Review of Economic Studies (2018) **85**, 999–1028 doi:10.1093/restud/rdx049

Advance access publication 28 August 2017

Fuzzy Differences-in-Differences

C. DE CHAISEMARTIN
University of California at Santa Barbara

and

X. D'HAULTFŒUILLE
Center for Research in Economics and Statistics

First version received June 2014*; Editorial decision June* 2017*; Accepted August* 2017 (*Eds.*)

Difference-in-differences (DID) is a method to evaluate the effect of a treatment. In its basic version, a "control group" is untreated at two dates, whereas a "treatment group" becomes fully treated at the second date. However, in many applications of the DID method, the treatment rate only increases more in the treatment group. In such fuzzy designs, a popular estimator of the treatment effect is the DID of the outcome divided by the DID of the treatment. We show that this ratio identifies a local average treatment effect only if the effect of the treatment is stable over time, and if the effect of the treatment is the same in the treatment and in the control group. We then propose two alternative estimands that do not rely on any assumption on treatment effects, and that can be used when the treatment rate does not change over time in the control group. We prove that the corresponding estimators are asymptotically normal. Finally, we use our results to reassess the returns to schooling in Indonesia.

Key words: Differences-in-differences, Control group, Local average treatment effects, Changes-in-changes, Partial identification, Returns to education.

JEL Codes: C21, C23

1. INTRODUCTION

Difference-in-differences (DID) is a method to estimate the effect of a treatment. In its basic version, a "control group" is untreated at two dates, whereas a "treatment group" becomes treated at the second date. If the trend on the mean outcome without treatment is the same in both groups, the so-called common trend assumption, one can estimate the effect of the treatment by comparing the evolution of the mean outcome in the two groups.

However, in many applications of the DID method, the share of treated units increases more in some groups than in others between the two dates, but no group experiences a sharp change in treatment, and no group remains fully untreated. In such fuzzy designs, a popular estimator of treatment effects is the DID of the outcome divided by the DID of the treatment, the so-called Wald-DID estimator. In de Chaisemartin and D'Haultfœuille (2016), we show that 10.1% of all papers published by the American Economic Review between 2010 and 2012 use fuzzy DID designs, and estimate either a simple Wald-DID or a weighted average of Wald-DIDs.

To our knowledge, no paper has studied treatment effect identification and estimation in fuzzy DID designs. This is the purpose of this article. Hereafter, let "switchers" refer to units that become treated at the second date. Our main parameter of interest is the local average treatment effect (LATE) of treatment group switchers.

First, we consider the case where the share of treated units does not change over time in the control group. In this case, we start by showing that the Wald-DID estimand identifies our parameter of interest if the usual common trend assumption holds, and if the average treatment effect of units treated at both dates is stable over time. This stable treatment effect assumption is not required for identification in sharp DID designs. It is often implausible: we review below some applications where the effect of the treatment is likely to change over time. Then, we show that two alternative estimands identify our parameter of interest even if the effect of the treatment changes over time. The first one, the time-corrected Wald ratio (Wald-TC), relies on common trend assumptions within subgroups of units sharing the same treatment at the first date. The second one, the changes-in-changes Wald ratio (Wald-CIC), generalizes the changes-in-changes (CIC) estimand introduced by Athey and Imbens (2006) to fuzzy designs. It relies on a "common change" assumption which is invariant to the scaling of the outcome, contrary to common trends, but which also imposes restrictions on the full distribution of the potential outcomes while common trends only imposes a restriction on their mean. We discuss below the respective advantages and drawbacks of the Wald-TC and Wald-CIC estimands.

Second, we consider the case where the share of treated units changes over time in the control group. In this case, we show that the Wald-DID estimand identifies the LATE of treatment group switchers if the aforementioned common trend and stable treatment effect assumptions are satisfied, and if the LATEs of treatment and control group switchers are equal. Here as well, this homogeneous treatment effect assumption is often implausible. We also show that under the assumptions underlying the Wald-TC and Wald-CIC estimands, our parameter of interest is partially identified. The smaller the change of the share of treated units in the control group, the tighter the bounds.

We extend these results in several directions. We start by showing how our results can be used in applications with more than two groups. We also show that our results extend to applications with a non-binary treatment variable. Finally, we consider estimators of the Wald-DID, Wald-TC, and Wald-CIC, and we derive their limiting distributions.[1]

We use our results to revisit findings in Duflo (2001) on returns to education in Indonesia. Years of schooling increased substantially in the control group used by the author. Hence, for her Wald-DID to identify the switchers' LATE, returns to schooling should be homogeneous in her treatment and control groups. As we argue in more detail later, this assumption might not be applicable in this context. The bounds we propose do not rely on this assumption, but they are wide and uninformative, here again because schooling increased in the author's control group. Therefore, we form a new control group where years of schooling did not change. The Wald-DID with our new groups is twice as large as the author's original estimate. But it is still likely to be biased, as it relies on the assumption that returns to schooling are stable between birth cohorts, which rules out decreasing returns to experience. The Wald-TC and Wald-CIC do not rely on such a restriction. They are very close to each other, and lie in-between the author's estimate and the Wald-DID.

Overall, our article shows that researchers who use the DID method with fuzzy groups can obtain estimates not resting on the assumption that treatment effects are stable and homogeneous, provided they can find a control group whose exposure to the treatment does not change over time. There are applications where such control groups are readily available (see Field, 2007; Gentzkow *et al*., 2011, which we revisit in our Supplementary Data). In other applications, the

1. A Stata package computing the estimators is available on the authors' webpages.

control groups need to be estimated. We propose a method to estimate the control groups that can be used when the number of groups is small relative to the size of each group. Studying how to estimate the control groups when the number of groups is large relative to their size, as is the case in Duflo (2001), is left for future work.

Though we are the first to study fuzzy DID estimators in models with heterogeneous treatment effects, our article is related to several other papers in the DID literature. Blundell *et al*. (2004) and Abadie (2005) consider a conditional version of the common trend assumption in sharp DID designs, and adjust for covariates using propensity score methods. Our Wald-DID estimator with covariates is related to their estimators. Bonhomme and Sauder (2011) consider a linear model allowing for heterogeneous effects of time, and show that in sharp designs it can be identified if the idiosyncratic shocks are independent of the treatment and of the individual effects. Our Wald-CIC estimator builds on Athey and Imbens (2006). In work posterior to ours, D'Haultfœuille *et al*. (2015) study the possibly non-linear effects of a continuous treatment, and propose an estimator related to our Wald-CIC estimator.

The remainder of the article is organized as follows. Section 2 presents our main identification results in a simple setting with two groups, two periods, and a binary treatment. Section 3 presents extensions of those main identification results. Section 4 presents estimation and inference. In Section 5 we revisit results from Duflo (2001). Section 6 concludes. The Appendix gathers the main proofs. For brevity, further identification and inference results, two additional empirical applications, and additional proofs are deferred to our Supplementary Data.

2. IDENTIFICATION

2.1. *Framework*

We are interested in measuring the effect of a treatment D on some outcome. For now, we assume that the treatment is binary. $Y(1)$ and $Y(0)$ denote the two potential outcomes of the same unit with and without treatment. The observed outcome is $Y = DY(1)+(1-D)Y(0)$.

Hereafter, we consider a model best suited for repeated cross sections. This model also applies to single cross sections where cohort of birth plays the role of time, as in Duflo (2001) for instance. The extension to panel data is sketched in Subsection 3.4 and developed in our Supplementary Data. We assume that the data can be divided into "time periods" represented by a random variable T, and into groups represented by a random variable G. In this section, we focus on the simplest possible case where there are only two groups, a "treatment" and a "control" group, and two periods of time. G is a dummy for units in the treatment group and T is a dummy for the second period.

We now introduce the notation we use throughout the article. For any random variable R, let $\mathcal{S}(R)$ denote its support. Let also R_{gt} and R_{dgt} be two other random variables such that $R_{gt} \sim R|G=g, T=t$ and $R_{dgt} \sim R|D=d, G=g, T=t$, where $\sim$ denotes equality in distribution. For instance, it follows from those definitions that $E(R_{11}) = E(R|G=1, T=1)$, while $E(R_{011}) = E(R|D=0, G=1, T=1)$. For any event or random variable A, let F_R and $F_{R|A}$ denote the cumulative distribution function (cdf) of R and its cdf conditional on A.[2] Finally, for any increasing function F on the real line, we denote by F^{-1} its generalized inverse, $F^{-1}(q) = \inf\{x \in \mathrm{IR} : F(x) \geq q\}$. In particular, F_R^{-1} is the quantile function of the random variable R.

Contrary to the standard "sharp" DID setting where $D = G \times T$, we consider a "fuzzy" setting where $D \neq G \times T$. Some units may be treated in the control group or at period 0, and some units

2. With a slight abuse of notation, $P(A)F_{R|A}$ should be understood as 0 when A is an event and $P(A)=0$.

may remain untreated in the treatment group at period 1. Still, we consider two assumptions on the evolution of the share of treated units in the treatment and control groups.

Assumption 1. (Fuzzy design). *$E(D_{11}) > E(D_{10})$, and $E(D_{11}) - E(D_{10}) > E(D_{01}) - E(D_{00})$.*

Assumption 2. (Stable percentage of treated units in the control group). $0 < E(D_{01}) = E(D_{00}) < 1$.

Assumption 1 is just a way to define the treatment and the control group in our fuzzy setting. The treatment group is the one experiencing the larger increase of its treatment rate. If the treatment rate decreases in both groups, one can redefine the treatment variable as $\tilde{D} = 1 - D$. Thus, Assumption 1 only rules out the case where the two groups experience the same evolution of their treatment rates. Assumption 2 corresponds to the special case where the percentage of treated units does not change between period 0 and 1 in the control group.

We consider the following treatment participation equation.

Assumption 3. (Treatment participation equation). *$D = \mathbb{1}\{V \geq v_{GT}\}$, with $V \perp\!\!\!\perp T | G$.*

Assumption 3 imposes a latent index model for the treatment (see, *e.g.*, Vytlacil, 2002), where V may be interpreted as a unit's propensity to be treated, and where the threshold for treatment participation depends both on time and group. This participation equation implies that within each group, units can switch treatment in only one direction. For instance, once combined with Assumption 1, Assumption 3 implies that in the treatment group there are no units switching from treatment to non treatment between period 0 and 1. Assuming that treatment is monotonous with respect to time is not necessary for our results to hold (see Subsection 3.4 for further detail on this point). However, this greatly simplifies the exposition.

We now define our parameters of interest. For that purpose, let us introduce

$$D(t) = \mathbb{1}\{V \geq v_{Gt}\}.$$

In repeated cross sections, $D(0)$ and $D(1)$ denote the treatment status of a unit at period 0 and 1, respectively, and only $D = D(T)$ is observed. In single cross sections where cohort of birth plays the role of time, $D(t)$ denotes instead the potential treatment of a unit had she been born at $T = t$. Here again, only $D = D(T)$ is observed. Let $S = \{D(0) < D(1), G = 1\}$. S stands for treatment group units going from non-treatment to treatment between period 0 and 1, hereafter referred to as the "treatment group switchers". Our parameters of interest are their LATE and Local Quantile Treatment Effects (LQTE), which are respectively defined by

$$\begin{aligned}\Delta &= E(Y_{11}(1) - Y_{11}(0)|S), \\ \tau_q &= F^{-1}_{Y_{11}(1)|S}(q) - F^{-1}_{Y_{11}(0)|S}(q), \quad q \in (0,1).\end{aligned}$$

We focus on these parameters for two reasons. First, there are instances where treatment group switchers are the only units affected by some policy, implying that they are the relevant subgroup one should consider to assess its effects. Consider for instance a policy whereby in $T = 1$, the treatment group becomes eligible to some treatment for which it was not eligible in $T = 0$ (see, *e.g.*, Field, 2007). In this example, treatment group switchers are all the units in that group treated in $T = 1$. Those units are affected by the policy: without it, they would have remained untreated. Moreover, nobody else is affected by the policy. Second, identifying treatment effects in the whole population would require additional conditions, on top of those we consider below. In the example above, the policy extension does not provide any information on treatment effects in the control group, because this group does not experience any change.

2.2. *The Wald-DID estimand*

We first investigate the commonly used strategy of running an IV regression of the outcome on the treatment with time and group as included instruments, and the interaction of the two as the excluded instrument. The estimand arising from this regression is the Wald-DID defined by $W_{\text{DID}} = \text{DID}_Y / \text{DID}_D$, where for any random variable R we let

$$\text{DID}_R = E(R_{11}) - E(R_{10}) - (E(R_{01}) - E(R_{00})).$$

Let also $S' = \{D(0) \neq D(1), G=0\}$ denote the control group switchers. Control group switchers are defined by $D(0) \neq D(1)$ because the treatment rate may decrease in this group, thus implying that switchers may go from being treated to being untreated between period 0 and 1.[3] Let $\Delta' = E\left(Y_{01}(1) - Y_{01}(0) | S'\right)$ denote their LATE. Finally, let $\alpha = (P(D_{11}=1) - P(D_{10}=1)) / \text{DID}_D$.

We consider the following assumptions, under which we can relate W_{DID} to Δ and Δ'.

Assumption 4. (Common trends). $E(Y(0)|G, T=1) - E(Y(0)|G, T=0)$ *does not depend on* G.

Assumption 5. (Stable treatment effect over time). *For all* $d \in \mathcal{S}(D)$,[4] $E(Y(d) - Y(0)|G, T=1, D(0)=d) = E(Y(d) - Y(0)|G, T=0, D(0)=d)$.

Assumption 6. (Homogeneous treatment effect between groups). $\Delta = \Delta'$.

Assumption 4 requires that the mean of $Y(0)$ follow the same evolution over time in the treatment and control groups. This assumption is not specific to the fuzzy settings we are considering here: DID in sharp settings also rely on this assumption (see, *e.g.*, Abadie, 2005). Assumption 5 requires that in both groups, the average effect of going from 0 to d units of treatment among units with $D(0)=d$ is stable over time. This is equivalent to assuming that among these units, the mean of $Y(d)$ and $Y(0)$ follow the same evolution over time:

$$\begin{aligned} & E(Y(d)|G, T=1, D(0)=d) - E(Y(d)|G, T=0, D(0)=d) \\ &= E(Y(0)|G, T=1, D(0)=d) - E(Y(0)|G, T=0, D(0)=d). \end{aligned} \tag{2.1}$$

This stable treatment effect condition, which is also equivalent to imposing a common trend condition for all potential outcomes $Y(d)$, is very different from Assumption 4, the only condition required for identification in sharp DID settings. Assumption 6 requires that in both groups, switchers have the same LATE. This assumption is also specific to fuzzy settings: it is not required for identification in sharp DID settings.

