Discussion Paper Series

RIEB

Kobe University

DP2024-22

Instrumented Difference-in-Differences with Heterogeneous Treatment Effects*

RIEB Junior Research Fellow

Sho MIYAJI

July 9, 2024

* This Discussion Paper won the Kanematsu Prize (FY 2023).

Research Institute for Economics and Business Administration

Kobe University

2-1 Rokkodai, Nada, Kobe 657-8501 JAPAN

Instrumented Difference-in-Differences with heterogeneous treatment effects*

Sho Miyaji†

July 8, 2024

Abstract

Many studies exploit variation in the timing of policy adoption across units as an instrument for treatment, and use instrumental variable techniques. This paper formalizes the underlying identification strategy as an instrumented difference-in-differences (DID-IV). In a simple setting with two periods and two groups, our DID-IV design mainly consists of a monotonicity assumption, and parallel trends assumptions in the treatment and the outcome. In this design, a Wald-DID estimand, which scales the DID estimand of the outcome by the DID estimand of the treatment, captures the local average treatment effect on the treated (LATET). In contrast to Fuzzy DID design considered in de Chaisemartin and D'Haultfœuille (2018), our DID-IV design does not *ex-ante* require strong restrictions on the treatment adoption behavior across units, and our target parameter, the LATET, is policy-relevant if the instrument is based on the policy change of interest to the researcher. We extend the canonical DID-IV design to multiple period settings with the staggered adoption of the instrument across units, which we call a staggered DID-IV design. We propose an estimation method in staggered DID-IV design that is robust to treatment effect heterogeneity. We illustrate our findings in the setting of Oreopoulos (2006), estimating returns to schooling in the United Kingdom. In this application, the two-way fixed effects instrumental variable regression, which is the conventional approach to implement a staggered DID-IV design, yields a negative estimate, whereas our estimation method indicates the substantial gain from schooling.

Keywords: difference-in-differences, instrumental variable, local average treatment effect, returns to education

JEL Classification: C21, C23, C26

*This paper won the Kanematsu Prize from the Research Institution for Economics and Business Administration, Kobe University, in 2023. This paper also won the Best Young Scholar Award at SETA (Symposium on Econometrics and Applications) Conference, Taipei, in 2024. I am grateful to my advisors, Daiji Kawaguchi and Ryo Okui for their continued guidance and support. I am also thankful to Alberto Abadie, Shakeeb Khan, Masamune Iwasawa, Shoya Ishimaru, Shosei Sakaguchi, Takahide Yanagi, and three anonymous referees of the Kanematsu Prize for their insightful comments. The author gratefully acknowledges the support of JSPS KAKENHI Grant Number JP 24KJ0817. All errors are my own.

†Graduate School of Economics, The University of Tokyo, 7-3-1 Hongo, Bunkyoku, Tokyo 113-0033, Japan; Email: shomiyaji-apple@g.ecc.u-tokyo.ac.jp.

1 Introduction

Instrumental variable (IV) strategies are widely used for causal inference across many fields in economics. Despite its popularity, empirical researchers commonly face two challenges in employing the IV method in practice. The first challenge is the internal validity of the IV estimate; in general, it is difficult to find an instrument that is randomized across units in practice, especially in observational studies. The second challenge is the external validity of the IV estimate; under heterogeneous treatment effects, a linear IV estimand captures the average treatment effect, for those who are affected by the instrument, and this local average treatment effect (LATE) may not be policy-relevant if the instrument does not represent the policy change of interest to the researcher (Heckman and Vytlacil (2001), Heckman and Vytlacil (2005)).

To enhance at least the external validity of the IV estimate, many studies have exploited variation in the timing of policy adoption across units as an instrument for treatment, and use instrumental variable techniques. For example, Duflo (2001) estimates returns to schooling in Indonesia, leveraging variation arising from the new school construction program across regions and cohorts as an instrument for education attainment. As another example, Black et al. (2005) estimates the causal link between parents' and children's education attainment, using variation occurring from the different timing of implementation of school reform across municipalities and cohorts as an instrument for parents' education attainment. The important point here is that the underlying identification strategy behind these studies is inconsistent with IV designs: the instrument based on these policy changes is not randomized across units in reality, thereby potentially threatening the internal validity of the IV estimate. Rather, the underlying identification strategy is similar to difference-in-differences (DID) designs, under which we identify the effect of the policy shock (treatment) on the outcome, exploiting the variation of that policy change across units and over time.

In this paper, we formalize the underlying identification strategy as an instrumented difference-in-differences (DID-IV). We define the target parameter and identifying assumptions in this design, and develop a credible estimation method that is robust to treatment effect heterogeneity in multiple time periods settings. Importantly, this design does not require the randomization assumption of the instrument. In this design, we can achieve the internal and external validity of the IV estimate if the parallel trends assumptions in the treatment and the outcome are plausible in a given application, and the instrument is based on the policy shift of interest to the researcher.

First, we consider a simple setting with two periods and two groups: some units are not exposed to the instrument over two periods (unexposed group), whereas some units become exposed at the second period (exposed group). In this setting, our target parameter is the local average treatment effect on the treated (LATET); this parameter measures the treatment effects, for those who belong to an exposed group and are induced to treatment by instrument in the second period. Our DID-IV design mainly consists of a monotonicity assumption, and parallel trends assumptions in the treatment and the outcome between exposed and unexposed groups. This framework comes from a simple and motivating observation: in the two-group/two-period (2×2) settings, a popular estimand, which we call the Wald-DID estimand, scales the DID estimand of the outcome by the DID estimand of the treatment between the two groups. Indeed, we show that the Wald-DID estimand captures the LATET under 2×2 DID-IV designs.

In DID-IV designs, the interpretation of the parallel trends assumption in the outcome is less clear than that of the parallel trends assumption in DID designs because it does not depend on the untreated outcome, but depends on the outcome under no exposure to the

instrument, which we call an unexposed outcome. Motivated by this observation, we next clarify the interpretation of the parallel trends assumption in the outcome in DID-IV designs. Specifically, we first uncover the selection mechanism behind DID-IV designs; we show that in DID-IV settings, time also plays the role of instrument, that is, time also affects one's treatment choices over time in the absence of the policy shock. Following to the terminology in Imbens and Angrist (1994), we divide the population into time groups defined by the potential treatment choices without instrument during the two periods: time always-takers, time never-takers, time compliers, and time defiers. For instance, the time compliers do not adopt the treatment in the first period, but become adopt it in the second period without instrument. We then show that the average time trend in the unexposed outcome equals a weighted average of the indirect effect of time on the outcome through treatment in each time group. We call this an expected time gain, and show that the parallel trends assumption in the outcome in DID-IV designs requires the same expected time gain between exposed and unexposed groups.

de Chaisemartin and D'Haultfœuille (2018) (hereafter, "dCDH" in this section) also consider the same setting in this article, but formalize 2×2 DID-IV designs differently, calling them Fuzzy DID designs. The main difference between this article and dCDH lies in the definition of target parameter: we focus on the LATET, whereas dCDH focus on the switcher local average treatment effect on the treated (SLATET); this parameter measures the treatment effects, for those who belong to an exposed group and start to receive the treatment in the second period. Accordingly, the identifying assumptions in dCDH are also different from those in this article. Furthermore, dCDH argue that the Wald-DID estimand identifies the SLATET under Fuzzy DID designs if the treatment effect is stable over time, and propose the alternative estimands that are free from the strong restriction on treatment effect heterogeneity.

Motivated by the disagreement between this article and dCDH, we next compare DID-IV designs to Fuzzy DID designs, and point out the issues inherent in Fuzzy DID designs. First, we show that dCDH's identifying assumptions *ex-ante* impose the strong restrictions on the treatment adoption behavior across units. It turns out that these restrictions are asymmetric between exposed and unexposed groups, and difficult to defend from the institutional knowledge of the problem in a given application. Next, under these restrictions, we show that dCDH's target parameter, the SLATET, can be decomposed into a weighted average of two different causal parameters: one parameter measures the treatment effects among the sub-population of the compliers and the other parameter measures the treatment effects among the time compliers, who are affected by time but not affected by instrument. This decomposition result for the SLATET has two concerning implications. First, the interpretation of the SLATET is less clear than that of the LATET in DID-IV designs. Second, the SLATET may not be policy-relevant even when the instrument is based on the policy change of interest to the researcher because this parameter is always contaminated by the treatment effects among the time compliers.

We also revisit the "issue" raised by dCDH regarding the use of the Wald-DID estimand, and demonstrate that while their argument is correct, it is misguided. Specifically, we first show that their argument stems from their reliance on the parallel trends assumption in DID designs. We then argue that this assumption is unsuitable for DID-IV settings for three reasons. First, this assumption generally fails to capture the average time trends of the outcome even in an unexposed group. This is because in DID-IV settings, units are allowed to receive the treatment in the absence of the policy shock (instrument). Second, in most applications of the DID-IV method, it is generally infeasible to impute the untreated potential outcomes (e.g., Duflo (2001), Black et al. (2005)). For instance, in Black et al. (2005), this assumption requires that the data contain the units with zero education attainment, which is unrealistic in practice. Finally, in DID-IV settings, we can not directly test this assumption using the pre-exposed

period data in general. Indeed, we show that the placebo test proposed by dCDH is incomplete for assessing the plausibility of this assumption.

Next, we extend the canonical DID-IV design to multiple period settings with the staggered adoption of the instrument across units. In this setting, units start exposed to the instrument at a particular point in time and remain exposed to that instrument afterward. Indeed, in most DID-IV applications, researchers exploit variation in the timing of policy adoption across units in more than two periods, instrumenting for the treatment with the natural variation. The instrument is constructed, for instance from the staggered adoption of school reforms across municipalities or across countries (e.g. Oreopoulos (2006), Lundborg et al. (2014), and Meghir et al. (2018)), the phase-in introduction of head starts across states (e.g. Johnson and Jackson (2019)), or the gradual adoption of broadband internet programs across municipalities (e.g. Akerman et al. (2015), Bhuller et al. (2013)). We refer to the underlying identification strategy as a staggered DID-IV design, and establish the target parameter and identifying assumptions. Specifically, in this design, we first partition units into mutually exclusive and exhaustive cohorts by the initial exposure date of the instrument. We then define our target parameter as the cohort specific local average treatment effect on the treated (CLATT); this parameter is a natural generalization of the LATET in 2×2 DID-IV designs, and measures the treatment effects, for those who belong to cohort e and are the compliers at a given relative period l from the initial exposure to the instrument. Finally, we extend the identification assumptions in 2×2 DID-IV designs to multiple period settings and state the interpretation of each assumption.

We extend our DID-IV designs in several directions. We first show that our DID-IV framework can be applied when we have a non-binary, ordered treatment. We also consider extensions to repeated cross sections, and triple DID-IV designs. Lastly, we consider the case when we introduce the treatment path in potential outcomes in 2×2 DID-IV designs.

Finally, we propose a regression-based method to consistently estimate our target parameter in staggered DID-IV designs under heterogeneous treatment effects. In practice, when researchers implicitly rely on a staggered DID-IV design, they usually implement this design via two-way fixed effects instrumental variable regressions. In companion paper (Miyaji (2024)), however, we show that in more than two periods, the TWFEIV estimand generally fails to summarize the treatment effects under staggered DID-IV designs. Specifically, Miyaji (2024) shows that under staggered DID-IV designs, the TWFEIV estimand is a weighted average of all possible CLATTs, but some weights can be negative if the effect of the instrument on the treatment or the outcome is not stable over time. Our proposed method avoids this issue and is robust to treatment effect heterogeneity. Our estimation procedure consists of two steps. First, we subset the data which contain only two cohorts and two periods. Next, in each data set, we run the TWFEIV regression. We call this a stacked two stage least squares (STS) regression and ensure its validity. Following to Callaway and Sant'Anna (2021), we also propose the weighting scheme to summarize the treatment effects under staggered DID-IV designs. In our weighting scheme, the weight reflects the share of compliers in a given relative period l in cohort e. We also discuss the procedure of pretrends tests to check the validity of the parallel trends assumptions in the treatment and the outcome in DID-IV designs.

We illustrate our findings with the setting of Oreopoulos (2006) who estimate returns to schooling in the United Kingdom, exploiting variation in the timing of implementation of school reforms between Britain and North Ireland as an instrument for education attainment. In this application, we first assess the plausibility of DID-IV identification assumptions implicitly imposed by Oreopoulos (2006). We then estimate the TWFEIV regression in the author's setting. We find that the TWFEIV estimate is strictly negative and not significantly different

from zero. Actually, the TWFEIV regression performs the "bad comparisons" (c.f. Goodman-Bacon (2021)) in the author's setting: we treat the already exposed units as controls during the periods after both regions were affected by the policy change. Finally, we use our estimation method to reassess returns to schooling in the U.K. We find that our STS estimates are all positive in each relative period after the school reform, and our weighting scheme yields a more plausible estimate than the TWFEIV estimate. Specifically, our weighting estimate indicates the roughly 20% gain from schooling in the U.K.

Overall, this paper provides a new econometric framework by combining the IV techniques with DID designs. Our DID-IV design can be applied to various empirical settings. First, it can cope with the general adoption process of the treatment when we are interested in the effect of the treatment on the outcome in DID designs. For instance, consider the case that some units already receive the treatment in period 0 before the policy shock, which distributes the treatment in period 1. In this setting, the canonical DID design is infeasible in practice unless we *ex-ante* discard these units from the data. On the other hand, our DID-IV design can be directly applied to this setting, allowing us to identify the LATET. In other words, we can view that DID-IV is DID with the noncompliance of the assigned treatment.

Second, our DID-IV design can yield valid counterfactuals to estimate the treatment effect when we have no control group under DID designs. For instance, suppose that the treatment is education attainment and we are interested in returns to schooling as in Duflo (2001). In this setting, the canonical DID design generally fails to construct the control group, as the data would not contain the units with zero education attainment in reality. On the other hand, our DID-IV design can partition units into exposed and unexposed groups depending on whether the unit is exposed to the policy shock (instrument). Once these groups are created under this design, one can identify the LATET by comparing the evolution of the treatment and the outcome between the two groups.

Finally, our DID-IV design can overcome the endogeneity issue in empirical work. In practice, due to the endogenous adoption of the treatment over time, researchers often fail to identify the treatment effect via simple linear regressions or two-way fixed effects regressions in panel data settings. In our DID-IV design, one can cope with the common identification challenge by leveraging variation in the timing of policy adoption across units as an instrument for treatment. Importantly, in this design, the identifying variation does not come from the random assignment of the policy shock across units; it mainly comes from the parallel trends assumption in the treatment and the outcome over time. This identification strategy would be less demanding than IV designs in some applications, enhancing the internal and external validity of the IV estimate in practice.

The rest of the paper is organized as follows. The next subsection discusses the related literature. Section 2 establishes DID-IV designs in a simple setting with two periods and two groups. Section 3 formalizes the target parameter and identifying assumptions in staggered DID-IV designs. Section 4 contains extensions. Section 5 presents our estimation method. Section 6 presents our empirical application. Section 7 concludes. All proofs are given in the Appendix.

1.1 Related literature

Our paper is related to a recent DID-IV literature (de Chaisemartin (2010); Hudson et al. (2017); de Chaisemartin and D'Haultfœuille (2018)). We contribute to this literature in three important ways.

The first contribution of this paper is to investigate the detailed connections between DID-IV and Fuzzy DID designs proposed by de Chaisemartin and D'Haultfœuille (2018) (henceforth,

"dCDH" in this section). In econometrics, a pioneering work formalizing 2×2 DID-IV designs is de Chaisemartin (2010), who shows that the Wald-DID estimand identifies the local average treatment effect on the treated (LATET) under parallel trends assumptions in the treatment and the outcome, and a monotonicity assumption. Hudson et al. (2017) also consider 2×2 DID-IV designs with non-binary, ordered treatment settings. Build on the work in de Chaisemartin (2010), however, de Chaisemartin and D'Haultfœuille (2018) formalize 2×2 DID-IV designs differently and call them Fuzzy DID designs; the settings are the same between the two designs, but the target parameter and identifying assumptions are different from each other.

In this paper, we first formalize 2×2 DID-IV designs and complement de Chaisemartin (2010) and Hudson et al. (2017). Specifically, we introduce the path of the instrument in 2×2 DID-IV designs and uncover the "hidden assumption" in the previous literature. This assumption requires the no anticipatory behavior of the treatment adoption before the exposure to the instrument. We also investigate the selection mechanism behind 2×2 DID-IV designs, and clarify the interpretation of the parallel trends assumption in the outcome.

Given our identification results and the terminology developed in this paper, we then compare DID-IV to Fuzzy DID designs, and point out the issues inherent in Fuzzy DID designs. Specifically, we show that the identifying assumptions in Fuzzy DID designs *ex-ante* requires the strong restrictions on the treatment adoption behavior across units, and the target parameter in Fuzzy DID designs can be decomposed into a weighted average of two different causal parameters. The discussion organizes the relationship between the two designs in the previous literature, which essentially treat the same setting, and indicate the relative advantages of DID-IV over Fuzzy DID.

The second contribution of this paper is to extend 2×2 DID-IV designs to multiple period settings with the staggered adoption of the instrument across units. In practice, in order to estimate the treatment effects, empirical researchers often leverage variation in the timing of policy adoption across units as an instrument for treatment in more than two periods (e.g., Black et al. (2005), Lundborg et al. (2014), and Johnson and Jackson (2019)). No previous studies, however, extend 2×2 DID-IV designs to such important settings. Our staggered DID-IV designs ensure the theoretical validity of the underlying identification strategy behind these studies, and allow the practitioners to estimate the local average treatment effect on the treated even when the treatment adoption is endogenous over time.

Finally, this paper provides a credible estimation method in staggered DID-IV designs under heterogeneous treatment effects. When empirical researchers implicitly rely on the staggered DID-IV design in practice, they commonly implement this design via TWFEIV regressions (e.g. Black et al. (2005), Lundborg et al. (2014)). In companion paper (Miyaji (2024)), however, we show that the TWFEIV estimand potentially fails to summarize the treatment effects under staggered DID-IV designs if the effect of the instrument on the treatment or the outcome evolves over time. Our proposed estimation method would serve as an alternative to the TWFEIV estimator and make DID-IV designs more credible in a given application.

Our paper is also related to a recent and growing DID literature. Several recent papers have pointed out the issue of implementing DID designs via two-way fixed effects regressions or its dynamic specifications in the presence of heterogeneous treatment effects (Athey and Imbens (2022); Borusyak et al. (2021); Callaway and Sant'Anna (2021); de Chaisemartin and D'Haultfœuille (2020); Goodman-Bacon (2021); Imai and Kim (2021); Sun and Abraham (2021)). All of these papers have documented that regression coefficients by the conventional approaches may fail to properly summarize treatment effects under DID designs. Some of these papers have proposed an alternative estimation method that delivers a sensible estimand and is robust to treatment effect heterogeneity. None of these papers, however, have explicitly consid-

ered the causal interpretation of the TWFEIV estimand when researchers exploit variation in the timing of policy adoption across units as an instrument for the treatment. In this paper, we formalize the rationale behind using the TWFEIV regression as an instrumented difference-in-differences design, and provide a reliable estimation method in this design in multiple periods settings. Built on the work in this paper, Miyaji (2024) studies the properties of the TWFEIV estimand under staggered DID-IV design and point out the issue of implementing this design via TWFEIV regressions.

2 DID-IV in two time periods

In this section, we formalize an instrumented difference-in-differences (DID-IV) in two-period/two-group settings. We first establish the target parameter and identifying assumptions in this design. We then uncover the selection mechanism behind this design, and clarify the interpretation of the parallel assumption in the outcome. We next describe the relationship between DID and DID-IV, indicating that DID-IV is DID with the noncompliance of the assigned treatment. At the end of this section, we also describe the connections between DID-IV and Fuzzy DID proposed by de Chaisemartin and D'Haultfœuille (2018).

2.1 Set up

We introduce the notation we use throughout section 2. We consider a panel data setting with two periods and N units. For any random variable R, we denote $\mathcal{S}(R)$ to be its support. For each $i \in \{1, \ldots, N\}$ and $t \in \{0, 1\}$, let $Y_{i,t}$ denote the outcome, and $D_{i,t} \in \{0, 1\}$ denote the treatment status: $D_{i,t} = 1$ if unit i receives the treatment in period t and $D_{i,t} = 0$ if unit i does not receives the treatment in period t. Let $Z_{i,t} \in \{0, 1\}$ denote the instrument status: $Z_{i,t} = 1$ if unit i is exposed to the instrument in period t and $Z_{i,t} = 0$ if unit i is not exposed to the instrument in period t. Throughout section 2, we assume that $\{Y_{i,0}, Y_{i,1}, D_{i,0}, D_{i,1}, Z_{i,0}, Z_{i,1}\}_{i=1}^{N}$ are independent and identically distributed (i.i.d).

We introduce the path of the treatment and the instrument. Let $D_i = (D_{i,0}, D_{i,1})$ be the treatment path and $Z_i = (Z_{i,0}, Z_{i,1})$ be the instrument path. We assume that no one is exposed to the instrument in period $t = 0$: $Z_{i,0} = 0$ for all i. Hereafter, we refer to this as a sharp assignment of the instrument. We denote $E_i \in \{0, 1\}$ to be the group variable: $E_i = 1$ if unit i is exposed to the instrument in period $t = 1$ (exposed group) and $E_i = 0$ if unit i is not exposed to the instrument in period $t = 1$ (unexposed group). In contrast to the sharp assignment of the instrument, we allow the general adoption process for the treatment: we assume that the treatment path can take four values with non-zero probability, that is, we have $\{(0, 0), (0, 1), (1, 0), (1, 1)\} \in \mathcal{S}(D)$.

In practice, researchers are interested in the effect of a treatment $D_{i,t}$ on an outcome $Y_{i,t}$, and the instrument $Z_{i,1}$ typically represents a program or a policy shock, which occurs in period $t = 1$ and encourages people to adopt the treatment. For instance, Duflo (2001) estimates returns to schooling in Indonesia, using a new school construction program across regions as an instrument for education attainment.

Next, we introduce the potential outcomes framework. Let $Y_{i,t}(d, z)$ denote the potential outcome in period t when unit i receives the treatment path $d \in \mathcal{S}(D)$ and the instrument path $z \in \mathcal{S}(Z)$. Similarly, let $D_{i,t}(z)$ denote the potential treatment status in period t when unit i receives the instrument path $z \in \mathcal{S}(Z)$. Hereafter, we refer to $D_{i,t}((0, 0))$ as unexposed treatment and $D_{i,t}((0, 1))$ as exposed treatment.

We make a no carryover assumption on the potential outcomes $Y_{i,t}(d, z)$.

Assumption 1 (No carryover assumption)**.**

$$\forall z \in \mathcal{S}(Z), \forall d \in \mathcal{S}(D), Y_{i,0}(d,z) = Y_{i,0}(d_0,z), Y_{i,1}(d,z) = Y_{i,1}(d_1,z),$$

where $d = (d_0, d_1)$ is the generic element of the treatment path D_i.

Assumption 1 states that potential outcomes $Y_{i,t}(d,z)$ depend only on the current treatment status d_t and the instrument path z. In the DID literature, several recent papers impose this assumption with settings of a non-staggered treatment; see, e.g., de Chaisemartin and D'Haultfœuille (2020) and Imai and Kim (2021).[1]

Next, we introduce the group variable $G_i^Z \equiv (D_{i,1}((0,0)), D_{i,1}((0,1)))$, which describes the type of unit i according to the response of $D_{i,1}$ on the instrument path z. Specifically, the first element $D_{i,1}((0,0))$ represents the potential treatment choice in period $t = 1$ if unit i belongs to an unexposed group, and the second element $D_{i,1}((0,1))$ represents the potential treatment choice in period $t = 1$ if unit i belongs to an exposed group. Following to the terminology in Imbens and Angrist (1994), we define $G_i^Z = (0,0) \equiv NT^Z$ to be the never-takers, $G_i^Z = (1,1) \equiv AT^Z$ to be the always-takers, $G_i^Z = (0,1) \equiv CM^Z$ to be the compliers, and $G_i^Z = (1,0) \equiv DF^Z$ to be the defiers.

Henceforth, we keep Assumption 1. In the next section, we use the notations developed so far to define our target parameter in 2×2 DID-IV designs.

2.2 Target parameter in two time periods

In 2×2 DID-IV designs, our target parameter is the local average treatment effect on the treated (LATET) in period $t = 1$ defined below.

Definition. The local average treatment effect on the treated (LATET) in period $t = 1$ is

$$\begin{aligned} LATET &\equiv E[Y_{i,1}(1) - Y_{i,1}(0) | E_i = 1, D_{i,1}((0,1)) > D_{i,1}((0,0))] \\ &= E[Y_{i,1}(1) - Y_{i,1}(0) | E_i = 1, CM^Z]. \end{aligned}$$

This parameter measures the treatment effects in period 1, for those who belong to an exposed group ($E_i = 1$), and are induced to treatment by instrument in period 1. In the DID-IV literature, de Chaisemartin (2010) consider the same target parameter, and Hudson et al. (2017) define the local treatment effect (LATE) in period $t = 1$ as their target parameter, which is unconditional on E_i. The LATET has been also proposed in heterogeneous effects IV models with binary instrument; see, e.g., Słoczyński et al. (2022) and Słoczyński (2020).

We define the LATET as our target parameter for two reasons. First, this parameter would be particularly of interest if the instrument reflects a policy change of interest (Heckman and Vytlacil (2001), Heckman and Vytlacil (2005)). Duflo (2001), for instance, are interested in "whether investments in infrastructure can cause an increase in education attainment, and whether an increase in education attainment causes an increase in earnings". To answer this question, the author exploits a new school construction program across regions as an instrument for education attainment. In this context, the LATET would be a policy-relevant parameter in the presence of heterogeneous treatment effects. Second, the LATET is a natural extension

[1] When treatment is binary and there exist only two periods, one can explicitly introduce the treatment path in potential outcomes because it takes only four values. We therefore consider this case in Appendix B as an extension. One can also weaken Assumption 1 in other cases such as non-binary, ordered treatments or multiple time periods, but this would requires the cumbersome notation and complicates the definition of our target parameter, thus is beyond scope of this paper.

of the target parameter in DID designs, the so-called average treatment effects on the treated (ATT). In section 2.5, we show that both causal parameters measure the treatment effects among the units who are affected by instrument (policy shock) and belong to an exposed group ($E_i = 1$).

Remark 1. de Chaisemartin and D'Haultfœuille (2018) define the switcher local average treatment effect on the treated (SLATET) as their target parameter in the same setting we are considering here[2].

Definition. The switcher local average treatment effect on the treated (SLATET) is

$$\begin{aligned} SLATET &\equiv E[Y_{i,1}(1) - Y_{i,1}(0)|E_i = 1, D_{i,1}((0,1)) > D_{i,0}((0,1))] \\ &= E[Y_{i,1}(1) - Y_{i,1}(0)|E_i = 1, SW], \end{aligned}$$

where the switcher (SW) is the units who become treated in period 1.

