$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 D_i \epsilon_i + \epsilon_i, \tag{B.2}$$

where α_0 is the constant, α_1, α_2 are the coefficients on the treatment and its interaction (capturing individual effect heterogeneity), and ϵ_i 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 a deterministic function of ϵ_i:

$$\gamma_i = \beta_0 + \rho \epsilon_i, \tag{B.3}$$

where β_0 is a constant and ρ isthe 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 is a monotonic function of ϵ_i (unless α_2 is exactly -1 and $D_i = 1$), the outcomes of the compliers and defiers do not overlap conditional on the treatment state so that LM is satisfied.

[9] We are indebted to Joshua Angrist for making valuable suggestions concerning potential models that fit our framework.

Secondly, we consider an extension of our setup to a Roy (1951) -type model, which implies that the probability of treatment increases with the gains it creates. To this end, we maintain the previous outcome equation (B.2), but modify the first stage:

$$D_i \quad = \quad I\left(Y_i(1) - Y_i(0) + \gamma_i Z_i + \nu_i > 0\right) = I\left(\beta_0 + \alpha_1 + (\alpha_2 + \rho Z_i)\epsilon_i + \nu_i > 0\right), \text{(B.4)}$$

where the individual level treatment effect, e.g., the returns to education or training, now influences the selection into treatment. In this case ν_i, if different from zero, may be interpreted as individual costs, disutility, or utility of the treatment not reflected by the treatment effect per se. The instrument Z exogenously shifts participation, but the direction depends again on ϵ_i as specified in (B.3). The expression left of the equality follows from substituting $Y_i(1) - Y_i(0)$ by $\alpha_1 + \alpha_2\epsilon_i$ and using (B.3). Again, this model implies a non-overlapping 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.

C Non-binary instruments

This section discusses the identification of LATEs in the presence of a multi-valued discrete instrument with bounded support.[10] Under (global) monotonicity, Frölich (2007) shows that if the support of Z is bounded so that $Z \in [z_{\min}, z_{\max}]$, where $z_{\min}$ and $z_{\max}$ are finite upper and lower bounds, it is possible to define and identify LATEs on the compliers with respect to any two distinct subsets of the support of Z. The proportion of compliers in general varies depending on the choice of subsets and is maximized when choosing the endpoints $z_{\min}$, $z_{\max}$. In our framework which allows for compliers and defiers, this result no longer holds in general without specifying LM more tightly. To see this, let z and z' $\in [z_{\min}, z_{\max}]$ denote two subsets such that $z \neq z'$. Define $\tilde{Z}$ as

$$\tilde{Z} = \begin{cases} 1 & \text{if } Z \in z \\ 0 & \text{if } Z \in z' \end{cases}. \tag{C.1}$$

Let there exist a random variable Z such that $Z \perp (D(1), D(0), Y(1), Y(0))$, where $\perp$ denotes independence.

As an example, consider the case that the instrument can take three values, e.g. $Z \in \{0, 1, 2\}$, such that instead of Assumption 1 we invoke the following independence assumption:

Assumption 1a: Let there exist a random variable Z such that $Z \perp (D(2), D(1), D(0), Y(1), Y(0))$, where $\perp$ denotes independence.

Without imposing any form of monotonicity, there now exist eight types according to $D(2), D(1), D(0)$, see Table 3. Positive monotonicity rules out types 3, 5, 6, and 7 so that only always takers (type 1), never takers (type 8) and compliers when switching the instrument from 0 to 1 (type 2) or from 1 to 2 (type 4) exist. In this framework, one could possibly think of five different definitions of z, z': (i) $z = \{0\}, z' = \{1\}$, (ii) $z = \{1\}, z' = \{2\}$, (iii) $z = \{0\}, z' = \{2\}$, (iv) $z = \{0, 1\}, z' = \{2\}$, (v) $z = \{0\}, z' = \{1, 2\}$. (iii) maximizes the complier proportion, namely the joint proportion of types 2 and 4. This is the case because it may induce individuals to react on the treatment that are otherwise always or never takers when the instrument has less asymptotic power, i.e., operates over a smaller support, such as in (i), which only covers type 2, and in (ii), which covers type 4. In contrast, (iv) and (v) may be chosen to maximize finite sample power, because these setups encounter at least as many observations as (iii), at the cost of a weakly lower complier proportion.

Identification becomes more complicated if we abandon (global) monotonicity. Without further restrictions, all eight types may exist, out of which two are pure compliers (types 2 and 4), two are pure defiers (types 5 and 7) and two even switch from compliance to defiance

[10] We thank Toru Kitagawa for very helpful comments concerning the case of non-binary instruments.

Table 3: Types for $Z \in \{0, 1, 2\}$

Type T	D(2)	D(1)	D(0)
1	1	1	1
2	1	1	0
3	1	0	1
4	1	0	0
5	0	1	1
6	0	1	0
7	0	0	1
8	0	0	0

(type 6) or vice versa (type 3). Clearly, if LM is imposed w.r.t. $D(1), D(0)$ only, which allows identifying LATEs within (i), or w.r.t. $D(2), D(1)$ only, which allows identifying LATEs within (ii), identification of LATEs in (iii) to (v) is generally not feasible. The reason is that the densities of compliers and defiers across (i) and (ii) may net each other out when coarsening the values of the instrument as in (iii) and (iv) or when considering endpoints only as in (v). I.e., some $y(1)$ and/or $y(0)$ might be inhabited by compliers in (i) and defiers in (ii) or vice versa such that any definition of z, z' not consisting of neighboring support points in Z does in general not identify LATEs. One possibility to establish identification is to assume that LM holds over all values in the support of the instrument.

Assumption 2a: For all subjects in the population, either $\Pr(D(2) \geq D(1) \geq D(0)|y(d)) = 1$ or $\Pr(D(0) \geq D(1) \geq D(2)|y(d)) = 1 \; \forall \; y(d) \in \mathcal{Y}$, where $d \in \{0, 1\}$.

Assumption 2a rules out types 3 and 6 globally, implying that no individuals switch their treatment state in opposite directions for distinct pairs of instrument values. Furthermore, either defying types 5 and 7 or complying types 2 and 4 must not exist locally for any $y(d)$, meaning that over the entire range of instrument values, the support of defiers and compliers never overlaps. Under Assumptions 1a and 2a, the LATEs on types 2, 4, 5, and 7 are identified. I.e., (i) identifies the LATEs on types 2 and 7 and (ii) those on types 4 and 5. Analogously to the setup under global monotonicity, (iii) now maximizes both the proportions of compliers and defiers by identifying the LATEs on the types 2 and 4 jointly as well as on 5 and 7 jointly.

D Proof of Theorem 1

In this section we provide a detailed proof of Theorem 1. First we provide a set of assumptions, A1 to A4, which restate Assumption 4 in the main text. Secondly, we present and proof three helpful lemmas. Thereafter, the results of Theorem 1 follows readily.

Assumption A1 The general nonnegative bounded kernel function $w(\cdot)$ satisfies (i) $\int w(v)\,dv = 1$, (ii) $w(v) = w(-v)$, and (iii) $\int v^2 w(v)\,dv = \kappa_2 > 0$.

Assumption A2 $\sup_{y\in\mathcal{Y}} B(y, D_i = d, Z_i = z) = M_{(d,z),y} < \infty$ for all $d = 0, 1$ and $z = 0, 1$ where $B(y, D = d, Z = z) = \frac{1}{2}\kappa_2 \frac{\nabla^2\left(f_{Y,D|Z}(y, D=d|Z=z)\right)}{(\nabla y)^2} \Big/ f_Z(Z = z)$.

Assumption A3 For all $d = 0, 1$ (i) $\inf_{y\in\mathcal{Y}} p_d(y) = M_{p_i,y} > 0$, and (ii) $\inf_{y\in\mathcal{Y}} q_d(y) = M_{q_i,y} > 0$.

Assumption A4 (Y_i, D_i, Z_i) for $i = 1, 2, ..., n$ is an *i.i.d.* distributed 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 $E\left(|Y|^2\right)$, $E_{p_d}\left(|Y|^2\right)$ and $E_{q_d}\left(|Y|^2\right)$ exist for $d = 0, 1$.

Assumptions A1-A4 are very commonly used in the kernel estimation literature, and hold for a wide range of data generating processes. In particular, Assumption A3 ensures that p and q are not truncated and rules out boundary effects. Assumption 3 simplifies the proof and can be relaxed by using boundary kernels (see, Li and Racine, 2007). Assumption A4 is stated in very general terms and implies the existence of moments like, for example, $E_{p_1}(Y_i|Y_i \in C_{q_1})$, $E_{q_1}(Y_i|Y_i \in C_{q_1})$, $E_{p_0}(Y_i|Y_i \in C_{p_0})$, $E_{q_0}(Y_i|Y_i \in C_{p_0})$, $E(|Y_i|)$ etc.