3. On the other hand, Assumptions 1 and 3 ensure that treatment group switchers can only go from being untreated to being treated, which is why they are defined by $D(0) < D(1)$.

4. When the treatment is binary, Assumption 5 only requires that the equation therein holds for $d=1$. Writing Assumption 5 this way ensures it carries through to the case of a non-binary treatment.

Theorem 1.

(1) If Assumptions 1 and 3–5 are satisfied, then

$$W_{\text{DID}} = \alpha\Delta + (1-\alpha)\Delta'.$$

(2) If Assumption 2 or 6 further holds, then

$$W_{\text{DID}} = \Delta.$$

When the treatment rate increases in the control group, $E(D_{01}) - E(D_{00}) > 0$, so $\alpha > 1$. Therefore, under Assumptions 1 and 3–5 the Wald-DID is equal to a weighted difference of the LATEs of treatment and control group switchers in period 1. In both groups, the evolution of the mean outcome between period 0 and 1 is the sum of three things: the change in the mean of $Y(0)$ for units untreated at $T=0$; the change in the mean of $Y(1)$ for units treated at $T=0$; the average effect of the treatment for switchers. Under Assumptions 4 and 5, changes in the mean of $Y(0)$ and $Y(1)$ in both groups cancel out. The Wald-DID is finally equal to the weighted difference between the LATEs of treatment and control group switchers. This weighted difference does not satisfy the no sign-reversal property: it may be negative even if the treatment effect is positive for everybody in the population. If one is ready to further assume that Assumption 6 is satisfied, this weighted difference simplifies into Δ.[5]

When the treatment rate diminishes in the control group, $E(D_{01}) - E(D_{00}) < 0$, so $\alpha < 1$. Therefore, under Assumptions 1 and 3–5 the Wald-DID is equal to a weighted average of the LATEs of treatment and control group switchers in period 1. This quantity satisfies the no sign-reversal property, but it still differs from Δ unless here as well one is ready to further assume that Assumption 6 is satisfied.

When the treatment rate is stable in the control group, $\alpha = 1$ so the Wald-DID is equal to Δ under Assumptions 1 and 3–5 alone. But even then, the Wald-DID relies on the assumption that in both groups, the average treatment effect among units treated at $T=0$ remains stable over time. This assumption is necessary. Under Assumptions 1 and 3–4 alone, the Wald-DID is equal to Δ plus a bias term involving several LATEs. Unless this combination of LATEs cancels out exactly, the Wald-DID differs from Δ. We give the formula of the bias term in the end of the proof of Theorem 1.

2.3. *The time-corrected Wald estimand*

In this section, we consider a first alternative estimand of Δ. Instead of relying on Assumptions 4 and 5, it is based on the following condition:

Assumption 4.' (Conditional common trends). *For all $d \in \mathcal{S}(D)$, $E(Y(d)|G, T=1, D(0)=d) - E(Y(d)|G, T=0, D(0)=d)$ does not depend on G.*

5. Under this assumption, the Wald-DID actually identifies the LATE of all switchers, not only of those in the treatment group. There are instances where this LATE measures the effect of the policy under consideration, because treatment and control group switchers are the only units affected by this policy. Consider for instance the case of a policy whereby a new treatment is introduced in both groups in $T=1$ (see Enikolopov *et al*., 2011). In this example, treatment and control group switchers are all the units treated in $T=1$. These units are affected by the policy (without it, they would have remained untreated) and nobody else is affected by it.

Assumption 4' requires that the mean of $Y(0)$ (resp. $Y(1)$) follows the same evolution over time among treatment and control group units that were untreated (resp. treated) at $T=0$.

Let $\delta_d = E(Y_{d01}) - E(Y_{d00})$ denote the change in the mean outcome between period 0 and 1 for control group units with treatment status d. Then, let

$$W_{\text{TC}} = \frac{E(Y_{11}) - E(Y_{10} + \delta_{D_{10}})}{E(D_{11}) - E(D_{10})}.$$

W_{TC} stands for "time-corrected Wald".

Theorem 2. *If Assumptions 1–3, and 4' are satisfied, then* $W_{\text{TC}} = \Delta$.

Note that

$$W_{\text{TC}} = \frac{E(Y|G=1,T=1) - E(Y+(1-D)\delta_0 + D\delta_1|G=1,T=0)}{E(D|G=1,T=1) - E(D|G=1,T=0)}.$$

This is almost the Wald ratio in the treatment group with time as the instrument, except that we have $Y+(1-D)\delta_0+D\delta_1$ instead of Y in the second term of the numerator. This difference arises because time is not a standard instrument: it can directly affect the outcome. When the treatment rate is stable in the control group, we can identify the trends on $Y(0)$ and $Y(1)$ by looking at how the mean outcome of untreated and treated units changes over time in this group. Under Assumption 4', these trends are the same in the two groups. As a result, we can add these changes to the outcome of untreated and treated units in the treatment group in period 0, to recover the mean outcome we would have observed in this group in period 1 if switchers had not changed their treatment between the two periods. This is what $(1-D)\delta_0+D\delta_1$ does. Therefore, the numerator of W_{TC} compares the mean outcome in the treatment group in period 1 to the counterfactual mean we would have observed if switchers had remained untreated. Once normalized, this yields the LATE of treatment group switchers.

2.4. *The CIC estimands*

In this section, we consider a second alternative estimand of Δ for continuous outcomes, as well as estimands of the LQTE. They rely on the following condition.

Assumption 7. (Monotonicity and time invariance of unobservables). $Y(d) = h_d(U_d, T)$, *with* $U_d \in \mathbb{R}$ *and* $h_d(u,t)$ *strictly increasing in* u *for all* $(d,t) \in \mathcal{S}((D,T))$. *Moreover,* $U_d \perp\!\!\!\perp T | G, D(0)$.

Assumptions 3 and 7 generalize the CIC model in Athey and Imbens (2006) to fuzzy settings. Assumptions 3 and 7 imply $U_d \perp\!\!\!\perp T | G$. Therefore, they require that at each period, both potential outcomes are strictly increasing functions of a scalar unobserved heterogeneity term whose distribution is stationary over time, as in Athey and Imbens (2006). But Assumption 7 also imposes $U_d \perp\!\!\!\perp T | G, D(0)$: the distribution of U_d must be stationary within subgroups of units sharing the same treatment status at $T=0$.

We also impose the assumption below, which is testable in the data.

Assumption 8. (Data restrictions).

(1) $\mathcal{S}(Y_{dgt}) = \mathcal{S}(Y)$ *for* $(d,g,t) \in \mathcal{S}((D,G,T))$, *and* $\mathcal{S}(Y)$ *is a closed interval of* $\mathbb{R}$.
(2) $F_{Y_{dgt}}$ *is continuous on* $\mathbb{R}$ *and strictly increasing on* $\mathcal{S}(Y)$, *for* $(d,g,t) \in \mathcal{S}((D,G,T))$.

The first condition requires that the outcome have the same support in each of the eight treatment $\times$ group $\times$ period cells. Athey and Imbens (2006) make a similar assumption.[6] Note that this condition does not restrict the outcome to have bounded support: for instance, $[0,+\infty)$ is a closed interval of IR. The second condition requires that the distribution of Y be continuous with positive density in each of the eight groups $\times$ periods $\times$ treatment status cells. With a discrete outcome, Athey and Imbens (2006) show that one can bound treatment effects under their assumptions. Similar results apply in fuzzy settings, but for the sake of brevity we do not present them here.

Let $Q_d(y)=F^{-1}_{Y_{d01}}\circ F_{Y_{d00}}(y)$ be the quantile-quantile transform of Y from period 0 to 1 in the control group conditional on $D=d$. This transform maps y at rank q in period 0 into the corresponding y' at rank q in period 1. Let also

$$F_{\text{CIC},d}(y) = \frac{P(D_{11}=d)F_{Y_{d11}}(y)-P(D_{10}=d)F_{Q_d(Y_{d10})}(y)}{P(D_{11}=d)-P(D_{10}=d)},$$

$$W_{\text{CIC}} = \frac{E(Y_{11})-E(Q_{D_{10}}(Y_{10}))}{E(D_{11})-E(D_{10})}.$$

Theorem 3. *If Assumptions 1–3, and 7–8 are satisfied, then* $W_{\text{CIC}}=\Delta$ *and* $F^{-1}_{CIC,1}(q)-F^{-1}_{\text{CIC},0}(q)=\tau_q$.

This result combines ideas from Imbens and Rubin (1997) and Athey and Imbens (2006). We seek to recover the distribution of, say, $Y(1)$ among switchers in the treatment group $\times$ period 1 cell. On that purpose, we start from the distribution of Y among all treated observations of this cell. Those include both switchers and units already treated at $T=0$. Consequently, we must "withdraw" from this distribution that of units treated at $T=0$, exactly as in Imbens and Rubin (1997). But this last distribution is not observed. To reconstruct it, we adapt the ideas in Athey and Imbens (2006) and apply the quantile-quantile transform from period 0 to 1 among treated observations in the control group to the distribution of Y among treated units in the treatment group in period 0.

Intuitively, the quantile-quantile transform uses a double-matching to reconstruct the unobserved distribution. Consider a treated unit in the treatment group $\times$ period 0 cell. She is first matched to a treated unit in the control group $\times$ period 0 cell with same y. Those two units are observed at the same period of time and are both treated. Therefore, under Assumption 7 they must have the same u_1. Second, the control $\times$ period 0 unit is matched to her rank counterpart among treated units of the control group $\times$ period 1 cell. We denote by y^* the outcome of this last observation. Because $U_1 \perp\!\!\!\perp T|G,D(0)=1$, under Assumption 2 those two observations must also have the same u_1. Consequently, $y^*=h_1(u_1,1)$, which means that y^* is the outcome that the treatment $\times$ period 0 cell unit would have obtained in period 1.

Note that

$$W_{\text{CIC}}=\frac{E(Y|G=1,T=1)-E((1-D)Q_0(Y)+DQ_1(Y)|G=1,T=0)}{E(D|G=1,T=1)-E(D|G=1,T=0)}.$$

6. Common support conditions might not be satisfied when outcome distributions differ in the treatment and control groups, the very situations where the Wald-CIC estimand we propose below might be more appealing than the Wald-DID or Wald-TC (see Subsection 2.5). Athey and Imbens (2006) show that in such instances, quantile treatment effects are still point identified over a large set of quantiles, while the average treatment effect can be bounded. Even though we do not present them here, similar results apply in fuzzy settings.

Here again, W_{CIC} is almost the standard Wald ratio in the treatment group with T as the instrument, except that we have $(1-D)Q_0(Y)+DQ_1(Y)$ instead of Y in the second term of the numerator. $(1-D)Q_0(Y)+DQ_1(Y)$ accounts for the fact that time directly affects the outcome, just as $(1-D)\delta_0+D\delta_1$ does in the W_{TC} estimand. Under Assumption 4', the trends affecting the outcome are identified by additive shifts, while under Assumptions 7–8 they are identified by possibly non-linear quantile-quantile transforms.

2.5. *Discussion*

There are many applications where Assumption 5 is implausible, because the effect of the treatment is likely to change over time. In Section 5 we review a specific example. In such instances, if the share of treated units is stable over time in the control group, one needs to choose between the Wald-TC and Wald-CIC estimands. This choice should be based on the suitability of Assumption 4' and 7 in the application under consideration. Assumption 4' is not invariant to the scaling of the outcome, but it only restricts its mean. Assumption 7 is invariant to the scaling of the outcome, but it restricts its entire distribution. When the treatment and control groups have different outcome distributions conditional on D in the first period, the scaling of the outcome might have a large effect on the Wald-TC. The Wald-CIC is much less sensitive to the scaling of the outcome, so using this estimand might be preferable. On the other hand, when the two groups have similar outcome distributions conditional on D in the first period, using the Wald-TC might be preferable as Assumption 4' only restricts the mean of the outcome. This choice should also be based on the parameters one seeks to identify. Under Assumption 7, both the LATE and LQTEs of treatment group switchers are identified; under Assumption 4', only the LATE is identified.

In applications where Assumption 5 is plausible, the Wald-DID may be appealing, especially when the treatment rate decreases in the control group. Indeed, in such instances, the assumptions underlying the Wald-TC and Wald-CIC only lead to partial identification (see Subsection 3.1). On the other hand, the Wald-DID identifies a weighted average of LATEs even if Assumption 6 fails to hold.[7]

3. EXTENSIONS

We now consider several extensions. We first show that when the treatment rate is not stable in the control group, Δ and τ_q can still be partially identified under our assumptions. We then consider applications with multiple groups. Next, we show that our results extend to ordered, non-binary treatments. Finally we sketch other extensions that are fully developed in the Supplementary Data.