This parameter measures the treatment effects among the units who belong to an exposed group ($E_i = 1$) and switch to receive the treatment in period 1 (SW). As we will show in Appendix E, it turns out that this parameter is different from the LATET and can be decomposed into a weighted average of two different causal parameters. In section 2.6, we provide a detailed discussion of the connections between this article and de Chaisemartin and D'Haultfœuille (2018).

2.3 Identification assumptions in two time periods

This section establishes DID-IV designs in a simple setting with two groups and two periods. In this setting, a popular estimand is the ratio between the DID estimand of the outcome and the DID estimand of the treatment (Duflo (2001), Field (2007)):

$$w_{DID} = \frac{E[Y_{i,1} - Y_{i,0}|E_i = 1] - E[Y_{i,1} - Y_{i,0}|E_i = 0]}{E[D_{i,1} - D_{i,0}|E_i = 1] - E[D_{i,1} - D_{i,0}|E_i = 0]}.$$

Following to the terminology in de Chaisemartin and D'Haultfœuille (2018), we call this the Wald-DID estimand. This estimand is obtained from the IV regression of the outcome on the treatment where the group and post-time dummies are included instruments and the interaction of the two is excluded instrument.

In practice, the Wald-DID estimand w_{DID} is commonly interpreted as measuring the local average treatment effect (LATE) in the presence of heterogeneous treatment effects, but the underlying identifying assumptions are not stated formally. Duflo (2001), for instance, notes that "if returns to education are not constant, the 2SLS estimates are a weighted average of the returns to education for people who are affected by the instruments (Angrist and Imbens (1995))".

We consider the following identifying assumptions for the Wald-DID estimand to capture the LATET. Henceforth, we assume that the denominator of the Wald-DID estimand is different from zero without loss of generality.

Assumption 2 (Exclusion restriction for potential outcomes)**.**

$$\forall z \in \mathcal{S}(Z), \forall d \in \mathcal{S}(D), \forall t \in \{0,1\}, Y_{i,t}(d,z) = Y_{i,t}(d).$$

[2] de Chaisemartin and D'Haultfœuille (2018) consider the same situation in repeated cross section settings. We therefore rewrite their target parameter by the notation we use in this article.

This assumption requires that the instrument path does not directly affect potential outcomes other than through treatment. This assumption is common in the IV literature; see e.g., Imbens and Angrist (1994) and Abadie (2003). In the DID-IV literature, de Chaisemartin (2010) and Hudson et al. (2017) impose a similar assumption without introducing the instrument path.

Given Assumption 1 and Assumption 2, we can write the observed outcomes $Y_{i,t}$ as

$$Y_{i,t} = D_{i,t}Y_{i,t}(1) + (1 - D_{i,t})Y_{i,t}(0).$$

Assumption 1 and Assumption 2 also allow us to introduce the notion of exposed and unexposed outcomes. For any $z \in \mathcal{S}(Z)$, let $Y_{i,t}(D_{i,t}(z))$ denote the potential outcome if the instrument path were z:

$$Y_{i,t}(D_{i,t}(z)) \equiv D_{i,t}(z)Y_{i,t}(1) + (1 - D_{i,t}(z))Y_{i,t}(0).$$

Hereafter, we refer to $Y_{i,t}(D_{i,t}((0,0)))$ as unexposed outcomes and $Y_{i,t}(D_{i,t}((0,1)))$ as exposed outcomes. Note that we can not observe different potential outcomes for the same unit at the same time. For instance, when unit i is assigned to $z = (0,0)$, that is, unit i is not exposed to the instrument during the two periods, we can observe only $Y_{i,t}(D_{i,t}((0,0)))$, and can not observe $Y_{i,t}(D_{i,t}((0,1)))$ for that unit.

Next, we make the following monotonicity assumption as in Imbens and Angrist (1994).

Assumption 3 (Monotonicity Assumption in period $t = 1$)**.**

$$Pr(D_{i,1}((0,1)) \geq D_{i,1}((0,0))) = 1 \quad \text{or} \quad Pr(D_{i,1}((0,1)) \leq D_{i,1}((0,0))) = 1.$$

This assumption requires that the instrument path affects the treatment choice at period $t = 1$ in a monotone (uniform) way. This assumption implies that the group variable $G_i^Z = (D_{i,1}((0,0)), D_{i,1}((0,1)))$ can take three values with non-zero probability. In the DID-IV literature, de Chaisemartin (2010) and Hudson et al. (2017) make the same assumption. Hereafter, we consider the type of monotonicity assumption which rules out the existence of the defiers DF^Z.

Assumption 4 (No anticipation in the first stage)**.**

$$D_{i,0}((0,1)) = D_{i,0}((0,0)) \ \ a.s. \ \ \text{for all units } i \text{ with } E_i = 1.$$

Assumption 4 requires that the potential treatment choice before the exposure to instrument is equal to the baseline treatment choice $D_{i,0}((0,0))$ in an exposed group. This assumption restricts the anticipatory behavior and would be plausible if the instrument path is *ex-ante* not known for all the units in an exposed group. This assumption is "hidden assumption" in the previous DID-IV literature: de Chaisemartin (2010) and Hudson et al. (2017) implicitly impose this assumption by writing observed treatment choice in period $t = 0$ as $D_{i,0}(0)$.

In the DID literature, recent studies impose the no anticipation assumption on untreated potential outcomes in several ways. Callaway and Sant'Anna (2021) and Sun and Abraham (2021) assume the average version of the no anticipation assumption, whereas Athey and Imbens (2022) assume the no anticipation for all units i. Roth et al. (2023) take the intermediate approach: they assume the no anticipation for the treated units. Our no anticipation assumption on potential treatment choices is in line with that of Roth et al. (2023).

Next, we impose the parallel trends assumptions in the treatment and the outcome. In the DID-IV literature, de Chaisemartin (2010) and Hudson et al. (2017) assume the similar assumptions.

Assumption 5 (Parallel Trends Assumption in the treatment)**.**

$$E[D_{i,1}((0,0)) - D_{i,0}((0,0))|E_i = 0] = E[D_{i,1}((0,0)) - D_{i,0}((0,0))|E_i = 1].$$

Assumption 5 is a parallel trends assumption in the treatment. This assumption requires that the expectation of the treatment between exposed and unexposed groups would have followed the same path if the assignment of the instrument had not occurred. For instance, in Duflo (2001), this assumption requires that the evolution of mean education attainment would have been the same between exposed and unexposed groups if the policy shock had not occurred during two periods.

Assumption 6 (Parallel Trends Assumption in the outcome)**.**

$$E[Y_{i,1}(D_{i,1}((0,0))) - Y_{i,0}(D_{i,0}((0,0)))|E_i = 0] = E[Y_{i,1}(D_{i,1}((0,0))) - Y_{i,0}(D_{i,0}((0,0)))|E_i = 1].$$

Assumption 6 is a parallel trends assumption in the outcome. This assumption requires that the evolution of the unexposed outcome is the same on average between exposed and unexposed groups. For instance, in Duflo (2001), this assumption requires that the expectation of log annual earnings would have followed the same path from period 0 to period 1 between exposed and unexposed groups in the absence of the policy shock.

When empirical researchers exploit variation arising from policy shock as an instrument for the treatment and use the Wald-DID estimand, they often refer to Assumption 5 and 6. For instance, Duflo (2001) estimates returns to schooling in Indonesia, relying on "the identification assumption that the evolution of wages and education across cohorts would not have varied systematically from one region to another in the absence of the program".

The theorem below shows that if Assumptions 1 - 6 hold, the Wald-DID estimand captures the local average treatment effects on the treated (LATET) in period 1.

Theorem 1. If Assumptions 1-6 hold, the Wald-DID estimand w_{DID} is equal to the LATET in period $t = 1$, that is,

$$w_{DID} = E[Y_{i,1}(1) - Y_{i,1}(0)|E_i = 1, CM^Z].$$

holds.

Proof. See Appendix A. □

We now provide the intuition behind 2×2 DID-IV designs. As we already noted, the Wald-DID estimand scales the DID estimand of the outcome by the DID estimand of the treatment. Motivated by this observation, under 2×2 DID-IV designs, we rely on the common DID identification strategy both in the numerator and the denominator of the Wald-DID estimand (see Figure 1). First, in the denominator of the Wald-DID estimand, we use the DID identification strategy to identify the direct effect of the instrument on the treatment: we recover the expectation of the unexposed treatment in period $t = 1$ in an exposed group ($z = (0,1)$), using Assumption 4 (no anticipation in the first stage) and Assumption 5 (parallel trends assumption in the treatment). Next, in the numerator of the Wald-DID estimand, we use the DID identification strategy in order to identify the indirect effect of the instrument on the outcome through treatment: we recover the expectation of the unexposed outcome in period $t = 1$ in an exposed group ($z = (0,1)$), using Assumption 6 (parallel trends assumption in the outcome). Finally, by scaling the indirect effect obtained from the numerator by the direct effect obtained from the denominator, we can capture the LATET in period $t = 1$.

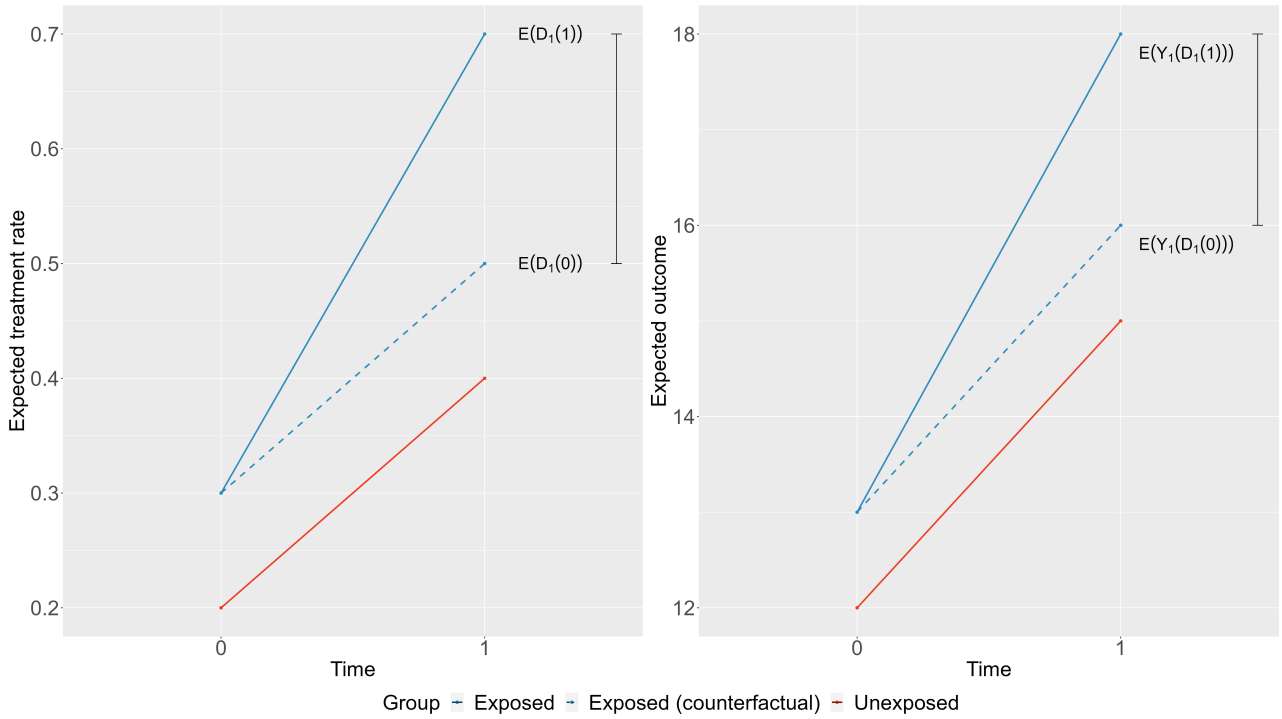

Fig. 1. DID-IV identification strategy in two periods and two groups. *Notes*: This figure plots the evolution of the expectation of the treatment (First stage DID) and the outcome (Reduced form DID) for exposed and unexposed groups respectively. The dotted line represents the counterfactual trends of the expectation of the treatment and the outcome in an exposed group. The effects of the instrument on the treatment and the outcome are 0.2 and 2 respectively, and the local average treatment effect on the treated is 10.

2.4 Interpreting the parallel trends assumption in the outcome

So far, we have formalized the target parameter and identifying assumptions in 2×2 DID-IV designs. Overall, in this design, we rely on the common DID identification strategy in both the denominator and the numerator of the Wald-DID estimand in order to identify the effects of the instrument on the treatment and the outcome. The interpretation of the parallel trends assumption in the outcome, however, is less clear than that of the parallel trends assumption in the canonical DID designs because it does not depend on the untreated outcome, but depends on the unexposed outcome.

In this section, we clarify the interpretation of the parallel trends assumption in the outcome in 2×2 DID-IV designs. To do so, we first uncover the selection mechanism behind the canonical DID-IV designs. Specifically, we show that in 2×2 DID-IV designs, time also plays the role of instrument, and investigate how the identifying assumptions in 2×2 DID-IV designs restrict the treatment adoption behavior across units in exposed and unexposed groups, respectively. We then show that the average time trend in the outcome is equal to a weighted average of the indirect effect of time on the outcome through treatment. We call this an expected time gain and show that the parallel trends assumption in the outcome requires the homogeneous expected time gain between exposed and unexposed groups.

First, we introduce the additional notation. Let $G_i^T = (D_{i,0}((0,0)), D_{i,1}((0,0)))$ denote the group variable for unit i, which represents the treatment adoption process from period 0 to period 1 if the instrument path were $z = (0,0)$: the first element $D_{i,0}((0,0))$ represents the potential treatment choice without instrument in period 0 and the second element $D_{i,1}((0,0))$ represents the potential treatment choice without instrument in period 1.

Table 1. Exposed group (assigned to $z = (0,1)$)

observed	counterfactual	
$(D_0((0,0)), D_1((0,1)))$	$D_1((0,0)) = 1$	$D_1((0,0)) = 0$
$D_0((0,0)) = 1, D_1((0,1)) = 1$	$AT^Z \wedge AT^T$	$CM^Z \wedge DF^T$
$D_0((0,0)) = 1, D_1((0,1)) = 0$	$DF^Z \wedge AT^T$	$NT^Z \wedge DF^T$
$D_0((0,0)) = 0, D_1((0,1)) = 1$	$AT^Z \wedge CM^T$	$CM^Z \wedge NT^T$
$D_0((0,0)) = 0, D_1((0,1)) = 0$	$DF^Z \wedge CM^T$	$NT^Z \wedge NT^T$

Table 2. Unexposed group (assigned to $z = (0,0)$)

observed	counterfactual	
$(D_0((0,0)), D_1((0,0)))$	$D_1((0,1)) = 1$	$D_1((0,1)) = 0$
$D_0((0,0)) = 1, D_1((0,0)) = 1$	$AT^Z \wedge AT^T$	$DF^Z \wedge AT^T$
$D_0((0,0)) = 1, D_1((0,0)) = 0$	$CM^Z \wedge DF^T$	$NT^Z \wedge DF^T$
$D_0((0,0)) = 0, D_1((0,0)) = 1$	$AT^Z \wedge CM^T$	$DF^Z \wedge CM^T$
$D_0((0,0)) = 0, D_1((0,0)) = 0$	$CM^Z \wedge NT^T$	$NT^Z \wedge NT^T$

We recall that in 2×2 DID-IV designs, the units are allowed to take the treatment without instrument during the two periods. In this setting, one can see that time also plays the role of instrument, that is, time also affects one's treatment choices during the two periods.

Similar to the group variable G_i^Z introduced in section 2.1, we define $G_i^T = (0,0) \equiv NT^T$ to be the time never-takers, $G_i^T = (1,1) \equiv AT^T$ to be the time always-takers, $G_i^T = (0,1) \equiv CM^T$ to be the time compliers, and $G_i^T = (1,0) \equiv DF^T$ to be the time defiers. The time never-takers NT^T and the time always-takers AT^T do not change their treatment status without instrument during the two periods. In other words, their treatment choices are time-invariant in the absence of the policy shock (instrument). On the other hand, the time compliers CM^T are induced to treatment in period 1 without instrument, and the time defiers DF^T leave the treatment in period 1 without instrument, that is, their treatment choices are affected by time in the absence of the policy shock.

By exploiting the group variables G_i^Z and G_i^T, we can describe how the identifying assumptions in 2×2 DID-IV designs restrict the treatment adoption behavior across units in exposed and unexposed groups, respectively. Table 1 and Table 2 show that we can partition units in exposed and unexposed groups into six mutually exclusive and exhaustive types under 2×2 DID-IV designs. Here, Assumption 3 (monotonicity assumption) excludes the defiers DF^Z in exposed and unexposed groups respectively, the types painted in gray color in both tables.

In an exposed group, we can observe $(D_0((0,0)), D_1((0,1)))$, but can not observe $D_1((0,0))$ for each unit (see Table 1). Here, we use Assumption 4 (no anticipation assumption in the first stage), replacing $D_0((0,1))$ with $D_0((0,0))$. In an unexposed group, we can observe $(D_0((0,0)), D_1((0,0)))$, but can not observe $D_1((0,1))$ for each unit (see Table 2). If, for instance, the observed treatment choices of unit i in an exposed group are $D_{i,0}((0,0)) = 0$ and $D_{i,1}((0,1)) = 1$, and the counterfactual treatment choice is $D_{i,1}((0,0)) = 0$, that unit belongs to the type $CM^Z \wedge NT^T$: unit i is induced to the treatment by instrument but not by time. If the observed treatment choices of unit i in an unexposed group are $D_{i,0}((0,0)) = 0$ and $D_{i,1}((0,0)) = 1$, and the counterfactual treatment choice is $D_{i,1}((0,1)) = 1$, that unit belongs to the type $AT^Z \wedge CM^T$: unit i is induced to the treatment by time but not by instrument.

Given these tables, we now consider the interpretation of the parallel trends assumption in the outcome. Let Δ_0 and Δ_1 denote the average time trends of the unexposed outcome in exposed and unexposed groups, respectively:

$$\Delta_0 \equiv E[Y_{i,1}(D_{i,1}((0,0))) - Y_{i,0}(D_{i,0}((0,0)))|E_i = 0],$$
$$\Delta_1 \equiv E[Y_{i,1}(D_{i,1}((0,0))) - Y_{i,0}(D_{i,0}((0,0)))|E_i = 1].$$

We note that the parallel trends assumption in the outcome requires $\Delta_0 = \Delta_1$. At the same time, we also note that Δ_0 is estimable by the observed change in outcomes in an unexposed group during the two periods, whereas Δ_1 is fundamentally unobservable because exposed group is assigned to the instrument path $z = (0, 1)$.

To make the interpretations of Δ_0 and Δ_1 clear, we introduce the additional notation. For each time group $g^t \in \{AT^T, NT^T, CM^T, DF^T\}$, let $w_{g^t,e} \equiv Pr[G_i^T = g^t|E_i = e]$ denote the population share in group $E_i = e \in \{0, 1\}$. Let $\Delta_{g^t,e}$ denote the expectation of the potential outcome path in the absence of the policy shock for time group g^t in group $E_i = e$:

$$\Delta_{AT^T,e} \equiv E[Y_{i,1}(1) - Y_{i,0}(1)|AT^T, E_i = e],$$
$$\Delta_{NT^T,e} \equiv E[Y_{i,1}(0) - Y_{i,0}(0)|NT^T, E_i = e],$$
$$\Delta_{CM^T,e} \equiv E[Y_{i,1}(1) - Y_{i,0}(0)|CM^T, E_i = e],$$
$$\Delta_{DF^T,e} \equiv E[Y_{i,1}(0) - Y_{i,0}(1)|DF^T, E_i = e].$$

For instance, $w_{CM^T,0} = Pr[CM^T|E_i = 0]$ and $\Delta_{CM^T,0} = E[Y_{i,1}(1) - Y_{i,0}(0)|CM^T, E_i = 0]$ represent the population share and the expectation of the outcome path without instrument for the time compliers in an unexposed group, respectively.

Each $w_{g^t,e}$ and $\Delta_{g^t,e}$ are estimable in an unexposed group ($e = 0$): we can identify the time group of unit i in an unexposed group and can observe the evolution of the outcome without instrument. For instance, when the observed treatment choices are $D_{i,0}((0,0)) = 0$ and $D_{i,1}((0,0)) = 1$ for unit i in an unexposed group, we can figure out that unit i is the time compliers CM^T, and can observe the potential outcome path $(Y_{i,0}(0), Y_{i,1}(1))$ without instrument for that unit. On the other hand, each $w_{g^t,e}$ and $\Delta_{g^t,e}$ are unobservables in an exposed group ($e = 1$) because we can not specify the time group of unit i in that group. For instance, when the observed treatment choices are $D_{i,0}((0,0)) = 0$ and $D_{i,1}((0,1)) = 1$ for unit i in an exposed group, we can not distinguish whether unit i belongs to the time compliers CM^T or the time never takers NT^T (see Table 1).

We can interpret each $\Delta_{g^t,e}$ as the indirect effect of time on the outcome through treatment for time group g^t in group $E_i = e$. First, $\Delta_{AT^T,e}$ and $\Delta_{NT^T,e}$ only reflect the expectation of the time effect for treated and untreated potential outcomes respectively:

$$\Delta_{AT^T,e} = \underbrace{E[Y_{i,1}(1) - Y_{i,0}(1)|AT^T, E_i = e]}_{\text{Time effect among } AT^T},$$
$$\Delta_{NT^T,e} = \underbrace{E[Y_{i,1}(0) - Y_{i,1}(0)|NT^T, E_i = e]}_{\text{Time effect among } NT^T}.$$

This is because the time always-takers AT^T and the time never-takers NT^T do not change their treatment status during the two periods. If the exclusion restriction of time on the outcome holds, that is, $Y_{i,1}(1) = Y_{i,0}(1)$ and $Y_{i,1}(0) = Y_{i,0}(0)$ hold for all units i (in other words, the stationary conditions hold), these quantities are equal to zero. Next, $\Delta_{CM^T,e}$ and $\Delta_{DF^T,e}$ can

be decomposed into two terms:

$$\Delta_{CM^T,e} = E[Y_{i,1}(0) - Y_{i,0}(0)|CM^T, E_i = e] + \underbrace{E[Y_{i,1}(1) - Y_{i,1}(0)|CM^T, E_i = e]}_{\text{Selection gain among } CM^T},$$

$$\Delta_{DF^T,e} = E[Y_{i,1}(1) - Y_{i,0}(1)|DF^T, E_i = e] + \underbrace{E[Y_{i,1}(0) - Y_{i,1}(1)|DF^T, E_i = e]}_{\text{Selection gain among } DF^T}.$$

The first terms in $\Delta_{CM^T,0}$ and $\Delta_{DF^T,0}$ represent the expectation of the time effect for untreated and treated potential outcomes in each time group in group $E_i = e$, respectively. The second term represents the selection gain in each time group in group $E_i = e$, respectively. The latter terms arise because the time compliers CM^T and the time defiers DF^T are affected by time and change their treatment status from period 0 to period 1.

The following theorem establishes the relationship between Δ_e and $\Delta_{g^t,e}$.

Theorem 2. Suppose Assumptions 1-4 hold, then

$$\Delta_e = \sum_{g^t} w_{g^t,e} \Delta_{g^t,e},$$

where the weight $w_{g^t,e}$ is non-negative and sum to one, that is, $\sum_{g^t} w_{g^t,e} = 1$ holds.

Proof. See Appendix A. □

Theorem 2 shows that the average time trend of the unexposed outcome in group e is equal to a weighted average of the indirect effect of time on the outcome through treatment for each time group g^t in group e. The weight assigned to each $\Delta_{g^t,e}$ is natural in that it reflects the population share in each time group g^t in group e. In this paper, we call Δ_e an expected time gain in group e hereafter.

In light of the view discussed so far, we can conclude that the parallel trends assumption in the outcome (Assumption 6), i.e., $\Delta_0 = \Delta_1$ requires the same expected time gain between exposed and unexposed groups, that is, the whole indirect effects of time on the outcome through treatment are the same on average between the two groups.

Assumption 6 is required for DID-IV identification strategy because in the numerator of the Wald-DID estimand, we should isolate the indirect effect of the instrument on the outcome through treatment from that of time in an exposed group. The underlying idea to capture the former effect is parallel to that in DID designs: we purge the latter effect by exploiting the observed time trend of the outcome in the group that did not receive the policy shock (instrument) in the second period. In practice, if there exist pre-exposed period data, we can indirectly check the validity of Assumption 6 as well as Assumption 5, comparing the evolution of the outcome between exposed and unexposed groups. The formal discussion for the pretrend test is provided in section 5.3.

2.5 Relationship between DID and DID-IV

In this section, we describe the relationship between DID and DID-IV designs, and provide the view that DID-IV is DID with the noncompliance of the assigned treatment.

According to section 2.1, we first introduce the notation in DID designs. Let $Z_i = (Z_{i,0}, Z_{i,1})$ denote the path of a program, which distributes the treatment $D_{i,t}$, and let $D_i = (D_{i,0}, D_{i,1})$ denote the path of the treatment for unit i. In the canonical DID set up, $Z_i = (Z_{i,0}, Z_{i,1})$ takes only two values: $Z_i = (0, 1)$ if unit i is assigned to the program in period 1 (treatment

Table 3. Exposed group (assigned to $z = (0,1)$)

observed	counterfactual
$(D_0((0,0)), D_1((0,1)))$	$D_1((0,0)) = 0$
$D_0((0,0)) = 0, D_1((0,1)) = 1$	$CM^Z \wedge NT^T$

Table 4. Treatment group (assigned to $z = (0,0)$)

observed	counterfactual
$(D_0((0,0)), D_1((0,0)))$	$D_1((0,1)) = 1$
$D_0((0,0)) = 0, D_1((0,0)) = 0$	$CM^Z \wedge NT^T$

group) and $Z_i = (0,0)$ if unit i is not assigned to the program in period 1 (control group). Let $E_i \in \{0,1\}$ denote the group variable: $E_i = 1$ if unit i belongs to the treatment group and $E_i = 0$ if unit i belongs to the control group. Hereafter, in this section, we refer to treatment and control groups as exposed and unexposed groups, respectively.

Note that in the common DID set up, the treatment status $D_{i,t}$ is equal to the exposure to the program $Z_{i,t}$ because we *ex ante* impose two restrictions on the treatment adoption behavior across units. First, we assume that all the units are not allowed to receive the treatment in period 0:

$$D_{i,0}((0,1)) = D_{i,0}((0,0)) = 0 \quad \text{for all } i. \tag{1}$$

Second, we assume that unit i should receive the treatment in period 1 if that unit belongs to an exposed group, and should not receive the treatment in period 1 if that unit belongs to an unexposed group:

$$1 = D_{i,1}((0,1)) > D_{i,1}((0,0)) = 0 \quad \text{for all } i. \tag{2}$$

These restrictions imply that we have Tables 3-4 under DID designs: there exists only the type $CM^Z \wedge NT^T$ in exposed and unexposed groups, respectively.

