Better Understanding Triple Differences Estimators*

Marcelo Ortiz-Villavicencio † Pedro H. C. Sant'Anna‡

May 15, 2025

Abstract

Triple Differences (DDD) designs are widely used in empirical work to relax parallel trends assumptions in Difference-in-Differences (DiD) settings. This paper shows that common DDD implementations—such as taking the difference between two DiDs or applying three-way fixed effects regressions—are generally invalid when identification requires conditioning on covariates. In staggered adoption settings, the common DiD practice of pooling all not-yet-treated units as a comparison group introduces additional bias, even when covariates are not required for identification. These insights challenge conventional empirical strategies and underscore the need for estimators tailored specifically to DDD structures. We develop regression adjustment, inverse probability weighting, and doubly robust estimators that remain valid under covariate-adjusted DDD parallel trends. For staggered designs, we show how to correctly leverage multiple comparison groups to get more informative inference. Simulations highlight substantial bias reductions and precision gains relative to standard approaches, offering a new framework for credible DDD estimation in empirical research.

JEL: C10; C14; C21; C23.
Keywords: Triple Differences; Difference-in-Differences; Difference-in-Difference-in-Differences; Parallel Trends; Doubly Robustness; Staggered Adoption.

*We thank participants from the 2023 SEA Conference, 2024 Midwest Econometrics Group Conference, 2025 Georgia Econometrics Workshop, and participants of the "Difference-in-Differences" Workshops from Causal Solutions in 2022 and 2023. This paper builds on and supersedes the material used for the 2022 Causal Solutions YouTube Lecture available here.

†Department of Economics, Emory University. Email: marcelo.ortiz@emory.edu
‡Department of Economics, Emory University. Email: pedro.santanna@emory.edu

1 Introduction

Over the last few years, we have seen a big Difference-in-Differences (DiD) "methodological revolution" with multiple DiD estimators being proposed to address the interpretability shortcomings associated with using more traditional two-way fixed-effects specifications in the presence of treatment effect heterogeneity.[1] Although these modern DiD estimators can capture richer notions of heterogeneity, in practice, they rely on parallel trends (PT) assumptions, and an important concern relates to how plausible these PT assumptions are. When such PT assumptions are not accurate approximations of reality, one may doubt the conclusions of DiD studies (Rambachan and Roth, 2023; Chiu, Lan, Liu and Xu, 2025).

In some setups, however, it is possible to naturally relax such DiD-type PT assumptions and retain the simplicity and empirical appeal of DiD-type analysis. This is particularly the case when a unit needs to fulfill two criteria to be treated, e.g., it belongs to (i) a group (e.g., a state) in which the treatment is already enabled, and (ii) a partition of the population that qualifies (or is *eligible*) for treatment (e.g., women). Such setups are often referred to as Triple Differences (DDD), and allow for group-specific and partition-specific violations of parallel trends. Since its introduction by Gruber (1994), DDD has become very popular among empirical researchers—see Olden and Møen (2022) for documentation.[2] Despite its empirical popularity, little attention has been devoted to better understanding DDD setups with covariates, multiple periods, and/or staggered treatment adoption.

This article aims to improve our understanding of Triple Difference (DDD) designs. We study identification, estimation, and inference procedures for DDD when covariates may be important for the reliability of the identification assumptions, multiple periods are available, and treatment adoption is potentially staggered over time. We tackle the DDD problem using causal inference first principles and uncover surprising results that challenge some conventional wisdom and common practices.

For instance, although DDD estimators can be understood as the difference between two DiD estimators in setups with two periods and no covariates (Olden and Møen, 2022), our results highlight that this is no longer the case when covariates are required to justify the plausibility of a DDD-type parallel trends assumption. As we illustrate via simulations, erroneously proceeding as if DDD was just the difference between 2 DiD estimators can lead to severely biased results. Such bias arises because this naive DDD strategy fails to integrate the covariate distribution over the correct reference group—the treated group. We show that it is straightforward to avoid these problems and propose regression adjustment, inverse probability weighting, and doubly robust DDD estimators. Our doubly robust DDD estimator builds on the efficient influence function

1 For an overview and a practitioner's guide, see Roth, Sant'Anna, Bilinski and Poe (2023) and Baker, Callaway, Cunningham, Goodman-Bacon and Sant'Anna (2025).