3.1. *Partial identification with a non-stable control group*

We first consider partial identification of Δ and τ_q when Assumption 2 does not hold. Let us introduce some additional notation. When the outcome is bounded, let $\underline{y}$ and $\overline{y}$ respectively denote the lower and upper bounds of its support. For any real number x, let $\bar{M}_{01}(x)=\min(1,\max(0,x))$. For any $g\in\mathcal{S}(G)$, let $\lambda_{gd}=P(D_{g1}=d)/P(D_{g0}=d)$ be the ratio of the shares of units in group g receiving treatment d in period 1 and period 0. For instance, $\lambda_{00}>1$ when the share of untreated

7. In our Supplementary Data, we also explain how to use placebo tests to assess the plausibility of Assumptions 4, 5, 4', and 7 when more than two periods of data are available.

observations increases in the control group between period 0 and 1. Let also

$$\underline{F}_{d01}(y)=M_{01}\left(1-\lambda_{0d}(1-F_{Y_{d01}}(y))\right)-M_{01}(1-\lambda_{0d})\mathbb{1}\{y<\overline{y}\},$$
$$\overline{F}_{d01}(y)=M_{01}\left(\lambda_{0d}F_{Y_{d01}}(y)\right)+(1-M_{01}(\lambda_{0d}))\mathbb{1}\{y\geq\underline{y}\},$$
$$\underline{\delta}_d=\int yd\overline{F}_{d01}(y)-E(Y_{d00}),\ \overline{\delta}_d=\int yd\underline{F}_{d01}(y)-E(Y_{d00}).$$

We define the bounds obtained under Assumption 4' (TC bounds hereafter) as follows:

$$\underline{W}_{\text{TC}}=\frac{E(Y_{11})-E(Y_{10}+\overline{\delta}_{D_{10}})}{E(D_{11})-E(D_{10})},\ \overline{W}_{\text{TC}}=\frac{E(Y_{11})-E(Y_{10}+\underline{\delta}_{D_{10}})}{E(D_{11})-E(D_{10})}.$$

Next, we define the bounds obtained under Assumptions 7–8 (CIC bounds hereafter). For $d\in\{0,1\}$ and any cdf T, let $H_d=F_{Y_{d10}}\circ F^{-1}_{Y_{d00}}$ and

$$G_d(T)\ =\ \lambda_{0d}F_{Y_{d01}}+(1-\lambda_{0d})T,$$
$$C_d(T)\ =\ \frac{P(D_{11}=d)F_{Y_{d11}}-P(D_{10}=d)H_d\circ G_d(T)}{P(D_{11}=d)-P(D_{10}=d)},$$
$$\underline{T}_d\ =\ M_{01}\left(\frac{\lambda_{0d}F_{Y_{d01}}-H_d^{-1}(\lambda_{1d}F_{Y_{d11}})}{\lambda_{0d}-1}\right),$$
$$\overline{T}_d\ =\ M_{01}\left(\frac{\lambda_{0d}F_{Y_{d01}}-H_d^{-1}(\lambda_{1d}F_{Y_{d11}}+(1-\lambda_{1d}))}{\lambda_{0d}-1}\right).$$

In the definition of $\underline{T}_d$ and $\overline{T}_d$, we use the convention that $F_R^{-1}(q)=\inf\mathcal{S}(R)$ for $q<0$, and $F_R^{-1}(q)=\sup\mathcal{S}(R)$ for $q>1$. Then let $\underline{F}_{\text{CIC},d}(y)=\sup_{y'\leq y}C_d\left(\underline{T}_d\right)(y')$ and $\overline{F}_{\text{CIC},d}(y)=\inf_{y'\geq y}C_d\left(\overline{T}_d\right)(y')$. We define the CIC bounds on Δ and τ_q by:

$$\underline{W}_{\text{CIC}}\ =\ \int yd\overline{F}_{\text{CIC},1}(y)-\int yd\underline{F}_{\text{CIC},0}(y),$$
$$\overline{W}_{\text{CIC}}\ =\ \int yd\underline{F}_{\text{CIC},1}(y)-\int yd\overline{F}_{\text{CIC},0}(y),$$
$$\underline{\tau}_q\ =\ \max(\overline{F}^{-1}_{\text{CIC},1}(q),\underline{y})-\min(\underline{F}^{-1}_{\text{CIC},0}(q),\overline{y}),$$
$$\overline{\tau}_q\ =\ \min(\underline{F}^{-1}_{\text{CIC},1}(q),\overline{y})-\max(\overline{F}^{-1}_{\text{CIC},0}(q),\underline{y}).$$

Finally, we introduce the two following conditions, which ensure that the CIC bounds are well-defined and sharp.

Assumption 9. (Existence of moments). *For* $d\in\{0,1\}$, $\int|y|d\overline{F}_{\text{CIC},d}(y)<+\infty$ *and* $\int|y|d\underline{F}_{\text{CIC},d}(y)<+\infty$.

Assumption 10. (Increasing bounds). *For* $(d,g,t)\in\mathcal{S}((D,G,T))$, $F_{Y_{dgt}}$ *is continuously differentiable, with positive derivative on the interior of* $\mathcal{S}(Y)$. *Moreover,* $\underline{T}_d,\overline{T}_d,G_d(\underline{T}_d),G_d(\overline{T}_d),C_d(\underline{T}_d)$ *and* $C_d(\overline{T}_d)$ *are increasing on* $\mathcal{S}(Y)$.

Theorem 4. *Assume that Assumptions 1 and 3 are satisfied and* $0 < P(D_{01}=1) \neq P(D_{00}=1) < 1$. *Then:*

(1) *If Assumption 4' holds and* $P(\underline{y} \leq Y(d) \leq \overline{y}) = 1$ *for* $d \in \{0,1\}$, $\underline{W}_{\text{TC}} \leq \Delta \leq \overline{W}_{\text{TC}}$.[8]

(2) *If Assumptions 7-9 hold,* $F_{Y_{11}(d)|S}(y) \in [\underline{F}_{\text{CIC},d}(y), \overline{F}_{\text{CIC},d}(y)]$ *for* $d \in \{0,1\}$, $\Delta \in [\underline{W}_{\text{CIC}}, \overline{W}_{\text{CIC}}]$ *and* $\tau_q \in [\underline{\tau}_q, \overline{\tau}_q]$. *These bounds are sharp if Assumption 10 holds.*

The reasoning underlying the TC bounds goes as follows. Assume for instance that the treatment rate increases in the control group. Then, the difference between $E(Y_{101})$ and $E(Y_{100})$ arises both from the trend on $Y(1)$, and from the fact the former expectation is for units treated at $T=0$ and switchers, while the latter is only for units treated at $T=0$. Therefore, we can no longer identify the trend on $Y(1)$ among units treated at $T=0$. But when the outcome has bounded support, this trend can be bounded, because we know the percentage of the control group switchers account for. A similar reasoning can be used to bound the trend on $Y(0)$ among units untreated at $T=0$. Eventually, Δ can also be bounded. The smaller the change of the treatment rate in the control group, the tighter the bounds.

The reasoning underlying the CIC bounds goes as follows. When $0 < P(D_{00}=1) \neq P(D_{01}=1) < 1$, the second matching described in Subsection 2.4 collapses, because treated (resp. untreated) observations in the control group are no longer comparable in period 0 and 1 as explained in the previous paragraph. Therefore, we cannot match period 0 and period 1 observations on their rank anymore. However, we know the percentage of the control group switchers account for, so we can match period 0 observations to their best- and worst-case rank counterparts in period 1.

If the support of the outcome is unbounded, $\underline{F}_{\text{CIC},0}$ and $\overline{F}_{\text{CIC},0}$ are proper cdfs when $\lambda_{00} > 1$, but they are defective when $\lambda_{00} < 1$. On the contrary, $\underline{F}_{\text{CIC},1}$ and $\overline{F}_{\text{CIC},1}$ are always proper cdfs. As a result, when $\mathcal{S}(Y)$ is unbounded and $\lambda_{00} > 1$, the CIC bounds we derive for Δ and τ_q are finite under Assumption 9. The TC bounds, on the other hand, are always infinite when $\mathcal{S}(Y)$ is unbounded.

3.2. *Multiple groups*

We now consider the case where there are more than two groups but only two time periods in the data. The case with multiple groups and time periods is considered in the Supplementary Data. Let $G \in \{0,1,...,\overline{g}\}$ denote the group a unit belongs to. For any $g \in \mathcal{S}(G)$, let $S_g = \{D(0) \neq D(1), G = g\}$ denote units of group g who switch treatment between $T=0$ and 1. Let $S^* = \cup_{g=0}^{\overline{g}} S_g$ be the union of all switchers. We can partition the groups depending on whether their treatment rate is stable, increases, or decreases. Specifically, let

$$\begin{aligned}
\mathcal{G}_s &= \{g \in \mathcal{S}(G) : E(D_{g1}) = E(D_{g0})\} \\
\mathcal{G}_i &= \{g \in \mathcal{S}(G) : E(D_{g1}) > E(D_{g0})\} \\
\mathcal{G}_d &= \{g \in \mathcal{S}(G) : E(D_{g1}) < E(D_{g0})\},
\end{aligned}$$

and let $G^* = \mathbb{1}\{G \in \mathcal{G}_i\} - \mathbb{1}\{G \in \mathcal{G}_d\}$.

Theorem 5 below shows that when there is at least one group in which the treatment rate is stable, our assumptions allow us to point identify $\Delta^* = E(Y(1) - Y(0)|S^*, T=1)$, the LATE of all

8. It is not difficult to show that these bounds are sharp. We omit the proof for brevity.

switchers. Before presenting this result, additional notation is needed. For any random variable R, $g \neq g' \in \{-1,0,1\}^2$, and $d \in \{0,1\}$, let

$$\begin{aligned}
\mathrm{DID}^*_R(g,g') &= E(R|G^*=g,T=1)-E(R|G^*=g,T=0)\\
&\quad -(E(R|G^*=g',T=1)-E(R|G^*=g',T=0)),\\
\delta^*_d &= E(Y|D=d,G^*=0,T=1)-E(Y|D=d,G^*=0,T=0),\\
Q^*_d(y) &= F^{-1}_{Y|D=d,G^*=0,T=1}\circ F_{Y|D=d,G^*=0,T=0}(y),\\
W^*_{\mathrm{DID}}(g,g') &= \frac{\mathrm{DID}^*_Y(g,g')}{\mathrm{DID}^*_D(g,g')},\\
W^*_{\mathrm{TC}}(g) &= \frac{E(Y|G^*=g,T=1)-E(Y+\delta^*_D|G^*=g,T=0)}{E(D|G^*=g,T=1)-E(D|G^*=g,T=0)},\\
W^*_{\mathrm{CIC}}(g) &= \frac{E(Y|G^*=g,T=1)-E(Q^*_D(Y)|G^*=g,T=0)}{E(D|G^*=g,T=1)-E(D|G^*=g,T=0)}.
\end{aligned}$$

We also define the following weight:

$$w_{10}=\frac{\mathrm{DID}^*_D(1,0)P(G^*=1)}{\mathrm{DID}^*_D(1,0)P(G^*=1)+\mathrm{DID}^*_D(0,-1)P(G^*=-1)}.$$

We finally define our estimands as $W^*_{\mathrm{DID}}=w_{10}W^*_{\mathrm{DID}}(1,0)+(1-w_{10})W^*_{\mathrm{DID}}(-1,0)$, $W^*_{\mathrm{TC}}=w_{10}W^*_{\mathrm{TC}}(1)+(1-w_{10})W^*_{\mathrm{TC}}(-1)$ and $W^*_{\mathrm{CIC}}=w_{10}W^*_{\mathrm{CIC}}(1)+(1-w_{10})W^*_{\mathrm{CIC}}(-1)$.

Theorem 5. *Assume that Assumption 3 is satisfied, that $\mathcal{G}_s \neq \emptyset$, and that $G \perp\!\!\!\perp T$.*

(1) If Assumptions 4 and 5 are satisfied, $W^*_{\mathrm{DID}}=\Delta^*$.
(2) If Assumption 4' is satisfied, $W^*_{\mathrm{TC}}=\Delta^*$.
(3) If Assumptions 7 and 8 are satisfied, $W^*_{\mathrm{CIC}}=\Delta^*$.

This theorem states that with multiple groups and two periods of time, treatment effects for switchers are identified if there is at least one group in which the treatment rate is stable over time. The estimands we propose can then be computed in four steps. First, we form three "supergroups", by pooling together the groups where treatment increases ($G^*=1$), those where it is stable ($G^*=0$), and those where it decreases ($G^*=-1$). While in some applications these three sets of groups are known to the analyst (see, *e.g.*, Gentzkow *et al.*, 2011), in other applications they must be estimated (see our application in Section 5). Second, we compute the Wald-DID, Wald-TC, or Wald-CIC estimand with $G^*=1$ and $G^*=0$ as the treatment and control groups. Third, we compute the Wald-DID, Wald-TC, or Wald-CIC estimand with $G^*=-1$ and $G^*=0$ as the treatment and control groups. Finally, we compute a weighted average of those two estimands.

Theorem 5 relies on the assumption that $G \perp\!\!\!\perp T$. This requires that the distribution of groups be stable over time. This will automatically be satisfied if the data are a balanced panel and G is time invariant. With repeated cross sections or cohort data, this assumption might fail to hold. However, when G is not independent of T, it is still possible to form Wald-DID and Wald-TC type of estimands identifying Δ^*. We give the formulas of these estimands in Subsection 1.2 in the Supplementary Data.

Two last comments on Theorem 5 are in order. First, groups where the treatment rate diminishes can be used as "treatment" groups, just as those where it increases. Indeed, it is easy to show that all the results from the previous section still hold if the treatment rate decreases in the treatment group and is stable in the control group. Second, when there are more than two groups where the treatment rate is stable, our three sets of assumptions become testable. Under each set of assumptions, using any subset of $\mathcal{G}_s$ as the control group should yield the same estimand for Δ^*.

3.3. *Non-binary, ordered treatment*

We now consider the case where treatment takes a finite number of ordered values, $D \in \{0, 1, ..., \overline{d}\}$. To accommodate this extension, Assumption 3 has to be modified as follows.

Assumption 3.' (Ordered treatment equation). *$D = \sum_{d=1}^{\overline{d}} \mathbb{1}\{V \geq v_{GT}^d\}$, with $-\infty = v_{GT}^0 < v_{GT}^1 ... < v_{GT}^{\overline{d}+1} = +\infty$ and $V \perp\!\!\!\perp T|G$. As before, let $D(t) = \sum_{d=1}^{\overline{d}} \mathbb{1}\{V \geq v_{Gt}^d\}$.*

Let $\succeq$ denote stochastic dominance between two random variables, and let $\sim$ denote equality in distribution. Let also $w_d = [P(D_{11} \geq d) - P(D_{10} \geq d)]/[E(D_{11}) - E(D_{10})]$.