Then, under these restrictions, we can show that DID-IV designs nest DID designs as a special case along with three dimensions. First, Assumptions 3-5 in DID-IV designs are automatically satisfied in DID designs. Assumption 3 follows from (2) and Assumption 4 follows from (1). Assumption 6 follows from combining (1) with (2).

Second, the ATT, the target parameter in DID designs, is equal to the LATET in the canonical DID set up. This follows from Tables 3 - 4:

$$\begin{aligned} ATT &= E[Y_{i,1}(1) - Y_{i,1}(0) | E_i = 1] \\ &= E[Y_{i,1}(1) - Y_{i,1}(0) | E_i = 1, CM^Z] \\ &= LATET. \end{aligned}$$

We note that both ATT and LATET measure the treatment effects, for those who are in an exposed group (assigned to $z = (0,1)$) and are induced to treatment by instrument (CM^Z). We therefore define the LATET as our target parameter in DID-IV designs in section 2.2.

Finally, a well-known parallel trends assumption in DID designs can also be interpreted as requiring the same expected time gain between exposed and unexposed groups. From Tables

3- 4, we can rewrite the parallel trends assumption in the untreated outcome as follows:

$$
\begin{aligned}
& E[Y_{i,1}(0) - Y_{i,0}(0)|E_i = 0] = E[Y_{i,1}(0) - Y_{i,0}(0)|E_i = 1] \\
\iff & E[Y_{i,1}(0) - Y_{i,0}(0)|E_i = 0, NT^T] = E[Y_{i,1}(0) - Y_{i,0}(0)|E_i = 1, NT^T] \\
\iff & \Delta_{NT^T,0} = \Delta_{NT^T,1} \\
\iff & \Delta_0 = \Delta_1,
\end{aligned}
$$

where the last equivalence holds because we have only the time never takers NT^T in exposed and unexposed groups respectively under DID designs:

$$w_{NT^T,0} = Pr(NT^T|E_i = e) = 1, \quad \text{for all} \quad e = 0, 1.$$

Overall, this section clarifies that DID-IV is a natural generalization of DID, and the differences between the two designs arise because DID-IV allows the noncompliance of the assigned treatment: $D_{i,t} \neq Z_{i,t}$. In the next section, we compare DID-IV and Fuzzy DID considered in de Chaisemartin and D'Haultfœuille (2018).

2.6 Comparing DID-IV to Fuzzy DID

de Chaisemartin and D'Haultfœuille (2018) (Henceforth, "dCDH" in this section) also investigate the identification assumptions for the Wald-DID estimand to capture some causal effects under the same setting in this article. However, they formalize 2×2 DID-IV designs differently, calling them Fuzzy DID designs. In this section, we compare DID-IV to Fuzzy DID and point out the issues embedded in Fuzzy DID. The detailed discussion is given in Appendix E, and this section provides the overview of that discussion.

The main difference between dCDH and this article is in the definition of target parameter: we focus on the LATET, whereas dCDH focus on the SLATET (see section 2.2 in this article). Accordingly, the identifying assumptions in Fuzzy DID designs are different from those in 2×2 DID-IV designs. Furthermore, dCDH argue that the Wald-DID estimand identifies the SLATET only if the treatment effect is stable over time, and propose the alternative estimands that do not require the strong restriction on treatment effect heterogeneity.

In Appendix E, we first show that the identifying assumptions in Fuzzy DID designs *ex-ante* impose the strong restrictions on the treatment adoption behavior across units and these restrictions are asymmetric between exposed and unexposed groups.

Specifically, we examine how dCDH's identifying assumptions exclude the types in the two groups and construct Tables 5-6[3], which are in line with Tables 1-2 in this article. From these tables, one can see that dCDH exclude the defiers DF^Z and the type $NT^Z \wedge DF^T$ in an exposed group, and the defiers DF^Z, the time defiers CM^T and the time compliers DF^T in an unexposed group. Note that in our DID-IV design, we exclude only the defiers DF^Z in exposed and unexposed groups by a monotonicity assumption. The difference between dCDH's restrictions and ours is critical for empirical application; in contrast to the monotonicity assumption, which excludes the defiers in both groups, dCDH's additional restrictions can be difficult to defend from the institutional knowledge of the problem in a given application. This is because these restrictions *ex-ante* exclude the units who change the treatment adoption in the absence of the policy shock (instrument).

Next, under Tables 5-6, we show that dCDH's target parameter, the SLATET, can be decomposed into a weighted average of two different causal parameters. The one parameter

[3]Because dCDH consider repeated cross section settings, we can observe either $D_0(0)$ or $D_1(1)$ for each unit in Tables 5-6.

Table 5. Exposed group ($z = 1$)

observed	counterfactual	
$D_0(0)$ or $D_1(1)$	$D_1(0) = 1$	$D_1(0) = 0$
$D_0(0) = 1, D_1(1) = 1$	$AT^Z \wedge AT^T$	$CM^Z \wedge DF^T$
$D_0(0) = 1, D_1(1) = 0$	$DF^Z \wedge AT^T$	$NT^Z \wedge DF^T$
$D_0(0) = 0, D_1(1) = 1$	$AT^Z \wedge CM^T$	$CM^Z \wedge NT^T$
$D_0(0) = 0, D_1(1) = 0$	$DF^Z \wedge CM^T$	$NT^Z \wedge NT^T$

Table 6. Unexposed group ($z = 0$)

observed	counterfactual	
$D_0(0)$ or $D_1(0)$	$D_1(1) = 1$	$D_1(1) = 0$
$D_0(0) = 1, D_1(0) = 1$	$AT^Z \wedge AT^T$	$DF^Z \wedge AT^T$
$D_0(0) = 1, D_1(0) = 0$	$CM^Z \wedge DF^T$	$NT^Z \wedge DF^T$
$D_0(0) = 0, D_1(0) = 1$	$AT^Z \wedge CM^T$	$DF^Z \wedge CM^T$
$D_0(0) = 0, D_1(0) = 0$	$CM^Z \wedge NT^T$	$NT^Z \wedge NT^T$

measures the treatment effects among the type $CM^Z \wedge AT^T$, who are the sub-population of the compliers. The other causal parameter measures the treatment effects among the time compliers CM^T. The time compliers are the units who switch from non-treatment to treatment in period 1 in the absence of the policy shock (instrument), that is, they are affected by time but not affected by the instrument. This decomposition result for the SLATET has two concerning implications. First, the interpretation of the SLATET is less clear than that of the LATET. Second, the SLATET may not be policy-relevant even when the instrument is based on the policy change of interest to the researcher because this parameter is always contaminated by the treatment effects among the time compliers.

Finally, we revisit the "issue" raised by dCDH regarding the use of the Wald-DID estimand, and show that while their argument is correct, it is misguided. During this discussion, we also clarify why dCDH *ex-ante* exclude so many types in exposed and unexposed groups for the Wald-DID estimand to identify the SLATET.

First, we show that dCDH's argument regarding the use of the Wald-DID estimand is based on their reliance on the parallel trends assumption in DID designs. Recall that in our DID-IV design, we do not assume the parallel trends assumption in the untreated outcome, but assume the parallel trends assumption in the unexposed outcome. Because some units are allowed to adopt the treatment without instrument in DID-IV settings, the former type of parallel trends assumption is generally not sufficient to capture the average time trend of the outcome in an unexposed group. We show that this fact leads dCDH to impose the stable treatment effect assumption for the Wald-DID estimand to identify the SLATET. In the proof of our argument, we also demonstrate that imposing the parallel trends assumption in DID designs yields the strong restrictions on the treatment adoption behavior across units under Fuzzy DID designs.

Next, after confirming that dCDH's argument is based on the parallel trends assumption in DID designs, we argue that this assumption is unsuitable for DID-IV settings for three reasons. First, as already noted, this assumption is not sufficient to identify the average time trends of the outcome even in an unexposed group. Second, in most applications of the DID-IV method, we can not impute the untreated potential outcomes in general (e.g., Duflo (2001), Black et al.

(2005)). For instance, in Black et al. (2005), this assumption requires the data to contain the parents with zero education attainment, which is unrealistic in practice. Finally, in most of the DID-IV settings, we can not test this assumption using pre-exposed period data, as some units can potentially adopt the treatment before period 0. Indeed, we show that dCDH's placebo test is incomplete for assessing the validity of this assumption. Surprisingly, we also show that their placebo test coincides with our pretrends test with only a slight modification, which we will introduce in section 5.3 for assessing the plausibility of Assumption 5 and Assumption 6.

3 DID-IV in multiple time periods

We now extend the 2×2 DID-IV design to multiple periods settings with the staggered adoption of the instrument across units. In reality, many studies exploit variation in the timing of policy adoption across units as an instrument for treatment in more than two periods (e.g. Black et al. (2005), Bhuller et al. (2013), Lundborg et al. (2014), and Meghir et al. (2018)). We refer to the underlying identification strategy as a staggered DID-IV design, and establish the target parameter and identifying assumptions.

3.1 Set up

We introduce the notation we use throughout section 3 to section 5. We consider a panel data setting with T periods and N units. For each $i \in \{1,\dots N\}$ and $t \in \{1,\dots,T\}$, let $Y_{i,t}$ denote the outcome and $D_{i,t} \in \{0,1\}$ denote the treatment status, and $Z_{i,t} \in \{0,1\}$ denote the instrument status. Let $D_i = (D_{i,1},\dots,D_{i,T})$ and $Z_i = (Z_{i,1},\dots,Z_{i,T})$ denote the path of the treatment and the instrument for unit i. Throughout section 3 to section 5, we assume that $\{Y_{i,t}, D_{i,t}, Z_{i,t}\}_{t=1}^{T}$ are independent and identically distributed (i.i.d).

We make the following assumption about the assignment process of the instrument.

Assumption 7 (Staggered adoption for $Z_{i,t}$)**.** For $s<t$, $Z_{i,s} \leq Z_{i,t}$ where $s,t \in \{1,\dots T\}$.

Assumption 7 requires that once units start exposed to the instrument, units remain exposed to that instrument afterward. In the DID literature, several papers make the similar assumption for the adoption process of the treatment and call it the "staggered treatment adoption"; see, e.g., Athey and Imbens (2022), Callaway and Sant'Anna (2021) and Sun and Abraham (2021).

Given Assumption 7, we can uniquely characterize one's instrument path by the initial exposure date of the instrument, which we denote $E_i = \min\{t : Z_{i,t} = 1\}$. If unit i is not exposed to the instrument for all time periods, we define $E_i = \infty$. Based on the initial exposure period E_i, we can uniquely partition units into mutually exclusive and exhaustive cohorts e for $e \in \{1,2,\dots,T,\infty\}$: all the units in cohort e start exposed to the instrument in period $E_i = e$.

Similar to the two periods setting in section 2.1, we allow the general adoption process for the treatment: the treatment can potentially turn on/off repeatedly over time. de Chaisemartin and D'Haultfœuille (2020) and Imai and Kim (2021) consider the same setting in the recent DID literature.

Next, we introduce the potential outcomes framework in multiple time periods. Let $Y_{i,t}(d,z)$ denote the potential outcome in period t when unit i receives the treatment path $d \in \mathcal{S}(D)$ and the instrument path $z \in \mathcal{S}(Z)$. Similarly, let $D_{i,t}(z)$ denote the potential treatment status in period t when unit i receives the instrument path $z \in \mathcal{S}(Z)$.

Assumption 7 allows us to rewrite $D_{i,t}(z)$ by the initial exposure date $E_i = e$. Let $D_{i,t}^{e}$ denote the potential treatment status in period t if unit i is first exposed to the instrument in period e. Let $D_{i,t}^{\infty}$ denote the potential treatment status in period t if unit i is never exposed

to the instrument. Hereafter, we call $D_{i,t}^{\infty}$ the "never exposed treatment". Since the adoption date of the instrument uniquely pins down one's instrument path, we can write the observed treatment status $D_{i,t}$ for unit i in period t as

$$D_{i,t} = D_{i,t}^{\infty} + \sum_{1 \leq e \leq T} (D_{i,t}^{e} - D_{i,t}^{\infty}) \cdot \mathbf{1}\{E_i = e\}.$$

We define $D_{i,t} - D_{i,t}^{\infty}$ to be the effect of an instrument on treatment for unit i in period t, which is the difference between the observed treatment status $D_{i,t}$ to the never exposed treatment status $D_{i,t}^{\infty}$. Hereafter, we refer to $D_{i,t} - D_{i,t}^{\infty}$ as the individual exposed effect in the first stage. In the DID literature, Callaway and Sant'Anna (2021) and Sun and Abraham (2021) define the effect of a treatment on an outcome in the same fashion.

Next, we introduce the group variable which describes the type of unit i in period t, based on the reaction of potential treatment choices in period t to the instrument path z. Let $G_{i,e,t} \equiv (D_{i,t}^{\infty}, D_{i,t}^{e})$ $(t \geq e)$ be the group variable in period t for unit i and the initial exposure date e. Specifically, the first element $D_{i,t}^{\infty}$ represents the treatment status in period t if unit i is never exposed to the instrument $E_i = \infty$ and the second element $D_{i,t}^{e}$ represents the treatment status in period t if unit i is first exposed to the instrument at $E_i = e$. Following to the terminology in section 2.1, we define $G_{i,e,t} = (0,0) \equiv NT_{e,t}$ to be the never-takers, $G_{i,e,t} = (1,1) \equiv AT_{e,t}$ to be the always-takers, $G_{i,e,t} = (0,1) \equiv CM_{e,t}$ to be the compliers and $G_{i,e,t} = (1,0) \equiv DF_{e,t}$ to be the defiers in period t and the initial exposure date e.

Finally, we make a no carryover assumption on potential outcomes $Y_{i,t}(d,z)$.

Assumption 8 (No carryover assumption)**.**

$$\forall z \in \mathcal{S}(Z), \forall d \in \mathcal{S}(D), \forall t \in \{1, \ldots, T\}, Y_{i,t}(d,z) = Y_{i,t}(d_t, z).$$

Henceforth, we keep Assumption 7 and 8. In the next section, we define the target parameter in staggered DID-IV designs.

3.2 Target parameter in staggered DID-IV designs

In staggered DID-IV designs, our target parameter is the cohort specific local average treatment effect on the treated (CLATT) defined below.

Definition. The cohort specific local average treatment effect on the treated (CLATT) at a given relative period l from the initial adoption of the instrument is

$$\begin{aligned} CLATT_{e,e+l} &= E[Y_{i,e+l}(1) - Y_{i,e+l}(0) | E_i = e, D_{i,e+l}^{e} > D_{i,e+l}^{\infty}] \\ &= E[Y_{i,e+l}(1) - Y_{i,e+l}(0) | E_i = e, CM_{e,e+l}]. \end{aligned}$$

Each CLATT is a natural generalization of the LATET in section 2.2 and suitable for the setting of the staggered instrument adoption. This parameter measures the treatment effects at a given relative period l from the initial exposure date $E_i = e$, for those who belong to cohort e, and are the compliers $CM_{e,e+l}$, that is, who are induced to treatment by instrument in period $e + l$. Each CLATT can potentially vary across cohorts and over time because it depends on cohort e, relative period l and the compliers $CM_{e,e+l}$. We note that the composition of the compliers can overlap over the post-exposure periods in cohort e. For instance, some units in cohort e may belong to $CM_{e,e+l}$ for all the relative periods $(l \geq e)$: they are affected by the instrument for all post-exposure periods.

This parameter would be policy-relevant if researchers leverage the policy change of their interest as an instrument for the treatment. Suppose for instance that the treatment represents participation in a job training program, the outcome represents real earnings, and the instrument represents a policy shock that encourages people to receive the treatment. In this example, if the instrument is an important policy lever for the researchers (e.g., nudge), each CLATT is of particular interest because the compliers in each relative period after a policy change represents the dynamic effect of that policy on the take up rates of the program, and this parameter can be interpreted as estimating the effect of participation in job training on real earnings among those affected by the policy shock in each relative period.

3.3 Identification assumptions in staggered DID-IV designs

In this section, we establish the identification assumptions in staggered DID-IV designs. These assumptions are the natural generalization of Assumptions 1-6 in 2×2 DID-IV designs.

Assumption 9 (Exclusion Restriction in multiple time periods)**.**

$$\forall z \in \mathcal{S}(Z), \forall d \in \mathcal{S}(D), \forall t \in \{1, \ldots, T\}, Y_{i,t}(d, z) = Y_{i,t}(d) \quad a.s.$$

Assumption 9 extends the exclusion restriction in two time periods (Assumption 2) to multiple period settings. This assumption requires that the instrument path does not directly affect the potential outcome for all time periods and its effects are only through treatment.

Given Assumption 8 and Assumption 9, we can write the potential outcome $Y_{i,t}(d, z)$ as $Y_{i,t}(d_t) = D_{i,t}Y_{i,t}(1) + (1 - D_{i,t})Y_{i,t}(0)$. Following to section 2.3, we introduce the potential outcomes in period t if unit i is assigned to the instrument path $z \in \mathcal{S}(Z)$:

$$Y_{i,t}(D_{i,t}(z)) \equiv D_{i,t}(z)Y_{i,t}(1) + (1 - D_{i,t}(z))Y_{i,t}(0).$$

Since the initial exposure date E_i completely characterizes the instrument path, we can write the potential outcomes for cohort e and cohort ∞ as $Y_{i,t}(D^e_{i,t})$ and $Y_{i,t}(D^\infty_{i,t})$ respectively. The potential outcome $Y_{i,t}(D^e_{i,t})$ represents the outcome status in period t if unit i is first exposed to the instrument in period e and the potential outcome $Y_{i,t}(D^\infty_{i,t})$ represents the outcome status in period t if unit i is never exposed to the instrument. Hereafter, we refer to $Y_{i,t}(D^\infty_{i,t})$ as the "never exposed outcome".

Assumption 10 (Monotonicity Assumption in multiple time periods)**.**

$$Pr(D^e_{i,e+l} \geq D^\infty_{i,e+l}) = 1 \text{ or } Pr(D^e_{i,e+l} \leq D^\infty_{i,e+l}) = 1 \text{ for all } e \in \mathcal{S}(E_i) \text{ and for all } l \geq 0.$$

This assumption requires that the instrument path affects the treatment adoption behavior in a monotone way for all relative periods after the initial exposure date $E_i = e$. We recall that we define $D_{i,t} - D^\infty_{i,t}$ to be the effect of an instrument on treatment for unit i in period t. Assumption 10 requires that the individual exposed effect in the first stage should be non-negative (or non-positive) during the periods after the initial exposure to the instrument for all i. This assumption implies that the group variable $G_{i,e,t} \equiv (D^\infty_{i,t}, D^e_{i,t})$ can take three values with non-zero probability for all e and all $t \geq e$. Hereafter, we consider the type of the monotonicity assumption that rules out the existence of the defiers $DF_{e,t}$ for all $t \geq e$ in any cohort e.

Assumption 11 (No anticipation in the first stage)**.**

$$D^e_{i,e+l} = D^\infty_{i,e+l} \ a.s. \ \text{ for all } i, \ \text{ for all } e \in \mathcal{S}(E_i) \text{ and for all } l < 0.$$

Assumption 11 requires that potential treatment choices in any l period before the initial exposure to the instrument is equal to the never exposed treatment. This assumption is a natural generalization of Assumption 4 to multiple period settings and restricts the anticipatory behavior before the initial exposure to the instrument.

Assumption 12 (Parallel Trends Assumption in the treatment in multiple time periods)**.**

$$\text{For all } s < t, \;\; E[D_{i,t}^{\infty} - D_{i,s}^{\infty} | E_i = e] \;\; \text{is same for all } e \in \mathcal{S}(E_i).$$

Assumption 12 is a parallel trends assumption in the treatment in multiple periods and multiple cohorts. This assumption requires that the trends of the treatment across cohorts would have followed the same path, on average, if there is no exposure to the instrument. Assumption 12 is analogous to that of Callaway and Sant'Anna (2021) and Sun and Abraham (2021) in DID designs: these studies impose the same type of parallel trends assumption on untreated outcomes with settings of multiple periods and multiple cohorts.

Assumption 13 (Parallel Trends Assumption in the outcome in multiple time periods)**.**

$$\text{For all } s < t, \;\; E[Y_{i,t}(D_{i,t}^{\infty}) - Y_{i,s}(D_{i,s}^{\infty}) | E_i = e] \;\; \text{is same for all } e \in \mathcal{S}(E_i).$$

Assumption 13 is a parallel trends assumption in the outcome with settings of multiple periods and multiple cohorts. This assumption requires that the expectation of the never exposed outcome across cohorts would have followed the same evolution if the assignment of the instrument had not occurred. From the discussion in section 2.4, we can interpret that this assumption requires the same expected time gain across cohorts and over time: the effects of time on the outcome through treatment are the same on average across cohorts and over time.

4 Extensions

This section contains extensions to non-binary, ordered treatments, triple DID-IV designs, and repeated cross sections. It also includes the case when we introduce the treatment path in potential outcomes in 2×2 DID-IV designs. For more details and the proofs in this section, see Appendix B.

Non-binary, ordered treatment

Up to now, we have considered only the case of a binary treatment. However the same idea can be applied when treatment takes a finite number of ordered values, that is, we have $D_{i,t} \in \{0, 1, \ldots, J\}$. When we have only two periods and treatment is non-binary, our target parameter is the average causal response on the treated (ACRT) defined below.

Definition. The average causal response on the treated (ACRT) is

$$ACRT \equiv \sum_{j=1}^{J} w_j \cdot E[Y_1(j) - Y_1(j-1) | D_{i,1}((0,1)) \geq j > D_{i,1}((0,0)), E_i = 1].$$

where the weight w_j is:

$$w_j = \frac{Pr(D_1((0,1)) \geq j > D_1((0,0)) | E_i = 1)}{\sum_{j=1}^{J} Pr(D_1((0,1)) \geq j > D_1((0,0)) | E_i = 1)}.$$

The average causal response on the treated (ACRT) is a weighted average of the effects of a unit increase in the treatment on the outcome, for those who belong to an exposed group and are induced to increase the treatment in period $t = 1$ by instrument. The ACRT is similar to the average causal response (ACR) considered in Angrist and Imbens (1995), but the difference here is that each weight w_j and the associated causal parameter in this parameter are conditional on $E_i = 1$.

Theorem 3 below shows that if we have a non-binary, ordered treatment, the Wald-DID estimand is equal to the average causal response on the treated (ACRT) under the same assumptions in Theorem 1.

Theorem 3. If Assumptions 1-6 hold, the Wald-DID estimand W_{DID} is equal to the ACRT.

$$\begin{aligned} W_{DID} &= \sum_{j=1}^{J} w_j \cdot E[Y_1(j) - Y_1(j-1) | D_{i,1}((0,1)) \geq j > D_{i,1}((0,0)), E_i = 1] \\ &= ACRT. \end{aligned}$$

Proof. See Appendix B.1. □

If we have a non-binary, ordered treatment in staggered DID-IV designs, our target parameter is the cohort specific average causal response on the treated (CACRT), which is the natural generalization of the ACRT. The identifying assumptions are the same in section 3.3.

Definition. The cohort specific average causal response on the treated (CACRT) at a given relative period l from the initial adoption of the instrument is

$$CACRT_{e,e+l} \equiv \sum_{j=1}^{J} w^{e}_{e+l,j} \cdot E[Y_{i,e+l}(j) - Y_{i,e+l}(j-1) | E_i = e, D^{e}_{i,e+l} \geq j > D^{\infty}_{i,e+l}]$$

where the weight $w^{e}_{e+l,j}$ is:

$$w^{e}_{e+l,j} = \frac{Pr(D^{e}_{i,e+l} \geq j > D^{\infty}_{i,e+l} | E_i = e)}{\sum_{j=1}^{J} Pr(D^{e}_{i,e+l} \geq j > D^{\infty}_{i,e+l} | E_i = e)}.$$

Triple DID-IV designs

As we have seen in section 2 and 3, our DID-IV designs mainly rely on the parallel trends assumption in the treatment and the outcome. In practice, if the data contain three different dimensions (e.g., cohort e, time t, and demographic group g), one can instead use the triple DID identification strategy to estimate the effects of the instrument on the treatment and the outcome, which we call a triple DID-IV design. In this design, one can replace Assumption 5 and 6 (Assumption 12 and 13 in multiple time periods) with common acceleration assumptions in the treatment and the outcome (Olden and Møen (2022), Fröhlich et al. (2019), and Wooldridge (2020)).[4]

Repeated cross sections

In some applications, researchers have no access to panel data and use repeated cross section data[5]. In Appendix B.2, we consider repeated cross section data settings and present the

[4] As an empirical example, Deschênes et al. (2017) implicitly use triple DID-IV designs to estimate the effects of NO_x (Nitrogen Oxides) emissions on medication purchases and mortality rates, exploiting the NO_x Budget Trading program as an instrument for NO_x emissions.

[5] It also includes the case that researchers use the cross section data, and exploit a policy shock across cohorts as an instrument for the treatment as in Duflo (2001).

identification assumptions in DID-IV designs.

Introducing the treatment path in 2×2 DID-IV designs

In section 2.1, we impose Assumption 1 (No carryover assumption) on potential outcomes $Y_{i,t}(d, z)$ as we consider the non-staggered adoption of the treatment. We can extend our results to the case where we introduce the treatment path in potential outcomes with binary treatment. In this case, our target parameter slightly changes with the cumbersome notation, formally defined below.

Definition. The local average treatment effect on the treated with the treatment path (LATET') is

$$LATET' \equiv w_0 \cdot \Delta_0 + w_1 \cdot \Delta_1$$

where Δ_0 and Δ_1 are:

$$\begin{aligned}\Delta_0 &\equiv E[Y_1((1,1)) - Y_1((1,0))|E_i = 1, (CM^Z \wedge DF^T)],\\ \Delta_1 &\equiv E[Y_1((0,1)) - Y_1((0,0))|E_i = 1, (CM^Z \wedge NT^T)].\end{aligned}$$

The weights w_0 and w_1 are:

$$w_0 = \frac{Pr((CM^Z \wedge DF^T)|E_i = 1)}{Pr((CM^Z \wedge DF^T)|E_i = 1) + Pr((CM^Z \wedge NT^T)|E_i = 1)}, \quad w_1 = 1 - w_0.$$

In Appendix B.3, we show that the similar assumptions as in Assumptions 1-6 in section 2.3 are sufficient for the Wald-DID to capture the LATET'.

5 Estimation and Inference

When researchers leverage variation in the timing of policy adoption across units as an instrument for treatment and implicitly rely on a staggered DID-IV design, they usually implement this design via TWFEIV regressions (e.g., Johnson and Jackson (2019), Lundborg et al. (2014), Black et al. (2005), Akerman et al. (2015), and Bhuller et al. (2013)).