2 Additional examples of papers using DDD strategy include Walker (2013), Garthwaite, Gross and Notowidigdo (2014), Antwi, Moriya and Simon (2013), and Hansen and Wingender (2023).

for the two-period DDD setup, allowing one to use flexible (potentially machine-learning-based) estimators for the nuisance functions. We also show that if one wants to cast DDD in terms of DiD, one would need three—and not two—DiD terms. Each of these DiD terms compares effectively treated units with a different type of untreated units, e.g., units in a treated state but ineligible for treatment, units that are eligible but are in an untreated state, or ineligible units in an untreated state.

In setups with staggered treatment adoption, our results once again challenge the interpretation of DDD as the difference between two DiDs. In DiD with staggered treatment adoptions, it is now common to pool all not-yet-treated units at a time period and use that aggregate set of units as a valid comparison group.[3] Thus, one may think that a similar strategy should work with DDD. However, our results highlight that this is generally not the case and that pooling all not-yet-treated units and proceeding as in staggered DiD procedures can lead to biased estimators for average treatment effects parameters, even when covariates do not play an important role. This arises because the proportion of units eligible for treatment may change across groups that enable treatment over time. As DDD allows for group-specific and partition-specific violations of DiD-type PT, these differential trends do not average out when pooling all not-yet-treated units, leading to potentially misleading estimates. We propose DDD estimators that bypass this drawback by using any specific not-yet-treated unit as a comparison group (e.g., the set of units in groups that never enabled treatment). As one can potentially use different not-yet-treated cohorts as comparison groups, we also discuss combining these to form more precise estimators. Our proposed DDD estimator that aggregates across different comparison groups can be understood as a two-step Generalized Method of Moments (GMM) procedure based on recentered influence functions. Importantly, our staggered DDD procedures can flexibly accommodate covariates using regression adjustment, inverse probability weighting, or doubly robust methods and can also be used to form event-study estimators that highlight how average treatment effects evolve with elapsed treatment time.

Related literature: This article contributes to the rapidly expanding literature on DiD-related methods. In particular, we contribute to the scarce literature on DDD procedures. Our paper is related to Olden and Møen (2022), though we cover substantially more general DDD setups with (a) multiple periods, (b) staggered treatment adoption, and (c) covariates potentially playing an important role for the plausibility of the identification assumptions. In this sense, our paper can be understood as the DDD analog of Callaway and Sant'Anna (2021). However, and in sharp contrast with Callaway and Sant'Anna (2021), our DDD procedures cannot pool all not-yet-treated units as an aggregated comparison group, highlighting some interesting differences between staggered DiD and DDD designs. Our paper is also related to Strezhnev (2023), who introduced a decomposition of the DDD estimators based on three-way fixed effects specifications, showing

3 See, e.g., Callaway and Sant'Anna (2021); De Chaisemartin and d'Haultfoeuille (2020); Wooldridge (2021); Borusyak, Jaravel and Spiess (2024).

when and why it fails to recover an easy-to-interpret causal parameter of interest when treatment effects are heterogeneous. To some extent, Strezhnev (2023) can be understood as the analog of Goodman-Bacon (2021) to DDD setups, while our paper is more closely related to Callaway and Sant'Anna (2021). As such, our results complement Strezhnev (2023), and as our estimators do not leverage a rigid three-way fixed effect specification, they avoid the issues highlighted in that study. Our paper is also related to Słoczyński (2022, 2024) in the sense that our proposed tools avoid issues related to potentially misleading weights related to model misspecifications.

In this paper, we use the term triple differences to qualify designs under which units need to satisfy two criteria to be (effectively) treated. However, we note that sometimes, different researchers use the term triple differences more broadly, for instance, when they are interested in analyzing treatment effect heterogeneity across subgroups. In such cases, it is important to highlight that the underlying identification assumptions and the parameters of interest would differ from those we study in this paper; see Caron (2025) for a recent example. Those procedures should be understood as a complement to the ones we discuss in this paper, as they can be used to answer different questions of interest.

Organization of the paper: The rest of the paper is organized as follows. In the next section, we present our framework. In Section 3, we challenge some standard empirical practices for DDD analyses in terms of interpreting it as a simple extension of DiD analysis, and we also highlight some important practical takeaway messages from our main results. Section 4 introduces our formal identification, estimation, and inference results. Section 5 presents a Monte Carlo study to demonstrate the finite sample properties of our estimator with conclusions drawn in Section 6. Detailed mathematical proofs and additional results can be found in the Supplemental Appendix.