Theorem 6. *Suppose that Assumptions 1 and 3' hold, that $D_{01} \sim D_{00}$, and that $D_{11} \succeq D_{10}$.*

(1) If Assumptions 4–5 are satisfied,

$$W_{\text{DID}} = \sum_{d=1}^{\overline{d}} w_d E(Y_{11}(d) - Y_{11}(d-1)|D(0) < d \leq D(1)).$$

(2) If Assumption 4' is satisfied,

$$W_{\text{TC}} = \sum_{d=1}^{\overline{d}} w_d E(Y_{11}(d) - Y_{11}(d-1)|D(0) < d \leq D(1)).$$

(3) If Assumptions 7 and 8 are satisfied,

$$W_{\text{CIC}} = \sum_{d=1}^{\overline{d}} w_d E(Y_{11}(d) - Y_{11}(d-1)|D(0) < d \leq D(1)).$$

Theorem 6 shows that with an ordered treatment, the estimands we considered in the previous sections are equal to the average causal response (ACR) parameter considered in Angrist and Imbens (1995). This parameter is a weighted average, over all values of d, of the effect of increasing treatment from $d-1$ to d among switchers whose treatment status goes from strictly below to above d over time.

For this theorem to hold, two conditions have to be satisfied. First, in the treatment group, the distribution of treatment in period 1 should dominate stochastically the corresponding distribution in period 0. Angrist and Imbens (1995) impose a similar stochastic dominance condition. Actually, this assumption is not necessary for our three estimands to identify a weighted sum of treatment effects. If it is not satisfied, one still has that W_{DID}, W_{TC}, or W_{CIC} identify

$$\sum_{d=1}^{\overline{d}} w_d E(Y_{11}(d)-Y_{11}(d-1)|D(0)<d\leq D(1)\cup D(1)<d\leq D(0)),$$

which is a weighted sum of treatment effects with some negative weights.

Second, the distribution of treatment should be stable over time in the control group. When it is not, one can still obtain some identification results. First, Theorem 1 generalizes to non-binary and ordered treatments. When treatment increases in the control group, the Wald-DID identifies a weighted difference of the ACRs in the treatment and in the control group; when treatment decreases in the control group, the Wald-DID identifies a weighted average of these two ACRs. The weights are the same as those in Theorem 1. Secondly, our partial identification results also generalize to non-binary and ordered treatments. When the distribution of treatment is not stable over time in the control group, the ACR in the treatment group can be bounded under Assumption 4', or Assumptions 7–8, as shown in Subsection 3.2 of the Supplementary Data.

Finally, Theorem 6 extends to a continuous treatment. In such instances, one can show that under an appropriate generalization of Assumption 3, the Wald-DID, Wald-TC, and Wald-CIC identify a weighted average of the derivative of potential outcomes with respect to treatment, a parameter that resembles that studied in Angrist *et al*. (2000).

3.4. *Other extensions*

In the Supplementary Data, we present additional extensions that we discuss briefly here.

3.4.1. Multiple groups and multiple periods. With multiple groups and periods, we show that one can gather groups into "supergroups" for each pair of consecutive dates, depending on whether their treatment increases, is stable, or decreases. Then, a properly weighted sum of the estimands for each pair of dates identifies a weighted average of the LATEs of units switching at any point in time.

3.4.2. Particular fuzzy designs. Up to now, we have considered both general fuzzy designs where the $P(D_{gt}=d)$s are restricted only by Assumption 1, and the special case where Assumption 2 is satisfied. In our Supplementary Data, we consider two other interesting special cases. First, we show that when $P(D_{00}=1)=P(D_{01}=1)=P(D_{10}=1)=0$, identification of the average treatment effect on the treated can be obtained under the same assumptions as those of the standard DID or CIC model. Second, we consider the case where $P(D_{00}=0)=P(D_{01}=0)\in\{0,1\}$. Such situations arise when a policy is extended to a previously ineligible group, or when a program or a technology previously available in some geographic areas is extended to others (see, *e.g.*, Field, 2007). One can show that Theorem 1 still holds in this special case. On the other hand, Theorems 2–3 do not hold. In such instances, we obtain identification by supposing that $Y(0)$ and $Y(1)$ change similarly over time.

3.4.3. Including covariates. We also propose Wald-DID, Wald-TC, and Wald-CIC estimands with covariates. Including covariates in the analysis has two advantages. First, our estimands with covariates rely on conditional versions of our assumptions, which might be more plausible than their unconditional counterparts. Second, there might be instances where $P(D_{00}=d)\neq P(D_{01}=d)$ but $P(D_{00}=d|X)=P(D_{01}=d|X)>0$ almost surely for some covariates

X, meaning that in the control group, the change in the treatment rate is driven by a change in the distribution of X. If that is the case, one can use our results with covariates to point identify treatment effects among switchers, while one would only obtain bounds without covariates.

3.4.4. Panel data. Finally, we discuss the plausibility of our assumptions when panel data are available. First, Assumption 3 is well suited for repeated cross sections or cohort data, but less so for panel data. Then, it implies that within each group units can switch treatment in only one direction, because V does not depend on time. Actually, this assumption is not necessary for our results to hold. For instance, Theorem 1 still holds if Assumptions 3 and 5 are replaced by

$$D_{it} = \mathbb{1}\{V_{it} \geq v_{G_i t}\}, \quad V_{i1}|G_i \sim V_{i0}|G_i \tag{3.2}$$

and

$$E(Y_{i1}(1) - Y_{i1}(0)|G_i, V_{i1} \geq v_{G_i 0}) = E(Y_{i0}(1) - Y_{i0}(0)|G_i, V_{i0} \geq v_{G_i 0}),$$

where we index random variables by i to distinguish individual effects from constant terms. The result applies to treatment and control group switchers, respectively defined as $S_i = \{V_{i1} \in [v_{11}, v_{10}), G_i = 1\}$ and $S'_i = \{V_{i1} \in [\min(v_{01}, v_{00}), \max(v_{01}, v_{00})), G_i = 0\}$. Theorems 2 and 3 also still hold if Assumption 3 is replaced by Equation (3.2), under modifications of Assumptions 4' and 7 that we present in our Supplementary Data.

Secondly, we discuss the validity of our estimands under Equation (3.2) and the following model:

$$Y_{it} = \Lambda(\alpha_i + \gamma_t + [\beta_i + \lambda_t]D_{it} + \varepsilon_{it}),$$

with

$$(\alpha_i, \beta_i)|G_i, V_{i1} \geq v_{G_i 0} \sim (\alpha_i, \beta_i)|G_i, V_{i0} \geq v_{G_i 0},$$
$$(\alpha_i, \beta_i)|G_i, V_{i1} < v_{G_i 0} \sim (\alpha_i, \beta_i)|G_i, V_{i0} < v_{G_i 0},$$

and $\Lambda(.)$ strictly increasing. We prove that Δ is identified by the Wald-DID, Wald-TC, or Wald-CIC estimand under alternative restrictions on $\Lambda(.)$, λ_t, and the distribution of ε_{it}.

4. ESTIMATION AND INFERENCE

In this section, we study the asymptotic properties of the estimators corresponding to the estimands introduced in Section 2, assuming we have an i.i.d. sample with the same distribution as (Y, D, G, T).

Assumption 11. (Independent and identically distributed observations). $(Y_i, D_i, G_i, T_i)_{i=1,...,n}$ *are i.i.d.*

Let $\mathcal{I}_{gt} = \{i : G_i = g, T_i = t\}$ (resp. $\mathcal{I}_{dgt} = \{i : D_i = d, G_i = g, T_i = t\}$) and n_{gt} (resp. n_{dgt}) denote the size of $\mathcal{I}_{gt}$ (resp. $\mathcal{I}_{dgt}$) for all $(d, g, t) \in \{0,1\}^3$. For $d \in \{0,1\}$, let

$$\widehat{\delta}_d = \frac{1}{n_{d01}} \sum_{i \in \mathcal{I}_{d01}} Y_i - \frac{1}{n_{d00}} \sum_{i \in \mathcal{I}_{d00}} Y_i.$$

Let

$$\widehat{W}_{\text{DID}} = \frac{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} Y_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} Y_i - \frac{1}{n_{01}}\sum_{i\in\mathcal{I}_{01}} Y_i + \frac{1}{n_{00}}\sum_{i\in\mathcal{I}_{00}} Y_i}{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} D_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} D_i - \frac{1}{n_{01}}\sum_{i\in\mathcal{I}_{01}} D_i + \frac{1}{n_{00}}\sum_{i\in\mathcal{I}_{00}} D_i},$$

$$\widehat{W}_{\text{TC}} = \frac{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} Y_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} \left[Y_i + \widehat{\delta}_{D_i}\right]}{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} D_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} D_i}$$

denote our Wald-DID and Wald-TC estimators.

Let $\widehat{F}_{Y_{dgt}}$ denote the empirical cdf of Y on the subsample $\mathcal{I}_{dgt}$, $\widehat{F}_{Y_{dgt}}(y) = \sum_{i\in\mathcal{I}_{dgt}} \mathbb{1}\{Y_i \leq y\}/n_{dgt}$. Let $\widehat{F}^{-1}_{Y_{dgt}}(q) = \inf\{y : \widehat{F}_{Y_{dgt}}(y) \geq q\}$ denote the empirical quantile of order $q \in (0,1)$ of Y_{dgt}. Let $\widehat{Q}_d = \widehat{F}^{-1}_{Y_{d01}} \circ \widehat{F}_{Y_{d00}}$ denote the estimator of the quantile-quantile transform. Let

$$\widehat{W}_{\text{CIC}} = \frac{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} Y_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} \widehat{Q}_{D_i}(Y_i)}{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} D_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} D_i}$$

denote the Wald-CIC estimator. Let $\widehat{P}(D_{gt} = d)$ denote the proportion of units with $D = d$ in the sample $\mathcal{I}_{gt}$, let $\widehat{H}_d = \widehat{F}_{Y_{d10}} \circ \widehat{F}^{-1}_{Y_{d00}}$, and let

$$\widehat{F}_{Y_{11}(d)|S} = \frac{\widehat{P}(D_{10} = d)\widehat{H}_d \circ \widehat{F}_{Y_{d01}} - \widehat{P}(D_{11} = d)\widehat{F}_{Y_{d11}}}{\widehat{P}(D_{10} = d) - \widehat{P}(D_{11} = d)}.$$

Finally, let

$$\widehat{\tau}_q = \widehat{F}^{-1}_{Y_{11}(1)|S}(q) - \widehat{F}^{-1}_{Y_{11}(0)|S}(q)$$

denote the estimator of the LQTE of order q for switchers.

We derive the asymptotic behaviour of our CIC estimators under the following assumption, which is similar to the one made by Athey and Imbens (2006) for the CIC estimators in sharp settings.

Assumption 12. (Regularity conditions for the CIC estimators). *$\mathcal{S}(Y)$ is a bounded interval $[\underline{y}, \overline{y}]$. Moreover, for all $(d,g,t) \in \{0,1\}^3$, $F_{Y_{dgt}}$ and $F_{Y_{11}(d)|S}$ are continuously differentiable with strictly positive derivatives on $[\underline{y}, \overline{y}]$.*

Theorem 7 shows that all our estimators are root-n consistent and asymptotically normal. We also derive the influence functions of our estimators. However, because these influence functions take complicated expressions, using the bootstrap might be convenient for inference. For any statistic T, we let T^b denote its bootstrap counterpart. For any root-n consistent statistic $\widehat{\theta}$ estimating consistently θ, we say that the bootstrap is consistent if with probability one and conditional on the sample, $\sqrt{n}(\widehat{\theta}^b - \widehat{\theta})$ converges to the same distribution as the limit distribution of $\sqrt{n}(\widehat{\theta} - \theta)$ (see, *e.g.*, van der Vaart, 2000, Section 23.2.1, for a formal definition of conditional convergence). Theorem 7 implies that bootstrap confidence intervals are asymptotically valid for all our estimators.

Theorem 7. *Suppose that Assumptions 1–3 and 11 hold. Then*

(1) If $E\left(Y^2\right) < \infty$ *and Assumptions 4–5 also hold,*

$$\sqrt{n}\left(\widehat{W}_{\text{DID}} - \Delta\right) \xrightarrow{L} \mathcal{N}(0, V(\psi_{\text{DID}})),$$

where ψ_{DID} *is defined in Equation* (A29) *in the Appendix. Moreover, the bootstrap is consistent for* $\widehat{W}_{\text{DID}}$.

(2) If $E\left(Y^2\right) < \infty$ *and Assumption 4' also holds,*

$$\sqrt{n}\left(\widehat{W}_{\text{TC}} - \Delta\right) \xrightarrow{L} \mathcal{N}(0, V(\psi_{\text{TC}})),$$

where ψ_{TC} *is defined in Equation* (A30) *in the Appendix. Moreover, the bootstrap is consistent for* $\widehat{W}_{\text{TC}}$.

(3) If Assumptions 7, 8 and 12 also hold,

$$\sqrt{n}\left(\widehat{W}_{\text{CIC}} - \Delta\right) \xrightarrow{L} \mathcal{N}(0, V(\psi_{\text{CIC}})),$$

$$\sqrt{n}\left(\widehat{\tau}_q - \tau_q\right) \xrightarrow{L} \mathcal{N}\left(0, V\left(\psi_{q,CIC}\right)\right),$$

where ψ_{CIC} *and* $\psi_{q,CIC}$ *are defined in Equations* (A31) *and* (A32) *in the Appendix. Moreover, the bootstrap is consistent for both estimators.*

The result is straightforward for $\widehat{W}_{\text{DID}}$ and $\widehat{W}_{\text{TC}}$. Regarding $\widehat{W}_{\text{CIC}}$ and $\widehat{\tau}_q$, our proof differs from the one of Athey and Imbens (2006). It is based on the weak convergence of the empirical cdfs of the different subgroups, and on a repeated use of the functional delta method. This approach can be readily applied to other functionals of $(F_{Y_{11}(0)|S}, F_{Y_{11}(1)|S})$.

In our Supplementary Data, we extend the asymptotic theory presented here in several directions. First, in applications with multiple groups, one sometimes needs to estimate the supergroups $\mathcal{G}_s, \mathcal{G}_i$, and $\mathcal{G}_d$ introduced in Subsection 3.2. We propose an estimation procedure, and show that when the number of groups is fixed, as is the case in the framework we considered in Subsection 3.2, this first-step estimation of the supergroups does not have any impact on the asymptotic variances of our estimators. Secondly, we show that we can allow for clustering. Even with repeated cross section or cohort data, independence is a strong assumption in DID analysis: clustering at the group level can induce both cross-sectional and serial correlation within clusters (see, *e.g.*, Bertrand *et al.*, 2004). Thirdly, we consider estimators of the bounds presented in Subsection 3.1 and we derive their limiting distributions. Finally, we consider estimators incorporating covariates and we also derive their limiting distributions.