In companion paper (Miyaji (2024)), however, we show that in more than two periods, the TWFEIV estimand potentially fails to summarize the treatment effects under staggered DID-IV designs in the presence of heterogeneous treatment effects. Specifically, Miyaji (2024) shows that under staggered DID-IV designs, the TWFEIV estimand is a weighted average of all possible CLATTs, but some weights can be negative if the effect of the instrument on the treatment or the outcome is not stable over time.[6]

In this section, we propose an alternative estimation method in staggered DID-IV designs that is robust to treatment effect heterogeneity. First, we propose a simple regression-based method for estimating each CLATT. Following to Callaway and Sant'Anna (2021), we then propose a weighting scheme to construct the summary causal parameters from each CLATT. At the end of this section, we also discuss the pretrends tests for checking the plausibility of parallel trends assumption in the treatment and the outcome in DID-IV designs.

[6] See also the discussions in de Chaisemartin and D'Haultfœuille (2020), who decompose the numerator and denominator in the TWFEIV estimand separately, and point out the issue of interpreting this estimand causally.

5.1 Stacked two stage least squares regression

In this section, we propose a regression-based method to consistently estimate our target parameter in staggered DID-IV designs. Recall that if we have a binary treatment, our target parameter is the $CLATT_{e,e+l}$ for each cohort e and a relative period $l \geq 0$:

$$CLATT_{e,e+l} = E[Y_{i,e+l}(1) - Y_{i,e+l}(0)|E_i = e, CM_{e,e+l}].$$

Our estimation method consists of two steps.

Step 1.

We create the data sets for each $CLATT_{e,e+l}$. Each data set includes the units of time $t = e-1$ and $t = e+l$, who are either in cohort $e \in \{2, \ldots, T\}$ or in the set of some unexposed cohorts, U ($e \notin U$).

Researchers should carefully specify the choice set U in practice. If there exists a never exposed cohort $E_i = \infty$, we can set $U = \{\infty\}$ and estimate $CLATT_{e,e+l}$ for all cohorts ($e \in \{2, \ldots, T\}$) and all relative time periods ($l \geq 0$). If there exists no never exposed cohort, we can set $U = \max\{E_i\}$, i.e., the last exposed cohort. In this case, we can estimate $CLATT_{e,e+l}$ for cohort $e \in \{2, \ldots, \max\{E_i\} - 1\}$ and relative time period $l \in \{0, \ldots, \max\{E_i\} - 1 - e\}$ because every unit will be exposed to the instrument after time $t = \max\{E_i\}$. We note that the units in an already exposed cohort $e = 1$ are not included in any data sets because we should include the pre-exposed period $e - 1$ for estimating each $CLATT_{e,e+l}$.

Step 2.

In each data set, we run the following IV regression and obtain the IV estimator $\hat{\beta}^{e,l}_{IV}$:

$$Y_{i,t} = \beta^{e,l}_0 + \beta^{e,l}_{i,\cdot}\mathbf{1}\{E_i = e\} + \beta^{e,l}_{\cdot,t}\mathbf{1}\{T_i = e + l\} + \beta^{e,l}_{IV}D_{i,t} + \epsilon^{e,l}_{i,t}.$$

The first stage regression is:

$$D_{i,t} = \pi^{e,l}_0 + \pi^{e,l}_{i,\cdot}\mathbf{1}\{E_i = e\} + \pi^{e,l}_{\cdot,t}\mathbf{1}\{T_i = e + l\} + \pi^{e,l}(\mathbf{1}\{E_i = e\} \cdot \mathbf{1}\{T_i = e + l\}) + \eta^{e,l}_{i,t},$$

where the group indicator $\mathbf{1}\{E_i = e\}$ and the post-period indicator $\mathbf{1}\{T_i = e + l\}$ are the included instruments and the interaction of the two is the excluded instrument.

Formally, our proposed estimator $\widehat{CLATT}_{e,e+l} \equiv \hat{\beta}^{e,l}_{IV}$ takes the following form:

$$\widehat{CLATT}_{e,e+l} \equiv \hat{\beta}^{e,l}_{IV} = \frac{\hat{\alpha}^{e,l}}{\hat{\pi}^{e,l}},$$

where $\hat{\alpha}^{e,l}$ and $\hat{\pi}^{e,l}$ are:

$$\begin{aligned}
\hat{\alpha}^{e,l} &= \frac{E_N[(Y_{i,e+l} - Y_{i,e-1}) \cdot \mathbf{1}\{E_i = e\}]}{E_N[\mathbf{1}\{E_i = e\}]} - \frac{E_N[(Y_{i,e+l} - Y_{i,e-1}) \cdot \mathbf{1}\{E_i \in U\}]}{E_N[\mathbf{1}\{E_i \in U\}]} \\
&\equiv \hat{\alpha}^1_{e,l} - \hat{\alpha}^2_{e,l}. \\
\hat{\pi}^{e,l} &= \frac{E_N[(D_{i,e+l} - D_{i,e-1}) \cdot \mathbf{1}\{E_i = e\}]}{E_N[\mathbf{1}\{E_i = e\}]} - \frac{E_N[(D_{i,e+l} - D_{i,e-1}) \cdot \mathbf{1}\{E_i \in U\}]}{E_N[\mathbf{1}\{E_i \in U\}]} \\
&\equiv \hat{\pi}^1_{e,l} - \hat{\pi}^2_{e,l}.
\end{aligned}$$

Here, $E_N[\cdot]$ is the sample analog of the conditional expectation. We note that the IV estimator $\widehat{CLATT}_{e,e+l}$ is the Wald-DID estimator for each $CLATT_{e,e+l}$, where the choice set U is an

unexposed group and time $t = e - 1$ is a pre-exposed period. From the estimation procedure above, we call this a stacked two-stage least squares (STS) estimator.

Our STS estimator is related to the DID estimators in staggered DID designs recently proposed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021). Specifically, the DID estimator of the treatment and the outcome in our STS estimator corresponds to that of Sun and Abraham (2021), and coincides with that of Callaway and Sant'Anna (2021) for the case when we have no covariates and use never treated units as a control group. Their proposed methods avoid the issue of two-way fixed effects estimators in staggered DID designs, and our STS estimator avoids the issue of two-way fixed effects instrumental variable estimators in staggered DID-IV designs.

Theorem 4 below guarantees the validity of our STS estimator.

Theorem 4. Suppose Assumptions 7-13 hold. Then, the STS estimator $\widehat{CLATT}_{e,e+l}$ is consistent and asymptotically normal:

$$\sqrt{n}(\widehat{CLATT}_{e,e+l} - CLATT_{e,e+l}) \xrightarrow{d} \mathcal{N}(0, V(\psi_{i,e,l})).$$

where $\psi_{i,e,l}$ is influence function for $\widehat{CLATT}_{e,e+l}$ and defined in Equation (25) in Appendix C.

Proof. See Appendix C. □

From Theorem 4, we can also construct the standard error of the STS estimator, using the sample analogue of the asymptotic variance $V(\psi_{i,e,l})$.

Remark 2. Our estimation procedure is identical if we have a non-binary, ordered treatment or use repeated cross section data. In Appendix C, we present the influence function of our SLS estimator in repeated cross section settings.

In triple DID-IV designs, one can replace the IV regression in step 2 with the following one ($g_i = A, B$):

$$Y_{i,t} = \beta_0^{e,l} + \beta_1^{e,l}\mathbf{1}_{E_i=e} + \beta_2^{e,l}\mathbf{1}_{T_i=e+l} + \beta_3^{e,l}\mathbf{1}_{g_i=A} + \beta_4^{e,l}\mathbf{1}_{E_i=e,T_i=e+l} + \beta_5^{e,l}\mathbf{1}_{E_i=e,g_i=A} + \beta_6^{e,l}\mathbf{1}_{T_i=e+l,g_i=A} + \beta_{IV}^{e,l}D_{i,t} + \epsilon_{i,t}^{e,l}.$$

The first stage regression is:

$$D_{i,t} = \pi_0^{e,l} + \pi_1^{e,l}\mathbf{1}_{E_i=e} + \pi_2^{e,l}\mathbf{1}_{T_i=e+l} + \pi_3^{e,l}\mathbf{1}_{g_i=A} + \pi_4^{e,l}\mathbf{1}_{E_i=e,T_i=e+l} + \pi_5^{e,l}\mathbf{1}_{E_i=e,g_i=A} + \pi_6^{e,l}\mathbf{1}_{T_i=e+l,g_i=A} + \pi_7^{e,l}\mathbf{1}_{E_i=e,T_i=e+l,g_i=A} + \eta_{i,t}^{e,l},$$

where $\mathbf{1}_A$ is the indicator function and takes one if A is true. In Appendix C, we also present the influence function of our triple DID-IV estimator $\beta_{IV}^{e,l}$.

5.2 Weighting scheme

The previous section provides the method to estimate each $CLATT_{e,e+l}$ in staggered DID-IV designs. In some applications, each $CLATT_{e,e+l}$ is of intrinsic interest if researchers are interested in the treatment effect heterogeneity across cohort e and relative period l. However, in many cases, researchers may also want to estimate summary causal parameters, which capture the overall effect of a treatment on an outcome. Actually, in staggered DID-IV designs, each $CLATT_{e,e+l}$ depends not only on cohort e, relative period l, but also on the compliers $CM_{e,e+l}$. In this case, each $CLATT_{e,e+l}$ may be difficult to interpret when the number of cohorts and the

relative period are fairly large, and when the members of compliers $CM_{e,e+l}$ in cohort e vary and overlap over time.

In this section, we explain how one can construct the summary causal parameters from each CLATT in staggered DID-IV designs, based on the weighting scheme proposed by Callaway and Sant'Anna (2021).

Specifically, we consider the following weighting scheme as in Callaway and Sant'Anna (2021):

$$\theta^{IV} = \sum_{e} \sum_{t=1}^{T} w(e,t) \cdot CLATT_{e,t},$$

where $w(e,t)$ are some reasonable weighting functions assigned to each $CLATT_{e,t}$.

To propose the weighting functions for a variety of summary causal parameters in staggered DID-IV designs, we first define the average effect of the instrument on the treatment at a given relative period l from the initial exposure to the instrument in cohort e. We call this the cohort specific average exposed effect on the treated in the first stage ($CAET^1_{e,e+l}$), formally defined below.

Definition. The cohort specific average exposed effect on the treated in the first stage ($CAET^1_{e,e+l}$) at a given relative period l from the initial adoption of the instrument is

$$CAET^1_{e,e+l} = E[D_{i,e+l} - D^{\infty}_{i,e+l} | E_i = e].$$

If treatment is binary, each $CETT^1_{e,l}$ is equal to the share of the compliers in cohort e in period $e+l$:

$$\begin{aligned} CAET^1_{e,e+l} &= E[D^{e}_{i,e+l} - D^{\infty}_{i,e+l} | E_i = e] \\ &= Pr(CM_{e,e+l} | E_i = e). \end{aligned}$$

In staggered DID designs, Callaway and Sant'Anna (2021) propose various aggregated measures along with different dimensions of treatment effect heterogeneity. We can employ their framework straightforwardly, but in staggered DID-IV designs, we should more carefully specify the weighting functions assigned to each summary measure.

For instance, to aggregate dynamic treatment effects in cohort e over time in staggered DID designs, Callaway and Sant'Anna (2021) consider the following summary measure:

$$\theta_{sel}(\tilde{e}) = \frac{1}{T - \tilde{e} + 1} \sum_{t=\tilde{e}}^{T} CATT(\tilde{e}, t),$$

where the $CATT(\tilde{e}, t)$ is the cohort specific average treatment effect on the treated at time t in cohort $\tilde{e}$ (see, e.g., Callaway and Sant'Anna (2021), Sun and Abraham (2021)). Callaway and Sant'Anna (2021) construct the summary measure $\theta_{sel}(\tilde{e})$ by equally weighting each $CATT(\tilde{e}, t)$ across all post-treatment periods because all the units in cohort e adopt the treatment after they start exposed to that treatment.

In staggered DID-IV designs, we propose the following summary measure $\theta_{sel}(\tilde{e})^{IV}$ that corresponds with $\theta_{sel}(\tilde{e})$:

$$\theta_{sel}(\tilde{e})^{IV} = \sum_{t=\tilde{e}}^{T} \frac{CAET^1_{\tilde{e},t}}{\sum_{t=\tilde{e}}^{T} CAET^1_{\tilde{e},t}} CLATT_{\tilde{e},t},$$

Table 7. Weights for a variety of summary causal parameters

Target Parameter	$w(e,t)$
$\theta^{IV}_{es(l)}$	$\mathbf{1}\{e+l \leq T\}\mathbf{1}\{t=e+l\}P(E=e\|E+l\leq T)\frac{CAET^1_{e,e+l}}{\sum_{e\in\mathcal{S}(E)} CAET^1_{e,e+l}}$
$\theta^{bal,IV}_{es(l,l')}$	$\mathbf{1}\{e+l' \leq T\}\mathbf{1}\{t=e+l\}P(E=e\|E+l'\leq T)\frac{CAET^1_{e,e+l}}{\sum_{e\in\mathcal{S}(E)} CAET^1_{e,e+l}}$
$\theta^{IV}_{sel(\tilde{e})}$	$\mathbf{1}\{t\geq e\}\mathbf{1}\{e=\tilde{e}\}\frac{CAET^1_{\tilde{e},t}}{\sum_{t=\tilde{e}}^{T} CAET^1_{\tilde{e},t}}$
$\theta^{IV}_{c(\tilde{t})}$	$\mathbf{1}\{t\geq e\}\mathbf{1}\{t=\tilde{t}\}P(E=e\|E\leq t)\frac{CAET^1_{e,t}}{\sum_{e\in\mathcal{S}(E)} CAET^1_{e,t}}$
$\theta^{cumm,IV}_{c(\tilde{t})}$	$\mathbf{1}\{t\geq e\}\mathbf{1}\{t\leq \tilde{t}\}P(E=e\|E\leq t)\frac{CAET^1_{e,t}}{\sum_{e\in\mathcal{S}(E)} CAET^1_{e,t}}$
$\theta^{o,IV}_{W}$	$\mathbf{1}\{t\geq e\}P(E=e\|E\leq T)/\sum_{e\in\mathcal{S}(E)}\sum_{t=1}^{T}\mathbf{1}\{t\geq e\}P(E=e\|E\leq T)$
$\theta^{o,IV}_{sel}$	$\mathbf{1}\{t\geq e\}P(E=e\|E\leq T)\frac{CAET^1_{e,t}}{\sum_{t=e}^{T} CAET^1_{e,t}}$

Notes: This table represents the specific expressions for the weights on each $CLATT(e,t)$ or each $CACRT(e,t)$ in a variety of summary causal parameters. Each target parameter corresponds with that of Callaway and Sant'Anna (2021). We superscript IV to associate each parameter with staggered DID-IV designs.

where the weights are non-negative and sum to one by construction. This parameter summarizes each $CLATT_{\tilde{e},t}$ in cohort $\tilde{e}$ across all post-exposed periods and the weight assigned to each $CLATT_{\tilde{e},t}$ reflects the relative share of the compliers $CM_{\tilde{e},t}$ in period t during the periods after the initial exposure to the instrument in cohort $\tilde{e}$. This weighting scheme would be reasonable in that it is designed to be larger at the period when the proportion of the compliers is relatively higher in cohort $\tilde{e}$.

By similar arguments, we can specify the weighting functions for various summary measures in staggered DID-IV designs, which correspond with those in Callaway and Sant'Anna (2021). In Table 7, we summarize the specific expressions for the weights assigned to each $CLATT_{\tilde{e},t}$ in each summary causal parameter. Here, we define each causal parameter by the notation analogous to that of Callaway and Sant'Anna (2021) and superscript IV to associate each parameter with staggered DID-IV designs. Note that our proposed weighting functions are the same under non-binary, ordered treatment settings, in which our target parameter is the $CACRT_{\tilde{e},t}$.

Estimation and Inference

We can construct the consistent estimator for the summary causal parameter θ^{IV} as follows:

$$\hat{\theta}^{IV} = \sum_{e}\sum_{t=1}^{T} \hat{w}(e,t)\cdot \widehat{CLATT}_{e,t},$$

where $\hat{w}(e,t)$ is the sample analog of each $w(e,t)$ and $\widehat{CLATT}_{e,t}$ is the consistent estimator for each $CLATT_{e,t}$, obtained from the stacked IV regression in section 5.1. Each $\hat{w}(e,t)$ is the regular asymptotically linear estimator:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta^{w}_{i,e,t} + o_p(1),$$

where $\zeta^w_{i,e,t}$ represents the influence function, and satisfies $E[\zeta^w_{i,e,t}] = 0$ and $E[\zeta^w_{i,e,t}{\zeta^w_{i,e,t}}^\top] < \infty$.

Corollary 1 below represents the asymptotic distribution of the plug-in estimator $\hat{\theta}_{IV}$ and ensures its validity.

Corollary 1. If the assumptions of Theorem 4 hold,

$$\sqrt{n}(\hat{\theta}_{IV} - \theta_{IV}) \xrightarrow{d} \mathcal{N}(0, V(l_i^{\theta_{IV}})),$$

where the influence function $l_i^{\theta_{IV}}$ takes the following form:

$$l_i^{\theta_{IV}} = \sum_e \sum_{t=1}^{T} \left(w(e,t) \cdot \psi_{i,e,t} + \zeta^w_{i,e,t} \cdot CLATT_{e,t} \right).$$

5.3 Pretrend test

In this section, we describe how one can conduct pretrend tests to check the validity of the parallel trends assumption in the outcome and the treatment.

Suppose that data are available for period $t = -1$. Then, one can assess the plausibility of Assumption 5 between period $t = -1$ and $t = 0$ by testing the following null hypothesis

$$\begin{aligned} & E[D_{i,0} - D_{i,-1}|E_i = 0] = E[D_{i,0} - D_{i,-1}|E_i = 1] \\ \iff & E[D_{i,0}((0,0)) - D_{i,-1}((0,0))|E_i = 0] = E[D_{i,0}((0,0)) - D_{i,-1}((0,0))|E_i = 1]. \end{aligned} \tag{3}$$

Similarly, one can assess the plausibility of Assumption 6 between period $t = -1$ and $t = 0$ by testing the following null hypothesis

$$\begin{aligned} & E[Y_{i,0} - Y_{i,-1}|E_i = 0] = E[Y_{i,0} - Y_{i,-1}|E_i = 1] \\ \iff & E[Y_{i,0}(D_{i,0}((0,0))) - Y_{i,-1}(D_{i,-1}((0,0)))|E_i = 0] = E[Y_{i,0}(D_{i,0}((0,0))) - Y_{i,-1}(D_{i,-1}((0,0)))|E_i = 1]. \end{aligned} \tag{4}$$

We can generalize these tests to multiple pre-exposed periods settings by using the pretrends tests recently developed in the context of DID designs (Callaway and Sant'Anna (2021), Sun and Abraham (2021), Borusyak et al. (2021)); we can apply these tools to the first stage and the reduced form respectively to confirm the plausibility of Assumption 12 and 13.

Remark 3. When researchers implement the staggered DID-IV design via TWFEIV regressions, they often use the dynamic specifications in the first stage and the reduced form:

$$Y_{i,t} = \mu_{i.} + \delta_{t.} + \sum_l \alpha_l \mathbf{1}\{t - E_i = l\} + \epsilon_{i,t}, \tag{5}$$

$$D_{i,t} = \gamma_{i.} + \zeta_{t.} + \sum_l \beta_l \mathbf{1}\{t - E_i = l\} + \eta_{i,t}, \tag{6}$$

where E_i represents the initial exposure date. They plot the coefficients α_l in equation (5) and β_l in equation (6) in each relative period l, and test whether the coefficients before the initial exposure to the instrument are significantly different from zero or not (e.g., Akerman et al. (2015) and Bhuller et al. (2013)).

However, such tests are problematic if the effects of the instrument on the treatment and outcome are heterogeneous. Sun and Abraham (2021) shows that in staggered DID designs, the coefficient for pre-treatment periods on event study regressions can be different from zero

even when the no anticipation assumption and the parallel trends assumption in the untreated outcome are plausible. Because we can view that $Y_{i,t}$ and $D_{i,t}$ are the outcomes and $Z_{i,t}$ is the binary absorbing treatment in equations (5)-(6), it is inappropriate to check whether the coefficients of pre-exposed periods are statistically different from zero or not as a way of testing Assumption 12 and Assumption 13.

6 Application

Oreopoulos (2006) estimates returns to schooling using a major education reform in the United Kingdom that increased the years of compulsory schooling from 14 to 15. Specifically, Oreopoulos (2006) exploit variation resulting from the different timing of implementation of school reforms between Britain (England, Scotland, and Wales) and Northern Ireland as an instrument for education attainment: the school-leaving age increased in Britain in 1947, while the reform was not implemented until 1957 in Northern Ireland. The data are a sample of individuals in Britain and Northern Ireland, who were aged 14 between 1936 and 1965, constructed from combining the series of U.K. General Household Surveys between 1984 and 2006; see Oreopoulos (2006), Oreopoulos (2008) for details.

In the first part of their analysis, Oreopoulos (2006) adopts regression discontinuity designs (RDD) and analyzes the data sets in Britain and Northern Ireland separately. Due to the imprecision of their standard errors[7], Oreopoulos (2006) then moves to "a difference-in-differences and instrumental-variables analysis by combining the two sets of U.K. data".

Specifically, Oreopoulos (2006) (more precisely, Oreopoulos (2008)) runs the following two-way fixed effects instrumental variable regression with the education reform as an excluded instrument for education attainment:

$$Y_{i,t} = \mu_{n.} + \delta_{.t} + \beta_{IV} D_{i,t} + \epsilon_{i,t}, \tag{7}$$

$$D_{i,t} = \gamma_{n.} + \zeta_{.t} + \pi Z_{i,t} + \eta_{i,t}. \tag{8}$$

Here, a cohort (a year when aged 14) plays a role of time as it determines the exposure to the policy change. The dependent variables $Y_{i,t}$ and $D_{i,t}$ are log annual earnings and education attainment for unit i and cohort t, respectively. Both the first stage and the reduced form regressions include a birth cohort fixed effect and a North Ireland fixed effect. The binary instrument $Z_{i,t} \in \{0, 1\}$ takes one if unit i in cohort t is exposed to the policy change.

The staggered introduction of the school reform can be viewed as a natural experiment, but is not randomized across regions in reality; Oreopoulos (2006) notes that the reform was implemented with political support, taking into account the cost and benefit. This indicates that Oreopoulos (2006) implicitly relies on a staggered DID-IV identification strategy instead of the IV design. Indeed, Oreopoulos (2006) presents the corresponding plots of British and Northern Irish average education attainment and average log earnings by cohort to illustrate the evolution of these variables before and after the policy shock.

In the author's setting, however, the effect of the school reform on education attainment and log annual earnings can differ between Britain and North Ireland, and potentially vary across cohorts. The author is also clearly conscious of the heterogeneous treatment effects: the title of this paper is "Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter". The TWFEIV estimator, therefore, can potentially fail to aggregate the treatment effects, yielding a misleading conclusion under staggered DID-IV designs.

[7] Kolesár and Rothe (2018) consider inference in regression discontinuity designs with a discrete running variable and apply their theory to Oreopoulos (2006).

In this section, we first assess the identifying assumptions in staggered DID-IV designs implicitly imposed by Oreopoulos (2006). We then estimate the TWFEIV regression in the author's setting. Finally, we estimate the target parameter and its summary measure by employing our proposed method and weighting scheme.

6.1 Assessing the identifying assumptions in staggered DID-IV designs

We first discuss the validity of the staggered DID-IV identification strategy in Oreopoulos (2006). In the author's setting, our target parameter is the cohort specific average causal response on the treated (CACRT), as education attainment is a non-binary, ordered treatment. We can identify each CACRT and its summary measure if the underlying DID-IV identifying assumptions seem plausible.

Exclusion restriction. It would be plausible, given that the policy reform did not affect log annual earnings other than by increasing education attainment. This assumption may be violated for instance if the reform affected both the quality and quantity of education.

Monotonicity assumption. It would be automatically satisfied in the author's setting: the policy change (instrument) increased the minimum schooling-leaving age from 14 to 15, which ensures that there are no defiers during the periods after the policy shock.

No anticipation in the first stage. It would be plausible that there is no anticipation, if the treatment adoption behavior is the same as the one in the absence of the policy change before its implementation in England. This assumption may be violated if units have private knowledge about the probability of extended education and manipulate their education attainment before the policy shock.

Next, we assess the validity of the parallel trends assumptions in the treatment and the outcome. To do so, we use the interacted two-way fixed effects regressions proposed by Sun and Abraham (2021) in the first stage and the reduced form, respectively. The results are shown in Figure 2.

Parallel trends assumption in the treatment. It requires that the expectation of education attainment would have followed the same path between England and North Ireland across cohorts in the absence of the school reform. Panel (a) in Figure 2 plots the results of the interacted two-way fixed effects regression in the first stage along with a 95% pointwise confidence interval. The pre-exposed estimates are not significantly different from zero and indicate the validity of Assumption 12.

Parallel trends assumption in the outcome. It requires that the expectation of log annual earnings would have followed the same evolution between England and North Ireland across cohorts in the absence of the school reform. Panel (b) in Figure 2 plots the results of the interacted two-way fixed effects regression in the reduced form along with a 95% pointwise confidence interval. The pre-exposed estimates seems consistent with Assumption 13: though there exists an upward pretrends, all the estimates before the initial exposure to the policy change are not significantly different from zero.

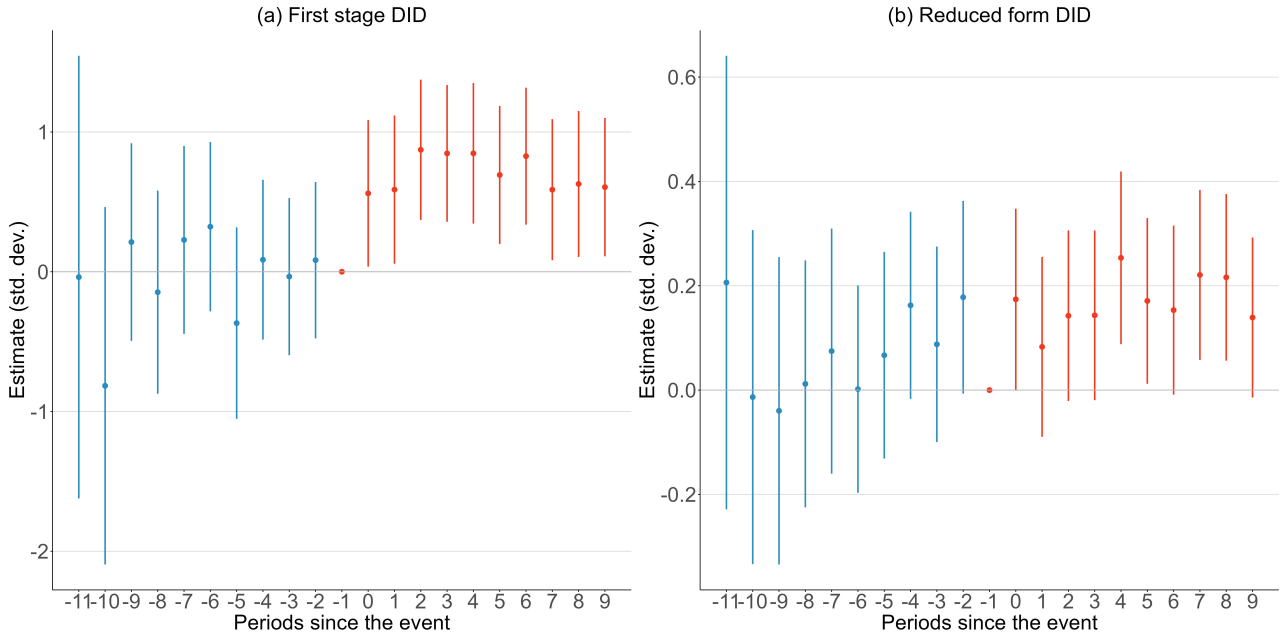

Fig. 2. The effect of the instrument in the first stage and reduced form in the setting of Oreopoulos (2006). *Notes*: The results for the effect of the school reform on education attainment (Panel (a)) and on log annual earnings (Panel (b)) under the staggered DID-IV identification strategy. The unexposed group U is a last-exposed cohort, North Ireland and the reference period is $t = 1946$. The blue lines represent the estimates with pointwise 95% confidence intervals for pre-exposed periods in both panels. These should be equal to zero under the null hypothesis that the parallel trends assumptions in the treatment and the outcome hold. The red lines represent the estimates with pointwise 95% confidence intervals for post-exposed periods in both panels.