We now establish the asymptotic distribution of $\widehat{\theta}_1$ by the following lemma:

Lemma 1 Let θ_1 be the true value of $\widehat{\theta}_1$ defined in Section Estimation. Under Assumptions A1, A2, A3, and A4, for $n \to \infty$, $h \to 0$, and $\sqrt{n}h^2 \to 0$, it follows

$$\sqrt{n}\left(\widehat{\theta}_1 - \theta_1 - a_1\right) \xrightarrow{d} N\left(0, \sigma^2_{p_1,(q_1)} + \sigma^2_{q_1,(q_1)} - 2\sigma_{p_1,q_1,(q_1)}\right),$$

where $a_1 = h^2 \int_{C_q} y\left(B\left(y, D=1, Z=1\right) - B\left(y, D=1, Z=0\right)\right) dy$, and

$$\begin{aligned}
\sigma^2_{p_1,(q_1)} &= E\left(\left(\frac{Y_i I\left(D_i=1\right) I\left(Z_i=1\right)}{\frac{1}{n}\sum_{j=1}^n I\left(Z_j=1\right)} 1\left(Y_i \in C_{q_1}\right) - E_{p_1}\left(Y_i | Y_i \in C_{q_1}\right)\right)^2\right), \\
\sigma^2_{q_1,(q_1)} &= E\left(\left(\frac{Y_i I\left(D_i=1\right) I\left(Z_i=0\right)}{\frac{1}{n}\sum_{j=1}^n I\left(Z_j=0\right)} 1\left(Y_i \in C_{q_1}\right) - E_{q_1}\left(Y_i | Y_i \in C_{q_1}\right)\right)^2\right), \\
\sigma_{p_1,q_1,(q_1)} &= E\left(\left(\frac{Y_i I\left(D_i=1\right) I\left(Z_i=1\right)}{\frac{1}{n}\sum_{j=1}^n I\left(Z_j=1\right)} 1\left(Y_i \in C_{q_1}\right) - E_{p_1}\left(Y_i | Y_i \in C_{q_1}\right)\right) \times \right. \\
&\quad \left. \left(\frac{Y_i I\left(D_i=1\right) I\left(Z_i=0\right)}{\frac{1}{n}\sum_{j=1}^n I\left(Z_j=0\right)} 1\left(Y_i \in C_{q_1}\right) - E_{q_1}\left(Y_i | Y_i \in C_{q_1}\right)\right)\right).
\end{aligned}$$

Proof of Lemma 1 We can write

$$\begin{aligned}
\sqrt{n}\left(\widehat{\theta}_1 - \theta_1\right) &= \sqrt{n}\int_{C_q} y\left(\widehat{p}_1(y) - p_1(y)\right) dy - \sqrt{n}\int_{C_q} y\left(\widehat{q}_1(y) - q_1(y)\right) dy \\
&= : A_{1n} + A_{2n},
\end{aligned}$$

Consider first the term A_{1n} given as

$$\begin{aligned}
A_{1n} &= \sqrt{n}\int_{C_q} y\left(\widehat{p}_1(y) - p_1(y)\right) dy \\
&= \sqrt{n}\int_{C_q} y\left(\widehat{p}_1(y) - E\left(\widehat{p}_1(y)\right)\right) dy + \sqrt{n}\int_{C_q} y\left(E\left(\widehat{p}_1(y)\right) - p_1(y)\right) dy \\
&= : A_{11n} + A_{12n}.
\end{aligned}$$

As $\widehat{p}_1(y)$ is a kernel estimator it is well known that under standard regularity conditions,

$$\left(E_{p_1}\left(\widehat{p}_1(y)\right) - p_1(y)\right) = h^2 B\left(y, D=1, Z=1\right) + O\left(h^3\right),$$

where $B\left(y, D=1, Z=1\right)$ is defined as in Assumption A2, see, e.g., Li and Racine (2007) pages 166 and 167. Given Assumptions A2 and A3 we can write

$$\begin{aligned}
A_{12n} &= \sqrt{n}\int_{C_q} y\left(E_{p_1}\left(\widehat{p}_1(y)\right) - p_1(y)\right)dy \\
&= \frac{1}{2}\sqrt{n}h^2\kappa_2\int_{C_q} yB\left(y, D=1, Z=1\right)dy + O\left(\sqrt{n}h^3\right)\int_{C_q} ydy \\
&\leq \frac{1}{2}\sqrt{n}h^2\kappa_2 M_{(1,1),u}\int_{C_q}\left|y\right|dy + O\left(\sqrt{n}h^3\right)\int_{C_q}\left|y\right|dy \\
&= \frac{1}{2}\sqrt{n}h^2\kappa_2 M_{(1,1),u}\int_{C_q}\frac{\left|y\right|}{p_1\left(y\right)}p_1\left(y\right)dy + O\left(\sqrt{n}h^3\right)\int_{C_q}\frac{\left|y\right|}{p_1\left(y\right)}p_1\left(y\right)dy \\
&\leq \frac{1}{2}\sqrt{n}h^2\kappa_2 M_{(1,1),u}\int_{\mathcal{Y}}\frac{\left|y\right|}{p_1\left(y\right)}p_1\left(y\right)dy + O\left(\sqrt{n}h^3\right)\int_{\mathcal{Y}}\frac{\left|y\right|}{p_1\left(y\right)}p_1\left(y\right)dy \\
&\leq \frac{1}{2}\sqrt{n}h^2\kappa_2 M_{(1,1),u}M_{p_1,l}^{-1}\int_{\mathcal{Y}}\left|y\right|p_1\left(y\right)dy + O\left(\sqrt{n}h^3\right)M_{p_1,l}^{-1}\int_{\mathcal{Y}}\left|y\right|p_1\left(y\right)dy \\
&= \frac{1}{2}\sqrt{n}h^2\kappa_2 M_{(1,1),u}M_{p_1,l}^{-1}E_{p_1}\left(\left|y\right|\right) + O\left(\sqrt{n}h^3\right)M_{p_1,l}^{-1}E_{p_1}\left(\left|y\right|\right) \\
&= O\left(\sqrt{n}h^2\right).
\end{aligned}$$

Next, consider A_{11n}:

$$\begin{aligned}
A_{11n} &= \sqrt{n}\int_{C_q} y\left(\widehat{p}_1(y) - E_{p_1}\left(\widehat{p}_1(y)\right)\right)dy = \sqrt{n}\int_{C_q} y\left(\widehat{p}_1(y) - E_{p_1}\left(\widehat{p}_1(y)\right)\right)dy \\
&= \sqrt{n}\int_{C_q} y\left(\widehat{p}_1(y) - p_1(y) - h^2B\left(y, D=1, Z=1\right) + o\left(h^2\right)\right)dy \\
&= \sqrt{n}\int_{C_q} y\widehat{p}_1(y)dy - \sqrt{n}\int_{C_q} yp_1(y)dy - \sqrt{n}\int_{C_q}\left(yh^2B\left(y, D=1, Z=1\right) + o\left(h^2\right)\right)dy.
\end{aligned}$$

Note that

$$\begin{aligned}
\sqrt{n}\int_{C_q} y\widehat{p_1}(y)dy &= \sqrt{n}\int_{C_q} y\frac{\frac{1}{nh}\sum_{i=1}^{n} I\left(D_i = d\right) I\left(Z_i = z\right) w\left(\frac{y-Y_i}{h}\right)}{\frac{1}{n}\sum_{i=1}^{n} I\left(Z_i = z\right)} dy \\
&= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \frac{I\left(D_i = d\right) I\left(Z_i = z\right)}{\frac{1}{n}\sum_{j=1}^{n} I\left(Z_j = z\right)} \int_{C_q} y\frac{1}{h} w\left(\frac{y-Y_i}{h}\right) dy \\
&= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \frac{I\left(D_i = d\right) I\left(Z_i = z\right)}{\frac{1}{n}\sum_{j=1}^{n} I\left(Z_j = z\right)} \int_{\widetilde{C}q} \left(Y_i + vh\right) w\left(v\right) dv \\
&= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i I\left(D_i = d\right) I\left(Z_i = z\right)}{\frac{1}{n}\sum_{j=1}^{n} I\left(Z_j = z\right)} \int_{\widetilde{C}q} w\left(v\right) dv + \\
&\quad \sqrt{n}h\sum_{i=1}^{n} \frac{I\left(D_i = d\right) I\left(Z_i = z\right)}{\frac{1}{n}\sum_{j=1}^{n} I\left(Z_j = z\right)} \int_{\widetilde{C}q} vw\left(v\right) dv \\
&= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i I\left(D_i = d\right) I\left(Z_i = z\right)}{\frac{1}{n}\sum_{j=1}^{n} I\left(Z_j = z\right)} \int_{\widetilde{C}q} w\left(v\right) dv
\end{aligned}$$

where

$$\widetilde{C}q_1 = \left\{ v \in \mathbb{R} : v = \frac{y - Y_i}{h} for\ \ (y, Y_i, h) \in C_{q_1} \times \mathcal{Y} \times \mathbb{R}_+ \right\}.$$

Importantly, note that if $Y_i < \min\left(C_{q_1}\right)$, it follows that $\lim_{h\to 0} \widetilde{C}q_1 = (\infty; \infty)$ and similarly, if $Y_i > \max\left(C_{q_1}\right)$, then $\lim_{h\to 0} \widetilde{C}q_1 = (-\infty; -\infty)$. Additionally, if $Y_i \in C_{q_1}$, then $\lim_{h\to 0} \widetilde{C}q_1 = (-\infty; \infty)$. This implies that

$$\int_{\widetilde{C}q} w\left(v\right) dv = 1\left(Y_i \in C_{q_1}\right), \quad \int_{\widetilde{C}q} vw\left(v\right) dv = 0.$$

Consequently,

$$\lim_{h\to 0}\int_{C_q} y\widehat{p_1}(y)dy = E_{\widehat{p_1}}(Y_i|Y_i \in C_{q_1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i I\left(D_i = 1\right) I\left(Z_i = 1\right)}{\frac{1}{n}\sum_{j=1}^{n} I\left(Z_j = 1\right)} 1\left(Y_i \in C_{q_1}\right),$$

and

$$\int_{C_q} yp_1(y)dy = E_{p_1}(Y_i|Y_i \in C_{q_1}), \quad \int_{C_q} \left(yh^2 B\left(y, D = 1, Z = 1\right) + o\left(h^2\right)\right) dy = O\left(h^2\right).$$

Given Assumption A4, we can apply the Lindeberg-Levy central limit theorem to the random variable $Y_i I(D_i = 1) I(Z_i = 1) / \frac{1}{n}\sum_{j=1}^{n} I(Z_j = 1)$ and write the limit of A_{11n} as

$$A_{11n} \rightarrow \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{Y_i I(D_i = 1) I(Z_i = 1)}{\frac{1}{n}\sum_{j=1}^{n} I(Z_j = 1)} - E_{p_1}(Y_i | Y_i \in C_{q_1})\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{p_1,(q_1)}\right),$$

for $n \rightarrow \infty$, $h \rightarrow 0$ and $\sqrt{n}h^2 \rightarrow 0$. Similary, we have

$$A_{21n} \rightarrow \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{Y_i I(D_i = 1) I(Z_i = 0)}{\frac{1}{n}\sum_{j=1}^{n} I(Z_j = 0)} - E_{q_1}(Y_i | Y_i \in C_{q_1})\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{q_1,(q_1)}\right).$$

Trivially, we define

$$Acov(A_{11n}, A_{21n}) = \sigma_{p_1,q_1,(q_1)},$$

which is guaranteed to exists by Cauchy-Schwarts and the existence of $\sigma^2_{p_1}$ and $\sigma^2_{q_1}$. This completes the proof of Lemma 1. □

Lemma 2 Let the assumptions of Lemma 1 hold. Then

$$\left(\frac{\widehat{\theta}_1}{\widehat{P}_{1|1} - \widehat{\lambda}_1}\right) \xrightarrow{d} N\left(\frac{\theta_1 - a_1}{P_{1|1} - \lambda_1}, \frac{\sigma^2_{p_1,(q_1)} + \sigma^2_{q_1,(q_1)} - 2\sigma_{p_1,q_1,(q_1)}}{n\left(P_{1|1} - \lambda_1\right)^2}\right),$$

for $n \rightarrow \infty$, $h \rightarrow 0$ and $\sqrt{n}h^2 \rightarrow 0$.