2 Framework

We start our analysis by discussing the specifics of our DDD research design, including potential outcomes, parameters of interest, and identification assumptions. We consider a setup with T time periods, $t = 1, 2, \ldots, T$. Units are indexed by i, with $i = 1, 2, \ldots, n$. We focus on setups where n is much larger than T, as our inference procedures are asymptotically justified using the "fixed-T, large-n" panel data framework.

Each unit may be exposed to a binary treatment in any time period $t > 1$. Treatment is an absorbing state such that once a unit is treated, it remains treated for the remainder of the panel. Each unit i belongs to a group (e.g., a state or a country) that enables treatment for the first time in period $g > 1$. Let $S_i \in \mathcal{S} \subseteq \{2, ..., T\} \cup \{\infty\}$ be a variable that indicates the first time the policy/treatment was enabled, with the notion that $S = \infty$ if the policy is not enabled by $t = T$. In addition, each unit belongs to a population partition that qualifies (or is eligible) for the treatment or not (e.g., being a woman, or an indicator for specific crops). We denote this variable by Q_i with $Q_i = 1$ if unit i (eventually) qualifies for the treatment and $Q_i = 0$ otherwise. For simplicity, we

assume that this population partition that eventually qualifies for treatment is time-invariant.

In our DDD setup, a unit i is treated in period t if $t \geqslant S_i$ and $Q_i = 1$, i.e., if it belongs to a group that has already enabled treatment by period t (i.e., $t \geqslant S_i$) and it qualifies for treatment (i.e., $Q_i = 1$). With this notation that makes it clear that a unit i is treated if it satisfies two criteria, let $D_{i,t} = 1\{t \geqslant S_i, Q_i = 1\}$ be an indicator for whether unit i receives treatment in period t, and let $G_i = \min\{t : D_{i,t} = 1\}$ be the earliest period at which unit i has received treatment. If i is never treated during the sample, then $G_i = \infty$. Here, we have that $G_i = S_i$ if $Q_i = 1$ and $G_i = \infty$ if $Q_i = 0$.[4] Let $\mathcal{G}$ denote the support of G_i and $\mathcal{G}_{\text{trt}} = \mathcal{G}\backslash\{\infty\}$. We assume that a group of "never-enabled" units always exists, i.e., $S_i = \infty$ for some units. In an application where all units belong to a group that eventually enables treatment, we remove all that data from all units from the time the last cohort enabled treatment onwards, i.e., we drop all observations from periods $t > \max S_i$, and retain the remaining data as the "effective" data to be used in our analysis, where the last-to-be-eligible group becomes the "never-eligible" group.[5] Finally, we also assume that a vector of pre-treatment covariates X_i, whose support is denoted by $\mathcal{X} \subseteq \mathbb{R}^d$ is available.

Regarding potential outcomes, we adopt the potential outcome framework of Robins (1986) with potential outcomes indexed by treatment sequences. Let $\mathbf{0}_s$ and $\mathbf{1}_s$ be s-dimensional vectors of zeros and ones, respectively, and denote the potential outcome for unit i at time t if unit i is first treated at time g by $Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$, and denote by $Y_{i,t}(\mathbf{0}_T)$ the outcome if untreated by time $t = T$. As we focus our attention on staggered treatment adoptions, we can simplify notation and index potential outcomes by the time treatment begins, g: $Y_{i,t}(g) = Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ and use $Y_{i,t}(\infty) = Y_{i,t}(\mathbf{0}_T)$ to denote never-treated potential outcomes. In practice, though, we observe,

$$Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g), \tag{2.1}$$

where $1\{A\}$ represents the indicator function, which equals one if A is true and zero otherwise. Additionally, we assume the observation of a random sample of $(Y_{t=1}, \ldots, Y_{t=T}, X', G, S, Q)'$.

Assumption S (Random Sampling)**.** $\{(Y_{i,t=1}, \ldots, Y_{i,t=T}, X_i', G_i, S_i, Q_i)'\}_{i=1}^n$ is a random sample from $(Y_{t=1}, \ldots, Y_{t=T}, X', G, S, Q)'$.

2.1 Parameters of interest

In this paper, we are interested in better understanding how average treatment effects vary across periods and different groups defined by treatment starting period. More specifically, we want to make inferences on functionals of the group-time average treatment effects, $ATT(g,t)$'s, defined

4 Note that when all units are eligible for treatment, we have $G_i = S_i$, getting us back to a (staggered) DiD setup; see, e.g., Callaway and Sant'Anna (2021).