Figure 2 also sheds light on the dynamic effects of the policy reform on education attainment and log annual earnings in post-exposed periods. In Panel (a), the estimated increase in years of schooling after the reform ranges from 0.56 to 0.87, and all the estimates are statistically significant. In Panel (b), the estimated increase in log annual earnings after the reform varies from 8% to 25%, and 5 out of 10 estimates are statistically significant.

Note that the post-exposed estimates in Panel (b) do not capture each CACRT in post-exposed periods: each estimate in the reduced form is not scaled by the corresponding estimate in the first stage. In the next section, we estimate each CACRT in post-exposed periods by employing our proposed estimation method.

6.2 Illustrating our estimation method

We start by estimating two-way fixed effects instrumental variable regression in the author's setting. To clearly illustrate the pitfalls of TWFEIV regression, in our estimation, we slightly modify the author's specification. Specifically, Oreopoulos (2006) (more precisely Oreopoulos (2008)) includes some covariates (survey year, sex, and a quartic in age) and runs the weighted regression in the main specification (see Oreopoulos (2006), Oreopoulos (2008) for details), whereas we exclude such covariates and do not apply their weights to our regression.

The result is shown in Table 8. The TWFEIV estimate is -0.009 and it is not significantly different from 0, which indicates that on the whole, the returns to schooling in the U.K. is nearly zero. However, this may be a misleading conclusion: we cannot interpret that the TWFEIV estimand captures the properly weighted average of each CACRT if the effect of the school reform on education attainment or log annual earnings is not stable across cohorts (Miyaji (2024)).

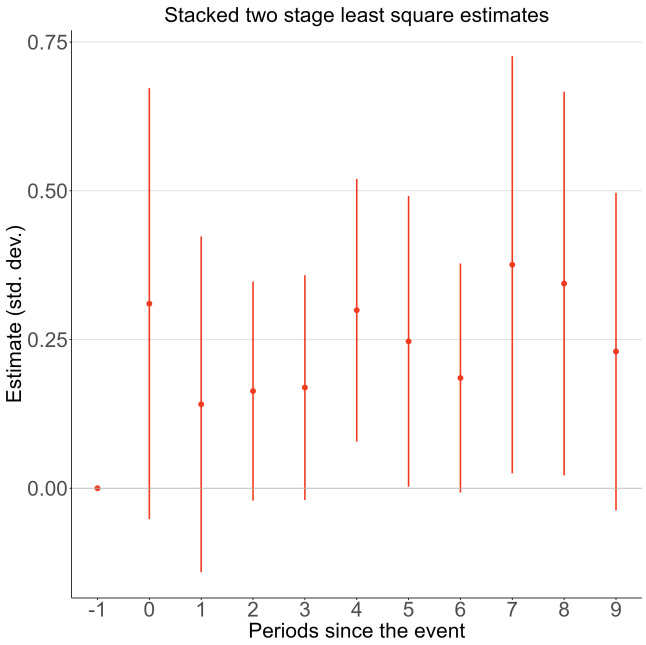

Fig. 3. The cohort specific average causal response on the treated in each relative period in the setting of Oreopoulos (2006). *Notes*: The results for returns to schooling under the staggered DID-IV identification strategy. The unexposed group U is North Ireland and the reference period is period $t = 1946$. The red lines represent the stacked two stage least squares estimates with pointwise 95% confidence intervals for post-exposed periods.

Table 8. Returns to schooling in Oreopoulos (2006)

	Estimate	Standard Error	95% CI
TSLS with fixed effects	-0.009	0.039	[-0.085, 0.068]
$\theta^{IV}_{sel}(e)$	0.240	0.098	[0.047, 0.433]

Notes: Sample size 82790 observations.

The TWFEIV regression performs the "bad comparisons" (c.f. Goodman-Bacon (2021)) in the author's setting: when we run the TWFEIV regression, we treat the set of already exposed units in North Ireland as controls between 1957 and 1965, as both regions are already exposed to the policy shock during these periods. In Appendix D, we quantify the bias terms of the TWFEIV estimand by using Lemma 7 in Miyaji (2024), and show that the TWFEIV estimand is negatively biased in Oreopoulos (2006). Specifically, we show that all the bias terms are positive, whereas all the assigned weights are negative, which yields the downward bias for the TWFEIV estimand.

We now employ our proposed method to estimate each CACRT. We first create data sets by cohort t ($t = 1947, \ldots, 1956$). Each data set only contains units of cohort t and cohort $t = 1946$ in England and North Ireland. We define North Ireland as an unexposed group U and discard the units who were aged at 14 between 1957 and 1965 in order to avoid the bad comparisons. We then run TSLS regression in each data set. The standard error is calculated by using the influence function shown in equation (26) in Appendix C.

Figure 3 plots the point estimates and the corresponding 95% confidence intervals in each relative period after the school reform. The estimates for each $CACRT_{e,t}$ range from 14% to 38% with wide confidence intervals, and 4 out of 10 estimates are statistically significant.

Finally, we estimate the summary causal measure by aggregating each $CACRT_{e,t}$. Specifically, we estimate the weighted average of each $CACRT_{e,t}$ during post-exposed periods in

England ($e = 1947$):

$$\theta^{IV}_{sel}(e) = \sum_{t=1947}^{1956} \frac{CAET^1_{e,t}}{\sum_{t=1947}^{1956} CAET^1_{e,t}} CACRT_{e,t}.$$

Here, each weight assigned to the $CACRT_{e,t}$ represents the relative amount of the effect of the policy reform on education attainment during post-exposed periods in England.

The result is shown in Table 8. The estimate is 0.24 and it is significantly different from zero. The estimated returns to schooling are substantial; it is likely because each $CACRT_{e,t}$ captures the returns to schooling among the compliers: such units may belong to relatively low-skilled labor or low-income family, and have potentially much gain from the school reform.

The estimates obtained from our proposed method and weighting scheme are significantly different from the TWFEIV estimate: our STS estimates and its summary measure are all positive, whereas the TWFEIV estimate is strictly negative. Overall, our results indicate that the economic returns of education are substantial in the U.K. and the estimation method matters in staggered DID-IV designs in practice.

7 Conclusion

Instrumental variable strategies are pervasive in economics to estimate the effect of a treatment on an outcome. In this design, however, we commonly face two challenges: the internal and external validity of the IV estimate. To overcome at least the external validity, many studies have leveraged variation occurring from the different timing of policy adoption across units as an instrument for treatment, and use instrumental variable techniques. The instrument based on these policy changes, however, generally fails to satisfy the randomization assumption in IV designs in practice, casting the doubt on the internal validity of the IV estimate. Rather, the underlying identification strategy is parallel to DID designs, under which we identify the effect of the policy shock on the outcome, leveraging the variation of the exposure to that policy shock across units and over time.

In this paper, we formalize the underlying identification strategy as an instrumented difference-in-differences (DID-IV) by combining the IV techniques with DID designs. We start by considering a simple setting with two periods and two groups. In this setting, our DID-IV design mainly consists of a monotonicity assumption, and parallel trends assumptions in the treatment and the outcome between the two groups. We show that the Wald-DID estimand captures the local average treatment effect on the treated (LATET) under 2×2 DID-IV designs. We also clarify the interpretation of the parallel trends assumption in the outcome. Specifically, we show that in DID-IV settings, time also plays the role of instrument, and the parallel trends assumption in the outcome requires the same expected time gain between exposed and unexposed groups. Finally, we compare DID-IV to Fuzzy DID designs considered in de Chaisemartin and D'Haultfœuille (2018), and point out the issues inherent in Fuzzy DID designs.

Next, we consider the DID-IV design in the case where there are more than two time periods and units become exposed to the instrument at different times. We call this a staggered DID-IV design, and formalize the target parameter and identifying assumptions. Specifically, our target parameter in this design is the cohort specific local average treatment effects on the treated (CLATT). The identifying assumptions are the natural generalization of those in 2×2 DID-IV designs.

We also provide the estimation method in staggered DID-IV designs, which does not require strong restrictions on treatment effect heterogeneity. In practice, most DID-IV applications implement the staggered DID-IV design by running TWFEIV regressions. In companion paper

(Miyaji (2024)), however, we show that the conventional approach is inadequate for summarizing the treatment effects if the effect of the instrument on the treatment or the outcome is not stable over time. Our estimation method carefully chooses the comparison groups and does not suffer from the bias arising from the time-varying exposed effects. We also propose the weighting scheme in staggered DID-IV designs, and explain how one can conduct pretrends test to assess the validity of the parallel trends assumption in the treatment and the outcome.

Finally, we illustrate the empirical relevance of our findings with the setting of Oreopoulos (2006) who estimate returns to schooling in the United Kingdom, exploiting the timing variation of the introduction of school reforms between British and North Ireland. First, we assess the underlying DID-IV identification strategy in the author's setting and check whether the identifying assumptions are plausible. Next, we estimate a TWFEIV regression and our stacked two stage least squares regressions in each dataset. We find that the TWFEV estimate turns out to be strictly negative, whereas our STS estimates are all positive. Finally, we estimate the summary causal parameter by exploiting our STS estimates. The estimate for the summary measure is also positive and indicates that returns to schooling are substantial in the U.K.

Overall, this paper provides a new econometric framework for estimating the causal effects when the treatment adoption is potentially endogenous over time, but researchers can use variation in the timing of policy adoption across units as an instrument for the treatment. To avoid the issue of using the TWFEIV estimator under this design, we also provide a reliable estimation method that is free from strong restrictions on treatment effect heterogeneity. Further developing alternative estimation methods and diagnostic tools will be a promising area for future research, facilitating the credibility of DID-IV research design in practice.

References

Abadie, A., 2003. Semiparametric instrumental variable estimation of treatment response models. J. Econom. 113 (2), 231–263.

Akerman, A., Gaarder, I., Mogstad, M., 2015. The Skill Complementarity of Broadband Internet. Q. J. Econ. 130 (4), 1781–1824.

Angrist, J. D., Imbens, G. W., 1995. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. J. Am. Stat. Assoc. 90 (430), 431–442.

Athey, S., Imbens, G. W., 2022. Design-based analysis in Difference-In-Differences settings with staggered adoption. J. Econom. 226 (1), 62–79.

Bhuller, M., Havnes, T., Leuven, E., Mogstad, M., 2013. Broadband Internet: An Information Superhighway to Sex Crime? Rev. Econ. Stud. 80 (4), 1237–1266.

Black, S. E., Devereux, P. J., Salvanes, K. G., 2005. Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital. Am. Econ. Rev. 95 (1), 437–449.

Borusyak, K., Jaravel, X., Spiess, J., 2021. Revisiting Event Study Designs: Robust and Efficient Estimation.

Callaway, B., Sant'Anna, P. H. C., 2021. Difference-in-Differences with multiple time periods. J. Econom. 225 (2), 200–230.

de Chaisemartin, 2010. A note on instrumented difference in differences. Unpublished Manuscript.

de Chaisemartin, C., D'Haultfœuille, X., 2018. Fuzzy Differences-in-Differences. Rev. Econ. Stud. 85 (2 (303)), 999–1028.

de Chaisemartin, C., D'Haultfœuille, X., 2020. Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. Am. Econ. Rev. 110 (9), 2964–2996.

Deschênes, O., Greenstone, M., Shapiro, J. S., 2017. Defensive Investments and the Demand for Air Quality: Evidence from the NOx Budget Program. Am. Econ. Rev. 107 (10), 2958–2989.

Duflo, E., 2001. Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. Am. Econ. Rev. 91 (4), 795–813.

Field, E., 2007. Entitled to Work: Urban Property Rights and Labor Supply in Peru. Q. J. Econ. 122 (4), 1561–1602.

Fröhlich, M., Frölich, M., Sperlich, S., 2019. Impact Evaluation.: Cambridge University Press.

Goodman-Bacon, A., 2021. Difference-in-differences with variation in treatment timing. J. Econom. 225 (2), 254–277.

Heckman, J. J., Vytlacil, E., 2001. Policy-Relevant Treatment Effects. Am. Econ. Rev. 91 (2), 107–111.

Heckman, J. J., Vytlacil, E., 2005. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. Econometrica. 73 (3), 669–738.

Hudson, S., Hull, P., Liebersohn, J., 2017. Interpreting Instrumented Difference-in-Differences. Available at http://www.mit.edu/~liebers/DDIV.pdf.

Imai, K., Kim, I. S., 2021. On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data. Polit. Anal. 29 (3), 405–415.

Imbens, G. W., Angrist, J. D., 1994. Identification and Estimation of Local Average Treatment Effects. Econometrica. 62 (2), 467–475.

Johnson, R. C., Jackson, C. K., 2019. Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public School Spending. Am. Econ. J. Econ. Policy. 11 (4), 310–349.

Kolesár, M., Rothe, C., 2018. Inference in regression discontinuity designs with a discrete running variable. Am. Econ. Rev. 108 (8), 2277–2304.

Lundborg, P., Nilsson, A., Rooth, D.-O., 2014. Parental education and offspring outcomes: Evidence from the Swedish compulsory school reform. Am. Econ. J. Appl. Econ. 6 (1), 253–278.

Meghir, C., Palme, M., Simeonova, E., 2018. Education and mortality: Evidence from a social experiment. Am. Econ. J. Appl. Econ. 10 (2), 234–256.

Miyaji, S., 2024. Two way fixed effects instrumental variable regressions in a staggered DID-IV design. Unpublished Manuscript.

Olden, A., Møen, J., 2022. The triple difference estimator. Econom. J. 25 (3), 531–553.

Oreopoulos, P., 2006. Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter. Am. Econ. Rev. 96 (1), 152–175.

Oreopoulos, P., 2008. Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter: Corrigendum. Available at https://www.aeaweb.org/articles?id=10.1257/000282806776157641.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., Poe, J., 2023. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. J. Econom.

Słoczyński, T., 2020. When Should We (Not) Interpret Linear IV Estimands as LATE?

Słoczyński, T., Uysal, S. D., Wooldridge, J. M., 2022. Doubly Robust Estimation of Local Average Treatment Effects Using Inverse Probability Weighted Regression Adjustment.: SSRN.

Sun, L., Abraham, S., 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. J. Econom. 225 (2), 175–199.

Vytlacil, E., 2002. Independence, Monotonicity, and Latent Index Models: An Equivalence Result. Econometrica. 70 (1), 331–341.

Wooldridge, J. M., 2020. Introductory Econometrics: A Modern Approach.: 7th edn. Boston, MA: Cengage.

A Proofs in section 2

Proof of Theorem 1.

Proof. First we consider the denominator of the Wald-DID estimand. For the denominator, we have

$$\begin{aligned}
&E[D_1 - D_0|E=1] - E[D_1 - D_0|E=0] \\
=&E[D_1((0,1)) - D_0((0,1))|E=1] - E[D_1((0,0)) - D_0((0,0))|E=0] \\
=&E[D_1((0,1)) - D_1((0,0))|E=1] \\
&+ \{E[D_1((0,0)) - D_0((0,1))|E=1] - E[D_1((0,0)) - D_0((0,0))|E=0]\} \\
=&E[D_1((0,1)) - D_1((0,0))|E=1] \\
&+ \{E[D_1((0,0)) - D_0((0,0))|E=1] - E[D_1((0,0)) - D_0((0,0))|E=0]\} \\
=&Pr(D_1((0,1)) > D_1((0,0))|E=1) \\
=&Pr(CM^Z|E=1).
\end{aligned} \tag{9} \tag{10}$$

where the second equality follows from the simple manipulation. The third equality follows from Assumption 4 (no anticipation in the first stage) and the forth equality follows from Assumption 3 (Monotonicity assumption) and Assumption 5 (Parallel Trends Assumption in the treatment).

Next, to show that the Wald-DID estimand equals to the LATET, it suffices to show that

$$\begin{aligned}
E[Y_1 - Y_0|E=1] - E[Y_1 - Y_0|E=0] &= E[Y_1(1) - Y_1(0)|E=1, CM^Z] \cdot Pr(CM^Z|E=1) \\
&= LATET \cdot Pr(CM^Z|E=1).
\end{aligned}$$

For the numerator, we have

$$\begin{aligned}
&E[Y_1 - Y_0|E=1] - E[Y_1 - Y_0|E=0] \\
=&E[Y_1(D_1((0,1))) - Y_0(D_0((0,1)))|E=1] - E[Y_1(D_1((0,0))) - Y_0(D_0((0,0)))|E=0] \\
=&E[Y_1(D_1((0,1))) - Y_1(D_1((0,0)))|E=1] \\
&+ \{E[Y_1(D_1((0,0))) - Y_0(D_0((0,1)))|E=1] - E[Y_1(D_1((0,0))) - Y_0(D_0((0,0)))|E=0]\} \\
=&E[Y_1(D_1((0,1))) - Y_1(D_1((0,0)))|E=1] \\
&+ \{E[Y_1(D_1((0,0))) - Y_0(D_0((0,0)))|E=1] - E[Y_1(D_1((0,0))) - Y_0(D_0((0,0)))|E=0]\} \\
=&E[(D_1((0,1)) - D_1((0,0))) \cdot (Y_1(1) - Y_1(0))|E=1] \\
=&E[Y_1(1) - Y_1(0)|E=1, CM^Z] \cdot Pr(CM^Z|E=1). \\
=&LATET \cdot Pr(CM^Z|E=1).
\end{aligned} \tag{11}$$

The first equality follows from Assumption 1 (No carryover assumption) and Assumption 2 (Exclusion restriction). The second equality follows from the simple manipulation. The third equality follows from Assumption 4. The forth equality follows from Assumption 6 (Parallel Trends Assumption in the outcome) and the simple calculation. The fifth equality follows from the Law of Iterated Expectations and Assumption 3.

Combining the result (11) with (10), we obtain

$$\begin{aligned}
\frac{E[Y_1 - Y_0|E=1] - E[Y_1 - Y_0|E=0]}{E[D_1 - D_0|E=1] - E[D_1 - D_0|E=1]} &= \frac{LATET \cdot Pr(CM^Z|E=1)}{Pr(CM^Z|E=1)} \\
&= LATET.
\end{aligned}$$

Completing the proof. □

Proof of Theorem 2.

Recall that we define Δ_e as follows:

$$\Delta_e \equiv E[Y_1(D_1((0,0))) - Y_0(D_0((0,0)))|E = e].$$

Then, we obtain:

$$\begin{aligned}
\Delta_e &= E[Y_1(1) - Y_0(1)|AT^T, E = e]Pr(AT^T|E = e) \\
&+ E[Y_1(0) - Y_0(0)|NT^T, E = e]Pr(NT^T|E = e) \\
&+ E[Y_1(1) - Y_0(0)|CM^T, E = e]Pr(CM^T|E = e) \\
&+ E[Y_1(0) - Y_0(1)|DF^T, E = e]Pr(DF^T|E = e) \\
&= \sum_{g^t} w_{g^t,e}\Delta_{g^t,e}.
\end{aligned}$$

Here the first equality follows from the Law of Iterated Expectation and the second equality follows from the definition of $w_{g^t,e}$ and $\Delta_{g^t,e}$. Clearly, the weight $w_{g^t,e}$ sum to one by construction.

Completing the proof.

B Extensions

B.1 Non-binary, ordered treatment

Proof of Theorem 3.

Proof. Let $\lambda^t_{Z,j} = \mathbf{1}(D_t((0,z)) \geq j)$ be the indicator function for $t \in \{0,1\}, z \in \{0,1\}$ and $j \in \{0,\ldots,J+1\}$. We note that $\lambda^t_{Z,0} = 1$ and $\lambda^t_{Z,J+1} = 0$ hold for all t and z by construction. Here, we can rewrite the observed outcome Y_t as follows:

$$\begin{aligned}
Y_t &= Z_1 \cdot Y_t(D_t((0,1))) + (1 - Z_1) \cdot Y_t(D_t((0,0))) \\
&= \left\{ Z_1 \cdot \sum_{j=0}^{J} Y_t(j) \cdot (\lambda^t_{1,j} - \lambda^t_{1,j+1}) \right\} + \left\{ (1 - Z_1) \cdot \sum_{j=0}^{J} Y_t(j) \cdot (\lambda^t_{0,j} - \lambda^t_{0,j+1}) \right\}.
\end{aligned}$$

where we use Assumption 1 and 2.

For the numerator in the Wald DID estimand, we have:

$$\begin{aligned}
&E[Y_1 - Y_0|E = 1] - E[Y_1 - Y_0|E = 0] \\
&= E[Y_1(D_1((0,1))) - Y_1(D_1((0,0)))|E = 1] \\
&+ \{E[Y_1(D_1((0,0))) - Y_0(D_0((0,0)))|E = 1] - E[Y_1(D_1((0,0))) - Y_0(D_0((0,0)))|E = 0]\} \\
&= E[Y_1(D_1((0,1))) - Y_1(D_1((0,0)))|E = 1].
\end{aligned}$$

Here the first equality follows from the simple manipulation and Assumption 4. The second equality follows from Assumption 6.

In terms of the $\lambda^t_{Z,j}$, we write the previous expression as:

$$
\begin{aligned}
&E[Y_1(D_1((0,1))) - Y_1(D_1((0,0)))|E=1] \\
&= E\left[\sum_{j=0}^{J} Y_1(j)\cdot[\lambda^1_{1,j} - \lambda^1_{1,j+1} - \lambda^1_{0,j} + \lambda^1_{0,j+1}]|E=1\right] \\
&= E\left[\sum_{j=1}^{J} (Y_1(j) - Y_1(j-1))\cdot[\lambda^1_{1,j} - \lambda^1_{0,j}] + Y_1(0)\cdot(\lambda^1_{1,0} - \lambda^1_{0,0})|E=1\right] \\
&= E\left[\sum_{j=1}^{J} (Y_1(j) - Y_1(j-1))\cdot[\lambda^1_{1,j} - \lambda^1_{0,j}]|E=1\right].
\end{aligned}
$$

where the second equality follows from the fact: $\lambda^1_{1,0} - \lambda^1_{0,0} = 1$.

We note that $\lambda^1_{1,j} \geq \lambda^1_{0,j}$ from Assumption 3 and that $\lambda^1_{1,j}$ and $\lambda^1_{0,j}$ can take only two values, zero or one. Therefore, we obtain:

$$
\begin{aligned}
&E\left[\sum_{j=1}^{J} (Y_1(j) - Y_1(j-1))\cdot[\lambda^1_{1,j} - \lambda^1_{0,j}]|E=1\right] \\
&= \sum_{j=1}^{J} E[(Y_1(j) - Y_1(j-1))|\lambda^1_{1,j} - \lambda^1_{0,j} = 1, E=1]Pr(\lambda^1_{1,j} - \lambda^1_{0,j} = 1|E=1) \\
&= \sum_{j=1}^{J} E[(Y_1(j) - Y_1(j-1))|D_1((0,1)) \geq j > D_1((0,0)), E=1] \\
&\times Pr(D_1((0,1)) \geq j > D_1((0,0))|E=1). \qquad (12)
\end{aligned}
$$

For the denominator in the Wald-DID estimand, by the similar argument, we first obtain:

$$
\begin{aligned}
&E[D_1 - D_0|E=1] - E[D_1 - D_0|E=0] \\
&= E[D_1((0,1)) - D_1((0,0))|E=1] \\
&+ \{E[D_1((0,0)) - D_0((0,0))|E=1] - E[D_1((0,0)) - D_0((0,0))|E=0]\} \\
&= E[D_1((0,1)) - D_1((0,0))|E=1].
\end{aligned}
$$

Here the first equality follows from the simple algebra and Assumption 4. The second equality follows from Assumption 5.

Then, we have:

$$
\begin{aligned}
&E[D_1((0,1)) - D_1((0,0))|E=1] \\
&= E\left[\sum_{j=0}^{J} j\cdot[\lambda^1_{1,j} - \lambda^1_{1,j+1} - \lambda^1_{0,j} + \lambda^1_{0,j+1}]|E=1\right] \\
&= E\left[\sum_{j=1}^{J} [\lambda^1_{1,j} - \lambda^1_{0,j}]|E=1\right] \\
&= \sum_{j=1}^{J} Pr(D_1((0,1)) \geq j > D_1((0,0))|E=1). \qquad (13)
\end{aligned}
$$

Combining the result (13) with (12), we obtain the desired result. □

B.2 Repeated cross sections

In this section, we present our identification results under repeated cross section settings.

B.2.1 Two time periods

Let Y_i and D_i denote the outcome and the treatment for unit i. Let Z_i denote the instrument path for unit i: $Z_i = (0,0)$ if unit i is not exposed to the instrument and $Z_i = (0,0)$ if unit i is exposed to the instrument. Let $E_i \in \{0,1\}$ denote the group indicator for unit i: $E_i = 1$ if $Z_i = (0,1)$ and $E_i = 0$ if $Z_i = (0,0)$. Let $T_i \in \{0,1\}$ denote the binary indicator for time. For all $z \in \mathcal{S}(Z)$, let $Y_i(0,z), Y_i(1,z)$, and $D_i(z)$ denote potential outcomes and potential treatment choices for unit i. Let $D_{i,t}(z)$ denote potential treatment choices under $T_i = t$. We assume that $\{Y_i, D_i, Z_i, E_i, T_i\}_{i=1}^n$ are independent and identically distributed (i.i.d).

In two periods and two groups settings, our target parameter is the local average treatment effect in period 1:

$$\begin{aligned} LATET &\equiv E[Y_i(1) - Y_i(0) | T_i = 1, E_i = 1, D_{i,1}((0,1)) > D_{i,1}((0,0))] \\ &= E[Y_i(1) - Y_i(0) | T_i = 1, E_i = 1, CM^Z]. \end{aligned}$$

We make the following identification assumptions for the Wald-DID estimand to capture the LATET. These assumptions are suitable for repeated cross section settings.