Proof of Lemma 2 First we establish the asymptotic properties of $P_{d|z}$ where $d = 0, 1$ and $z = 0, 1$. Note that

$$\Pr(D = d | Z = z) = \frac{\Pr(D = d, Z = z)}{\Pr(Z = z)}.$$

By analogy, the estimator we consider is given by

$$\widehat{P}_{d|z} = \frac{\frac{1}{n}\sum_{i=1}^{n} I(D_i = d) I(Z_i = z)}{\frac{1}{n}\sum_{i=1}^{n} I(Z_i = z)}.$$

Under Assumption A4, this expression is the ratio of two consistent maximum likelihood estimators (MLE). Specifically, the numerator is the MLE of a multinomial distribution and the denominator is the MLE of a Bernoulli distribution. Hence, by Slutsky it follows that

$$\widehat{P}_{d|z} - \Pr\left(D = d|Z = z\right) \xrightarrow{p} 0,$$

in probability as $n \to \infty$ and that the rate of convergence is $\sqrt{n}$. Furthermore, Schmid and Schmidt (2006) show that under our assumptions $\widehat{\lambda}_i$ for $i = 0, 1$ converges in probability to $\lambda_i = \int_{\mathcal{Y}} \min\left(p_i(y), q_i\left(y\right)\right) dy$. By application of Slutsky and the result of Lemma 1, Lemma 2 follows trivially. This completes the proof of Lemma 2. □

Lemma 3 Let the assumptions of Lemma 1 hold. Then, for $n \to \infty$, $h \to 0$ and $\sqrt{n}h^2 \to 0$,

$$\left(\frac{\widehat{\theta}_0}{\widehat{P}_{0|0} - \widehat{\lambda}_0}\right) \xrightarrow{d} N\left(\frac{\theta_0 - a_0}{P_{0|0} - \lambda_0}, \frac{\sigma^2_{p0,(p0)} + \sigma^2_{q0,(p0)} - 2\sigma_{p0,q0,(p0)}}{n\left(P_{0|0} - \lambda_0\right)^2}\right),$$

where $a_0 = h^2 \int_{C_p} y\left(B\left(y, D = 0, Z = 1\right) - B\left(y, D = 0, Z = 0\right)\right) dy$, and

$$\begin{aligned}
\sigma^2_{p0,(p0)} &= E\left(\left(\frac{Y_i I\left(D_i = 0\right) I\left(Z_i = 1\right)}{\frac{1}{n}\sum_{j=1}^n I\left(Z_j = 1\right)} 1\left(Y_i \in C_{p0}\right) - E_{p0}\left(Y_i | Y_i \in C_{p0}\right)\right)^2\right), \\
\sigma^2_{q0,(p0)} &= E\left(\left(\frac{Y_i I\left(D_i = 0\right) I\left(Z_i = 0\right)}{\frac{1}{n}\sum_{j=1}^n I\left(Z_j = 0\right)} 1\left(Y_i \in C_{p0}\right) - E_{q0}\left(Y_i | Y_i \in C_{p0}\right)\right)^2\right), \\
\sigma_{p0,q0,(p0)} &= E\left(\left(\frac{Y_i I\left(D_i = 0\right) I\left(Z_i = 1\right)}{\frac{1}{n}\sum_{j=1}^n I\left(Z_j = 1\right)} 1\left(Y_i \in C_{p0}\right) - E_{p0}\left(Y_i | Y_i \in C_{p0}\right)\right)\right. \\
&\quad \times \left.\left(\frac{Y_i I\left(D_i = 0\right) I\left(Z_i = 0\right)}{\frac{1}{n}\sum_{j=1}^n I\left(Z_j = 0\right)} 1\left(Y_i \in C_{p0}\right) - E_{q0}\left(Y_i | Y_i \in C_{p0}\right)\right)\right).
\end{aligned}$$

Proof of Lemma 3 The proof of Lemma 3 is symmetric to the proof of Lemma 2, and therefore omitted. □

Proof of Theorem 1 Let Assumptions A1, A2, A3, and A4 hold. The result regarding the asymptotic distribution of $\sqrt{n}\left(\widehat{\mu}_c - \mu_c - b_c\right)$ follows directly from applying Lemmas 1, 2, and 3 for $b_c = a_0 - a_1$. Furthermore, notice that by Lemmas 2 and 3, the asymptotic variance-covariance matrix Ω_c is given by

$$\begin{aligned}
\Omega_c &= Avar\left(\widehat{\mu}_c\right) \\
&= Avar\left(\frac{\widehat{\theta}_1}{\widehat{P}_{1|1}-\widehat{\lambda}_1}\right)+Avar\left(\frac{\widehat{\theta}_0}{\widehat{P}_{0|0}-\widehat{\lambda}_0}\right)-2*Acov\left(\frac{\widehat{\theta}_1}{\widehat{P}_{1|1}-\widehat{\lambda}_1},\frac{\widehat{\theta}_0}{\widehat{P}_{0|0}-\widehat{\lambda}_0}\right) \\
\Omega_c &= \left(\frac{\sigma^2_{p_0,(p_0)}+\sigma^2_{q_0,(p_0)}-2\sigma_{p_0,q_0,(p_0)}}{\left(P_{0|0}-\lambda_0\right)^2}\right)+\left(\frac{\sigma^2_{p_1,(q_1)}+\sigma^2_{q_1,(q_1)}-2\sigma_{p_1,q_1,(q_1)}}{\left(P_{1|1}-\lambda_1\right)^2}\right)- \\
&\quad 2\frac{Acov\left(\widehat{\theta}_0,\widehat{\theta}_1\right)}{\left(P_{0|0}-\lambda_0\right)\left(P_{1|1}-\lambda_1\right)}.
\end{aligned}$$

Proceeding similarly, the proofs of the asymptotic distributions of $\sqrt{n}\left(\widehat{\mu}_{c,d}-\mu_{c,d}-b_{c,d}\right)$ and $\sqrt{n}\left(\widehat{\mu}_d-\mu_d-b_d\right)$ follow straightforwardly and are therefore omitted. It can be shown that

$$\begin{aligned}
\Omega_{c,d} &= \left(\frac{\sigma^2_{p_1,(q_1)}+\sigma^2_{q_1,(q_1)}+\sigma^2_{p_1,(p_1)}+\sigma^2_{q_1,(p_1)}-2\left(\sigma_{p_1,q_1,(q_1)}+\sigma_{p_1,q_1,(p_1)}\right)}{\left(P_{1|1}+P_{1|0}-2\lambda_1\right)^2}\right)+ \\
&\quad \left(\frac{\sigma^2_{p_0,(p_0)}+\sigma^2_{q_0,(p_0)}+\sigma^2_{p_0,(q_0)}+\sigma^2_{q_0,(q_0)}-2\left(\sigma_{p_0,q_0,(p_0)}+\sigma_{p_0,q_0,(q_0)}\right)}{\left(P_{0|0}+P_{0|1}-2\lambda_0\right)^2}\right)- \\
&\quad 2\frac{Acov\left(\widehat{\theta}_1+\widehat{\theta}_2,\widehat{\theta}_0+\widehat{\theta}_3\right)}{\left(P_{1|1}+P_{1|0}-2\lambda_1\right)\left(P_{0|0}+P_{0|1}-2\lambda_0\right)}, \\
\Omega_d &= \left(\frac{\sigma^2_{p_0,(p_1)}+\sigma^2_{q_0,(p_1)}-2\sigma_{p_0,q_0,(p_1)}}{\left(P_{0|1}-\lambda_0\right)^2}\right)+\left(\frac{\sigma^2_{p_1,(q_0)}+\sigma^2_{q_1,(q_0)}-2\sigma_{p_1,q_1,(q_0)}}{\left(P_{1|0}-\lambda_1\right)^2}\right)- \\
&\quad 2\frac{Acov\left(\widehat{\theta}_2,\widehat{\theta}_3\right)}{\left(P_{1|0}-\lambda_1\right)\left(P_{0|1}-\lambda_0\right)}.
\end{aligned}$$

This completes the proof of Theorem 1. □

E Alternative estimators

E.1 Two alternative estimators

We subsequently discuss two alternative estimation approaches to the one presented in the main text, even though all 3 are equivalent asymptotically. Lemmas 2 and 3 above show that an alternative approach to $\widehat{\mu}_c$ (and $\widehat{\mu}_{c,d}, \widehat{\mu}_d$) in the main text can be based on

$$\widehat{\theta}_0^a = \frac{1}{n}\sum_{i=1}^{n} Y_i \left(\frac{1(D_i=0)\,1(Z_i=0)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=0)} - \frac{1(D_i=0)\,1(Z_i=1)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=1)} \right) 1(Y_i \in C_{p_0}), \quad \text{(E.1)}$$

$$\widehat{\theta}_1^a = \frac{1}{n}\sum_{i=1}^{n} Y_i \left(\frac{1(D_i=1)\,1(Z_i=1)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=1)} - \frac{1(D_i=1)\,1(Z_i=0)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=0)} \right) 1\left(Y_i \in C_{q_1}\right), \quad \text{(E.2)}$$

$$\widehat{\theta}_2^a = \frac{1}{n}\sum_{i=1}^{n} Y_i \left(\frac{1(D_i=1)\,1(Z_i=1)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=1)} - \frac{1(D_i=1)\,1(Z_i=0)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=0)} \right) 1\left(Y_i \in C_{p_1}\right), \quad \text{(E.3)}$$

$$\widehat{\theta}_3^a = \frac{1}{n}\sum_{i=1}^{n} Y_i \left(\frac{1(D_i=1)\,1(Z_i=1)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=1)} - \frac{1(D_i=1)\,1(Z_i=0)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=0)} \right) 1\left(Y_i \in C_{q_0}\right). \quad \text{(E.4)}$$