5 If needed, we also update the notion of support of all variables to reflect this change. In particular, T here denotes the number of available periods in the subset of the data that we will use in our analysis.

as

$$ATT(g,t) \equiv \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty)|G_i = g] = \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty)|S_i = g, Q_i = 1], \tag{2.2}$$

By exploring that in our context $G_i = g$ if and only if $S_i = g$ and $Q_i = 1$, we have that $ATT(g,t) = \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty)|S_i = g, Q_i = 1]$. Note that $ATT(g,t)$ captures how the average treatment effects evolve over time for each treatment group g (Callaway and Sant'Anna, 2021). As such, one can use $ATT(g,t)$ to construct group-g-specific event studies by analyzing how average treatment effects vary with elapsed treatment time $e = t - g$.

In some setups with multiple groups g, researchers may want to summarize over the many $ATT(g,t)$'s into a more aggregate parameter. A natural summary parameter that still allows one to understand treatment effect dynamics with respect to elapsed treatment time is the aggregated event study parameter $ES(e)$, defined as

$$ES(e) \equiv \mathbb{E}\big[ATT\left(G, G+e\right)\big|G + e \in [2,T]\big] = \sum_{g \in \mathcal{G}_{\text{trt}}} \mathbb{P}(G = g|G + e \in [2,T])ATT(g, g+e). \tag{2.3}$$

One may also want to aggregate the event study coefficients further to recover a scalar summary measure. Let $\mathcal{E}$ denote the support of post-treatment event time $E = t - G$, $t \geqslant G$, and let N_E denote its cardinality. Then,

$$ES_{\text{avg}} \equiv \frac{1}{N_E} \sum_{e \in \mathcal{E}} ES(e), \tag{2.4}$$

provides a simple average of all post-treatment event study coefficients. Many other summary parameters are possible; see Callaway and Sant'Anna (2021) for discussions of several alternatives.

2.2 Identification assumptions

To identify the $ATT(g,t)$'s and their functionals $ES(e)$ and ES_{avg}, we impose the following assumptions.

Assumption SO (Strong Overlap)**.** For every $(g,q) \in \mathcal{S} \times \{0,1\}$ and for some $\epsilon > 0$, $\mathbb{P}[S = g, Q = q|X] > \epsilon$ with probability one.

Assumption SO is an overlap condition that ensures that for any value of X, there are units with any combination $(g,q) \in \mathcal{S}$ that have comparable X values. Heuristically, this condition guarantees that we cannot perfectly predict which (g,q)-partition a unit belongs to using information from X. This assumption also rules out irregular identification (Khan and Tamer, 2010).[6]

We also impose the following no-anticipation assumptions.

Assumption NA (No-Anticipation)**.** For every $g \in \mathcal{G}_{\text{trt}}$, and every pre-treatment period $t < g$, $\mathbb{E}[Y_{i,t}(g)|S = g, Q = 1, X] = \mathbb{E}[Y_{i,t}(\infty)|S = g, Q = 1, X]$ with probability one.

6 As our focus is on $ATT(g,t)$-type parameters, it is possible to relax Assumption SO to hold only over X in the support of the covariates among the (eventually) treated units. To simplify the discussion, we abstract from these subtle points.

Assumption NA rules out anticipatory effects among treated units as in, e.g., Abbring and van den Berg (2003), Callaway and Sant'Anna (2021), and Sun and Abraham (2021). This assumption is important as it allows us to consider observations in pre-treatment periods $t < g$ as effectively untreated. If units are expected to anticipate some treatments—for example, if treatment is announced in advance—it is important to adjust the definition of the treatment date to account for it; see Malani and Reif (2015) for a discussion.

Next, we impose our final identification assumption that restricts the evolution of average untreated potential outcomes across groups.