5. APPLICATION: RETURNS TO SCHOOLING IN INDONESIA

5.1. *Results with the same control and treatment groups as in Duflo (2001)*

Duflo (2001) uses the 1995 intercensal survey of Indonesia to measures returns to education among men. In 1973–74, the Indonesian government launched a major primary school construction program, the so-called INPRES program. In the author's analysis, year of birth plays the role of time, as it determines exposure to the program. The author defines men born between 1957 and 1962 as her cohort 0, as they had finished primary school by the time the program was launched. She defines men born between 1968 and 1972 as her cohort 1, as they entered primary school after

TABLE 1
Returns to education using the groups in Duflo (2001)

	Estimate	95% CI
Wald-DID	0.195	[−0.102, 0.491]
2SLS with fixed effects	0.073	[−0.011, 0.157]
TC bounds	[−3.70, 2.18]	[−5.29, 3.00]
CIC bounds	[−5.60, 3.36]	[−8.00, 4.63]

Notes: Sample size 30,828 observations. Confidence intervals account for clustering at the district level.

the program was launched. The author constructs two "supergroups" of districts, by regressing the number of primary schools constructed on the number of school-age children in each district. She defines treatment districts as those with a positive residual in that regression. She starts by using a simple Wald-DID with her two groups of districts and cohorts to estimate returns to education. She also estimates a 2SLS regression of wages on cohort dummies, district dummies, and years of schooling, using the interaction of cohort 1 and schools constructed in one's district of birth as the instrument for years of schooling.

As an alternative, we apply our results to the author's data. Because years of schooling changed between cohorts 0 and 1 in her control group, we estimate bounds for returns to schooling. These bounds are similar to those presented in Subsection 3.1, but account for the fact that schooling is not binary (see Subsection 3.2 of the Supplementary Data for more details). We estimate TC bounds relying on Assumptions 1, 3, and 4', and CIC bounds relying on Assumptions 1, 3, and 7–8. Because the support of wages does not have natural boundaries, we use the lowest and highest wage in the sample.

Results are shown in Table 1.[9] The Wald-DID is large. However, it is not significantly different from 0, which is the reason why the author turns to a 2SLS regression with cohort and district dummies. The estimate of returns to schooling in that regression is equal to 7.3% and is more precisely estimated.[10]

To identify switchers' LATE, the Wald-DID and 2SLS estimands rely on the assumption that returns to education are homogeneous between districts. The INPRES program explains a small fraction of the differences in increases in schooling between districts. A district-level regression of the increase in years of schooling between cohort 0 and 1 on primary schools constructed has an R^2 of 0.03 only. Accordingly, years of schooling increased almost as much in the author's control group than in her treatment group: while the average of years of schooling increased by 0.47 between cohort 0 and 1 in her treatment group, it increased by 0.36 in her control group (see Table 3 in Duflo, 2001). Therefore, one can show that under Assumptions 1 and 3–5, the author's Wald-DID is equal to $0.47/0.11 \times ACR - 0.36/0.11 \times ACR'$, where ACR and ACR' respectively denote the ACR parameters we introduced in Section 3.3 in her treatment and control groups. If $ACR \neq ACR'$, this Wald-DID could lie far from both ACR and ACR'.

However, returns to schooling might differ across districts. In cohort 0, years of schooling were substantially higher in control than in treatment districts (see Table 3 in Duflo, 2001). This difference in years of schooling might for instance indicate a higher level of economic

9. Our Wald-DID and 2SLS estimates differ slightly from those of in Duflo (2001) because we were not able to obtain exactly her sample of 31,061 observations.

10. This point estimate was significant at the 5% level in the original paper (see the 3rd line and 1st column of Table 7 in Duflo, 2001). But once clustering standard errors at the district level, which has become standard practice in DID analysis since Bertrand *et al*. (2004), it loses some statistical significance.

development in control districts, in which case demand for skilled labour and returns to education could be higher there.

Our TC and CIC bounds do not rely on this assumption. But because years of schooling changed substantially in the authors' control group, they turn out not to be informative. One could argue that this is due to outliers or measurement error, because we use the minimum and maximum wages in the sample as estimates of the boundaries of the support of wages. Using instead the first and third quartile of wages still yields very uninformative bounds: our TC and CIC bounds are respectively equal to $[-0.79, 0.36]$ and $[-0.97, 0.35]$.

Overall, using the treatment and control groups of districts defined in Duflo (2001) either yields point estimates relying on a questionable assumption, or wide and uninformative bounds.

5.2. *Results with new control and treatment groups*

In this subsection, we propose a method to point estimate returns to schooling without assuming that returns are homogeneous between districts or over time. On that purpose, we form three supergroups of districts depending on whether their average years of schooling increased, remained stable, or decreased between cohorts 0 and 1. Though this approach is inspired from the results of Subsection 3.2, a difficulty here is that the supergroups are not known and need to be estimated. Indeed, years of schooling vary at the individual level, and therefore there is no group where years of schooling are perfectly stable between cohort 0 and 1. In Subsection 2.1 in our Supplementary Data, we propose a method to estimate the supergroups, and we show that when the number of groups is fixed, this first-step estimation does not have any impact on the asymptotic variances of our estimators. We expect this asymptotic framework to provide a good approximation of the finite sample behaviour of our estimators when the size of each group is large compared to the total number of groups. However, in Duflo (2001) there are 284 groups with 109 units on average, so this asymptotic framework is not appropriate. Studying how the first-step estimation of the supergroups should be accounted for in an asymptotic framework where the number of groups goes to infinity is beyond the scope of this article. Therefore, what follows is tentative.[11]

The procedure we use to estimate the supergroups should classify as controls only districts with a stable distribution of education. Any classification method leads us to make two types of errors: classify some districts where the distribution of education remained constant as treatments (type 1 error); and classify some districts where this distribution changed as controls (type 2 error). Type 1 errors are innocuous. For instance, if Assumptions 4 and 5 are satisfied, all control districts have the same evolution of their expected outcome. Misclassifying some as treatment districts leaves our estimators unchanged, up to sampling error. On the other hand, type 2 errors are a more serious concern. They lead us to include districts where the true distribution of education was not stable in our super control group, thus violating one of the requirements of Theorem 5. We therefore choose a method based on chi-squared tests with very liberal level. Specifically, we assign a district to our control group if the p-value of a chi-squared test comparing the distribution of education between the two cohorts in that district is greater than 0.5.

We end up with estimated control ($\widehat{G}^* = 0$) and treatment groups ($\widehat{G}^* = 1$ and $\widehat{G}^* = -1$) respectively made up of 64, 123, and 97 districts. Table 2 shows that in treatment districts where schooling increased, cohort 1 completed one more year of schooling than cohort 0. In treatment

11. There are many other applications of the fuzzy DID method where the set of groups where treatment is stable is known and does not need to be estimated. Examples include Field (2007) or Gentzkow *et al*. (2011), which we revisit in our Supplementary Data.

TABLE 2
Years of schooling completed in the new groups of districts

	Cohort 0	Cohort 1	Evolution	(s.e.)
Districts where schooling increased	8.65	9.64	0.99	(0.082)
Control districts	9.60	9.55	−0.05	(0.097)
Districts where schooling decreased	10.17	9.43	−0.74	(0.080)

Notes: Sample size 30,828 observations. Standard errors are clustered at the district level.

TABLE 3
Returns to education using our new groups

	W_{DID}	W_{TC}	W_{CIC}
Returns to education	0.140	0.101	0.099
	(0.015)	(0.017)	(0.017)

Notes: Sample size 30,828 observations. Standard errors are clustered at the district level.

districts where schooling decreased, cohort 1 completed nine months less of schooling than cohort 0. Finally, in control districts, the number of years of schooling did not change.

We now follow results from Theorem 5 to estimate returns to education in Indonesia, treating our estimated treatment and control groups as if they were the true treatment and control groups. In Table 3, we report the Wald-DID, Wald-TC, and Wald-CIC estimates we obtain.[12] The Wald-DID is large, and suggests returns of 14% per year of schooling. Our Wald-TC and Wald-CIC estimators are substantially smaller, around 10% per year of schooling. They significantly differ from the Wald-DID, with t-stats respectively equal to −4.27 and −4.61.

In this application, Assumption 5 is implausible, because it is incompatible with decreasing returns to experience, while an extensive literature has shown that returns to experience tend to be decreasing (see Mincer and Jovanovic, 1979, and Willis, 1986, for a survey). The data used in this application are a single cross-section, the 1995 intercensal survey of Indonesia, where cohort of birth plays the role of time. Then, Equation (2.1) shows that Assumption 5 implies, for instance, that the wage gap between high-school graduates in cohort 0 and 1 should remain the same if they had only completed primary school. Had they only completed primary school, high-school graduates would have joined the labour market earlier, and would have had more labor market experience in 1995, the year when their wages are measured. If returns to experience are decreasing, the wage gap between the two cohorts would then have been lower.[13] Therefore, using the Wald-TC or Wald-CIC seems preferable to using the Wald-DID estimator, as those two estimators do not rule out decreasing returns to experience.

We also conduct placebo tests, which are presented in Subsection 3.3.2 in the Supplementary Data. Overall, they lend slightly stronger support to our Wald-TC and Wald-CIC estimators, though we lack power to make definitive conclusions. In Subsection 3.3.2 of

12. Our Wald-DID estimate is actually a weighted average of the Wald-DID estimate comparing $\widehat{G}^*=1$ and $\widehat{G}^*=0$ and of the Wald-DID estimate comparing $\widehat{G}^*=-1$ and $\widehat{G}^*=0$. The same holds for the Wald-TC and Wald-CIC estimates. The formulas of the corresponding estimands are given in Theorem 5. Also, to estimate the numerators of the Wald-CICs, we group schooling into five categories (did not complete primary school, completed primary school, completed middle school, completed high school, completed college). Thus, we avoid estimating the Q_ds on small numbers of units. To be consistent, we also use this definition to estimate the numerators of the Wald-TCs. Using years of schooling hardly changes our Wald-TC estimate.

13. In Subsection 3.3.1 of our Supplementary Data , we prove this point formally.

the Supplementary Data, we show that decreasing returns to experience can explain why our placebo DID tests may have low power. Finally, in Subsection 3.4 of the Supplementary Data, we explore the effect of the first-stage estimation of the supergroups on our final estimators. While the corresponding theoretical analysis goes beyond the scope of this article, our robustness checks suggest that our estimates of returns to schooling are not sensitive to this first-step estimation.

6. CONCLUSION

In many applications of the DID method, the treatment increases more in the treatment group, but some units are also treated in the control group, and some units remain untreated in the treatment group. In such fuzzy designs, a popular estimand of treatment effects is the DID of the outcome divided by the DID of the treatment, the so-called Wald-DID estimand. We start by showing that the Wald-DID identifies the LATE of treatment group switchers only if two restrictions on treatment effects are satisfied, in addition to the usual common trend assumption. The average treatment effect of units treated at both dates must not change over time. Moreover, when the share of treated units varies in the control group, the LATEs of treatment and control group switchers must be equal. Second, we propose two new estimands that can be used when the share of treated units in the control group is stable, and that do not rely on any assumption on treatment effects. We use our results to revisit Duflo (2001).

Overall, our article shows that researchers who use the DID method with fuzzy groups can obtain estimates not resting on the assumption that treatment effects are stable and homogeneous, provided they can find a control group whose exposure to the treatment does not change over time. When a policy is extended to a previously ineligible subgroup or when the treatment is assigned at the group level, such control groups are usually readily available. Examples include Field (2007) or Gentzkow *et al*. (2011), which we revisit in our Supplementary Data. When the treatment is assigned at the individual level and no policy rule warrants that treatment remains stable in some groups, such control groups usually still exist but they need to be estimated. We propose a method to estimate the control groups that can be used when the number of groups is small relative to the size of each group. Studying how to estimate the control groups when the number of groups is large relative to their size, as is the case in Duflo (2001), is left for future work. Finally, when researchers cannot find a control group whose exposure to the treatment is stable over time, our results show that their conclusions will rest on the assumption that treatment effects are stable and homogeneous. These assumptions should then be discussed.

APPENDIX

A. MAIN PROOFS

The lemmas prefixed by S are stated and proven in our Supplementary Data (see de Chaisemartin and D'Haultfœuille, 2017). To simplify the notation, throughout the proofs we adopt the following normalization, which is without loss of generality: $v_{10}=v_{00}$. For any $(d,g,t)\in\mathcal{S}((D,G,T))$, we also let $p_{gt}=P(G=g,T=t)$, $p_{dgt}=P(D=d,G=g,T=t)$, $p_{d|gt}=P(D_{gt}=d)$, and $F_{dgt}=F_{Y_{dgt}}$. Finally, for any $\Theta\subset\mathrm{IR}^k$, let $\overset{\circ}{\Theta}$ denote its interior and let $\mathcal{C}^0(\Theta)$ and $\mathcal{C}^1(\Theta)$ denote respectively the set of continuous functions and the set of continuously differentiable functions with strictly positive derivative on Θ. We often use this notation with $\Theta=\mathcal{S}(Y)$, in which case we respectively denote these sets by $\mathcal{C}^0$ and $\mathcal{C}^1$.

Proof of Theorem 1

Proof of 1 when $p_{1|01}\geq p_{1|00}$ It follows from Assumptions 1 and 3 that

$$\begin{aligned} p_{1|11}-p_{1|10} &= P(V\geq v_{11}|G=1,T=1)-P(V\geq v_{00}|G=1,T=0) \\ &= P(S|G=1). \end{aligned} \tag{A3}$$

Moreover,

$$
\begin{aligned}
& E(Y_{11})-E(Y_{10}) \\
= \ & E(Y_{11}(1)[\mathbb{1}\{V_{11}\in[v_{11},v_{00})\}+\mathbb{1}\{V_{11}\geq v_{00}\}])+E(Y_{11}(0)\mathbb{1}\{V_{11}<v_{11}\}) \\
- \ & E(Y_{10}(1)\mathbb{1}\{V_{10}\geq v_{00}\})-E(Y_{10}(0)\mathbb{1}\{V_{10}<v_{00}\}) \\
= \ & E(Y_{11}(1)-Y_{11}(0)|S)P(S|G=1)+E(Y_{11}(0))-E(Y_{10}(0)) \\
+ \ & E((Y_{11}(1)-Y_{11}(0))\mathbb{1}\{V_{11}\geq v_{00}\})-E((Y_{10}(1)-Y_{10}(0))\mathbb{1}\{V_{10}\geq v_{00}\}) \\
= \ & \Delta P(S|G=1)+E(Y_{11}(0))-E(Y_{10}(0)).
\end{aligned} \tag{A4}
$$

The first equality follows from Assumptions 1 and 3, the second follows from Assumption 3, and the third follows from Assumptions 3 and 5.