In addition one can exploit that $\min(a,b) = \frac{1}{2}(a+b) - \frac{1}{2}|a-b|$ and write the expression for λ_d, for $d = 0, 1$, as follows:

$$\begin{aligned}
\lambda_d &= \int_{\mathcal{Y}} \min(p_d(y), q_d(y))\,dy = \frac{1}{2}\int_{\mathcal{Y}} (p_d(y) + q_d(y))\,dy - \frac{1}{2}\int_{\mathcal{Y}} |p_d(y) - q_d(y)|\,dy \\
&= \frac{1}{2}\int_{\mathcal{Y}} (p_d(y) + q_d(y))\,dy - \frac{1}{2}\int_{C_{p_d}} (q_d(y) - p_d(y))\,dy - \frac{1}{2}\int_{C_{q_d}} (p_d(y) - q_d(y))\,dy,
\end{aligned}$$

since $\frac{1}{2}\int_{C_{p_d,q_d}} |p_d(y) - q_d(y)|\,dy = 0$. Alternative estimators of λ_0 and λ_1 can therefore be obtained as

$$\begin{aligned}
\widehat{\lambda}_d^a = {} & \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} \left(\frac{1(D_i=d)\,1(Z_i=1)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=1)} + \frac{1(D_i=d)\,1(Z_i=0)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=0)} \right) \right) - \qquad \text{(E.5)} \\
& \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} \left(\frac{1(D_i=d)\,1(Z_i=0)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=0)} - \frac{1(D_i=d)\,1(Z_i=1)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=1)} \right) 1(Y_i \in C_{p_d}) \right) - \\
& \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} \left(\frac{1(D_i=d)\,1(Z_i=1)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=1)} - \frac{1(D_i=d)\,1(Z_i=0)}{\frac{1}{n}\sum_{i=1}^n 1(Z_i=0)} \right) 1\left(Y_i \in C_{q_d}\right) \right),
\end{aligned}$$

for $d = 0, 1$. We denote by $\widehat{\mu}^a_{c,d}, \widehat{\mu}^a_c$, and $\widehat{\mu}^a_d$ the LATE estimators making use of $\widehat{\theta}^a_i$ for $i = 0, .., 3$ and $\widehat{\lambda}^a_d$ for $d = 0, 1$ instead of the expressions offered in the main text.

A further set of alternative and asymptotically equivalent LATE estimators can be derived using a computationally convenient approximation of the integrals involved in all our parameters. This approximation uses the fact that a proper density must integrate to 1. Let $Y^{\min} = \min(Y_1, \ldots, Y_n)$ and $Y^{\max} = \max(Y_1, \ldots, Y_n)$ the minimum and the maximum sample values of Y, respectively, $\Delta = \frac{Y^{\max} - Y^{\min}}{k}$, and $Y^*_i = Y^{\min} + i\Delta$. It can be shown (a proof i sketched in the next subsection) that as $k \to \infty$, estimators based on the parameters defined below are equivalent to the plug-in estimator:

$$\widehat{\theta}^b_0 = \frac{1}{k}\sum_{i=1}^{k} Y^*_i \left(\frac{\widehat{f}\left(Y^*_i, D=0, Z=0\right) \widehat{P}_{0|0}}{\sum_{i=1}^{n} \widehat{f}\left(Y^*_i, D=0, Z=0\right)} - \frac{\widehat{f}\left(Y^*_i, D=0, Z=1\right) \widehat{P}_{0|1}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=0, Z=1\right)} \right) 1\left(Y^*_i \in C_{p_0}\right), \tag{E.6}$$

$$\widehat{\theta}^b_1 = \frac{1}{k}\sum_{i=1}^{k} Y^*_i \left(\frac{\widehat{f}\left(Y^*_i, D=1, Z=1\right) \widehat{P}_{1|1}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=1, Z=1\right)} - \frac{\widehat{f}\left(Y^*_i, D=1, Z=0\right) \widehat{P}_{1|0}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=1, Z=0\right)} \right) 1\left(Y^*_i \in C_{q_1}\right), \tag{E.7}$$

$$\widehat{\theta}^b_2 = \frac{1}{k}\sum_{i=1}^{k} Y^*_i \left(\frac{\widehat{f}\left(Y^*_i, D=1, Z=0\right) \widehat{P}_{1|0}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=1, Z=0\right)} - \frac{\widehat{f}\left(Y^*_i, D=1, Z=1\right) \widehat{P}_{1|1}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=1, Z=1\right)} \right) 1\left(Y^*_i \in C_{p_1}\right), \tag{E.8}$$

$$\widehat{\theta}^b_3 = \frac{1}{k}\sum_{i=1}^{k} Y^*_i \left(\frac{\widehat{f}\left(Y^*_i, D=0, Z=1\right) \widehat{P}_{0|1}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=0, Z=1\right)} - \frac{\widehat{f}\left(Y^*_i, D=0, Z=0\right) \widehat{P}_{0|0}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=0, Z=0\right)} \right) 1\left(Y^*_i \in C_{q_0}\right), \tag{E.9}$$

with

$$\widehat{f}_{Y,D,Z}\left(Y^*_i, D=d, Z=z\right) \equiv \widehat{f}\left(Y^*_i, D=d, Z=z\right),$$

$$\widehat{P}_{d|z} = \frac{1}{n}\sum_{i=1}^{n} \frac{I\left(D_i = d\right) I\left(Z_i = z\right)}{\frac{1}{n}\sum_{i=1}^{n} I\left(Z_i = z\right)}, \quad d, z = 0, 1,$$

and

$$\begin{aligned}
\widehat{\lambda}^b_d = {} & \frac{1}{2}\sum_{i=1}^{k} \left(\frac{\widehat{f}\left(Y^*_i, D=d, Z=1\right) \widehat{P}_{d|1}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=d, Z=1\right)} + \frac{\widehat{f}\left(Y^*_i, D=d, Z=0\right) \widehat{P}_{d|0}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=d, Z=0\right)} \right) - \\
& \frac{1}{2}\sum_{i=1}^{k} \left(\frac{\widehat{f}\left(Y^*_i, D=d, Z=1\right) \widehat{P}_{d|1}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=d, Z=1\right)} - \frac{\widehat{f}\left(Y^*_i, D=d, Z=0\right) \widehat{P}_{d|0}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=d, Z=0\right)} \right) 1\left(Y^*_i \in C_{q_d}\right) - \\
& \frac{1}{2}\sum_{i=1}^{k} \left(\frac{\widehat{f}\left(Y^*_i, D=d, Z=0\right) \widehat{P}_{d|0}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=d, Z=0\right)} - \frac{\widehat{f}\left(Y^*_i, D=d, Z=1\right) \widehat{P}_{d|1}}{\sum_{i=1}^{k} \widehat{f}\left(Y^*_i, D=d, Z=1\right)} \right) 1\left(Y^*_i \in C_{p_d}\right),
\end{aligned} \tag{E.10}$$

for $d = 0, 1$. We denote by $\widehat{\mu}^b_{c,d}, \widehat{\mu}^b_c$, and $\widehat{\mu}^b_d$ the LATE estimators that make use of $\widehat{\theta}^b_i$ for $i = 0, .., 3$ and $\widehat{\lambda}^b_d$ for $d = 0, 1$.

E.2 Alternative approximation of the integral of the minimum of two density functions

Consider the integral

$$\beta = \int_a^b \min(f_x(z), f_y(z))dz.$$

Let $\Delta = \frac{b-a}{k}$, using the fact that such an integral can be approximated with a Riemman sum we can write

$$\beta_1 = \sum_{i=0}^{k} \min\left(f_x\left(a + i\Delta\right), f_y\left(a + i\Delta\right)\right)\Delta$$

Let $\theta_j = \sum_{i=0}^{k} f_j(a + i\Delta)$ for j=x,y. Note that for $j = x, y$ $\theta_j\Delta \approx 1$ provided that f_x and f_y are proper pdfs. Then we can write

$$\begin{aligned} \beta_2 &= \sum_{i=0}^{k} \min\left(\frac{f_x(a+i\Delta)}{\theta_x\Delta}, \frac{f_y(a+i\Delta)}{\theta_y\Delta}\right)\Delta \\ &= \sum_{i=0}^{k} \min\left(\frac{f_x(a+i\Delta)}{\theta_x}, \frac{f_y(a+i\Delta)}{\theta_y}\right) \end{aligned}$$

This shows that β_2 is also a valid approximation for e as $k \to \infty$. Using this result and simple algebra it is easy to derive the parameters defined at the end of the previous subsection.

F Simulations

The simulation study will be divided into two main parts. In the first part, the limited sample distributions of the LATE estimators are simulated and compared to the asymptotic distributions of Theorem 1. In the second part, simulations are used to illustrate the basic properties of the finite sample LATE point estimates in terms of bias and efficiency.

For all the simulation results presented, the data is assumed to be generated from the following model:

$$\begin{aligned}
Y(1)|T=a &\sim \mathcal{N}(0,1), \quad Y(0)|T=n \sim \mathcal{N}(0,1),\\
Y(d)|T=c &\sim d\cdot\mathcal{N}_T(2,.25,0.5,5)+(1-d)\cdot\mathcal{N}_T(-2,.25,-5,.5),\\
Y(d)|T=d &\sim d\cdot\mathcal{N}_T(-2,.25,-5,-.5)+(1-d)\cdot\mathcal{N}_T(2,.25,.5,5),\\
\Pr(T=a) &= 0.5\cdot(1-\Pr(T=c))\,, \quad \Pr(T=n)=\Pr(T=a)-\Pr(T=d), \quad Z\sim \text{Bernoulli}(0.5).
\end{aligned}$$

$\mathcal{N}(0,1)$ denotes a standard normal distribution, while $\mathcal{N}_T(\mu,\sigma^2,A,B)$ is the truncated normal distribution on the interval $[A,B]$.