Assumption 14 (Exclusion restriction)**.**

$$\forall z \in \mathcal{S}(Z), Y_i(0,z) = Y_i(0), Y_i(1,z) = Y_i(1).$$

Assumption 15 (Monotonicity)**.**

$$Pr(D_{i,1}((0,1)) \geq D_{i,1}((0,0))) = 1.$$

Assumption 16 (No anticipation in the first stage)**.**

$$D_{i,0}((0,1)) = D_{i,0}((0,0)) \ \ a.s. \ \text{ for all units } i \text{ with } E_i = 1.$$

Assumption 17 (Parallel Trends assumption in the treatment)**.**

$$\begin{aligned} &E[D_i((0,0)) | E_i = 1, T_i = 1] - E[D_i((0,0)) | E_i = 1, T_i = 0] \\ =&E[D_i((0,0)) | E_i = 0, T_i = 1] - E[D_i((0,0)) | E_i = 0, T_i = 0]. \end{aligned}$$

Assumption 18 (Parallel Trends assumption in the outcome)**.**

$$\begin{aligned} &E[Y_i(D_i(0,0)) | E_i = 1, T_i = 1] - E[Y_i(D_i(0,0)) | E_i = 1, T_i = 0] \\ =&E[Y_i(D_i(0,0)) | E_i = 0, T_i = 1] - E[Y_i(D_i(0,0)) | E_i = 0, T_i = 0]. \end{aligned}$$

Theorem 5 below shows that the Wald-DID estimand is equal to the local average treatment effect on the treated (LATET) in period 1 under Assumptions 14-18. Here, we assume the binary treatment, but non-binary, ordered treatment case is easy to extend as in Theorem 3 (thus we omit it for brevity). In the proof, we omit the index i to ease the notation.

Theorem 5. Suppose Assumptions 14-18 hold. If we assume a binary treatment, the Wald-DID estimand corresponds to the local average treatment effect on the treated (LATET) in period 1:

$$\begin{aligned} w_{DID} &\equiv \frac{E[Y|E=1,T=1]-E[Y|E=1,T=0]-\{E[Y|E=0,T=1]-E[Y|E=0,T=0]\}}{E[D|E=1,T=1]-E[D|E=1,T=0]-\{E[D|E=0,T=1]-E[D|E=0,T=0]\}} \\ &= E[Y(1)-Y(0)|T=1,E=1,CM^Z] \\ &= LATET. \end{aligned}$$

Proof. First we consider the denominator of the Wald-DID estimand. For the denominator, we have

$$\begin{aligned} & E[D|E=1,T=1]-E[D|E=1,T=0]-\{E[D|E=0,T=1]-E[D|E=0,T=0]\} \\ =& E[D((0,1))|E=1,T=1]-E[D((0,1))|E=1,T=0] \\ & -\{E[D((0,0))|E=0,T=1]-E[D((0,0))|E=0,T=0]\} \\ =& E[D((0,1))-D((0,0))|E=1,T=1] \\ & +\Big\{E[D((0,0))|E=1,T=1]-E[D((0,1))|E=1,T=0] \\ & -\{E[D((0,0))|E=0,T=1]-E[D((0,0))|E=0,T=0]\}\Big\} \\ =& E[D((0,1))-D((0,0))|E=1,T=1] \\ & +\Big\{E[D((0,0))|E=1,T=1]-E[D((0,0))|E=1,T=0] \\ & -\{E[D((0,0))|E=0,T=1]-E[D((0,0))|E=0,T=0]\}\Big\} \\ =& Pr(D_1((0,1))>D_1((0,0))|E=1,T=1) \\ =& Pr(CM^Z|E=1,T=1). \end{aligned} \tag{14}$$

The second equality follows from the simple manipulation. The third equality follows from Assumption 16 and the forth equality follows from Assumption 15 and Assumption 17.

Next, for the numerator, we have

$$
\begin{aligned}
&E[Y|E=1,T=1]-E[Y|E=1,T=0]-\{E[Y|E=0,T=1]-E[Y|E=0,T=0]\}\\
=&E[Y(D((0,1)))|E=1,T=1]-E[Y(D((0,1)))|E=1,T=0]\\
&-\{E[Y(D((0,0)))|E=0,T=1]-E[Y(D((0,0)))|E=0,T=0]\}\\
=&E[Y(D((0,1)))-Y(D((0,0)))|E=1,T=1]\\
&+\Big\{E[Y(D((0,0)))|E=1,T=1]-E[Y(D((0,1)))|E=1,T=0]\\
&-\{E[Y(D((0,0)))|E=0,T=1]-E[Y(D((0,0)))|E=0,T=0]\Big\}\\
=&E[Y(D((0,1)))-Y(D((0,0)))|E=1,T=1]\\
&+\Big\{E[Y(D((0,0)))|E=1,T=1]-E[Y(D((0,0)))|E=1,T=0]\\
&-\{E[Y(D((0,0)))|E=0,T=1]-E[Y(D((0,0)))|E=0,T=0]\Big\}\\
=&E[Y(D((0,1)))-Y(D((0,0)))|E=1,T=1]\\
=&E[(D((0,1))-D((0,0)))\cdot(Y(1)-Y(0))|E=1,T=1]\\
=&E[Y(1)-Y(0)|T=1,E=1,CM^Z]\cdot Pr(CM^Z|E=1,T=1).\\
=&LATET\cdot Pr(CM^Z|E=1,T=1). \qquad (15)
\end{aligned}
$$

The second equality follows from the simple manipulation. The third equality follows from Assumption 16. The forth equality follows from Assumption 18 and the simple calculation. The fifth equality follows from the Law of Iterated Expectations and Assumption 15.

Combining the result (14) with (15), we obtain

$$
\begin{aligned}
w_{DID} &= \frac{LATET\cdot Pr(CM^Z|E=1,T=1)}{Pr(CM^Z|E=1,T=1)}\\
&= LATET.
\end{aligned}
$$

Completing the proof. □

B.2.2 Multiple time periods

Let $T_i \in \{1,\dots,T\}$ denote the time period when unit i is observed and let $Z_{i,t}$ denote the instrument status under $T_i = t$. We assume Assumption 7 in section 3, that is, we assume the staggered assignment of the instrument across units. In addition, we assume Assumption 14 (Exclusion restriction), which allows us to write $Y_i(D_i, z) = Y_i(D_i)$. Let $E_i \in \{1,\dots,T,\infty\}$ denote the cohort which unit i belongs to. Let $Y_i(D_i^e)$ and D_i^e denote potential outcomes and potential treatment choices when unit i belongs to cohort e. Let $D_{i,t}^e$ denote potential treatment choices under $T_i = t$ when unit i belongs to cohort e. We assume that $\{Y_i, D_i, Z_i, E_i, T_i\}_{i=1}^n$ are independent and identically distributed (i.i.d).

In multiple time periods and multiple cohorts settings, our target parameter is the cohort specific local average treatment effect on the treated in period $T = e + l$:

$$
\begin{aligned}
CLATT_{e,e+l} &\equiv E[Y_i(1)-Y_i(0)|T_i=e+l,E_i=e,D_{i,e+l}^e > D_{i,e+l}^\infty]\\
&= E[Y_i(1)-Y_i(0)|T_i=e+l,E_i=e,CM_{e,e+l}].
\end{aligned}
$$

When repeated cross section data are available and we have multiple time periods, our staggered DID-IV designs consist of Assumptions 7, 10, 11 (in section 3), Assumption 14 (Exclusion restriction), and the following parallel trends assumptions.

Assumption 19 (Parallel Trends Assumption in the treatment in multiple time periods)**.**

$$\text{For all } s < t, \;\; E[D_i^{\infty}|E_i = e, T_i = t] - E[D_i^{\infty}|E_i = e, T_i = s] \;\; \text{is same for all } e \in \mathcal{S}(E_i).$$

Assumption 20 (Parallel Trends Assumption in the outcome in multiple time periods)**.**

$$\text{For all } s < t, \;\; E[Y_i(D_i^{\infty})|E_i = e, T_i = t] - E[Y_i(D_i^{\infty})|E_i = e, T_i = s] \;\; \text{is same for all } e \in \mathcal{S}(E_i).$$

B.3 Introducing the treatment path in 2×2 DID-IV designs

In this section, we discuss the DID-IV identification strategy in two periods with a binary treatment, introducing the treatment path in potential outcomes $Y_{i,t}(d, z)$. We first consider the two types of the no anticipation assumption on potential outcomes $Y_{i,t}(d, z)$ instead of imposing Assumption 1 (No carryover assumption). We then define the *expected time gain* and formalize the parallel trends assumption in the outcome as in Assumption 6 in this setting. Finally, we show that a Wald-DID estimand identifies the LATET' (see section 4) under the suitable assumptions similar to the ones in section 2.3. Henceforth, we consider the same setting as in section 2.1 except that we do not adopt Assumption 1.

First, we make the following no anticipation assumptions on the outcome.

Assumption 21 (No anticipation on the outcome among the $CM^Z \wedge NT^T$ in an exposed group)**.**

$$\forall z \in \mathcal{S}(Z), Y_{i,0}((0,1), z) = Y_{i,0}((0,0), z) \;\; \text{for all the units with } E_i = 1 \text{ and } (CM^Z \wedge NT^T).$$

This assumption requires that the potential outcome at period 0 is same as the baseline outcome $Y_{i,0}((0,0), z)$ given the instrument path z, for those who are assigned to the instrument $z = (0,1)$ and belongs to the type $CM^Z \wedge NT^T$. We note that this assumption corresponds to the no anticipation assumption usually made in the common DID set up as in Roth et al. (2023) (see Table 3 and Table 4 in section 2.5)[8].

Assumption 22 (No anticipation on the outcome among the $CM^Z \wedge DF^T$ in an exposed group)**.**

$$\forall z \in \mathcal{S}(Z), Y_{i,0}((1,1), z) = Y_{i,0}((1,0), z) \;\; \text{for all the units with } E_i = 1 \text{ and } (CM^Z \wedge DF^T).$$

This assumption requires that the potential outcome at period 0 is same as the baseline outcome $Y_{i,0}((1,0), z)$ given the instrument path z, for those who are assigned to the instrument $z = (0,1)$ and belongs to the type $CM^Z \wedge DF^T$. Assumption 22 is inherent to the DID-IV identification strategy when we introduce the treatment path.

We can rationalize Assumption 21 and 22 by noting that both the two Assumptions and the no anticipation assumption proposed by the recent DID literature restrict the anticipatory behavior of the units who belong to the exposed group and the compliers CM^Z.

[8]Here, it is inconsequential whether the instrument path z is included or not in potential outcomes because we have $d = z$ in the common DID set up.

We now make the parallel trends assumption in the outcome that is suitable for the set up we are considering here. We first define the similar notation as in section 2.4. Let $\bar{\Delta}_{g^t,e}$ denote the expectation of the time trends in each groups in group $E_i = e$:

$$\begin{aligned}
\bar{\Delta}_{AT^T,e} &\equiv E[Y_{i,1}((1,1)) - Y_{i,0}((1,1))|AT^T, E_i = e], \\
\bar{\Delta}_{NT^T,e} &\equiv E[Y_{i,1}((0,0)) - Y_{i,0}((0,0))|NT^T, E_i = e], \\
\bar{\Delta}_{CM^T,e} &\equiv E[Y_{i,1}((0,1)) - Y_{i,0}((0,1))|CM^T, E_i = e], \\
\bar{\Delta}_{DF^T,e} &\equiv E[Y_{i,1}((1,0)) - Y_{i,0}((1,0))|DF^T, E_i = e].
\end{aligned}$$

Let $\bar{\Delta}_0$ and $\bar{\Delta}_1$ denote the *expected time gain* in unexposed and exposed groups respectively:

$$\begin{aligned}
\bar{\Delta}_0 &\equiv \sum_{g^t} w_{g^t,0} \bar{\Delta}_{g^t,0}, \\
\bar{\Delta}_1 &\equiv \sum_{g^t} w_{g^t,1} \bar{\Delta}_{g^t,1},
\end{aligned}$$

where the definition of the weight $w_{g^t,e}$ is same as in section 2.4.

As we previously noted in section 2.4, the parallel trends assumption in the outcome can be interpreted as requiring the homogeneous expected time gain between exposed and unexposed groups. The following assumption indeed make a such restriction in the current set up.

Assumption 23 (Parallel trend assumption in the outcome)**.** The expected time gain between exposed and unexposed groups is same, that is, we have $\bar{\Delta}_0 = \bar{\Delta}_1$.

The theorem below shows that a Wald-DID estimand is equal to the target parameter $LATET'$ under Assumption 2-5 and Assumption 21-23.

Theorem 6. Suppose Assumption 2-5 and Assumption 21-23 holds. Then, a Wald-DID estimand corresponds to $LATET'$, that is, we have

$$\frac{E[Y_1 - Y_0|E = 1] - E[Y_1 - Y_0|E = 0]}{E[D_1 - D_0|E = 1] - E[D_1 - D_0|E = 0]} = LATET'.$$

Proof. First, we consider the denominator of the Wald-DID estimand. From the similar argument in the proof of Theorem 1, one can show the following result:

$$\begin{aligned}
E[D_1 - D_0|E = 1] - E[D_1 - D_0|E = 0] &= Pr(CM^Z|E = 1) \\
&= Pr(CM^Z \wedge DF^T|E = 1) + Pr(CM^Z \wedge NT^T|E = 1).
\end{aligned} \tag{16}$$

Here the second equality holds from the fact that we can partition the compliers CM^Z in an exposed group into $(CM^Z \wedge DF^T)$ and $(CM^Z \wedge NT^T)$ (see Table 1).

Next, for the numerator, we first decompose $E[Y_1 - Y_0|E=1]$ as follows.

$$
\begin{aligned}
E[Y_1 - Y_0|E=1] &= \bar{\Delta}_{CM^T,1} w_{CM^T,1} + \bar{\Delta}_{AT^T,1} w_{AT^T,1} \\
&+ E[Y_1((0,1)) - Y_0((0,1))|E=1,(CM^Z \wedge NT^T)]Pr((CM^Z \wedge NT^T)|E=1) \\
&+ E[Y_1((1,1)) - Y_0((1,1))|E=1,(CM^Z \wedge DF^T)]Pr((CM^Z \wedge DF^T)|E=1) \\
&+ E[Y_1((1,0)) - Y_0((1,0))|E=1,(NT^Z \wedge DF^T)]Pr((NT^Z \wedge DF^T)|E=1) \\
&+ E[Y_1((0,0)) - Y_0((0,0))|E=1,(NT^Z \wedge NT^T)]Pr((NT^Z \wedge NT^T)|E=1) \\
&= \bar{\Delta}_{CM^T,1} w_{CM^T,1} + \bar{\Delta}_{AT^T,1} w_{AT^T,1} \\
&+ E[Y_1((0,1)) - Y_1((0,0))|E=1,(CM^Z \wedge NT^T)]Pr((CM^Z \wedge NT^T)|E=1) \\
&+ E[Y_1((0,0)) - Y_0((0,0))|E=1,(CM^Z \wedge NT^T)]Pr((CM^Z \wedge NT^T)|E=1) \\
&+ E[Y_1((1,1)) - Y_1((1,0))|E=1,(CM^Z \wedge DF^T)]Pr((CM^Z \wedge DF^T)|E=1) \\
&+ E[Y_1((1,0)) - Y_0((1,0))|E=1,(CM^Z \wedge DF^T)]Pr((CM^Z \wedge DF^T)|E=1) \\
&+ E[Y_1((1,0)) - Y_0((1,0))|E=1,(NT^Z \wedge DF^T)]Pr((NT^Z \wedge DF^T)|E=1) \\
&+ E[Y_1((0,0)) - Y_0((0,0))|E=1,(NT^Z \wedge NT^T)]Pr((NT^Z \wedge NT^T)|E=1) \\
&= \bar{\Delta}_{CM^T,1} w_{CM^T,1} + \bar{\Delta}_{AT^T,1} w_{AT^T,1} \\
&+ \bar{\Delta}_{NT^T,1} w_{NT^T,1} + \bar{\Delta}_{DF^T,1} w_{DF^T,1} \\
&+ E[Y_1((0,1)) - Y_1((0,0))|E=1,(CM^Z \wedge NT^T)]Pr((CM^Z \wedge NT^T)|E=1) \\
&+ E[Y_1((1,1)) - Y_1((1,0))|E=1,(CM^Z \wedge DF^T)]Pr((CM^Z \wedge DF^T)|E=1) \\
&= \bar{\Delta}_1 \\
&+ E[Y_1((0,1)) - Y_1((0,0))|E=1,(CM^Z \wedge NT^T)]Pr((CM^Z \wedge NT^T)|E=1) \\
&+ E[Y_1((1,1)) - Y_1((1,0))|E=1,(CM^Z \wedge DF^T)]Pr((CM^Z \wedge DF^T)|E=1)
\end{aligned} \tag{17}
$$

The first equality follows from Table 1 and Assumption 2. The second equality follows from Assumption 21, Assumption 22 and simple manipulation. The third equality follows from Table 1. The final equality follows from the definition of the expected time gain $\bar{\Delta}_1$.

We next decompose $E[Y_1 - Y_0|E=0]$ as follows.

$$
\begin{aligned}
&E[Y_1 - Y_0|E=0] \\
&= E[Y_1((1,1)) - Y_0((1,1))|E=0,AT^T]Pr(AT^T|E=0) \\
&+ E[Y_1((0,1)) - Y_0((0,1))|E=0,CM^T]Pr(CM^T|E=0) \\
&+ E[Y_1((0,0)) - Y_0((0,0))|E=0,NT^T]Pr(NT^T|E=0) \\
&+ E[Y_1((1,0)) - Y_0((1,0))|E=0,DF^T]Pr(DF^T|E=0) \\
&= \bar{\Delta}_0.
\end{aligned} \tag{18}
$$

Here the first equality follows from Table 2 and the second equality follows from the definition of the expected time gain $\bar{\Delta}_0$.

From the result (18) with (17) and Assumption 23, we obtain

$$
\begin{aligned}
&E[Y_1 - Y_0|E=1] - E[Y_1 - Y_0|E=0] \\
&= E[Y_1((0,1)) - Y_1((0,0))|E=1,(CM^Z \wedge NT^T)]Pr((CM^Z \wedge NT^T)|E=1) \\
&+ E[Y_1((1,1)) - Y_1((1,0))|E=1,(CM^Z \wedge DF^T)]Pr((CM^Z \wedge DF^T)|E=1).
\end{aligned} \tag{19}
$$

Combining the result (19) with (16), we obtain the desired result. □

C Proof in section 4

C.1 Proof of Theorem 4

First, we prove the consistency of our SLS estimator. Recall that $\hat{\alpha}^{e,l}$ and $\hat{\pi}^{e,l}$ take the following form:

$$\begin{aligned}
\hat{\alpha}^{e,l} &= \frac{E_N[(Y_{i,e+l} - Y_{i,e-1}) \cdot \mathbf{1}\{E_i = e\}]}{E_N[\mathbf{1}\{E_i = e\}]} - \frac{E_N[(Y_{i,e+l} - Y_{i,e-1}) \cdot \mathbf{1}\{E_i \in U\}]}{E_N[\mathbf{1}\{E_i \in U\}]} \\
&\equiv \hat{\alpha}^1_{e,l} - \hat{\alpha}^2_{e,l}. \\
\hat{\pi}^{e,l} &= \frac{E_N[(D_{i,e+l} - D_{i,e-1}) \cdot \mathbf{1}\{E_i = e\}]}{E_N[\mathbf{1}\{E_i = e\}]} - \frac{E_N[(D_{i,e+l} - D_{i,e-1}) \cdot \mathbf{1}\{E_i \in U\}]}{E_N[\mathbf{1}\{E_i \in U\}]} \\
&\equiv \hat{\pi}^1_{e,l} - \hat{\pi}^2_{e,l}.
\end{aligned}$$

We assume that $\hat{\alpha}^{e,l}$ and $\hat{\pi}^{e,l}$ are well defined. More precisely, we assume that cohort e is non-empty, that is, $\sum_{i=1}^n \mathbf{1}\{E_i = e\} > 0$ and there exist some sets of the unexposed cohort $U \subseteq \{u : e - 1 < u \leq T\}$ that are non-empty, i.e., $\sum_{i=1}^n \mathbf{1}\{E_i \in U\} > 0$. This assumption guarantees that each Wald-DID estimator

$$\begin{aligned}
\widehat{CLATT}_{e,e+l} &= \frac{\hat{\alpha}^{e,l}}{\hat{\pi}^{e,l}} \\
&= \frac{\hat{\alpha}^1_{e,l} - \hat{\alpha}^2_{e,l}}{\hat{\pi}^1_{e,l} - \hat{\pi}^2_{e,l}}.
\end{aligned}$$

is also well defined.

First we derive the probability limit of $\widehat{CLATT}_{e,e+l}$. We note that by the Law of Large Numbers and Slutsky's theorem, we have

$$\begin{aligned}
\hat{\alpha}^1_{e,l} &\xrightarrow{p} \frac{E[(Y_{i,e+l} - Y_{i,e-1}) \cdot \mathbf{1}\{E_i = e\}]}{Pr(E_i = e)} \\
&= E[Y_{i,e+l} - Y_{i,e-1} | E_i = e],
\end{aligned}$$

where the second equality follows from the Law of Iterated Expectations.

By the similar arguments, we have

$$\begin{aligned}
\hat{\alpha}^2_{e,l} &\xrightarrow{p} E[Y_{i,e+l} - Y_{i,e-1} | E_i \in U] \\
\hat{\pi}^1_{e,l} &\xrightarrow{p} E[D_{i,e+l} - D_{i,e-1} | E_i = e] \\
\hat{\pi}^2_{e,l} &\xrightarrow{p} E[D_{i,e+l} - D_{i,e-1} | E_i \in U]
\end{aligned}$$

From the Slutsky's theorem, we have

$$\widehat{CLATT}_{e,e+l} \xrightarrow{p} \frac{E[Y_{i,e+l} - Y_{i,e-1} | E_i = e] - E[Y_{i,e+l} - Y_{i,e-1} | E_i \in U]}{E[D_{i,e+l} - D_{i,e-1} | E_i = e] - E[D_{i,e+l} - D_{i,e-1} | E_i \in U]}. \tag{20}$$

Lastly, it remains to show that the right hand side of (20) is equal to $CLATT_{e,e+l}$.

For the numerator, we have

$$
\begin{aligned}
&E[Y_{i,e+l} - Y_{i,e-1}|E_i = e] - E[Y_{i,e+l} - Y_{i,e-1}|E_i \in U] \\
=&E[Y_{i,e+l} - Y_{i,e-1}|E_i = e] - \sum_{e \in U} E[Y_{i,e+l} - Y_{i,e-1}|E_i = e] Pr(E_i = e|E_i \in U) \\
=&E[Y_{i,e+l}(D^e_{i,e+l}) - Y_{i,e-1}(D^\infty_{i,e-1})|E_i = e] - \sum_{e \in U} E[Y_{i,e+l}(D^\infty_{i,e+l}) - Y_{i,e-1}(D^\infty_{i,e-1})|E_i \in U] Pr(E_i = e|E_i \in U) \\
=&E[Y_{i,e+l}(D^e_{i,e+l}) - Y_{i,e+l}(D^\infty_{i,e+l})|E_i = e] + E[Y_{i,e+l}(D^\infty_{i,e+l}) - Y_{i,e-1}(D^\infty_{i,e-1})|E_i = e] \\
&- \sum_{e \in U} E[Y_{i,e+l}(D^\infty_{i,e+l}) - Y_{i,e-1}(D^\infty_{i,e-1})|E_i \in U] Pr(E_i = e|E_i \in U) \\
=&E[Y_{i,e+l}(D^e_{i,e+l}) - Y_{i,e+l}(D^\infty_{i,e+l})|E_i = e] \\
&+ E[Y_{i,e+l}(D^\infty_{i,e+l}) - Y_{i,e-1}(D^\infty_{i,e-1})] - E[Y_{i,e+l}(D^\infty_{i,e+l}) - Y_{i,e-1}(D^\infty_{i,e-1})] \\
=&E[Y_{i,e+l}(1) - Y_{i,e+l}(0)|E_i = e, CM_{e,e+l}] Pr(CM_{e,e+l}|E_i = e). \qquad (21)
\end{aligned}
$$

The third equality follows from Assumption 8, 9 and 11. The forth equality follows from Assumption 13 and the fifth equality follows from Assumption 10.

For the denominator, by the similar calculations,

$$
\begin{aligned}
&E[D_{i,e+l} - D_{i,e-1}|E_i = e] - E[D_{i,e+l} - D_{i,e-1}|E_i \in U] \\
&= E[D^e_{i,e+l} - D^\infty_{i,e+l}|E_i = e] + E[D^\infty_{i,e+l} - D^\infty_{i,e-1}|E_i = e] - \sum_{e \in U} E[D^\infty_{i,e+l} - D^\infty_{i,e-1}|E_i = e] Pr(E_i = e|E_i \in U) \\
&= Pr(CM_{e,e+l}|E_i = e), \qquad (22)
\end{aligned}
$$

where the final equality follows from Assumption 10 and Assumption 12.

Combining the result (21) with (22), we obtain the desirable result:

$$
\begin{aligned}
&\frac{E[Y_{i,e+l} - Y_{i,e-1}|E_i = e] - E[Y_{i,e+l} - Y_{i,e-1}|E_i \in U]}{E[D_{i,e+l} - D_{i,e-1}|E_i = e] - E[D_{i,e+l} - D_{i,e-1}|E_i \in U]} \\
&= \frac{E[Y_{i,e+l}(1) - Y_{i,e+l}(0)|E_i = e, CM_{e,e+l}] Pr(CM_{e,e+l}|E_i = e)}{Pr(CM_{e,e+l}|E_i = e)} \\
&= CLATT_{e,e+l}.
\end{aligned}
$$

Next, we prove that our STS estimator is asymptotically normal, deriving its influence function.

We define $\theta_{e,l}$ to be

$$
\begin{aligned}
\theta_{e,l} &\equiv \frac{E[Y_{i,e+l} - Y_{i,e-1}|E_i = e] - E[Y_{i,e+l} - Y_{i,e-1}|E_i \in U]}{E[D_{i,e+l} - D_{i,e-1}|E_i = e] - E[D_{i,e+l} - D_{i,e-1}|E_i \in U]} (= CLATT_{e,e+l}) \\
&\equiv \frac{\alpha^1_{e,l} - \alpha^2_{e,l}}{\pi^1_{e,l} - \pi^2_{e,l}}.
\end{aligned}
$$

We use the following fact repeatedly in the derivation. This fact is also found in de Chaisemartin and D'Haultfœuille (2018).