Similarly, one can show that

$$p_{1|01}-p_{1|00} = P(S'|G=0) \tag{A5}$$

$$E(Y_{01})-E(Y_{00}) = \Delta' P(S'|G=0)+E(Y_{01}(0))-E(Y_{00}(0)). \tag{A6}$$

Taking the difference between Equations (A4) and (A6), and using Assumption 4, we obtain

$$\text{DID}_Y=\Delta P(S|G=1)-\Delta' P(S'|G=0).$$

Dividing each side by DID_D and using Equations (A3) and (A5) yields the result.

Proof of 1 when $p_{1|01}<p_{1|00}$ In this case, reasoning similarly as in the derivation of Equation (A3) yields

$$p_{1|00}-p_{1|01} = P(S'|G=0). \tag{A7}$$

Moreover,

$$
\begin{aligned}
& E(Y_{01})-E(Y_{00}) \\
= \ & E(Y_{01}(1)[\mathbb{1}\{V_{01}\geq v_{00}\}-\mathbb{1}\{V_{01}\in[v_{00},v_{01})\}])+E(Y_{01}(0)\times \\
& [\mathbb{1}\{V_{01}\in[v_{00},v_{01})\}+\mathbb{1}\{V_{01}<v_{00}\}])-E(Y_{00}(1)\mathbb{1}\{V_{00}\geq v_{00}\})-E(Y_{00}(0)\mathbb{1}\{V_{00}<v_{00}\}) \\
= \ & -\Delta' P(S'|G=1)+E(Y_{01}(0)-Y_{00}(0)) \\
+ \ & E((Y_{01}(1)-Y_{01}(0)\mathbb{1}\{V_{01}\geq v_{00}\})-E((Y_{00}(1)-Y_{00}(0)\mathbb{1}\{V_{00}\geq v_{00}\}) \\
= \ & -\Delta' P(S'|G=1)+E(Y_{01}(0)-Y_{00}(0)).
\end{aligned} \tag{A8}
$$

The first equality follows from Assumption 3 and $p_{1|01}<p_{1|00}$, the second follows from Assumption 3, and the third follows from Assumptions 3 and 5. Taking the difference between Equations (A4) and (A8), and using Assumption 4, we obtain

$$\text{DID}_Y=\Delta P(S|G=1)+\Delta' P(S'|G=0).$$

Dividing each side of the previous display by DID_D and using Equations (A3) and (A7) yields the result.

Proof of 2 The result follows directly from the first point of the theorem.

Bias term without Assumption 5 Using the Equations above (A4) and (A8), we obtain that when Assumption 2 holds but Assumption 5 does not,

$$W_{\text{DID}}=\Delta+\frac{1}{\text{DID}_D}\left((\Delta_{11}-\Delta_{10})p_{1|10}-(\Delta_{01}-\Delta_{00})p_{1|00}\right),$$

where $\Delta_{gt}=E\left(Y_{gt}(1)-Y_{gt}(0)|D(0)=1\right)$.

Proof of Theorem 2

Following the same steps as those used to derive Equation (A4), we obtain

$$\begin{aligned} & E(Y_{11})-E(Y_{10}) \\ = {} & E(Y_{11}(1)-Y_{11}(0)|S)P(S|G=1)+E(Y_{11}(1)-Y_{10}(1)|G=1,V\geq v_{00})P(V\geq v_{00}|G=1) \\ + {} & E(Y_{11}(0)-Y_{10}(0)|G=1,V<v_{00})P(V<v_{00}|G=1). \end{aligned} \tag{A9}$$

Then,

$$\begin{aligned} \delta_1 &= E(Y_{01}(1)|G=0,V\geq v_{01})-E(Y_{00}(1)|G=0,V\geq v_{00}) \\ &= E(Y_{01}(1)-Y_{00}(1)|G=0,V\geq v_{00}). \end{aligned} \tag{A10}$$

The first equality follows from Assumption 3. The second one follows from the fact that $p_{1|01}=p_{1|00}$ combined with Assumption 3 implies that $\{G=0,V\leq v_{01}\}=\{G=0,V\leq v_{00}\}$.

Similarly,

$$\delta_0=E(Y_{01}(0)-Y_{00}(0)|G=0,V<v_{00}). \tag{A11}$$

Finally, the result follows combining Equations (A9), (A10), (A11), and Assumption 4', once noted that $p_{1|10}=P(V\geq v_{00}|G=1)$ and $P(S|G=1)=p_{1|11}-p_{1|10}$ □

Proof of Theorem 3

We first prove the following result, which holds irrespective of whether Assumption 2 holds:

$$F_{Y_{11}(d)|S}=\frac{p_{d|11}F_{d11}-p_{d|10}H_d\circ\left(\lambda_{0d}F_{d01}+(1-\lambda_{0d})F_{Y_{01}(d)|S'}\right)}{p_{d|11}-p_{d|10}}, \tag{A12}$$

where $\lambda_{gd}=p_{d|g1}/p_{d|g0}$. We establish (A12) for $d=0$ only, the reasoning is similar for $d=1$.

First,

$$\begin{aligned} P(S|G=1,T=1,V<v_{00}) &= \frac{P(S|G=1)}{P(V<v_{00}|G=1,T=0)} \\ &= \frac{p_{0|10}-p_{0|11}}{p_{0|10}}. \end{aligned}$$

The first equality stems from Assumption 3, the second from Equation (A3) and Assumption 3. Moreover,

$$\begin{aligned} F_{Y_{11}(0)|V<v_{00}} &= P(S|G=1,T=1,V<v_{00})F_{Y_{11}(0)|S} \\ &+ (1-P(S|G=1,T=1,V<v_{00}))F_{Y_{11}(0)|V<v_{11}} \\ &= \frac{p_{0|10}-p_{0|11}}{p_{0|10}}F_{Y_{11}(0)|S}+\frac{p_{0|11}}{p_{0|10}}F_{011}. \end{aligned}$$

Therefore,

$$F_{Y_{11}(0)|S}=\frac{p_{0|11}F_{011}-p_{0|10}F_{Y_{11}(0)|V<v_{00}}}{p_{0|11}-p_{0|10}}. \tag{A13}$$

Then, we show that for all $y\in\mathcal{S}(Y_{11}(0)|V<v_{00})$,

$$F_{Y_{11}(0)|V<v_{00}}=F_{010}\circ F_{000}^{-1}\circ F_{Y_{01}(0)|V<v_{00}}. \tag{A14}$$

For all $(g,t)\in\{0,1\}^2$,

$$\begin{aligned} F_{Y_{gt}(0)|V<v_{00}}(y) &= P(h_0(U_0,t)\leq y|G=g,T=t,V<v_{00}) \\ &= P(U_0\leq h_0^{-1}(y,t)|G=g,V<v_{00}) \\ &= F_{U_0|G=g,V<v_{00}}(h_0^{-1}(y,t)), \end{aligned}$$

where the first and second equalities follow from Assumption 7. Assumptions 7 and 8 imply that $F_{U_0|G=g,V<v_{00}}$ is strictly increasing. Hence, its inverse exists and for all $q\in(0,1)$,

$$F_{Y_{gt}(0)|V<v_{00}}^{-1}(q)=h_0\left(F_{U_0|G=g,V<v_{00}}^{-1}(q),t\right).$$

This implies that for all $y\in\mathcal{S}(Y_{g1}(0)|V<v_{00})$,

$$F^{-1}_{Y_{g0}(0)|V<v_{00}}\circ F_{Y_{g1}(0)|V<v_{00}}(y)=h_0(h_0^{-1}(y,1),0). \tag{A15}$$

By Assumption 8, we have

$$\begin{aligned}&\mathcal{S}(Y_{010})=\mathcal{S}(Y_{000})\\ &\Rightarrow \mathcal{S}(Y_{10}(0)|V<v_{00})=\mathcal{S}(Y_{00}(0)|V<v_{00})\\ &\Rightarrow \mathcal{S}(h_0(U_0,0)|V<v_{00},G=1,T=0)=\mathcal{S}(h_0(U_0,0)|V<v_{00},G=0,T=0)\\ &\Rightarrow \mathcal{S}(U_0|V<v_{00},G=1)=\mathcal{S}(U_0|V<v_{00},G=0)\\ &\Rightarrow \mathcal{S}(h_0(U_0,1)|V<v_{00},G=1,T=1)=\mathcal{S}(h_0(U_0,1)|V<v_{00},G=0,T=1)\\ &\Rightarrow \mathcal{S}(Y_{11}(0)|V<v_{00})=\mathcal{S}(Y_{01}(0)|V<v_{00}),\end{aligned}$$

where the third and fourth implications follow from Assumption 7. Once combined with Equation (A15), the previous display implies that for all $y\in\mathcal{S}(Y_{11}(0)|V<v_{00})$,

$$F^{-1}_{Y_{10}(0)|V<v_{00}}\circ F_{Y_{11}(0)|V<v_{00}}(y)=F^{-1}_{Y_{00}(0)|V<v_{00}}\circ F_{Y_{01}(0)|V<v_{00}}(y).$$

This proves Equation (A14), because $\{V<v_{00},G=g,T=0\}=\{D=0,G=g,T=0\}$.

Finally, we show that

$$F_{Y_{01}(0)|V<v_{00}}=\lambda_{00}F_{001}+(1-\lambda_{00})F_{Y_{01}(0)|S'}. \tag{A16}$$

Suppose first that $\lambda_{00}\le 1$. Then, $v_{01}\le v_{00}$ and $S'=\{V\in[v_{01},v_{00}),G=0\}$. Moreover, reasoning as for $P(S|G=1,V<v_{00})$, we get

$$\begin{aligned}\lambda_{00} &= \frac{P(V<v_{01}|G=0)}{P(V<v_{00}|G=0)}=P(V<v_{01}|G=0,V<v_{00}),\\ F_{Y_{01}(0)|V<v_{00}} &= \lambda_{00}F_{001}+(1-\lambda_{00})F_{Y_{01}(0)|S'}.\end{aligned}$$

If $\lambda_{00}>1$, $v_{01}>v_{00}$ and $S'=\{V\in[v_{00},v_{01}),G=0\}$. We then have

$$\begin{aligned}1/\lambda_{00} &= P(V<v_{00}|G=0,V<v_{01}),\\ F_{001} &= 1/\lambda_{00}F_{Y_{01}(0)|V<v_{00}}+(1-1/\lambda_{00})F_{Y_{01}(0)|S'},\end{aligned}$$

so Equation (A16) is also satisfied. (A12) follows by combining (A13), (A14), and (A16).

Now, under Assumption 2, $\lambda_{00}=\lambda_{01}=1$. This and the fact that $H_d\circ F_{d01}=F_{Q_d(Y_{d10})}$ shows that $F_{Y_{11}(d)|S}=F_{\text{CIC},d}$. This proves that $\tau_q=F^{-1}_{\text{CIC},1}(q)-F^{-1}_{\text{CIC},0}(q)$. Moreover,

$$\begin{aligned}W_{\text{CIC}} &= \frac{p_{1|11}\int ydF_{Y_{111}}(y)-p_{1|10}\int ydF_{Q_1(Y_{101})}(y)}{p_{1|11}-p_{1|10}}\\ &- \frac{p_{0|11}\int ydF_{Y_{011}}(y)-p_{0|10}\int ydF_{Q_0(Y_{001})}(y)}{p_{0|11}-p_{0|10}}\\ &= \int ydF_{\text{CIC},1}(y)-\int ydF_{\text{CIC},0}(y)\\ &= E(Y_{11}(1)-Y_{11}(0)|S).\end{aligned}$$

Proof of Theorem 4

Proof of 1. We only prove that $\underline{W}_{\text{TC}}$ is a lower bound when $\lambda_{00}>1$. The proofs for the upper bound and when $\lambda_{00}<1$ are symmetric.

We have

$$\begin{aligned}&E(Y_{11}(1)-Y_{11}(0)|S)P(S|G=1)\\ =\ & E(Y_{11})-E(Y_{10})-E(Y_{11}(1)-Y_{10}(1)|V\ge v_{00})P(V\ge v_{00}|G=1)\\ -\ & E(Y_{11}(0)-Y_{10}(0)|V<v_{00})P(V<v_{00}|G=1)\\ =\ & E(Y_{11})-E(Y_{10})-E(Y_{01}(1)-Y_{00}(1)|V\ge v_{00})p_{1|10}\\ -\ & E(Y_{01}(0)-Y_{00}(0)|V<v_{00})p_{0|10}\\ =\ & E(Y_{11})-E(Y_{10})-(E(Y_{01}(1)|V\ge v_{00})-E(Y_{100}))p_{1|10}\\ -\ & (E(Y_{01}(0)|V<v_{00})-E(Y_{000}))p_{0|10}.\end{aligned}$$

The first equality follows from Equation (A9), the second from Assumptions 3 and 4'. Thus, the proof will be complete if we can show that $\overline{\delta}_1$ and $\overline{\delta}_0$ are respectively upper bounds for $E(Y_{01}(1)|V \geq v_{00})-E(Y_{100})$ and $E(Y_{01}(0)|V<v_{00})-E(Y_{000})$.

When $\lambda_{00}>1$, Assumption 3 implies that $v_{00}<v_{01}$. Then, reasoning as in the proof of Theorem 3, we obtain $P(V \geq v_{01}|G=0,T=1,V \geq v_{00})=\lambda_{01}$ and

$$\begin{aligned} E(Y_{01}(1)|V \geq v_{00}) &= \lambda_{01}E(Y_{01}(1)|V \geq v_{01})+(1-\lambda_{01})E(Y_{01}(1)|S') \\ &\leq \lambda_{01}E(Y_{101})+(1-\lambda_{01})\overline{y}=\int yd\underline{F}_{101}(y). \end{aligned}$$

This proves that $\overline{\delta}_1$ is an upper bound for $E(Y_{01}(1)|V \geq v_{00})-E(Y_{100})$.