In this setup, the shares of always and never takers are defined as functions of the shares of compliers and defiers, which are to be defined. Note that the random assignment of Z implies the satisfaction of Assumption 1. Furthermore, Assumption 2 is satisfied because the supports of compliers' and defiers' potential outcomes do not overlap in either treatment state. Finally, it follows from the definition of types that $D=Z$ if $T=c$, $D=1-Z$ if $T=d$, $D=1$ if $T=a$, and $D=0$ if $T=n$. In our benchmark scenario we fix the complier and defier shares to $\Pr(T=c)=0.20$ and $\Pr(T=d)=0.15$, respectively. Consequently, the LATE parameters of primary interest in the population are $\mu_{c,d}=0.57$, $\mu_c=4.00$ and $\mu_d=-4.00$. The level sets associated with the model are given as $C_{p_0}=(-5;-.5]\,, C_{q_0}=(.5;5]\,, C_{p_1}=(-5;-.5]$ and $C_{q_1}=(.5;5]$.

F.1 One-sided coverage probabilities

According to Theorem 1 the asymptotic or nominal one-sided coverage probability is given by

$$P\left(\left(\widehat{\mu}_n - z_q\left(\widehat{\sigma}_\infty/\sqrt{n}\right)\right)\leq \mu\right)=q, \tag{F.1}$$

where $P(\cdot)$ denotes the standard normal distribution and z_q is the assocated q'th quantile and $\widehat{\sigma}_\infty$ equals the estimated asymptotic variance of $\widehat{\mu}_n$. Equation (F.1) implies that if the sample size is sufficiently large, then the proportion of times the quantity $(\widehat{\mu}_n - z_q(\widehat{\sigma}_\infty/\sqrt{n}))$ is observed to be smaller than μ will equal q. This suggests that in order to evaluate how well the asymptotic distributions of Theorem 1 approximate the finite sample distributions of the LATE estimators, the following simulated quantity can be considered

$$\widetilde{F}_{n,m}\left(\mu|q\right)=\frac{1}{m}\sum_{j=1}^{m}I\left(\left(\widehat{\mu}_{n}^{j}-z_{q}\left(\widehat{\sigma}_{\infty}/\sqrt{n}\right)\right)\leq\mu\right), \tag{F.2}$$

for $q\in(0,1)$, where $I(A)$ is the indicator function that equals unity if A is true and is zero otherwise. m denotes the number of Monte Carlo replications, n is the sample size, while $\widehat{\mu}_n^j$ denotes the LATE estimator based on simulation sample j. The asymptotic variance is computed only once based on an independently simulated sample of 20000 observations. If we observe that $\widetilde{F}_{n,m}\left(\mu|q\right)\approx q$ for all $q\in(0,1)$ we take this as evidence that the asymptotic distribution provides a satisfying approximation for the finite sample distribution of $\widehat{\mu}_n$ for a sample size equal to n.

In Tables 4 - 6 results on the simulated coverage probabilities under the assumption that the level sets $C_{p_0}, C_{q_0}, C_{p_1}$, and C_{q_1} are known are reported for sample sizes $n=(3200,6400,12000)$ and quantiles $q=(0.025,0.05,0.10,0.,25,0.5,0.75,0.95,0.0975)$. Tables 4 - 6 differ only in the choice of bandwidths, denoted by h. In Table 4, results are based on the Silverman (1986) rule-of-thumb bandwidth (denoted by b), i.e., $h=1.06\cdot s\cdot n^{-\frac{1}{5}}=b$, where s is the sample standard deviation. Even in large samples, the simulated coverage probabilities of the estimators that depend on the choice of a bandwidth, ($\widehat{\mu}_{c,d}$, $\widehat{\mu}_c$, $\widehat{\mu}_d$ (see the main text) and $\widehat{\mu}_{c,d}^b$), $\widehat{\mu}_c^b$), $\widehat{\mu}_d^b$) (see Section E in this appendix) do not always closely match the nominal rates. This is not too surprising, since the rule-of-thumb bandwidth does not satisfy the condition

$$\underset{n\to\infty}{lim}\sqrt{n}h^2=0, \tag{F.3}$$

which is one of the key assumptions of Theorem 1. For the bandwidth-free LATE estimators $\widehat{\mu}_{c,d}^b, \widehat{\mu}_c^b$ and $\widehat{\mu}_d^b$ (see Section E in this appendix), the results of Table 4 are encouraging, particularly for samples sizes $n=(6400,12800)$, as the simulated coverage probabilities are generally close to the nominal ones. Notice that the results for the bandwidth-free estimators are by construction the same across Tables 4 - 6 because bluntly speaking, $h=0$ for all n. Therefore, the bandwidth-free estimators can be interpreted as limits of the estimators of Theorem 1 such that condition (F.3) is always satisfied for $\widehat{\mu}_{c,d}^b, \widehat{\mu}_c^b$ and $\widehat{\mu}_d^b$.

In Table 5, the results are reported for $b^{3/2}$ as in Anderson, Linton, and Whang (2012) and for this choice the condition in (F.3) is satisfied. Indeed, the correspondence of nominal and simulated coverage rates are improving for all sample sizes for the two bandwidth-dependent LATE estimators. When undersmooting even more severly by setting the bandwidth equal to $b^{5/2}$, the results improve further, as illustrated in Table 6. This suggests that at least in our simulations, the asymptotic distribution of the LATE estimators approximates their finite sample behavior decently if the plug-in densities are sufficiently undersmoothed when taking the Silverman rule-of-thumb bandwidth as reference.

Table 4: Simulated coverage probabilities under known level sets using the Silverman rule-of-thumb bandwidth (b)

	LATEcd			**LATEc**			**LATEd**		
	$\widehat{F}(\widehat{\mu}_{c,d})$	$\widehat{F}(\widehat{\mu}^a_{c,d})$	$\widehat{F}(\widehat{\mu}^b_{c,d})$	$\widehat{F}(\widehat{\mu}_c)$	$\widehat{F}(\widehat{\mu}^a_c)$	$\widehat{F}(\widehat{\mu}^b_c)$	$\widehat{F}(\widehat{\mu}_d)$	$\widehat{F}(\widehat{\mu}^a_d)$	$\widehat{F}(\widehat{\mu}^b_d)$
n=3200									
q=0.025	0.021	0.023	0.024	0.012	0.023	0.012	0.045	0.053	0.045
q=0.050	0.042	0.046	0.045	0.024	0.048	0.025	0.068	0.082	0.068
q=0.250	0.226	0.241	0.232	0.199	0.226	0.198	0.287	0.279	0.285
q=0.500	0.495	0.503	0.506	0.466	0.465	0.466	0.540	0.538	0.538
q=0.750	0.745	0.758	0.754	0.723	0.735	0.724	0.771	0.756	0.767
q=0.950	0.947	0.952	0.948	0.943	0.923	0.945	0.964	0.955	0.964
q=0.975	0.975	0.970	0.975	0.970	0.962	0.970	0.984	0.979	0.983
n=6400									
q=0.025	0.029	0.028	0.028	0.014	0.028	0.014	0.035	0.028	0.035
q=0.050	0.061	0.060	0.062	0.037	0.060	0.037	0.067	0.059	0.069
q=0.250	0.271	0.270	0.278	0.212	0.256	0.213	0.276	0.255	0.275
q=0.500	0.511	0.520	0.506	0.442	0.497	0.443	0.537	0.477	0.537
q=0.750	0.769	0.776	0.769	0.719	0.749	0.720	0.783	0.740	0.781
q=0.950	0.952	0.949	0.952	0.933	0.940	0.932	0.966	0.946	0.966
q=0.975	0.971	0.973	0.972	0.969	0.970	0.970	0.990	0.979	0.990
n=12800									
q=0.025	0.027	0.028	0.029	0.012	0.024	0.012	0.046	0.035	0.046
q=0.050	0.053	0.054	0.054	0.032	0.051	0.032	0.078	0.054	0.078
q=0.250	0.243	0.252	0.243	0.165	0.256	0.164	0.293	0.242	0.294
q=0.500	0.506	0.506	0.502	0.419	0.507	0.419	0.542	0.464	0.542
q=0.750	0.750	0.762	0.752	0.682	0.730	0.682	0.802	0.754	0.801
q=0.950	0.954	0.945	0.955	0.917	0.949	0.917	0.970	0.960	0.970
q=0.975	0.978	0.977	0.978	0.954	0.977	0.954	0.989	0.981	0.989

Note: The nominal coverage probabilities (quantiles of the asymptotic distribution according to Theorem 1) are given in the first column. $\widehat{F}(\widehat{\mu}_{c,d})$, $\widehat{F}(\widehat{\mu}_c)$, $\widehat{F}(\widehat{\mu}_d)$ denote the estimated distributions based on the respective LATE estimators on compliers and/or defiers in the main text. $\widehat{F}(\widehat{\mu}^a_{c,d})$, $\widehat{F}(\widehat{\mu}^a_c)$, $\widehat{F}(\widehat{\mu}^a_d)$ denote the estimates for the bandwidth-free estimators, while $\widehat{F}(\widehat{\mu}^b_{c,d})$, $\widehat{F}(\widehat{\mu}^b_c)$, $\widehat{F}(\widehat{\mu}^b_d)$ denote the estimates for the modified plug-in estimators, see Section E in this appendix. The number of replications equals 6000.