Fact 1. *If*

$$
\sqrt{n}(\hat{A} - A) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} a_i + o_p(1), \quad \sqrt{n}(\hat{B} - B) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_i + o_p(1),
$$

we have

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

First we derive the influence function of $\sqrt{n}(\hat{\alpha}^1_{e,l}+\hat{\alpha}^2_{e,l}-(\alpha^1_{e,l}+\alpha^2_{e,l}))$. We note that

$$\sqrt{n}(E_N[\mathbf{1}\{E_i=e\}]-E[\mathbf{1}\{E_i=e\}])=\frac{1}{\sqrt{n}}\sum_{i}^{n}(\mathbf{1}\{E_i=e\}-E[\mathbf{1}\{E_i=e\}]).$$

and

$$\begin{aligned}&\sqrt{n}(E_N[(Y_{i,e+l}-Y_{i,e-1})\mathbf{1}\{E_i=e\}]-E[(Y_{i,e+l}-Y_{i,e-1})\mathbf{1}\{E_i=e\}])\\&=\frac{1}{\sqrt{n}}\sum_{i}^{n}((Y_{i,e+l}-Y_{i,e-1})\cdot\mathbf{1}\{E_i=e\}-E[(Y_{i,e+l}-Y_{i,e-1})\cdot\mathbf{1}\{E_i=e\}]).\end{aligned}$$

From the observations and using Fact 1, we obtain

$$\begin{aligned}\sqrt{n}(\hat{\alpha}^1_{e,l}-\alpha^1_{e,l})&=\frac{1}{\sqrt{n}}\sum_{i}^{n}\frac{\mathbf{1}\{E_i=e\}\cdot[(Y_{i,e+l}-Y_{i,e-1})-E[Y_{i,e+l}-Y_{i,e-1}|E_i=e]]}{E[\mathbf{1}\{E_i=e\}]}+o_p(1).\\&\equiv\frac{1}{\sqrt{n}}\sum_{i}^{n}\zeta_{i,e,l}+o_p(1).\end{aligned}$$

By symmetry, we also obtain the influence function of $\sqrt{n}(\hat{\alpha}^2_{e,l}-\alpha^2_{e,l})$:

$$\begin{aligned}\sqrt{n}(\hat{\alpha}^2_{e,l}-\alpha^2_{e,l})&=\frac{1}{\sqrt{n}}\sum_{i}^{n}\frac{\mathbf{1}\{E_i=U\}\cdot[(Y_{i,e+l}-Y_{i,e-1})-E[Y_{i,e+l}-Y_{i,e-1}|E_i=U]]}{E[\mathbf{1}\{E_i=U\}]}+o_p(1).\\&\equiv\frac{1}{\sqrt{n}}\sum_{i}^{n}\phi_{i,e,l}+o_p(1).\end{aligned}$$

This implies that

$$\begin{aligned}\sqrt{n}(\hat{\alpha}^1_{e,l}-\hat{\alpha}^2_{e,l}-(\alpha^1_{e,l}-\alpha^2_{e,l}))&=\frac{1}{\sqrt{n}}\sum_{i}^{n}(\zeta_{i,e,l}-\phi_{i,e,l})+o_p(1)\\&\equiv\frac{1}{\sqrt{n}}\sum_{i}^{n}a_{i,e,l}+o_p(1).\end{aligned}\tag{23}$$

Next we derive the influence function of $\sqrt{n}(\hat{\pi}^1_{e,l}+\hat{\pi}^2_{e,l}-(\pi^1_{e,l}+\pi^2_{e,l}))$. From the symmetry, by replacing Y with D, we obtain the following:

$$\begin{aligned}\sqrt{n}(\hat{\pi}^1_{e,l}-\hat{\pi}^2_{e,l}-(\pi^1_{e,l}-\pi^2_{e,l}))&=\frac{1}{\sqrt{n}}\sum_{i}^{n}(\eta_{i,e,l}-\xi_{i,e,l})+o_p(1)\\&\equiv\frac{1}{\sqrt{n}}\sum_{i}^{n}b_{i,e,l}+o_p(1),\end{aligned}\tag{24}$$

where we define

$$\eta_{i,e,l} \equiv \frac{\mathbf{1}\{E_i = e\} \cdot [(D_{i,e+l} - D_{i,e-1}) - E[D_{i,e+l} - D_{i,e-1}|E_i = e]]}{E[\mathbf{1}\{E_i = e\}]},$$
$$\xi_{i,e,l} \equiv \frac{\mathbf{1}\{E_i = U\} \cdot [(D_{i,e+l} - D_{i,e-1}) - E[D_{i,e+l} - D_{i,e-1}|E_i = U]]}{E[\mathbf{1}\{E_i = U\}]}.$$

Using (23), (24) and Fact 1, by some calculations, we obtain

$$\begin{aligned}\sqrt{n}(\widehat{CLATT}_{e,e+l} - CLATT_{e,e+l}) &= \sqrt{n}(\frac{\hat{\alpha}^1_{e,l} - \hat{\alpha}^2_{e,l}}{\hat{\pi}^1_{e,l} - \hat{\pi}^2_{e,l}} - \frac{\alpha^1_{e,l} - \alpha^2_{e,l}}{\pi^1_{e,l} - \pi^2_{e,l}}) \\ &= \frac{1}{\sqrt{n}} \sum_i^n \psi_{i,e,l} + o_p(1),\end{aligned}$$

where the influence function $\psi_{i,e,l}$ is:

$$\begin{aligned}&\psi_{i,e,l} \\ &= \frac{1}{\pi^1_{e,l} - \pi^2_{e,l}} \left(\zeta_{i,e,l} - \phi_{i,e,l} - \frac{\alpha^1_{e,l} - \alpha^2_{e,l}}{\pi^1_{e,l} - \pi^2_{e,l}} (\eta_{i,e,l} - \xi_{i,e,l}) \right) \\ &= \frac{1}{\pi^1_{e,l} - \pi^2_{e,l}} \left(\frac{\mathbf{1}\{E_i = e\} \left[\delta^p_{i,e,l} - E[\delta^p_{i,e,l}|E_i = e]\right]}{E[\mathbf{1}\{E_i = e\}]} - \frac{\mathbf{1}\{E_i = U\} \left[\delta^p_{i,e,l} - (E[\delta^p_{i,e,l}|E_i = U])\right]}{E[\mathbf{1}\{E_i = U\}]} \right),\end{aligned} \tag{25}$$

where we define $\delta^p_{i,e,l} = (Y_{i,e+l} - Y_{i,e-1}) - \theta_{e,l} \cdot (D_{i,e+l} - D_{i,e-1})$.

Therefore, the asymptotic variance $V(\psi_{i,e,l})$ is

$$\begin{aligned}&V(\psi_{i,e,l}) \\ &= \frac{1}{(\pi^1_{e,l} - \pi^2_{e,l})^2} \left(\frac{E\left[\mathbf{1}\{E_i = e\} \left[\delta^p_{i,e,l} - E[\delta^p_{i,e,l}|E_i = e]\right]^2\right]}{E[\mathbf{1}\{E_i = e\}]^2} + \frac{E\left[\mathbf{1}\{E_i = U\} \left[\delta^p_{i,e,l} - (E[\delta^p_{i,e,l}|E_i = U]\right]^2\right]}{E[\mathbf{1}\{E_i = U\}]^2} \right).\end{aligned}$$

C.2 Repeated cross sections

We also present the influence function of our STS estimator when repeated cross section data are available.[9]. Let $\theta^r_{e,l}$ define

$$\theta^r_{e,l} \equiv \frac{\alpha_{e,e+l} - \alpha_{e,e-1} - [\beta_{U,e+l} - \beta_{U,e-1}]}{\pi_{e,e+l} - \pi_{e,e-1} - [\gamma_{U,e+l} - \gamma_{U,e-1}]},$$

where

$$\begin{aligned}&\alpha_{e,t} = E[Y_i|E_i = e, T_i = t], \quad \beta_{U,t} = E[Y_i|E_i \in U, T_i = t], \\ &\pi_{e,t} = E[D_i|E_i = e, T_i = t], \quad \gamma_{U,t} = E[D_i|E_i \in U, T_i = t] \ \ (t = e + l, e - 1).\end{aligned}$$

In repeated cross section settings, our STS estimator is the sample analog of $\theta^r_{e,l}$, which we denote $\hat{\theta}^r_{e,l}$. By the similar arguments in the proof of Theorem 4, one can show that $\hat{\theta}^r_{e,l}$ is

[9]The derivation here is essentially the same in de Chaisemartin and D'Haultfœuille (2018), who present the influence function of the Wald-DID estimator in repeated cross section settings with two periods and two groups.

consistent for $CLATT_{e,e+l}$ under Assumptions 7, 10, 11, 14, 19, and 20, and have the following influence function:

$$\psi^r_{i,e,l} = \frac{1}{\pi^1_{e,l} - \pi^2_{e,l} - [\gamma^1_{e,l} - \gamma^2_{e,l}]} \times \left(\frac{\mathbf{1}_{e,e+l} \cdot [\delta_{i,e,l} - E[\delta_{i,e,l}|E_i = e, T_i = e+l]]}{E[\mathbf{1}_{e,e+l}]} - \frac{\mathbf{1}_{e,e-1} \cdot [\delta_{i,e,l} - E[\delta_{i,e,l}|E_i = e, T_i = e-1]]}{E[\mathbf{1}_{e,e-1}]} - \frac{\mathbf{1}_{U,e+l} \cdot [\delta_{i,e,l} - E[\delta_{i,e,l}|E_i \in U, T_i = e+l]]}{E[\mathbf{1}_{U,e+l}]} + \frac{\mathbf{1}_{U,e-1} \cdot [\delta_{i,e,l} - E[\delta_{i,e,l}|E_i \in U, T_i = e-1]]}{E[\mathbf{1}_{U,e-1}]} \right), \tag{26}$$

where $\delta^r_{i,e,l} = Y_i - \theta^r_{e,l} \cdot D_i$.

C.3 Triple DID-IV designs

We also present the influence function of our triple DID-IV estimator when panel or repeated cross section data are available. Let $\theta^p_{e,l,g}$ and $\theta^r_{e,l,g}$ define

$$\theta^p_{e,l,g} \equiv \frac{\alpha_{e,A} - \alpha_{e,B} - [\beta_{U,A} - \beta_{U,B}]}{\pi_{e,A} - \pi_{e,B} - [\gamma_{U,A} - \gamma_{U,B}]},$$
$$\theta^r_{e,l,g} \equiv \frac{[(\alpha_{e,e+l,A} - \alpha_{e,e-1,A}) - (\alpha_{e,e+l,B} - \alpha_{e,e-l,B})] - [(\beta_{U,e+l,A} - \beta_{U,e-1,A}) - (\beta_{U,e+l,B} - \beta_{U,e-1,B})]}{[(\pi_{e,e+l,A} - \pi_{e,e-1,A}) - (\pi_{e,e+l,B} - \pi_{e,e-l,B})] - [(\gamma_{U,e+l,A} - \gamma_{U,e-1,A}) - (\gamma_{U,e+l,B} - \gamma_{U,e-1,B})]},$$

where

$$\alpha_{e,g'} = E[Y_{i,e+l} - Y_{i,e-1}|E_i = e, g_i = g'], \quad \beta_{U,g'} = E[Y_{i,e+l} - Y_{i,e-1}|E_i \in U, g_i = g'],$$
$$\pi_{e,g'} = E[D_{i,e+l} - D_{i,e-1}|E_i = e, g_i = g'], \quad \gamma_{U,g'} = E[D_{i,e+l} - D_{i,e-1}|E_i \in U, g_i = g'] \ \ (g' = A, B),$$

and

$$\alpha_{e,t,g'} = E[Y_i|E_i = e, T_i = t, g_i = g'], \quad \beta_{U,t,g'} = E[Y_i|E_i \in U, T_i = t, g_i = g'],$$
$$\pi_{e,t,g'} = E[D_i|E_i = e, T_i = t, g_i = g'], \quad \gamma_{U,t,g'} = E[D_i|E_i \in U, T_i = t, g_i = g'] \ \ (g' = A, B, \ \ t = e+l, e-1).$$

In panel data settings, our triple DID-IV estimator is the sample analog of $\theta^p_{e,l,g}$ and has the following influence function:

$$\psi^p_{i,e,l,g} = \frac{1}{DID^p_{e,l,g}}(\zeta_{i,e,A} - \zeta_{i,e,B} - \zeta_{i,U,A} + \zeta_{i,U,B}),$$

where $DID^p_{e,l,g} = \pi_{e,A} - \pi_{e,B} - [\gamma_{U,A} - \gamma_{U,B}]$, $\delta^p_{i,e,l,g} = (Y_{i,e+l} - Y_{i,e-1}) - \theta^p_{e,l,g} \cdot (D_{i,e+l} - D_{i,e-1})$, and

$$\zeta_{i,e,g'} = \frac{\mathbf{1}_{e,g'} \cdot \left[\delta^p_{i,e,l,g} - E[\delta^p_{i,e,l,g}|E_i = e, g_i = g']\right]}{E[\mathbf{1}_{e,g'}]}, \quad \zeta_{i,U,g'} = \frac{\mathbf{1}_{U,g'} \cdot \left[\delta^p_{i,e,l,g} - E[\delta^p_{i,e,l,g}|E_i \in U, g_i = g']\right]}{E[\mathbf{1}_{U,g'}]}.$$

In repeated cross section data settings, our triple DID-IV estimator is the sample analog of $\theta^r_{e,l,g}$ and has the following influence function:

$$\psi^r_{i,e,l,g} = \frac{1}{DID^r_{e,l,g}}([\zeta_{i,e,e+l,A} - \zeta_{i,e,e-1,A} - \zeta_{i,e,e+l,B} + \zeta_{i,e,e-1,B}] - [\zeta_{i,U,e+l,A} - \zeta_{i,U,e-1,A} - \zeta_{i,U,e+l,B} + \zeta_{i,U,e-1,B}]),$$

where $DID^r_{e,l,g} = [(\pi_{e,e+l,A} - \pi_{e,e-1,A}) - (\pi_{e,e+l,B} - \pi_{e,e-l,B})] - [(\gamma_{U,e+l,A} - \gamma_{U,e-1,A}) - (\gamma_{U,e+l,B} - \gamma_{U,e-1,B})]$, $\delta^r_{i,e,l,g} = Y_i - \theta^r_{e,l,g} \cdot D_i$, and

$$\zeta_{i,e,t,g'} = \frac{\mathbf{1}_{e,t,g'} \cdot \left[\delta^r_{i,e,l,g} - E[\delta^r_{i,e,l,g} | E_i = e, T_i = t, g_i = g']\right]}{E[\mathbf{1}_{e,t,g'}]},$$
$$\zeta_{i,U,t,g'} = \frac{\mathbf{1}_{U,t,g'} \cdot \left[\delta^r_{i,e,l,g} - E[\delta^r_{i,e,l,g} | E_i \in U, T_i = t, g_i = g']\right]}{E[\mathbf{1}_{U,t,g'}]}.$$

D Decomposing the TWFEIV estimand in Oreopoulos (2006)

In this appendix we quantify the bias terms in the TWFEIV estimand, arising from the bad comparisons in Oreopoulos (2006). Hereafter, let $e = 1947$ and $U = 1957$ denote England and North Ireland respectively, and let $N_{e,t}$ denotes the sample size for England in cohort t. Let $R_{e,t}$ denote the sample mean of random variable $R_{i,t}$ in region e ($e = 1947, 1957$) and cohort t:

$$R_{e,t} \equiv \frac{1}{N_{e,t}} \sum_{i}^{N_{e,t}} R_{e(i),t}.$$

First, we decompose the TWFEIV estimator $\hat{\beta}_{IV}$ as follows[10]:

$$\begin{aligned}\hat{\beta}_{IV} &= \frac{\sum_t N_{e,t} \hat{Z}_{e,t} \left[Y_{e,t} - Y_{e,1946} - (Y_{U,t} - Y_{U,1946})\right]}{\sum_t N_{e,t} \hat{Z}_{e,t} \left[D_{e,t} - D_{e,1946} - (D_{U,t} - D_{U,1946})\right]} \\ &= \frac{\sum_t N_{e,t} \hat{Z}_{e,t} \left[D_{e,t} - D_{e,1946} - (D_{U,t} - D_{U,1946})\right] \cdot \widehat{WDID}_{e,t}}{\sum_t N_{e,t} \hat{Z}_{e,t} \left[D_{e,t} - D_{e,1946} - (D_{U,t} - D_{U,1946})\right]},\end{aligned}$$

where we define:

$$\widehat{WDID}_{e,t} \equiv \frac{\left[Y_{e,t} - Y_{e,1946} - (Y_{U,t} - Y_{U,1946})\right]}{\left[D_{e,t} - D_{e,1946} - (D_{U,t} - D_{U,1946})\right]},$$

and $\hat{Z}_{e,t}$ is the residuals from regressing $Z_{i,t}$ on the cohort and North Ireland fixed effects[11].

By the similar arguments in the proof of Theorem 4, one can show that in staggered DID-IV designs, we have

$$\widehat{WDID}_{e,t} \xrightarrow{p} \begin{cases} 0 & (t < e) \\ CLATT_{e,t} & (U - 1 \geq t \geq e) \\ \dfrac{CAET^1_{e,t} \cdot CLATT_{e,t} - CAET^1_{U,t} \cdot CLATT_{U,t}}{CAET^1_{e,t} - CAET^1_{U,t}} & (t \geq U) \end{cases} \tag{27}$$

where $CAET^1_{e,t} = E[D_i - D^\infty_i | T_i = t, E_i = e]$. Similarly, we have

$$\left[D_{e,t} - D_{e,1946} - (D_{U,t} - D_{U,1946})\right] \xrightarrow{p} \begin{cases} 0 & (t < e) \\ CAET^1_{e,t} & (U - 1 \geq t \geq e) \\ CAET^1_{e,t} - CAET^1_{U,t} & (t \geq U) \end{cases} \tag{28}$$

[10] For the detailed calculation steps, see the proof of Lemma 7 in Miyaji (2024). In our decomposition, the reference period ($t = 1$) in the proof is $t = 1946$.

[11] We can write $\hat{Z}_{i,t} = \hat{Z}_{e,t}$ because $Z_{i,t}$ only varies across cohorts and regions.

Combining the result (27) with (28) and by the Slutsky's theorem, we obtain the following decomposition result for the TWFEIV estimand in Oreopoulos (2006):

$$\beta_{IV} = \sum_{U-1 \geq t \geq e} w^1_{e,t} \cdot CLATT_{e,t} + \sum_{t \geq U} w^2_{e,t} \cdot \Delta_{e,t}, \tag{29}$$

where $\Delta_{e,t}$ is:

$$\frac{CAET^1_{e,t} \cdot CLATT_{e,t} - CAET^1_{U,t} \cdot CLATT_{U,t}}{CAET^1_{e,t} - CAET^1_{U,t}},$$

and the weights $w^1_{e,t}$ and $w^2_{e,t}$ are:

$$w^1_{e,t} = \frac{E[\hat{Z}_{i,t}|E_i = e] \cdot n_{e,t} \cdot CAET^1_{e,t}}{\sum_{U-1 \geq t \geq e} E[\hat{Z}_{i,t}|E_i = e] \cdot n_{e,t} \cdot CAET^1_{e,t} + \sum_{t \geq U} E[\hat{Z}_{i,t}|E_i = e] \cdot n_{e,t} \cdot (CAET^1_{e,t} - CAET^1_{U,t})}, \tag{30}$$

$$w^2_{e,t} = \frac{E[\hat{Z}_{i,t}|E_i = e] \cdot n_{e,t} \cdot (CAET^1_{e,t} - CATT^1_{U,t})}{\sum_{U-1 \geq t \geq e} E[\hat{Z}_{i,t}|E_i = e] \cdot n_{e,t} \cdot CAET^1_{e,t} + \sum_{t \geq U} E[\hat{Z}_{i,t}|E_i = e] \cdot n_{e,t} \cdot (CAET^1_{e,t} - CAET^1_{U,t})}, \tag{31}$$

where $n_{e,t}$ and $E[\hat{Z}_{i,t}|E_i = e]$ represent the population share and population residuals for England in cohort t respectively.

In Oreopoulos (2006), we can only identify $CLATT_{e,t}$ between 1947 and 1956 because the policy change occurred in England in 1947 and in North Ireland in 1957. This implies that each $\Delta_{e,t}$ in equation (29) is the bias term arising from the bad comparisons between 1957 and 1965 performed by the TWFEIV regression.

Figure 4 plots the weight and the corresponding estimate for each $CLATT_{e,t}$ and $\Delta_{e,t}$ (bias term). The consistent estimators for each $w^1_{e,t}$, $w^2_{e,t}$, $CLATT_{e,t}$ and $\Delta_{e,t}$ are constructed from the sample analogue, using equations (27) - (28).

Figure 4 shows that all the CLATTs are positive and positively weighted, whereas all the bias terms are positive and negatively weighted. This indicates that the TWFEIV estimand is negatively biased in the setting of Oreopoulos (2006).

E Comparing DID-IV to Fuzzy DID

In this appendix we compare DID-IV and Fuzzy DID proposed by de Chaisemartin and D'Haultfœuille (2018) (henceforth, "dCDH"). We assume that the reader has briefly read dCDH. We first explain how dCDH's identifying assumptions are related to ours in this article. We then point out the issues inherent in Fuzzy DID designs. First, we show that dCDH's assumptions *ex ante* impose strong and asymmetric restrictions on treatment adoption behavior across units between exposed and unexposed groups. Next, under these restrictions, we show that dCDH's target parameter, the SLATET, can be decomposed into a weighted average of two different causal parameters. At the end of this appendix, we also revisit the "issue" pointed out by dCDH regarding the use of the Wald-DID estimand, and show that while their argument is correct, it is misguided.

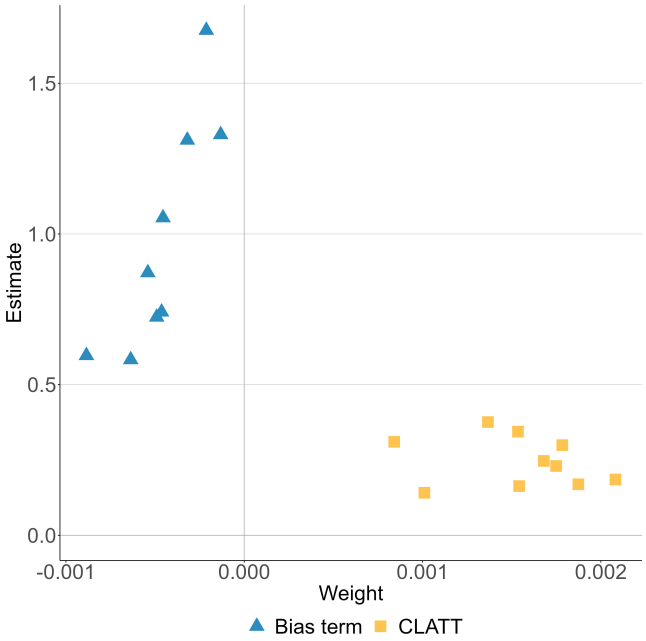

Fig. 4. Decomposition result for the TWFEIV estimand in Oreopoulos (2006). *Notes*: The figure plots the estimated weights and the corresponding estimates for each $CLATT_{e,t}$ and $\Delta_{e,t}$ (bias term). The closed squares are the weights and the corresponding estimates for each $CLATT_{e,t}$. The closed triangles are the weights and the corresponding estimates for each $\Delta_{e,t}$ (bias term).

E.1 Comparing dCDH's identifying assumptions to ours

In this section, we investigate the detailed connections between dCDH's identifying assumptions and ours. Because dCDH consider the repeated cross section settings, when we compare the identifying assumptions between the two papers, we will refer to Assumptions 14-18 in this article. For now, we ignore the difference in the definition of target parameter between dCDH and this article.

dCDH consider the following identification assumptions for the Wald-DID estimand to capture their target parameter, the SLATET[12]. We note that in dCDH, a group variable G plays the role of instrument Z in our notation[13]. In the following, we therefore replace G with Z and call the groups $Z = 1$ and $Z = 0$ exposed and unexposed groups, respectively.[14] For the detail explanations for each assumption, see dCDH.

Assumption 24 (Fuzzy design)**.**

$$E[D|Z = 1, T = 1] > E[D|Z = 1, T = 0], \tag{32}$$

$$E[D|Z = 1, T = 1] - E[D|Z = 1, T = 0] > E[D|Z = 0, T = 1] - E[D|Z = 0, T = 0]. \tag{33}$$

Assumption 25 (Stable percentage of treated units in an unexposed group)**.**

$$0 < E[D|Z = 0, T = 1] = E[D|Z = 0, T = 0] < 1.$$

[12]In repeated cross section settings, the formal definition is provided in Assumption 29.

[13]This follows from two observations in dCDH. First, dCDH call the groups $G = 1$ and $G = 0$ treatment and control groups, respectively. Second, the group variable G is included in Assumption 26 (Treatment participation equation).

[14]Note that in Appendix B.2 in this article, we construct the group variable E from the instrument path Z in order to uncover Assumption 16 (No anticipation in the first stage), which is the hidden assumption in the previous literature.

Assumption 26 (Treatment participation equation)**.**

$$D = \mathbf{1}\{V \geq v_{ZT}\}, \text{ with } V \perp\!\!\!\perp T|Z.$$

Assumption 27 (Parallel Trends assumption in the untreated outcome)**.**

$$E[Y(0)|Z=1,T=1] - E[Y(0)|Z=1,T=0] = E[Y(0)|Z=0,T=1] - E[Y(0)|Z=0,T=0].$$

Assumption 28 (Stable treatment effect over time)**.**

$$E[Y(1)-Y(0)|Z,T=1,D_0(Z)=1] = E[Y(1)-Y(0)|Z,T=0,D_0(Z)=1].$$

In Assumption 28, we modify the dCDH's notation $D(t)$ in order to ease its interpretation: we define $D_t(Z) = \mathbf{1}\{V \geq v_{Zt}\}$ to be the treatment status under time $T=t$ in group Z instead of defining it as $D(t)$.

We now state the relationships between dCDH's assumptions and ours. We first explain the similarities. First, both dCDH and this article impose the condition (33) in Assumption 24 (Fuzzy design), as Assumptions 15-17 in this article imply this condition[15].

Next, the treatment participation equation in Assumption 26 corresponds to Assumption 15 (Monotonicity) in this article. The equivalence between the treatment participation equation and the potential treatment choices framework with Assumption 15 (Monotonicity) follows from the result of Vytlacil (2002). Here, we point out that dCDH implicitly assume the sharp assignment of the instrument Z in their proofs: they assume $v_{10} = v_{00}$ in Assumption 26, that is, they assume that the thresholds in the treatment participation equation are the same between exposed and unexposed groups in period 0.

Finally, dCDH implicitly impose Assumption 14 (Exclusion Restriction) and Assumption 16 (No anticipation in the first stage) in this article when they formalize potential outcomes and treatment participation equations. Specifically, dCDH do not introduce the instrument path in their framework and define potential outcomes as $Y(D)$, which implies Assumption 14. As one can see in Assumption 26, dCDH assume that the treatment participation equation only depends on the current instrument status, which implies Assumption 16 since dCDH assume the sharp assignment of the instrument: $v_{10} = v_{00}$.