Similarly, $P(V<v_{00}|G=0,T=1,V<v_{01})=1/\lambda_{00}$ and

$$F_{Y_{001}} = 1/\lambda_{00}F_{Y_{01}(0)|V<v_{00}}+(1-1/\lambda_{00})F_{Y_{01}(0)|S'}.$$

Following Horowitz and Manski (1995), the last display implies that

$$E(Y_{01}(0)|V<v_{00}) \leq \int yd\underline{F}_{001}(y).$$

This proves that $\overline{\delta}_0$ is an upper bound for $E(Y_{01}(0)|V<v_{00})-E(Y_{000})$.

Proof of 2. Construction of the bounds.

We only establish the validity of the bounds for $F_{Y_{11}(0)|S}(y)$. The reasoning is similar for $F_{Y_{11}(1)|S}(y)$. Bounds for Δ and τ_q directly follow from those for the cdfs. Hereafter, to simplify the notation, we let $T_0=F_{Y_{01}(0)|S'}$. Following the notation introduced in Subsection 3.1 combined with Equations (A12) and (A16), we then have $G_0(T_0)=F_{Y_{01}(0)|V<v_{00}}$ and $C_0(T_0)=F_{Y_{11}(0)|S}$.

We start by considering the case where $\lambda_{00}<1$. We first show that in such instances, $0 \leq T_0, G_0(T_0), C_0(T_0) \leq 1$ if and only if

$$\underline{T_0} \leq T_0 \leq \overline{T_0}. \tag{A17}$$

$G_0(T_0)$ is included between 0 and 1 if and only if

$$\frac{-\lambda_{00}F_{001}}{1-\lambda_{00}} \leq T_0 \leq \frac{1-\lambda_{00}F_{001}}{1-\lambda_{00}},$$

while $C_0(T_0)$ is included between 0 and 1 if and only if

$$\frac{H_0^{-1}(\lambda_{10}F_{011})-\lambda_{00}F_{001}}{1-\lambda_{00}} \leq T_0 \leq \frac{H_0^{-1}(\lambda_{10}F_{011}+(1-\lambda_{10}))-\lambda_{00}F_{001}}{1-\lambda_{00}}.$$

Since $-\lambda_{00}F_{001}/(1-\lambda_{00}) \leq 0$ and $(1-\lambda_{00}F_{001})/(1-\lambda_{00}) \geq 1$, T_0, $G_0(T_0)$ and $C_0(T_0)$ are all included between 0 and 1 if and only if

$$M_0\left(\frac{H_0^{-1}(\lambda_{10}F_{011})-\lambda_{00}F_{001}}{1-\lambda_{00}}\right) \leq T_0 \leq m_1\left(\frac{H_0^{-1}(\lambda_{10}F_{011}+(1-\lambda_{10}))-\lambda_{00}F_{001}}{1-\lambda_{00}}\right), \tag{A18}$$

where $M_0(x)=\max(0,x)$ and $m_1(x)=\min(1,x)$. Composing each term of these inequalities by $M_0(.)$ and then by $m_1(.)$ yields Equation (A17), since $M_0(T_0)=m_1(T_0)=T_0$ and $M_0 \circ m_1=m_1 \circ M_0$.

Now, when $\lambda_{00}<1$, $G_0(T_0)$ is increasing in T_0, so $C_0(T_0)$ as well is increasing in T_0. Combining this with (A17) implies that for every y',

$$C_0(\underline{T_0})(y') \leq C_0(T_0)(y') \leq C_0(\overline{T_0})(y'). \tag{A19}$$

Because $C_0(T_0)(y)$ is a cdf,

$$C_0(T_0)(y)=\inf_{y' \geq y} C_0(T_0)(y') \leq \inf_{y' \geq y} C_0(\overline{T_0})(y')=\overline{F}_{\mathrm{CIC},0}(y).$$

This proves the result for the upper bound. The result for the lower bound follows similarly.

Let us now turn to the case where $\lambda_{00}>1$. Using the same reasoning as above, we get that $G_0(T_0)$ and $C_0(T_0)$ are included between 0 and 1 if and only if

$$\frac{\lambda_{00}F_{001}-1}{\lambda_{00}-1} \leq T_0 \leq \frac{\lambda_{00}F_{001}}{\lambda_{00}-1},$$

$$\frac{\lambda_{00}F_{001}-H_0^{-1}(\lambda_{10}F_{011}+(1-\lambda_{10}))}{\lambda_{00}-1} \leq T_0 \leq \frac{\lambda_{00}F_{001}-H_0^{-1}(\lambda_{10}F_{011})}{\lambda_{00}-1}.$$

The inequalities in the first line are not binding since they are implied by those on the second line. Thus, we also get (A18). Hence, $0 \leq T_0, G_0(T_0), C_0(T_0) \leq 1$ if and only if

$$\overline{T_0} \leq T_0 \leq \underline{T_0}. \tag{A20}$$

Besides, when $\lambda_{00} > 1$, $G_0(T_0)$ is decreasing in T_0, so $C_0(T_0)$ is also decreasing in T_0. Combining this with Equation (A20) implies that for every y, Equation (A19) holds as well. This proves the result.

Sketch of the proof of sharpness.

The full proof is in the Supplementary Data (see de Chaisemartin and D'Haultfœuille, 2017). We only consider the sharpness of $\underline{F}_{\text{CIC},0}$, the reasoning being similar for the upper bound. The proof is also similar and actually simpler for $d=1$. The corresponding bounds are proper cdfs, so we do not have to consider converging sequences of cdfs as we do in case b) below.

a. $\lambda_{00} > 1$. We show that if Assumptions 8–10 hold, then $\underline{F}_{\text{CIC},0}$ is sharp. For that purpose, we construct $\widetilde{h}_0, \widetilde{U}_0, \widetilde{V}$ such that:

(i) $Y = \widetilde{h}_0(\widetilde{U}_0, T)$ when $D=0$ and $D = \mathbb{1}\{\widetilde{V} \geq v_{GT}\}$;

(ii)$(\widetilde{U}_0, \widetilde{V}) \perp\!\!\!\perp T|G$;

(iii)$\widetilde{h}_0(., t)$ is strictly increasing for $t \in \{0,1\}$;

(iv)$F_{\widetilde{h}_0(\widetilde{U}_0,1)|G=0,T=1,\widetilde{V}\in[v_{00},v_{01})} = \underline{T}_0$.

Because we can always define $\widetilde{Y}(0)$ as $\widetilde{h}_0(\widetilde{U}_0, T)$ when $D=1$ without contradicting the data, (i)–(iii) ensures that Assumptions 3 and 7 (for $d=0$) are satisfied with $\widetilde{h}_0, \widetilde{U}_0$ and $\widetilde{V}$. (iv) ensures that the DGP corresponding to $(\widetilde{h}_0, \widetilde{U}_0, \widetilde{V})$ rationalizes the bound.

The construction of $\widetilde{h}_0$, $\widetilde{U}_0$, and $\widetilde{V}$ is long, so its presentation is deferred to the Supplementary Data.

b. $\lambda_{00} < 1$. The idea is similar as in the previous case. A difference, however, is that when $\lambda_{00} < 1$, $\underline{T}_0$ is not a proper cdf, but a defective one, since $\lim_{y \to \overline{y}} \underline{T}_0(y) < 1$. As a result, we cannot define a DGP such that $\widetilde{T}_0 = \underline{T}_0$, However, by Lemma S2, there exists a sequence $(\underline{T}_0^k)_{k\in\mathbb{N}}$ of cdfs such that $\underline{T}_0^k \to \underline{T}_0$, $G_0(\underline{T}_0^k)$ is an increasing bijection from $\mathcal{S}(Y)$ to $(0,1)$ and $C_0(\underline{T}_0^k)$ is increasing and onto $(0,1)$. We can then construct a sequence of DGP $(\widetilde{h}_0^k(.,0), \widetilde{h}_0^k(.,1), \widetilde{U}_0^k, \widetilde{V}^k)$ such that Points (i) to (iii) listed above hold for every k, and such that $\widetilde{T}_0^k = \underline{T}_0^k$. Since $\underline{T}_0^k(y)$ converges to $\underline{T}_0(y)$ for every y in $\overset{\circ}{\mathcal{S}}(Y)$, we thus define a sequence of DGP such that $\widetilde{T}_0^k$ can be arbitrarily close to $\underline{T}_0$ on $\overset{\circ}{\mathcal{S}}(Y)$ for sufficiently large k. Since $C_0(.)$ is continuous, this proves that $\underline{F}_{\text{CIC},0}$ is sharp on $\overset{\circ}{\mathcal{S}}(Y)$. Again, this construction is long and its exposition is deferred to the Supplementary Data.

Proof of Theorem 5

We prove the first statement, the second and third following from similar arguments. Under the assumptions of the theorem, Assumptions 1–5 are satisfied for the treatment and control groups $G^*=1$ and $G^*=0$. Therefore, it follows from Theorem 1 that

$$W_{\text{DID}}^*(1,0) = E(Y(1)-Y(0)|S^*, G^*=1, T=1). \tag{A21}$$

Similarly, one can show that

$$W_{\text{DID}}^*(-1,0) = E(Y(1)-Y(0)|S^*, G^*=-1, T=1). \tag{A22}$$

Moreover, by Assumption 3 and $G \perp\!\!\!\perp T$,

$$\begin{aligned}\text{DID}_D^*(1,0)P(G^*=1) &= \big[E(D|G^*=1,T=1)-E(D|G^*=1,T=0)\big]P(G^*=1|T=1)\\ &= P(S^*, G^*=1|T=1).\end{aligned}$$

Similarly, $\text{DID}_D^*(0,-1)P(G^*=-1) = P(S^*, G^*=-1|T=1)$. Combining both equalities, we obtain

$$w_{10} = \frac{P(S^*, G^*=1|T=1)}{P(S^*, G^*=1|T=1)+P(S^*, G^*=-1|T=1)} = P(G^*=1|S^*, T=1). \tag{A23}$$

The result follows combining Equations (A21)–(A23).

Proof of Theorem 6

We only prove the first statement, the second and third statements follow from similar arguments.

$D_{01} \sim D_{00}$ and $D_{11} \succeq D_{10}$ combined with Assumption 3' imply that for all $d \in \{1,...,\overline{d}\}$,

$$v_{01}^d = v_{00}^d, \tag{A24}$$

$$v_{11}^d \leq v_{10}^d. \tag{A25}$$

Then, it follows from Assumption 3' and Equation (A25) that for every $d \in \{1,...,\overline{d}\}$,

$$\begin{aligned} P(D_{11} \geq d) - P(D_{10} \geq d) &= P(V \geq v_{11}^d | T=1, G=1) - P(V \geq v_{10}^d | T=0, G=1) \\ &= P(V \in [v_{11}^d, v_{10}^d) | G=1) \\ &= P(D(0) < d \leq D(1) | G=1). \end{aligned} \tag{A26}$$

Then, for every $g \in \{0,1\}$,

$$\begin{aligned} & E(Y_{g1}) - E(Y_{g0}) \\ &= \sum_{d=0}^{\overline{d}} E(Y_{g1}(d) | G=g, V \in [v_{g1}^d, v_{g1}^{d+1})) P(V \in [v_{g1}^d, v_{g1}^{d+1}) | G=g) \\ &- \sum_{d=0}^{\overline{d}} E(Y_{g0}(d) | G=g, V \in [v_{g0}^d, v_{g0}^{d+1})) P(V \in [v_{g0}^d, v_{g0}^{d+1}) | G=g) \\ &= \sum_{d=1}^{\overline{d}} E(Y_{g1}(d) - Y_{g1}(d-1) | G=g, V \in [v_{g1}^d, v_{g0}^d)) P(V \in [v_{g1}^d, v_{g0}^d) | G=g) \\ &+ \sum_{d=0}^{\overline{d}} E(Y_{g1}(d) - Y_{g0}(d) | G=g, V \in [v_{g0}^d, v_{g0}^{d+1})) P(V \in [v_{g0}^d, v_{g0}^{d+1}) | G=g) \\ &= \sum_{d=1}^{\overline{d}} E(Y_{g1}(d) - Y_{g1}(d-1) | D(0) < d \leq D(1)) P(D(0) < d \leq D(1) | G=g) \\ &+ E(Y_{g1}(0)) - E(Y_{g0}(0)). \end{aligned} \tag{A27}$$

The first, second and third equalities respectively follow from Assumption 3', Equations (A24) and (A25) combined with Assumption 3', and Assumptions 3' and 5.

Combining Equation (A27) with Equation (A24) and Assumption 4 imply that

$$\text{DID}_Y = \textstyle\sum_{d=1}^{\overline{d}} E(Y_{11}(d) - Y_{11}(d-1) | D(0) < d \leq D(1)) P(D(0) < d \leq D(1) | G=1).$$

The result follows from Equation (A26), after dividing each side of the previous display by DID_D.