Table 5: Simulated coverage probabilities under known level sets using bandwidth $b^{3/2}$

	LATEcd			**LATEc**			**LATEd**		
	$\widehat{F}(\widehat{\mu}_{c,d})$	$\widehat{F}(\widehat{\mu}^a_{c,d})$	$\widehat{F}(\widehat{\mu}^b_{c,d})$	$\widehat{F}(\widehat{\mu}_c)$	$\widehat{F}(\widehat{\mu}^a_c)$	$\widehat{F}(\widehat{\mu}^b_c)$	$\widehat{F}(\widehat{\mu}_d)$	$\widehat{F}(\widehat{\mu}^a_d)$	$\widehat{F}(\widehat{\mu}^b_d)$
n=3200									
q=0.025	0.021	0.023	0.023	0.011	0.023	0.011	0.046	0.053	0.046
q=0.050	0.046	0.046	0.045	0.034	0.048	0.035	0.067	0.082	0.067
q=0.250	0.235	0.241	0.232	0.218	0.226	0.217	0.273	0.279	0.268
q=0.500	0.504	0.503	0.509	0.489	0.465	0.489	0.523	0.538	0.525
q=0.750	0.762	0.758	0.760	0.737	0.735	0.737	0.766	0.756	0.767
q=0.950	0.948	0.952	0.950	0.941	0.923	0.941	0.961	0.955	0.960
q=0.975	0.976	0.970	0.976	0.971	0.962	0.972	0.982	0.979	0.982
n=6400									
q=0.025	0.028	0.028	0.029	0.023	0.028	0.023	0.031	0.028	0.031
q=0.050	0.061	0.060	0.063	0.043	0.060	0.043	0.056	0.059	0.056
q=0.250	0.278	0.270	0.279	0.251	0.256	0.251	0.248	0.255	0.246
q=0.500	0.519	0.520	0.517	0.490	0.497	0.490	0.491	0.477	0.492
q=0.750	0.772	0.776	0.778	0.753	0.749	0.755	0.747	0.740	0.747
q=0.950	0.949	0.949	0.951	0.945	0.940	0.945	0.952	0.946	0.952
q=0.975	0.972	0.973	0.972	0.982	0.970	0.983	0.982	0.979	0.982
n=12800									
q=0.025	0.028	0.028	0.027	0.021	0.024	0.021	0.041	0.035	0.041
q=0.050	0.055	0.054	0.056	0.048	0.051	0.049	0.065	0.054	0.065
q=0.250	0.246	0.252	0.252	0.230	0.256	0.230	0.247	0.242	0.248
q=0.500	0.504	0.506	0.505	0.498	0.507	0.498	0.480	0.464	0.480
q=0.750	0.760	0.762	0.761	0.732	0.730	0.730	0.762	0.754	0.763
q=0.950	0.949	0.945	0.950	0.942	0.949	0.942	0.964	0.960	0.963
q=0.975	0.979	0.977	0.977	0.968	0.977	0.968	0.982	0.981	0.982

Note: The nominal coverage probabilities (quantiles of the asymptotic distribution according to Theorem 1) are given in the first column. $\widehat{F}(\widehat{\mu}_{c,d})$, $\widehat{F}(\widehat{\mu}_c)$, $\widehat{F}(\widehat{\mu}_d)$ denote the estimated distributions based on the respective LATE estimators on compliers and/or defiers in the main text. $\widehat{F}(\widehat{\mu}^a_{c,d})$, $\widehat{F}(\widehat{\mu}^a_c)$, $\widehat{F}(\widehat{\mu}^a_d)$ denote the estimates for the bandwidth-free estimators, while $\widehat{F}(\widehat{\mu}^b_{c,d})$, $\widehat{F}(\widehat{\mu}^b_c)$, $\widehat{F}(\widehat{\mu}^b_d)$ denote the estimates for the modified plug-in estimators, see Section E in this appendix. The number of replications equals 6000.

Table 6: Simulated coverage probabilities under known level sets using bandwidth $b^{5/2}$

	LATEcd			**LATEc**			**LATEd**		
	$\widehat{F}(\widehat{\mu}_{c,d})$	$\widehat{F}(\widehat{\mu}^a_{c,d})$	$\widehat{F}(\widehat{\mu}^b_{c,d})$	$\widehat{F}(\widehat{\mu}_c)$	$\widehat{F}(\widehat{\mu}^a_c)$	$\widehat{F}(\widehat{\mu}^b_c)$	$\widehat{F}(\widehat{\mu}_d)$	$\widehat{F}(\widehat{\mu}^a_d)$	$\widehat{F}(\widehat{\mu}^b_d)$
n=3200									
q=0.025	0.019	0.023	0.021	0.017	0.023	0.017	0.051	0.053	0.051
q=0.050	0.045	0.046	0.046	0.050	0.048	0.050	0.074	0.082	0.074
q=0.250	0.236	0.241	0.236	0.220	0.226	0.220	0.270	0.279	0.268
q=0.500	0.507	0.503	0.507	0.493	0.465	0.493	0.525	0.538	0.527
q=0.750	0.760	0.758	0.760	0.738	0.735	0.738	0.761	0.756	0.762
q=0.950	0.949	0.952	0.950	0.939	0.923	0.939	0.957	0.955	0.957
q=0.975	0.972	0.970	0.974	0.965	0.962	0.965	0.979	0.979	0.979
n=6400									
q=0.025	0.025	0.028	0.025	0.028	0.028	0.028	0.025	0.028	0.027
q=0.050	0.063	0.060	0.062	0.054	0.060	0.054	0.051	0.059	0.051
q=0.250	0.271	0.270	0.271	0.271	0.256	0.271	0.248	0.255	0.247
q=0.500	0.522	0.520	0.522	0.511	0.497	0.512	0.472	0.477	0.474
q=0.750	0.775	0.776	0.778	0.760	0.749	0.760	0.737	0.740	0.736
q=0.950	0.949	0.949	0.950	0.950	0.940	0.949	0.945	0.946	0.944
q=0.975	0.975	0.973	0.975	0.979	0.970	0.979	0.976	0.979	0.977
n=12800									
q=0.025	0.028	0.028	0.028	0.026	0.024	0.026	0.035	0.035	0.036
q=0.050	0.054	0.054	0.052	0.054	0.051	0.054	0.060	0.054	0.059
q=0.250	0.246	0.252	0.248	0.257	0.256	0.259	0.240	0.242	0.240
q=0.500	0.512	0.506	0.511	0.509	0.507	0.510	0.453	0.464	0.454
q=0.750	0.757	0.762	0.758	0.738	0.730	0.739	0.743	0.754	0.743
q=0.950	0.947	0.945	0.947	0.944	0.949	0.944	0.961	0.960	0.960
q=0.975	0.980	0.977	0.980	0.973	0.977	0.973	0.980	0.981	0.980

Note: The nominal coverage probabilities (quantiles of the asymptotic distribution according to Theorem 1) are given in the first column. $\widehat{F}(\widehat{\mu}_{c,d})$, $\widehat{F}(\widehat{\mu}_c)$, $\widehat{F}(\widehat{\mu}_d)$ denote the estimated distributions based on the respective LATE estimators on compliers and/or defiers in the main text. $\widehat{F}(\widehat{\mu}^a_{c,d})$, $\widehat{F}(\widehat{\mu}^a_c)$, $\widehat{F}(\widehat{\mu}^a_d)$ denote the estimates for the bandwidth-free estimators, while $\widehat{F}(\widehat{\mu}^b_{c,d})$, $\widehat{F}(\widehat{\mu}^b_c)$, $\widehat{F}(\widehat{\mu}^b_d)$ denote the estimates for the modified plug-in estimators, see Section E in this appendix. The number of replications equals 6000.

F.2 Estimation bias and efficiency

Tables 7-9 report results on bias and efficiency under the various sample sizes of the various LATE estimators for bandwidths b, $b^{3/2}$, $b^{5/2}$, respectively, and of 2SLS. Common to all three tables is that the level sets are again assumed to be known. In terms of simulated average bias (**BIAS**) and simulated median bias (**MBIAS**) the bandwidth-free LATE estimators $\widehat{\mu}^b_{c,d}$, $\widehat{\mu}^b_c$, and $\widehat{\mu}^b_d$ perform very well and the bias measures are very close to zero even for a moderate sample size of $n = 800$. In fact, the bandwidth-free LATE estimators outperform the alternative estimators with bandwidth choices b and $b^{3/2}$ uniformly over all sample sizes. When setting the bandwidth to $b^{5/2}$ all LATE estimators perform similarly well and for all sample sizes the bias measures are close to zero. In terms of efficiency, measured by the standard deviation of the estimators (**SD**), and root mean squared error (**RMSE**), the differences across the three types of estimators are minor across bandwidths for all sample sizes. For all methods the simulated bias terms and standard deviations approach zero as the sample size increases, confirming they are consistent as implied by Theorem 1. In contrast, the 2SLS estimator remains severely biased and imprecise in all simulations even when $n = 12800$.

Next, we investigate the finite sample behavior of the methods when the level sets are assumed to be unknown and estimated by kernel methods according to equations (42) in the main text. In Table 10, the results are reported when setting the bandwidth to $b^{3/2}$ in any of the kernel procedures. Not surprisingly, the simulated biases and standard deviations of all of our suggested LATE estimators increase for small to moderate sample sizes when the level sets need to be estimated. However, the effects of relying on estimated rather than the true level sets appears to vanish in large samples. For $n = 12800$, the biases are all close to zero and of similar magnitudes as in Table 8, while the simulated standard errors are are only slightly larger than those in Table 8.

Finally, we consider the case that there are no defiers, $P(T = d) = 0$, such that global monotonicity holds. When the level sets are known, our LATE estimators on the compliers are about four times more precise than the (also consistent) 2SLS estimator and interestingly generally also less biased, see Table 11. When the level sets are unknown, there estimation introduces some bias to our methods as can be seen from Table 12. However, our LATE estimators for the compliers still clearly dominate 2SLS, whose standard deviations and RMSEs are at least twice as high as those of our methods for any sample size.

Table 7: Simulation results on the performance of LATE estimators under known level sets using the Silverman rule-of-thumb bandwidth (b)

	LATEcd			**LATEc**			**LATEd**			**LATE**
	$\widehat{\mu}_{c,d}$	$\widehat{\mu}^a_{c,d}$	$\widehat{\mu}^b_{c,d}$	$\widehat{\mu}_c$	$\widehat{\mu}^a_c$	$\widehat{\mu}^b_c$	$\widehat{\mu}_d$	$\widehat{\mu}^a_d$	$\widehat{\mu}^b_d$	2SLS
RMSE										
n=800	0.329	0.356	0.328	0.142	0.185	0.142	0.187	0.257	0.186	722.496
n=1600	0.230	0.254	0.229	0.098	0.128	0.097	0.125	0.168	0.125	483.395
n=3200	0.163	0.175	0.162	0.071	0.089	0.071	0.097	0.121	0.097	263.379
n=6400	0.116	0.124	0.116	0.051	0.062	0.051	0.068	0.081	0.068	28.642
n=12800	0.081	0.087	0.081	0.037	0.042	0.037	0.048	0.056	0.048	25.609
BIAS										
n=800	0.013	0.017	0.024	−0.036	0.011	−0.037	0.026	−0.026	0.022	0.425
n=1600	−0.003	0.000	0.002	0.001	0.017	0.000	0.001	−0.017	0.000	47.853
n=3200	0.003	0.002	0.005	0.009	0.008	0.009	−0.010	−0.011	−0.011	35.703
n=6400	−0.004	−0.006	−0.004	0.007	−0.001	0.007	−0.007	0.001	−0.007	26.534
n=12800	0.000	−0.002	0.000	0.008	0.000	0.008	−0.008	0.001	−0.008	24.982
MBIAS										
n=800	0.016	0.019	0.031	−0.046	−0.011	−0.048	0.035	−0.002	0.033	19.258
n=1600	−0.009	−0.008	−0.006	−0.005	0.009	−0.004	0.004	−0.006	0.001	23.995
n=3200	−0.001	−0.001	0.001	0.007	0.007	0.006	−0.011	−0.011	−0.013	24.292
n=6400	−0.003	−0.007	−0.003	0.007	0.000	0.008	−0.006	0.004	−0.006	24.056
n=12800	−0.001	−0.002	−0.001	0.006	−0.001	0.007	−0.005	0.006	−0.005	24.058
SD										
n=800	0.329	0.356	0.327	0.138	0.185	0.137	0.185	0.256	0.185	722.858
n=1600	0.230	0.254	0.229	0.098	0.127	0.097	0.125	0.167	0.125	481.261
n=3200	0.163	0.175	0.162	0.070	0.089	0.070	0.096	0.121	0.096	261.079
n=6400	0.116	0.124	0.116	0.051	0.062	0.050	0.067	0.081	0.067	10.789
n=12800	0.081	0.087	0.081	0.036	0.042	0.036	0.047	0.056	0.047	5.633