We next describe the differences between dCDH's assumptions and ours. First, dCDH assume the parallel trends assumption in the untreated outcome (Assumption 27), whereas we assume the parallel trends assumption in the unexposed outcome. Second, dCDH assume the stable treatment rate assumption in an unexposed group (Assumption 25), whereas we assume the parallel trends assumption in the treatment between exposed and unexposed groups. Third, dCDH impose the independence assumption between unobserved heterogeneity V and time variable T given Z in Assumption 26, whereas we do not impose such restriction. Forth, dCDH argue that the Wald-DID estimand requires the stable treatment effect assumption during the two periods (Assumption 28) to capture the SLATET, whereas we do not require such restriction in order for this estimand to identify the LATET. Finally, dCDH assume the condition (32) in Assumption 24 (Fuzzy design), which requires that the treatment rate in an exposed group increases between the two periods, whereas we do not assume this condition.

[15]From Assumptions 15-17, we have:

$$\begin{aligned}
&E[D|E=1,T=1] - [D|E=1,T=0] - (E[D|E=0,T=1] - E[D|E=0,T=0]) \\
=&E[D_1((0,1)) - D_1((0,0))|E=1,T=1] \\
=&Pr(CM^Z|E=1,T=1) > 0.
\end{aligned}$$

Remark E.1.1. When Assumption 25 is violated, dCDH consider the following assumption in order for the Wald-DID estimand to capture the SLATET.

Assumption 29 (Homogeneous treatment effect between exposed and unexposed groups)**.**

$$SLATET = SLATET',$$

where we define[16]:

$$SLATET = E[Y(1) - Y(0)|Z = 1, T = 1, D_1(1) > D_0(0)],$$
$$SLATET' = E[Y(1) - Y(0)|Z = 0, T = 1, D_1(0) \neq D_0(0)].$$

This assumption requires that the treatment effects among the switchers should be the same between exposed and unexposed groups, which dCDH call the "homogeneous" treatment effect assumption. In Appendix E.4, we present the decomposition result for the SLATET, and show that we cannot straightforwardly interpret this assumption as requiring the "homogeneous" treatment effect between the two groups. In the following discussion, we therefore treat Assumptions 24-28 as dCDH's main identifying assumptions.

E.2 Strong and asymmetric restrictions in Fuzzy DID designs

In this section, we show that dCDH's identifying assumptions implicitly impose strong and asymmetric restrictions on treatment adoption behavior across units between exposed and unexposed groups. First, we prove some lemmas implied by dCDH's assumptions. Based on these lemmas, we then create the tables which present all the types in exposed and unexposed groups under Fuzzy DID designs. These tables show that dCDH's assumptions *ex-ante* exclude many types in both groups and these restrictions are heavily assigned on an unexposed group. At the end of this section, we provide the empirical implication of the distinction between dCDH's restrictions and ours in DID-IV designs.

First, we show that Assumption 26 (Treatment Participation equation) implies the following assumption.

Assumption 30 (Monotone change in treatment in exposed and unexposed groups)**.** For all $z \in \{0, 1\}$,

$$Pr(D_0(z) \geq D_1(z)|Z = z) = 1 \text{ or } Pr(D_0(z) \leq D_1(z)|Z = z) = 1$$

holds.

Here, recall that we define $D_t(Z) = \mathbf{1}\{V \geq v_{Zt}\}$ to be the treatment status under $T = t$ in group Z.

Lemma 1. Assumption 26 implies Assumption 30.

Proof. Fix $z \in \{0, 1\}$. Then we have $v_{z1} \leq v_{z0}$ or $v_{z1} \geq v_{z0}$.

Therefore, conditioning $Z = z$, we have

$$D_1(z) = \mathbf{1}\{V \geq v_{z1}\} \leq D_0(z) = \mathbf{1}\{V \geq v_{z0}\} \text{ a.s. or } D_1(z) = \mathbf{1}\{V \geq v_{z1}\} \geq D_0(z) = \mathbf{1}\{V \geq v_{z0}\} \text{ a.s.}$$

because we assume that the unobserved heterogeneity V is independent of time variable T conditional on the instrument Z.

Completing the proof. □

[16] dCDH denote SLATET and SLATET' as Δ and Δ' respectively. When we define the SLATET, we use $D_0(0) = D_0(1)$, which follows from the sharp assignment of the instrument: $v_{10} = v_{00}$.

Lemma 1 shows that in Fuzzy DID designs, the treatment status changes between period 0 and period 1 in a monotone (uniform) way in exposed and unexposed groups, respectively. We note that the treatment participation equation itself does not imply Assumption 30: the independence between the unobserved heterogeneity V and time variable T given Z yields this assumption. Indeed, in this article, we do not assume this restriction.

As we have seen in section 2.4 in this article, time also plays the role of instrument in DID-IV settings. From this observation, we can view that in an unexposed group ($z = 0$), Assumption 30 corresponds to a monotonicity assumption with respect to time, formally defined below.

Assumption 31 (Monotonicity Assumption w.r.t. time in an unexposed group)**.**

$$Pr(D_0(0) \leq D_1(0)|Z = 0) \text{ or } Pr(D_0(0) \geq D_1(0)|Z = 0).$$

Assumption 31 requires that time affects the treatment adoption process from period 0 to period 1 in a monotone (uniform) way in an unexposed group $z = 0$. In terms of the notation introduced in section 2.4 in this article, this assumption implies that the group variable G^T can take three values with non-zero probability. Hereafter, we consider the type of Assumption 31 which rules out the existence of the time-defiers DF^T in an unexposed group.

Next, we show that dCDH's assumptions imply that there are no time compliers CM^T and time defiers DF^T in an unexposed group ($z = 0$).

Lemma 2. Suppose Assumption 25, Assumption 26 and Assumption 31 hold. Then, there are no time compliers CM^T and time defiers DF^T in an unexposed group ($z = 0$).

Proof. Without loss of generality, suppose that Assumption 31 excludes the time defiers DF^T in an unexposed group ($z = 0$).

Then, Assumption 25 implies:

$$\begin{aligned} & E[D|T = 1, Z = 0] = E[D|T = 0, Z = 0] \\ \iff & Pr(D_1(0) > D_0(0)|Z = 0) = 0 \\ \iff & Pr(CM^T|Z = 0) = 0 \end{aligned} \tag{34}$$

The first equivalence follows from $V \perp\!\!\!\perp T|Z$ in Assumption 26. Equation (34) implies that there are no time comliers CM^T in an unexposed group ($z = 0$). □

Finally, we show that the condition (32) in Assumption 24 (Fuzzy design) and Assumption 26 imply that in an exposed group ($z = 1$), there are no units switching from treatment to non treatment between period 0 and period 1. In other words, in Assumption 30, we should have:

$$Pr(D_0(1) \leq D_1(1)|Z = 1) = 1.$$

Lemma 3. The condition (32) in Assumption 24 and Assumption 26 imply $Pr(D_0(1) \leq D_1(1)|Z = 1)$=1.

Proof. From the condition (32) in Assumption 24, we have

$$\begin{aligned} & E[D|Z = 1, T = 1] > E[D|Z = 1, T = 0] \\ \iff & E[D_1(1) - D_0(1)|Z = 1] > 0 \\ \iff & Pr(D_1(1) > D_0(1)|Z = 1) - Pr(D_0(1) > D_1(1)|Z = 1) > 0, \end{aligned} \tag{35}$$

where the first equivalence follows from $V \perp\!\!\!\perp T|Z$ in Assumption 26. Recall that Assumption 26 implies Assumption 30 from Lemma 1. Suppose that $Pr(D_0(1) \geq D_1(1)|Z = 1) = 1$ holds. Then, we have $Pr(D_1(1) > D_0(1)|Z = 1) = 0$. This contradicts with the condition (35). □

Table 9. Exposed group ($z = 1$)

observed	counterfactual	
$D_0(0)$ or $D_1(1)$	$D_1(0) = 1$	$D_1(0) = 0$
$D_0(0) = 1, D_1(1) = 1$	$AT^Z \wedge AT^T$	$CM^Z \wedge DF^T$
$D_0(0) = 1, D_1(1) = 0$	$DF^Z \wedge AT^T$	$NT^Z \wedge DF^T$
$D_0(0) = 0, D_1(1) = 1$	$AT^Z \wedge CM^T$	$CM^Z \wedge NT^T$
$D_0(0) = 0, D_1(1) = 0$	$DF^Z \wedge CM^T$	$NT^Z \wedge NT^T$

Table 10. Unexposed group ($z = 0$)

observed	counterfactual	
$D_0(0)$ or $D_1(0)$	$D_1(1) = 1$	$D_1(1) = 0$
$D_0(0) = 1, D_1(0) = 1$	$AT^Z \wedge AT^T$	$DF^Z \wedge AT^T$
$D_0(0) = 1, D_1(0) = 0$	$CM^Z \wedge DF^T$	$NT^Z \wedge DF^T$
$D_0(0) = 0, D_1(0) = 1$	$AT^Z \wedge CM^T$	$DF^Z \wedge CM^T$
$D_0(0) = 0, D_1(0) = 0$	$CM^Z \wedge NT^T$	$NT^Z \wedge NT^T$

We now clarify how dCDH's identifying assumptions restrict the treatment adoption behavior across units in exposed and unexposed groups, respectively. Tables 9 - 10 show that their identifying assumptions exclude many types in both groups and the restrictions are heavily imposed on an unexposed group[17]: there exist five and three types in exposed and unexposed groups respectively, where the types painted in gray color are excluded in both groups.

The distinction between dCDH's restrictions and ours is apparent when we compare Tables 9 - 10 to Table 1 - 2 in this article[18]: in both groups, we only exclude the defiers DF^Z by Assumption 15 (Monotonicity), whereas dCDH additionally exclude the type $NT^Z \wedge DF^T$ in an exposed group by Lemma 3, and additionally exclude the time defiers DF^T and the time compliers CM^T in an unexposed group by Lemma 2.

These differences between dCDH's restrictions and ours seem critical in many empirical settings: unlike the monotonicity assumption, which eliminate the compliers CM^Z in both groups, dCDH's additional restrictions seem difficult to assess from the institutional knowledge in a given application, thus the practitioners may want to avoid imposing these restrictions in empirical work. For instance, in some cases, we may have no valid reason for excluding the specific type $NT^Z \wedge DF^T$ in an exposed group, who are not affected by instrument but affected by time. Similarly, in other cases, we may have no justification to exclude the time defiers DF^T and the time compliers CM^T in an unexposed group because of the steady growth of the treatment rate.

Although these restrictions in dCDH seem strong, one might argue that each assumption, such as the condition (32) in Assumption 24 and Assumption 25 (Stable percentage of treated units in an unexposed group), can be justifiable in certain applications. However, these assumptions give rise to two additional issues. First, when we impose the condition (32) in Assumption

[17] In both tables, we can only observe either $D_0(z)$ or $D_1(z)$ in group z as we consider repeated cross section settings.

[18] We note that Tables 1-2 in this article is constructed under panel data settings, but the types represented in these tables remain consistent even when considering repeated cross section settings. The only modification is that as in Tables 9-10, one can only observe either $D_0(z)$ or $D_1(z)$ for unit i in group z in each table.

24, we should test the plausibility of this condition from the data in practice because this condition excludes the case that the treatment rate in an exposed group is stable during the two periods: $E[D|Z=1,T=1]=E[D|Z=1,T=0]$. Although dCDH argue that "Assumption 1 (Assumption 24 in this article) is just a way to define the treatment and the control group (exposed and unexposed group in this article) in our fuzzy setting", they do not explain how we should proceed when this condition is violated in practice.

Second, when we impose Assumption 25, we should find or estimate the groups whose treatment rates are stable over time. Although dCDH propose the two-step estimation method when we should estimate the stable treatment group from the data, its performance and feasibility is not clear, especially when we have multiple periods and multiple groups.

E.3 Decomposing the SLATET

In the previous section, we investigate how dCDH's identifying assumptions *ex-ante* restrict the treatment adoption behavior across units between exposed and unexposed groups. In this section, under these restrictions, we show that dCDH's target parameter, the SLATET, can be decomposed into a weighted average of two different causal parameters.

The following theorem presents the decomposition result for the SLATET under Fuzzy DID designs, and clarifies the difference between dCDH's target parameter (SLATET) and ours (LATET).

Theorem 7. Suppose Assumptions 24 - 28 hold. Then, we can decompose the SLATET into two different causal parameters:

$$\begin{aligned} SLATET &\equiv E[Y(1)-Y(0)|Z=1,T=1,D_1(1)>D_0(0)] \\ &= E[Y(1)-Y(0)|Z=1,T=1,CM^T]Pr(CM^T|Z=1,SW) \\ &+ E[Y(1)-Y(0)|Z=1,T=1,CM^Z \wedge NT^T]Pr(CM^Z \wedge NT^T|Z=1,SW), \quad (36) \end{aligned}$$

where the switcher (SW) in an exposed group ($Z=1$) is the units who become treated in time $T=1$.

Proof. The proof directly follows from Table 9 in an exposed group. □

Theorem 7 shows that we can decompose the SLATET into a weighted average of two different causal parameters under Fuzzy DID designs: one parameter measures the treatment effects among the time compliers CM^T in an exposed group and the other parameter measures the treatment effects among the type $(CM^Z \wedge NT^T)$ in an exposed group.

This theorem has two concerning implications. First, the interpretation of the SLATET is less clear than that of the LATET because it consists of two different causal parameters. Although the SLATET indeed captures the "local" average treatment effect, it may be difficult to interpret economically in a given application because one of the two causal parameters captures the treatment effects, for those who are induced to treatment by time.

Second, the SLATET may not be policy-relevant even when the researchers exploit the policy change of their interest as an instrument for treatment. Suppose that the instrument represents the important policy lever for researchers, and they are intrinsically interested in the treatment effects among the compliers CM^Z in an exposed group. In this situation, our target parameter, the LATET, is a policy-relevant parameter by construction. On the other hand, the SLATET only captures the treatment effects among the type $(CM^Z \wedge NT^T)$, which is the sub-population of the compliers CM^Z, even when the proportion of the time compliers CM^T is zero within that group. As the fraction of the time compliers CM^T increases in that

Table 11. Unexposed group ($z = 0$)

observed	counterfactual	
$D_0(0)$ or $D_1(0)$	$D_1(1) = 1$	$D_1(1) = 0$
$D_0(0) = 1, D_1(0) = 1$	$AT^Z \wedge AT^T$	$DF^Z \wedge AT^T$
$D_0(0) = 1, D_1(0) = 0$	$CM^Z \wedge DF^T$	$NT^Z \wedge DF^T$
$D_0(0) = 0, D_1(0) = 1$	$AT^Z \wedge CM^T$	$DF^Z \wedge CM^T$
$D_0(0) = 0, D_1(0) = 0$	$CM^Z \wedge NT^T$	$NT^Z \wedge NT^T$

group, the SLATET puts more weight on the treatment effects among the time compliers, which contaminates the interpretation of the magnitude of this parameter even in such policy-relevant contexts.

Using the decomposition result for the SLATET, we finally clarify the interpretation of Assumption 29. When Assumption 25 (Stable percentage of treated units in an unexposed group) is violated, dCDH additionally consider Assumption 29 in order for the Wald-DID estimand to identify the SLATET, and call this the "homogeneous" treatment effect assumption between exposed and unexposed groups.

We show that this assumption can not be interpreted as requiring the "homogeneous" treatment effect among the *same* sub-population between exposed and unexposed groups. To do so, we first note that if we do not assume Assumption 25, we have Table 11 in an unexposed group instead of Table 10 in Appendix E.2. Here, we assume that Assumption 26 and 31 exclude the defiers DF^Z and the time defiers DF^T in an unexposed group, respectively.

Then, from Table 11, we can interpret the SLATET' (see Assumption 29) as the treatment effects among the time compliers in an unexposed group in period 1:

$$\begin{aligned} SLATET' &\equiv E[Y(1) - Y(0)|Z = 0, T = 1, D_1(0) \neq D_0(0)] \\ &= E[Y(1) - Y(0)|Z = 0, T = 1, CM^T]. \end{aligned} \tag{37}$$

By combining equations (36) and (37), we can conclude that Assumption 29 ($STATET = SLATET'$) does not require the homogeneous treatment effects in the *same* sub-population between the two groups. Rather, this assumption requires that a weighted average of the treatment effects among the time compliers CM^T and the type ($CM^Z \wedge NT^T$) in an exposed group should be equal to the treatment effect among the time compliers CM^T in an unexposed group. This interpretation implies that the requirement of the "homogeneous" treatment effect in Assumption 29 is less clear than one might expect.

E.4 Revisiting the "issue" regarding the use of the Wald-DID estimand

Finally, we revisit the "issue" raised by dCDH regarding the use of the Wald-DID estimand in Fuzzy DID designs, and show that while their argument is correct, it is misguided. During this discussion, we also clarify why dCDH exclude so many types in exposed and unexposed groups as we see in Appendix E.2. Hereafter, we assume Assumptions 24 - 28 in Fuzzy DID designs.

dCDH argue that the Wald-DID estimand requires Assumption 28 (stable treatment effect over time) in order to identify the SLATET, and propose alternative estimands that do not rely on this assumption. We first show that their argument stems from their reliance on the parallel trends assumption in the untreated outcome, which is common in DID designs, instead of on the parallel trends assumption in the unexposed outcome as in this article. We next argue that the former type of parallel trends assumption is unsuitable for DID-IV settings.

First, we clarify the interpretation of Assumption 28 by introducing the additional notation. This assumption is actually conditional on $D_0(Z) = 1$ in addition to Z and T, but dCDH do not provide the reason why this strong restriction is required only for the sub-population. For each time group g^t, let $w_{g^t,z} = Pr(G^T = g^t|Z = z)$ denote the population share conditional on group $Z = z$, and let $w_{g^t,z,D_0(0)=1} = Pr(G^T = g^t|Z = z, D_0(0) = 1)$ denote the population share conditional on $D_0(0) = 1$ in addition to group $Z = z$. Additionally, let $\delta_{z,t,g^t} = E[Y(1) - Y(0)|Z = z, T = t, G^T = g^t]$ denote the average treatment effect conditional on group $Z = z$, time $T = t$, and time group $G^T = g^t$.

Then, from Table 9 and Table 10, we can decompose Assumption 28 in the following:

$$\begin{aligned} &E[Y(1) - Y(0)|Z = 0, T = 1, D_0(0) = 1] = E[Y(1) - Y(0)|Z = 0, T = 0, D_0(0) = 1] \\ &\qquad \iff \delta_{0,1,AT^T} = \delta_{0,0,AT^T}, \end{aligned} \tag{38}$$

$$\begin{aligned} &E[Y(1) - Y(0)|Z = 1, T = 1, D_0(1) = 1] = E[Y(1) - Y(0)|Z = 1, T = 0, D_0(1) = 1] \\ \iff &\delta_{1,1,AT^T} \cdot w_{AT^T,1,D_0(0)=1} + \delta_{1,1,DF^T} \cdot w_{DF^T,1,D_0(0)=1} \\ &= \delta_{1,0,AT^T} \cdot w_{AT^T,1,D_0(0)=1} + \delta_{1,0,DF^T} \cdot w_{DF^T,1,D_0(0)=1}, \end{aligned} \tag{39}$$

where we use the Law of Iterated Expectation, $D_0(0) = D_0(1)$ (implied by the sharp assignment of the instrument), and $V \perp\!\!\!\perp T|Z$ in Assumption 26. Equation (38) shows that in an unexposed group ($z = 0$), Assumption 28 requires the stable treatment effect in the time always-takers AT^T. Equation (39) shows that in an exposed group ($z = 1$), this assumption requires that a weighted average of the treatment effects among the time defiers DF^T and the time always takers AT^T is stable during the two periods.

Conditions (38) and (39) are required for Fuzzy DID designs because dCDH rely on the parallel trends assumption in the untreated outcomes. In DID-IV settings, we can not observe the average time trends of untreated potential outcomes for the time always takers AT^T in an unexposed group and the time always takers AT^T and the time defiers DF^T in an exposed group because they always adopt the treatment during the two periods (see Table 9 and Table 10). Therefore, if we assume the parallel trends assumption in the untreated outcomes, we should recover them by using conditions (38) and (39).

To see this formally, we use our notation to briefly sketch the proof of dCDH on the Wald-DID estimand identifying SLATET under fuzzy DID designs. First, consider the conditional expectation $E[Y|Z = 0, T = 1] - E[Y|Z = 0, T = 0]$, which appears in the numerator of the Wald-DID estimand. We can decompose this quantity into two terms:

$$\begin{aligned} &E[Y|Z = 0, T = 1] - E[Y|Z = 0, T = 0] \\ &= (\delta_{0,1,AT^T} - \delta_{0,0,AT^T}) \cdot w_{AT^T,0} + \{E[Y(0)|Z = 0, T = 1] - E[Y(0)|Z = 0, T = 0]\}. \end{aligned}$$

The first term arises because we can not observe the average time trends of the untreated potential outcome for the time always-takers AT^T in an unexposed group, and we should add and subtract this quantity to construct the second term. The first term disappears if we assume condition (38).

Similarly, one can decompose the conditional expectation $E[Y|Z = 1, T = 1] - E[Y|Z =$

$1, T = 0]$ as follows:

$$
\begin{aligned}
&E[Y|Z=1,T=1] - E[Y|Z=1,T=0] \\
=&\Big\{\delta_{1,1,AT^T} \cdot w_{AT^T,1,D_0(0)=1} + \delta_{1,1,DF^T} \cdot w_{DF^T,1,D_0(0)=1} \\
&- \delta_{1,0,AT^T} \cdot w_{AT^T,1,D_0(0)=1} - \delta_{1,0,DF^T} \cdot w_{DF^T,1,D_0(0)=1}\Big\} \\
&+\{E[Y(0)|Z=1,T=1] - E[Y(0)|Z=1,T=0]\} \\
&+E[Y(1)-Y(0)|SW,Z=1,T=1]Pr(SW|Z=1),
\end{aligned}
$$

where the type SW (switcher) includes the units who switch from non treatment to treatment during the two periods. The first term arises because we can not observe the average time trends of the untreated potential outcome for the time always-takers AT^T and the time never-takers NT^T in an exposed group, and we should add and subtract this quantity to construct the second term. The first term disappears if we assume condition (39). Here, we note that we do not require the stable treatment assumption for the time compliers CM^T and the type $CM^Z \wedge NT^T$ in an exposed group: by adding and subtracting the expectation of the untreated potential outcome path for these types, we have the third term.

Then, by noticing that the denominator of the Wald-DID estimand identifies $Pr(SW|Z=1)$ under Assumption 25 and Assumption 26, the Wald-DID estimand captures the SLATET under Assumption 27 (parallel trends assumption in the untreated outcome) and conditions (38) and (39). From the discussions so far, one can conclude that the parallel trends assumption in the untreated outcome leads dCDH to impose conditions (38) and (39).

Here, we add few remarks to clarify why dCDH *ex ante* impose strong restrictions on treatment adoption behavior across units in order for the Wald-DID estimand to identify the SLATET. It turns out that this also stems from their reliance on the parallel trends assumption in the untreated outcome.

Remark E.4.1. dCDH exclude the type $NT^Z \wedge DF^T$ in an exposed group and the time defiers DF^T in an unexposed group because if we add and subtract the average time trends of the untreated outcome for these types, we additionally identify the treatment effects among them in period 0 (not in period 1), as they adopt the treatment in period 0, but leave it in period 1 (see Tables 9-10). This identification result seems not intuitive because the identification variation comes from the policy shock (instrument) in period 1.

Remark E.4.2. dCDH exclude the time compliers CM^T in an unexposed group because if we add and subtract the average time trends of the untreated outcome for that type, we additionally capture the treatment effects among the time compliers CM^T from the conditional expectation $E[Y|Z=0,T=1] - E[Y|Z=0,T=0]$, which implies that the SLATET in an exposed group is offset by the SLATET' in an unexposed group.

So far, we have confirmed that dCDH's argument regarding the use of the Wald-DID estimand rests on the parallel trends assumption in the untreated outcome in DID designs. Finally, we argue that this assumption is unsuitable for DID-IV settings for three reasons.

First, this assumption is not sufficient to capture the average time trends of the outcome even in an unexposed group. This is because in DID-IV settings, the units are allowed to take the treatment in the absence of the instrument, and we can not observe the untreated outcomes for these types in an unexposed group.

Second, in most of the DID-IV applications, we can not impute the untreated potential outcomes in general. For instance, Black et al. (2005) estimates the causal link between parents'

and children's education attainment, exploiting variation in the timing of the implementation of the school reform across municipalities and cohorts as an instrument for parents' education attainment. In the author's setting, the parallel trends assumption in the untreated outcome requires the data to contain the parents with zero education attainment, which is unrealistic and the authors would not have in mind in practice.

Finally, this assumption is generally not testable using the pre-exposed period data in a given application. This seems apparent, as some units can already adopt the treatment before period 0, and we can not compare the pretrends of the untreated outcome between exposed and unexposed groups.

dCDH propose the placebo test in their appendix, but we show that it is incomplete for checking the plausibility of the parallel trends assumption in the untreated outcome. dCDH's test consists of two steps. First, we test whether the share of treated units is zero in both groups from $T = -1$ to $T = 0$:

$$E[D|Z = 1, T = 0] - E[D|Z = 1, T = -1] = 0, \tag{40}$$
$$E[D|Z = 0, T = 0] - E[D|Z = 0, T = -1] = 0. \tag{41}$$

Second, we test the following null hypothesis:

$$E[Y|Z = 1, T = 0] - E[Y|Z = 1, T = -1] = E[Y|Z = 0, T = 0] - E[Y|Z = 0, T = -1]. \tag{42}$$

Although dCDH "assume" the conditions (40) and (41), dCDH also note that when one of the two conditions fails, the second test can "no longer be used to test" Assumption 27. We, therefore, describe the dCDH's test in two steps.

dCDH's test has some drawbacks. First, the conditions (40) and (41), which we should check in the first step, are strong, but dCDH do not explain how we should proceed when these conditions are violated in practice. Second, in this test, the second step does not directly test Assumption 27 in the pre-exposed periods, as it does not compare the evolution of the mean untreated outcome between exposed and unexposed group before period $T = 0$. Indeed, dCDH describe the equality (42) as a necessary condition for Assumption 27 and Assumption 28 to be satisfied under the conditions (40) and (41). Finally, it is unclear how we should conduct this test when we have multiple pre-exposed period data. dCDH assume "for instance that data is available for the period $T = -1$", but they do not indicate how we should perform this test in multiple pre-exposed period settings.

Surprisingly, dCDH's test corresponds to our pretrends test in section 5.3 if we replace the conditions (40)- (41) in the first step with the weaker one:

$$E[D|Z = 1, T = 0] - E[D|Z = 1, T = -1] = E[D|Z = 0, T = 0] - E[D|Z = 0, T = -1]. \tag{43}$$

Under repeated cross section settings, one can figure out that the conditions (42) - (43) coincide with the conditions (3) -(4) in this article.