Proof of Theorem 7

Proof of 1 and 2 Asymptotic normality is obvious by the central limit theorem and the delta method. Consistency of the bootstrap follows by consistency of the bootstrap for sample means (see, *e.g.*, van der Vaart, 2000, Theorem 23.4) and the delta method for bootstrap (van der Vaart, 2000, Theorem 23.5). A convenient way to obtain the asymptotic variance is to use repeatedly the following argument. If

$$\sqrt{n}\left(\widehat{A} - A\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n a_i + o_P(1), \; \sqrt{n}\left(\widehat{B} - B\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n b_i + o_P(1),$$

then Lemma S3 ensures that

$$\sqrt{n}\left(\frac{\widehat{A}}{\widehat{B}} - \frac{A}{B}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{a_i - (A/B) b_i}{B} + o_P(1). \tag{A28}$$

This implies for instance that

$$\sqrt{n}\left(\widehat{E}(Y_{11}) - E(Y_{11})\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{G_i T_i (Y_i - E(Y_{11}))}{p_{11}} + o_P(1),$$

and similarly for $\widehat{E}(D_{11})$. Applying repeatedly this argument, we obtain, after some algebra,

$$\sqrt{n}\left(\widehat{W}_{\text{DID}}-\Delta\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{DID,i}+o_P(1),$$

where, omitting the index i, ψ_{DID} is defined by

$$\psi_{\text{DID}}=\frac{1}{\text{DID}_D}\left[\frac{GT(\varepsilon-E(\varepsilon_{11}))}{p_{11}}-\frac{G(1-T)(\varepsilon-E(\varepsilon_{10}))}{p_{10}}-\frac{(1-G)T(\varepsilon-E(\varepsilon_{01}))}{p_{01}}\right.$$
$$\left.+\frac{(1-G)(1-T)(\varepsilon-E(\varepsilon_{00}))}{p_{00}}\right] \tag{A29}$$

and $\varepsilon=Y-\Delta D$. Similarly,

$$\sqrt{n}\left(\widehat{W}_{\text{TC}}-\Delta\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{TC,i}+o_P(1),$$

where ψ_{TC} is defined by

$$\psi_{\text{TC}} = \frac{1}{E(D_{11})-E(D_{10})}\left\{\frac{GT(\varepsilon-E(\varepsilon_{11}))}{p_{11}}-\frac{G(1-T)(\varepsilon+(\delta_1-\delta_0)D-E(\varepsilon_{10}+(\delta_1-\delta_0)D_{10}))}{p_{10}}\right.$$
$$-E(D_{10})D(1-G)\left[\frac{T(Y-E(Y_{101}))}{p_{101}}-\frac{(1-T)(Y-E(Y_{100}))}{p_{100}}\right]$$
$$\left.-(1-E(D_{10}))(1-D)(1-G)\left[\frac{T(Y-E(Y_{001}))}{p_{001}}-\frac{(1-T)(Y-E(Y_{000}))}{p_{000}}\right]\right\}. \tag{A30}$$

Proof of 3. We first show that $(\widehat{F}_{Y_{11}(0)|S},\widehat{F}_{Y_{11}(1)|S})$ tends to a continuous gaussian process. Let $\widetilde{\theta}=(F_{000},F_{001},...,F_{111},\lambda_{10},\lambda_{11})$. By Lemma S4, $\widehat{\widetilde{\theta}}=(\widehat{F}_{000},\widehat{F}_{001},...,\widehat{F}_{111},\widehat{\lambda}_{10},\widehat{\lambda}_{11})$ converges to a continuous gaussian process. Let

$$\pi_d:(F_{000},F_{001},...,F_{111},\lambda_{10},\lambda_{11})\mapsto(F_{d10},F_{d00},F_{d01},F_{d11},1,\lambda_{1d}),\quad d\in\{0,1\},$$

so that $(\widehat{F}_{Y_{11}(0)|S},\widehat{F}_{Y_{11}(1)|S})=\left(R_1\circ\pi_0(\widetilde{\theta}),R_1\circ\pi_1(\widetilde{\theta})\right)$, where R_1 is defined as in Lemma S5. π_d is Hadamard differentiable as a linear continuous map. Because $F_{d10},F_{d00},F_{d01},F_{d11}$ are continuously differentiable with strictly positive derivative by Assumption 12, $\lambda_{1d}>0$, and $\lambda_{1d}\neq 1$ under Assumption 8, R_1 is also Hadamard differentiable at $(F_{d10},F_{d00},F_{d01},F_{d11},1,\lambda_{1d})$ tangentially to $(\mathcal{C}^0)^4\times\text{IR}^2$. By the functional delta method (see, *e.g.*, van der Vaart and Wellner, 1996, Lemma 3.9.4), $(\widehat{F}_{Y_{11}(0)|S},\widehat{F}_{Y_{11}(1)|S})$ tends to a continuous gaussian process.

Now, by integration by parts for Lebesgue-Stieljes integrals,

$$\Delta=\int_{\underline{y}}^{\overline{y}}F_{Y_{11}(0)|S}(y)-F_{Y_{11}(1)|S}(y)dy.$$

Moreover, the map $\varphi_1:(F_1,F_2)\mapsto\int_{\mathcal{S}(Y)}(F_2(y)-F_1(y))dy$, defined on the domain of bounded càdlàg functions, is linear. Because $\mathcal{S}(Y)$ is bounded by Assumption 12, φ_1 is also continuous with respect to the supremum norm. It is thus Hadamard differentiable. Because $\widehat{\Delta}=\varphi_1\left(\widehat{F}_{Y_{11}(1)|S},\widehat{F}_{Y_{11}(0)|S}\right)$, $\widehat{\Delta}$ is asymptotically normal by the functional delta method. The asymptotic normality of $\widehat{\tau}_q$ follows along similar lines. By Assumption 12, $F_{Y_{11}(d)|S}$ is differentiable with strictly positive derivative on its support. Thus, the map $(F_1,F_2)\mapsto F_2^{-1}(q)-F_1^{-1}(q)$ is Hadamard differentiable at $(F_{Y_{11}(0)|S},F_{Y_{11}(1)|S})$ tangentially to the set of functions that are continuous at $(F^{-1}_{Y_{11}(0)|S}(q),F^{-1}_{Y_{11}(1)|S}(q))$ (see Lemma 21.3 in van der Vaart, 2000). By the functional delta method, $\widehat{\tau}_q$ is asymptotically normal.

The validity of the bootstrap follows along the same lines. By Lemma S4, the bootstrap is consistent for $\widehat{\widetilde{\theta}}$. Because both the LATE and LQTE are Hadamard differentiable functions of $\widehat{\widetilde{\theta}}$, as shown above, the result simply follows by the functional delta method for the bootstrap (see, *e.g.*, van der Vaart, 2000, Theorem 23.9).

Finally, we compute the asymptotic variance of both estimators. The functional delta method also implies that both estimators are asymptotically linear. To compute their asymptotic variance, it suffices to provide their asymptotic linear approximation. For that purpose, let us first linearize $F_{Y_{11}(d)|S}(y)$, for all y. It follows from the proof of the first point of Lemma S5 that the mapping $\phi_1:(F_1,F_2,F_3)\mapsto F_1\circ F_2^{-1}\circ F_3$ is Hadamard differentiable at $(F_{d10},F_{d00},F_{d01})$, tangentially to $(\mathcal{C}^0)^3$. Moreover applying the chain rule, we obtain

$$d\phi_1(h_1,h_2,h_3)=h_1\circ Q_d^{-1}+H_d'\circ F_{d01}\times\left[-h_2\circ Q_d^{-1}+h_3\right].$$

Applied to $(F_1,F_2,F_3)=(F_{d10},F_{d00},F_{d01})$, this and the functional delta method once more imply that

$$\sqrt{n}\left(\widehat{H}_d\circ\widehat{F}_{d01}-H_d\circ F_{d01}\right)=d\phi_1(h_{1n},h_{2n},h_{3n})+o_P(1),$$

where the $o_P(1)$ is uniform over y and $h_{1n}=\sqrt{n}(\widehat{F}_{d10}-F_{d10})$. h_{2n} and h_{3n} are defined similarly. Furthermore, applying Lemma S3 yields, uniformly over y,

$$h_{1n}(y)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mathbb{1}\{D_i=d\}G_i(1-T_i)(\mathbb{1}\{Y_i\leq y\}-F_{d10}(y))}{p_{d10}}+o_P(1).$$

A similar expression holds for h_{2n} and h_{3n}. Hence, by continuity of $d\phi_1$, we obtain, after some algebra,

$$\begin{aligned}&\sqrt{n}\left(\widehat{H}_d\circ\widehat{F}_{d01}(y)-H_d\circ F_{d01}(y)\right)\\ &=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{1}\{D_i=d\}\left\{\frac{G_i(1-T_i)(\mathbb{1}\{Q_d(Y_i)\leq y\}-H_d\circ F_{d01}(y))}{p_{d10}}+(1-G_i)H_d'\circ F_{d01}(y)\right.\\ &\quad\left.\times\left[-\frac{(1-T_i)(\mathbb{1}\{Q_d(Y_i)\leq y\}-F_{d01}(y))}{p_{d00}}+\frac{T_i(\mathbb{1}\{Y_i\leq y\}-F_{d01}(y))}{p_{d01}}\right]\right\}+o_P(1),\end{aligned}$$

which holds uniformly over y. Applying repeatedly Lemma S3, we then obtain, after some algebra,

$$\sqrt{n}\left(\widehat{F}_{Y_{11}(d)|S}(y)-F_{Y_{11}(d)|S}(y)\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_{di}(y)+o_P(1),$$

where, omitting the index i,

$$\begin{aligned}\Psi_d(y)=&\ \frac{1}{p_{d|11}-p_{d|10}}\left\{\frac{GT}{p_{11}}\left[\mathbb{1}\{D=d\}\mathbb{1}\{Y\leq y\}-p_{d|11}F_{d11}(y)-F_{Y_{11}(d)|S}(y)\left(\mathbb{1}\{D=d\}-p_{d|11}\right)\right]\right.\\ &+\frac{G(1-T)}{p_{10}}\left[-\mathbb{1}\{D=d\}(\mathbb{1}\{Q_d(Y)\leq y\}-H_d\circ F_{d01}(y))+\left(\mathbb{1}\{D=d\}-p_{d|10}\right)\left(F_{Y_{11}(d)|S}(y)\right.\right.\\ &\left.\left.-H_d\circ F_{d01}(y)\right)\right]+p_{d|10}(1-G)\mathbb{1}\{D=d\}H_d'\circ F_{d01}(y)\left[\frac{(1-T)(\mathbb{1}\{Q_d(Y)\leq y\}-F_{d01}(y))}{p_{d00}}\right.\\ &\left.\left.-\frac{T(\mathbb{1}\{Y\leq y\}-F_{d01}(y))}{p_{d01}}\right]\right\}.\end{aligned}$$

By the functional delta method, this implies that we can also linearize $\widehat{W}_{\text{CIC}}$ and $\widehat{\tau}_q$. Moreover, we obtain by the chain rule the following influence functions:

$$\psi_{\text{CIC}}=\int\Psi_0(y)-\Psi_1(y)dy, \tag{A31}$$

$$\psi_{q,CIC}=\left[\frac{\Psi_1}{f_{Y_{11}(1)|S}}\right]\circ F^{-1}_{Y_{11}(1)|S}(q)-\left[\frac{\Psi_0}{f_{Y_{11}(0)|S}}\right]\circ F^{-1}_{Y_{11}(0)|S}(q). \tag{A32}$$

Acknowledgments. We thank Yannick Guyonvarch for outstanding research assistance, and are very grateful to Esther Duflo for sharing her data with us. We also want to thank the editor, five anonymous referees, Alberto Abadie, Joshua Angrist, Marc Gurgand, Guido Imbens, Rafael Lalive, Thierry Magnac, Blaise Melly, Roland Rathelot, Bernard Salanié, Frank Vella, Fabian Waldinger, Yichong Zhang, and participants at various conferences and seminars for their helpful comments. This article is a merged and revised version of "Fuzzy Changes-in-Changes" by the two authors, and "A note on instrumented Differences-in-Differences" by de Chaisemartin. Xavier D'Haultfoeuille gratefully acknowledges financial support from the research grant Labex Ecodec: Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

Supplementary Data

Supplementary Data are available at *Review of Economic Studies* online.

REFERENCES

ABADIE, A. (2005), "Semiparametric Difference-in-Differences Estimators", *Review of Economic Studies*, **72**, 1–19.

ANGRIST, J. D., GRADDY, K. and IMBENS, G. W. (2000), "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish", *The Review of Economic Studies*, **67**, 499–527.

ANGRIST, J. D. and IMBENS, G. W. (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity", *Journal of the American Statistical Association*, **90**, 431–442.

ATHEY, S. and IMBENS, G. W. (2006), "Identification and Inference in Nonlinear Difference-in-Differences Models", *Econometrica*, **74**, 431–497.

BERTRAND, M., DUFLO, E. and MULLAINATHAN, S. (2004), "How Much Should We Trust Differences-in-Differences Estimates?", *The Quarterly Journal of Economics*, **119**, 249–275.

BLUNDELL, R., DIAS, M. C., MEGHIR, C., *et al*. (2004), "Evaluating the Employment Impact of a Mandatory Job Search Program", *Journal of the European Economic Association*, **2**, 569–606.

BONHOMME, S. and SAUDER, U. (2011), "Recovering Distributions in Difference-in-Differences Models: A Comparison of Selective and Comprehensive Schooling", *Review of Economics and Statistics*, **93**, 479–494.

DE CHAISEMARTIN, C. and D'HAULTFŒUILLE, X. (2016), "Double Fixed Effects Estimators with Heterogeneous Treatment Effects" (Working paper).

—— (2017), "Supplement to "Fuzzy Differences-in-Differences" " (Working paper).

D'HAULTFŒUILLE, X., HODERLEIN, S. and SASAKI, Y. (2015): "Nonlinear Difference-in-Differences in Repeated Cross Sections with Continuous Treatments" (Working paper).

DUFLO, E. (2001), "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment", *American Economic Review*, **91**, 795–813.

ENIKOLOPOV, R., PETROVA, M. and ZHURAVSKAYA, E. (2011), "Media and Political Persuasion: Evidence from Russia", *The American Economic Review*, **101**, 3253.

FIELD, E. (2007), "Entitled to Work: Urban Property Rights and Labor Supply in Peru", *The Quarterly Journal of Economics*, **122**, 1561–1602.

GENTZKOW, M., SHAPIRO, J. M. and SINKINSON, M. (2011), "The Effect of Newspaper Entry and Exit on Electoral Politics", *The American Economic Review*, **101**, 2980.

HOROWITZ, J. L. and MANSKI, C. F. (1995): "Identification and Robustness with Contaminated and Corrupted Data", *Econometrica*, **63**, 281–302.

IMBENS, G. W. and RUBIN, D. B. (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variables Models", *Review of Economic Studies*, **64**, 555–574.

MINCER, J. and JOVANOVIC, B. (1979), "Labor Mobility and Wages" (Working paper).

VAN DER VAART, A. W. and WELLNER, J. A. (1996), *Weak Convergence and Empirical Processes* (New York: Springer).

VAN DER VAART, A. W. (2000), *Asymptotic Statistics* (Cambridge: Cambridge Series in Statistical and Probabilistic Mathematics).

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

WILLIS, R. J. (1986), "Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions", *Handbook of labor economics*, **1**, 525–602.