Note: **BIAS**, **MBIAS**, **SD**, and **RMSE** provide the simulated average bias, simulated median bias, standard deviation, and root mean squared error of the LATE estimators, respectively. $\widehat{\mu}_{c,d}$, $\widehat{\mu}_c$, $\widehat{\mu}_d$ denote the LATE estimators on compliers and/or defiers discussed in the main text. $\widehat{\mu}^a_{c,d}$, $\widehat{\mu}^a_c$), $\widehat{\mu}^a_d$ denote the bandwidth-free estimators, while $\widehat{\mu}^b_{c,d}$, $\widehat{\mu}^b_c$, $\widehat{\mu}^b_d$ denote the bias corrected estimators, see Section E in this appendix. The number of replications equals 6000.

Table 8: Simulation results on the performance of LATE estimators under known level sets using the bandwidth $b^{3/2}$

Performance: Known level sets	**LATEcd**			**LATEc**			**LATEd**			**LATE**
	$\widehat{\mu}_{c,d}$	$\widehat{\mu}^a_{c,d}$	$\widehat{\mu}^b_{c,d}$	$\widehat{\mu}_c$	$\widehat{\mu}^a_c$	$\widehat{\mu}^b_c$	$\widehat{\mu}_d$	$\widehat{\mu}^a_d$	$\widehat{\mu}^b_d$	2SLS
RMSE										
n=800	0.333	0.356	0.332	0.151	0.185	0.150	0.204	0.257	0.203	722.496
n=1600	0.236	0.254	0.235	0.107	0.128	0.106	0.138	0.168	0.138	483.395
n=3200	0.167	0.175	0.167	0.076	0.089	0.076	0.106	0.121	0.106	263.379
n=6400	0.120	0.124	0.120	0.054	0.062	0.054	0.073	0.081	0.073	28.642
n=12800	0.084	0.087	0.084	0.039	0.042	0.039	0.051	0.056	0.051	25.609
BIAS										
n=800	0.013	0.017	0.020	−0.022	0.011	−0.022	0.011	−0.026	0.009	0.425
n=1600	−0.003	0.000	0.000	0.004	0.017	0.004	−0.003	−0.017	−0.003	47.853
n=3200	0.003	0.002	0.004	0.006	0.008	0.006	−0.008	−0.011	−0.008	35.703
n=6400	−0.006	−0.006	−0.006	0.000	−0.001	0.000	−0.001	0.001	−0.001	26.534
n=12800	−0.001	−0.002	−0.001	0.002	0.000	0.002	−0.001	0.001	−0.001	24.982
MBIAS										
n=800	0.008	0.019	0.023	−0.033	−0.011	−0.034	0.026	−0.002	0.025	19.258
n=1600	−0.014	−0.008	−0.007	−0.001	0.009	−0.001	0.007	−0.006	0.007	23.995
n=3200	−0.003	−0.001	−0.002	0.002	0.007	0.002	−0.008	−0.011	−0.010	24.292
n=6400	−0.007	−0.007	−0.007	0.001	0.000	0.001	0.001	0.004	0.001	24.056
n=12800	−0.003	−0.002	−0.002	0.000	−0.001	0.000	0.003	0.006	0.003	24.058
SD										
n=800	0.333	0.356	0.332	0.149	0.185	0.148	0.203	0.256	0.203	722.858
n=1600	0.236	0.254	0.236	0.106	0.127	0.106	0.138	0.167	0.138	481.261
n=3200	0.167	0.175	0.167	0.076	0.089	0.076	0.105	0.121	0.105	261.079
n=6400	0.120	0.124	0.120	0.055	0.062	0.055	0.073	0.081	0.073	10.789
n=12800	0.084	0.087	0.084	0.039	0.042	0.039	0.051	0.056	0.051	5.633

Note: **BIAS**, **MBIAS**, **SD**, and **RMSE** provide the simulated average bias, simulated median bias, standard deviation, and root mean squared error of the LATE estimators, respectively. $\widehat{\mu}_{c,d}$, $\widehat{\mu}_c$, $\widehat{\mu}_d$ denote the LATE estimators on compliers and/or defiers discussed in the main text. $\widehat{\mu}^a_{c,d}$, $\widehat{\mu}^a_c$), $\widehat{\mu}^a_d$ denote the bandwidth-free estimators, while $\widehat{\mu}^b_{c,d}$, $\widehat{\mu}^b_c$, $\widehat{\mu}^b_d$ denote the bias corrected estimators, see Section E in this appendix. The number of replications equals 6000.

Table 9: Simulation results on the performance of LATE estimators under known level sets using the bandwidth $b^{5/2}$

Performance: Known level sets	**LATEcd**			**LATEc**			**LATEd**			**LATE**
	$\widehat{\mu}_{c,d}$	$\widehat{\mu}^a_{c,d}$	$\widehat{\mu}^b_{c,d}$	$\widehat{\mu}_c$	$\widehat{\mu}^a_c$	$\widehat{\mu}^b_c$	$\widehat{\mu}_d$	$\widehat{\mu}^a_d$	$\widehat{\mu}^b_d$	2SLS
RMSE										
n=800	0.343	0.356	0.342	0.167	0.185	0.167	0.229	0.257	0.229	722.496
n=1600	0.246	0.254	0.245	0.118	0.128	0.118	0.155	0.168	0.155	483.395
n=3200	0.172	0.175	0.172	0.084	0.089	0.084	0.116	0.121	0.116	263.379
n=6400	0.123	0.124	0.123	0.059	0.062	0.059	0.079	0.081	0.079	28.642
n=12800	0.086	0.087	0.086	0.041	0.042	0.041	0.055	0.056	0.055	25.609
BIAS										
n=800	0.011	0.017	0.016	−0.008	0.011	−0.009	−0.004	−0.026	−0.005	0.425
n=1600	−0.004	0.000	−0.001	0.009	0.017	0.009	−0.008	−0.017	−0.008	47.853
n=3200	0.003	0.002	0.003	0.005	0.008	0.005	−0.009	−0.011	−0.009	35.703
n=6400	−0.007	−0.006	−0.007	−0.002	−0.001	−0.002	0.002	0.001	0.002	26.534
n=12800	−0.002	−0.002	−0.002	0.000	0.000	0.000	0.001	0.001	0.001	24.982
MBIAS										
n=800	0.008	0.019	0.011	−0.023	−0.011	−0.023	0.010	−0.002	0.011	19.258
n=1600	−0.009	−0.008	−0.009	0.006	0.009	0.006	−0.001	−0.006	−0.002	23.995
n=3200	−0.002	−0.001	−0.002	0.001	0.007	0.001	−0.010	−0.011	−0.009	24.292
n=6400	−0.006	−0.007	−0.005	−0.001	0.000	−0.001	0.004	0.004	0.004	24.056
n=12800	−0.004	−0.002	−0.003	−0.001	−0.001	−0.001	0.006	0.006	0.006	24.058
SD										
n=800	0.343	0.356	0.342	0.167	0.185	0.166	0.229	0.256	0.229	722.858
n=1600	0.246	0.254	0.245	0.117	0.127	0.117	0.154	0.167	0.155	481.261
n=3200	0.172	0.175	0.172	0.084	0.089	0.084	0.116	0.121	0.116	261.079
n=6400	0.123	0.124	0.123	0.059	0.062	0.059	0.079	0.081	0.079	10.789
n=12800	0.086	0.087	0.086	0.041	0.042	0.041	0.055	0.056	0.055	5.633

Note: **BIAS**, **MBIAS**, **SD**, and **RMSE** provide the simulated average bias, simulated median bias, standard deviation, and root mean squared error of the LATE estimators, respectively. $\widehat{\mu}_{c,d}$, $\widehat{\mu}_c$, $\widehat{\mu}_d$ denote the LATE estimators on compliers and/or defiers discussed in the main text. $\widehat{\mu}^a_{c,d}$, $\widehat{\mu}^a_c$), $\widehat{\mu}^a_d$ denote the bandwidth-free estimators, while $\widehat{\mu}^b_{c,d}$, $\widehat{\mu}^b_c$, $\widehat{\mu}^b_d$ denote the bias corrected estimators, see Section E in this appendix. The number of replications equals 6000.

Table 10: Simulation results on the performance of LATE estimators under estimated level sets using the Silverman rule-of-thumb bandwidth

Performance: Estimated level sets	**LATEcd**			**LATEc**			**LATEd**			**LATE**
	$\widehat{\mu}_{c,d}$	$\widehat{\mu}^a_{c,d}$	$\widehat{\mu}^b_{c,d}$	$\widehat{\mu}_c$	$\widehat{\mu}^a_c$	$\widehat{\mu}^b_c$	$\widehat{\mu}_d$	$\widehat{\mu}^a_d$	$\widehat{\mu}^b_d$	2SLS
RMSE										
n=800	0.313	0.335	0.314	0.189	0.209	0.189	0.199	0.219	0.198	Inf
n=1600	0.243	0.257	0.244	0.138	0.151	0.138	0.143	0.158	0.143	461.141
n=3200	0.158	0.167	0.158	0.103	0.114	0.103	0.117	0.127	0.117	57.078
n=6400	0.115	0.121	0.115	0.080	0.086	0.080	0.087	0.096	0.087	27.864
n=12800	0.080	0.082	0.080	0.060	0.065	0.060	0.064	0.069	0.064	25.806
BIAS										
n=800	−0.009	−0.014	−0.003	−0.071	−0.087	−0.072	0.027	0.054	0.026	Inf
n=1600	0.013	0.007	0.015	−0.055	−0.056	−0.055	0.024	0.036	0.024	51.680
n=3200	−0.003	−0.004	−0.002	−0.034	−0.035	−0.034	0.027	0.031	0.027	29.403
n=6400	−0.002	−0.003	−0.002	−0.023	−0.023	−0.023	0.020	0.020	0.020	26.194
n=12800	−0.005	−0.004	−0.004	−0.010	−0.010	−0.010	0.011	0.011	0.011	25.186
MBIAS										
n=800	−0.028	−0.032	−0.013	−0.053	−0.072	−0.052	0.007	0.038	0.007	21.671
n=1600	0.014	−0.001	0.014	−0.045	−0.044	−0.045	0.013	0.025	0.013	21.863
n=3200	−0.003	−0.005	−0.004	−0.030	−0.033	−0.030	0.017	0.021	0.017	24.264
n=6400	−0.001	−0.001	−0.001	−0.020	−0.020	−0.020	0.017	0.018	0.018	23.973
n=12800	−0.007	−0.007	−0.006	−0.006	−0.009	−0.006	0.009	0.007	0.009	24.055
SD										
n=800	0.313	0.335	0.314	0.175	0.190	0.175	0.197	0.212	0.197	.
n=1600	0.243	0.257	0.244	0.126	0.141	0.126	0.141	0.154	0.141	458.466
n=3200	0.158	0.167	0.158	0.097	0.109	0.097	0.114	0.123	0.113	48.946
n=6400	0.115	0.121	0.115	0.076	0.083	0.076	0.085	0.094	0.085	9.507
n=12800	0.080	0.082	0.080	0.059	0.065	0.059	0.063	0.068	0.063	5.627

Note: **BIAS**, **MBIAS**, **SD**, and **RMSE** provide the simulated average bias, simulated median bias, standard deviation, and root mean squared error of the LATE estimators, respectively. $\widehat{\mu}_{c,d}$, $\widehat{\mu}_c$, $\widehat{\mu}_d$ denote the LATE estimators on compliers and/or defiers discussed in the main text. $\widehat{\mu}^a_{c,d}$, $\widehat{\mu}^a_c$), $\widehat{\mu}^a_d$ denote the bandwidth-free estimators, while $\widehat{\mu}^b_{c,d}$, $\widehat{\mu}^b_c$, $\widehat{\mu}^b_d$ denote the bias corrected estimators, see Section E in this appendix. The number of replications equals 6000.

Table 11: Simulation results on the performance of LATE estimators under known level sets and no defiers using the bandwidth $b^{3/2}$

Performance: Known level sets	**LATEc**			**LATE**
	$\widehat{\mu}_c$	$\widehat{\mu}_c^a$	$\widehat{\mu}_c^b$	2SLS
RMSE				
n=800	0.175	0.202	0.174	0.800
n=1600	0.123	0.139	0.123	0.535
n=3200	0.086	0.096	0.086	0.368
n=6400	0.065	0.069	0.065	0.264
n=12800	0.045	0.048	0.045	0.180
BIAS				
n=800	−0.009	0.026	−0.013	0.141
n=1600	−0.005	0.009	−0.006	0.036
n=3200	0.004	0.010	0.003	0.012
n=6400	0.003	0.003	0.003	0.018
n=12800	0.002	0.001	0.002	0.004
MBIAS				
n=800	−0.022	0.008	−0.026	0.049
n=1600	−0.013	−0.004	−0.014	−0.013
n=3200	−0.002	0.004	−0.003	−0.009
n=6400	0.002	0.000	0.001	0.000
n=12800	−0.001	−0.001	−0.001	−0.001
SD				
n=800	0.175	0.201	0.173	0.788
n=1600	0.123	0.139	0.123	0.534
n=3200	0.086	0.096	0.086	0.368
n=6400	0.065	0.069	0.065	0.263
n=12800	0.045	0.048	0.045	0.180

Note: **BIAS**, **MBIAS**, **SD**, and **RMSE** provide the simulated average bias, simulated median bias, standard deviation, and root mean squared error of the LATE estimators, respectively. $\widehat{\mu}_{c,d}$, $\widehat{\mu}_c$, $\widehat{\mu}_d$ denote the LATE estimators on compliers and/or defiers discussed in the main text. $\widehat{\mu}^a_{c,d}$, $\widehat{\mu}^a_c$), $\widehat{\mu}^a_d$ denote the bandwidth-free estimators, while $\widehat{\mu}^b_{c,d}$, $\widehat{\mu}^b_c$, $\widehat{\mu}^b_d$ denote the bias corrected estimators, see Section E in this appendix. The number of replications equals 6000.

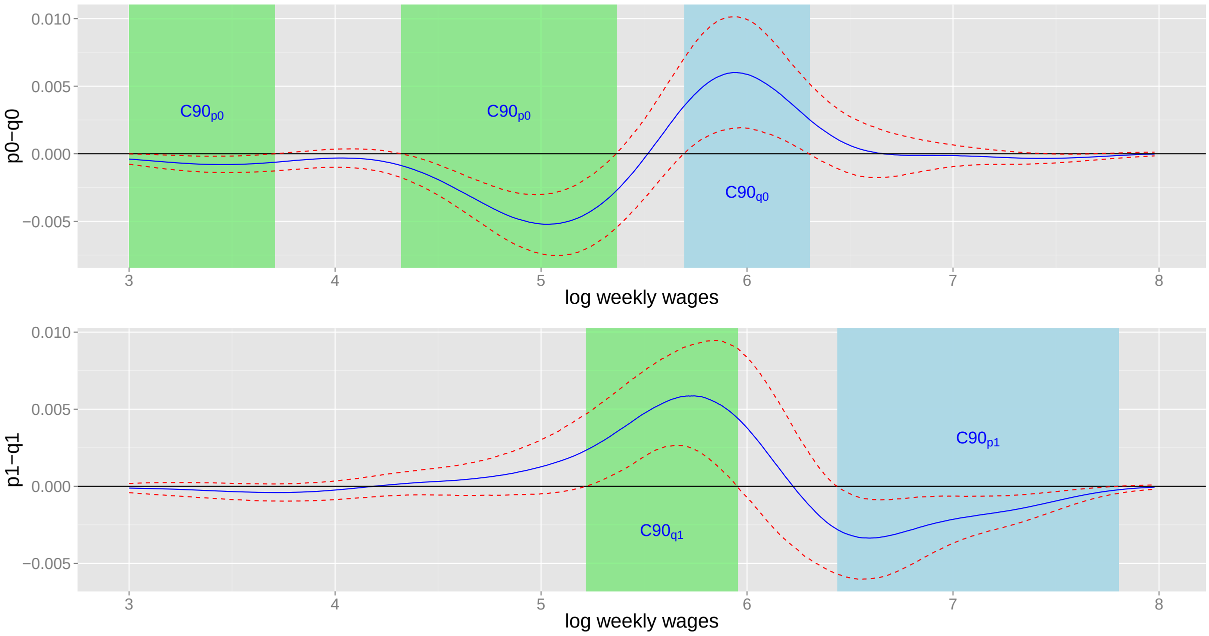
Table 12: Simulation results on the performance of LATE estimators under estimated level sets and no defiers using the bandwidth $b^{3/2}$

Performance: Estimated level sets	**LATEc**			**LATE**
	$\widehat{\mu}_c$	$\widehat{\mu}^a_c$	$\widehat{\mu}^b_c$	2SLS
RMSE				
n=800	0.264	0.278	0.267	0.807
n=1600	0.189	0.201	0.190	0.552
n=3200	0.163	0.169	0.163	0.365
n=6400	0.117	0.121	0.117	0.247
n=12800	0.092	0.096	0.092	0.177
BIAS				
n=800	−0.115	−0.119	−0.118	0.138
n=1600	−0.068	−0.069	−0.070	0.088
n=3200	−0.066	−0.067	−0.067	0.007
n=6400	−0.040	−0.041	−0.040	0.007
n=12800	−0.030	−0.030	−0.030	0.006
MBIAS				
n=800	−0.074	−0.081	−0.077	0.021
n=1600	−0.042	−0.045	−0.042	0.034
n=3200	−0.041	−0.041	−0.042	−0.017
n=6400	−0.028	−0.028	−0.029	−0.008
n=12800	−0.017	−0.017	−0.017	0.006
SD				
n=800	0.238	0.252	0.239	0.795
n=1600	0.176	0.188	0.177	0.546
n=3200	0.149	0.156	0.149	0.365
n=6400	0.110	0.114	0.110	0.247
n=12800	0.087	0.092	0.087	0.177

Note: **BIAS**, **MBIAS**, **SD**, and **RMSE** provide the simulated average bias, simulated median bias, standard deviation, and root mean squared error of the LATE estimators, respectively. $\widehat{\mu}_{c,d}$, $\widehat{\mu}_c$, $\widehat{\mu}_d$ denote the LATE estimators on compliers and/or defiers discussed in the main text. $\widehat{\mu}^a_{c,d}$, $\widehat{\mu}^a_c$), $\widehat{\mu}^a_d$ denote the bandwidth-free estimators, while $\widehat{\mu}^b_{c,d}$, $\widehat{\mu}^b_c$, $\widehat{\mu}^b_d$ denote the bias corrected estimators, see Section E in this appendix. The number of replications equals 6000.

G Level set figure for C90

Figure 4: Estimated differences in densities of log weekly wages (w), i.e., $\hat{p}_d\left(w\right) - \hat{q}_d\left(w\right)$ (blue solid line) under treatment ($d = 1$, lower panel) and non-treatment ($d = 0$, upper panel). The red dashed lines indicate 90% confidence bands. Estimation of the curves are based on 6000 bootstrap replications using a Gaussian Kernel and cross-validated bandwidth selection. Compliers are classified/estimated to exists on the green areas whereas defiers exists on the blue areas.


p0−q0
$C90_{p0}$
$C90_{p0}$
$C90_{q0}$
log weekly wages
p1−q1
$C90_{p1}$
$C90_{q1}$
log weekly wages