Assumption DDD-CPT (DDD-Conditional Parallel Trends)**.** For each $g \in \mathcal{G}_{\text{trt}}$, $g' \in \mathcal{S}$ and time periods t such that $t \geqslant g$ and $g' > \max\{g, t\}$, with probability one,

$$\begin{aligned} \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g, Q = 1, X\right] &- \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g, Q = 0, X\right] \\ &= \\ \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g', Q = 1, X\right] &- \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g', Q = 0, X\right]. \end{aligned}$$

Assumption DDD-CPT is a conditional parallel trends assumption for DDD setups that generalizes the unconditional DDD parallel trend assumption for the two-period setup of Olden and Møen (2022) to setups with multiple periods, staggered treatment adoption, and when assumptions are only plausible after conditioning on X. Assumption DDD-CPT can also be understood as an extension of the conditional PT assumption based on not-yet-treated units from the DiD setup of Callaway and Sant'Anna (2021) to our DDD setup— i.e., we can use any unit from groups that either never enabled treatment or those that will eventually enable treatment. Moreover, if covariates do not play any important identification role in the analysis, one can take $X = 1$ for all units, so Assumption DDD-CPT would hold unconditionally.

Several remarks about Assumption DDD-CPT are worth making. First, if all units in a group S are eligible for treatment, Assumption DDD-CPT reduces to Assumption 5 of Callaway and Sant'Anna (2021) under the no-anticipation condition in Assumption NA. However, this case is not appealing to us, as that would not qualify as a DDD design. Second, as Assumption DDD-CPT only holds after conditioning on covariates, it does not restrict the evolution of untreated potential outcomes across covariate-strata, i.e., it allows for covariate-specific trends, which can be very important in applications. Third, and perhaps the most empirically relevant, Assumption DDD-CPT does not impose DiD-type parallel trends among units with $S = g$—i.e., it does not impose that $\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g, Q = 1, X\right] = \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g, Q = 0, X\right]$—nor impose DiD-type parallel trends across treated groups—i.e., it does not impose that $\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g, Q = 1, X\right] = \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g', Q = 1, X\right]$. As such, Assumption DDD-CPT allows for violations of traditional DiD-based PT, as long as these violations are stable across groups. This observation is arguably what makes DDD appealing in setups in which it can be used.

3 Implications for Empirical Practices

Before introducing our formal results on identification, estimation, and inference for average treatment effects in DDD designs, we challenge some standard empirical practices for DDD analyses and highlight some important practical takeaway messages from our paper.

We first start with a simple setup with only two periods, $t = 1$ and $t = 2$, and two eligibility groups, $S_i = 2$ (who enabled treatment in period 2) and $S_i = \infty$ (who have not enabled treatment by period two). As before, units are either eligible ($Q_i = 1$) or ineligible ($Q_i = 0$) for the treatment, and we let $D_{i,t}$ be a treatment indicator for unit i in time period t, i.e., $D_{i,t} = 1\{t \geqslant S_i, Q_i = 1\}$. Since there are only two eligibility groups and two time periods, the relevant group-time ATT in such a scenario is $ATT(2,2)$. As discussed in Olden and Møen (2022), when covariates are not important for the analysis, one can use ordinary least squares (OLS) based on the following three-way fixed effects linear regression specification to recover the $ATT(2,2)$:

$$Y_{i,t} = \gamma_i + \gamma_{s,t} + \gamma_{q,t} + \beta_{\text{3wfe}} D_{i,t} + \varepsilon_{i,t}, \tag{3.1}$$

where γ_i are unit fixed effects, $\gamma_{s,t}$ and $\gamma_{q,t}$ are enabled-group-by-time and qualified-group-by-time fixed effects, and β_{3wfe} is the parameter of interest. Indeed, in this particular setup, under Assumptions S, SO, NA, and DDD-CPT with $X = 1$ a.s., it is straightforward to show that

$$\begin{aligned}
\beta_{\text{3wfe}} = &\Bigg[\underbrace{\Big(\mathbb{E}\left[Y_2 - Y_1 | S = 2, Q = 1\right]\Big) - \Big(\mathbb{E}\left[Y_2 - Y_1 | S = 2, Q = 0\right]\Big)}_{\text{DiD estimand among } S=2} \Bigg] \\
&- \Bigg[\underbrace{\Big(\mathbb{E}\left[Y_2 - Y_1 | S = \infty, Q = 1\right]\Big) - \Big(\mathbb{E}\left[Y_2 - Y_1 | S = \infty, Q = 0\right]\Big)}_{\text{DiD estimand among } S=\infty} \Bigg] \\
&= ATT(2,2).
\end{aligned} \tag{3.2}$$

The observation that $\beta_{\text{3wfe}} = ATT(2,2)$ in this simple setup has two implications: (i) one can use a simple three-way fixed effects (3WFE) regression specification and use OLS to estimate $ATT(2,2)$ in the DDD design, and (ii) DDD estimates can be understood as the difference between two DiD estimates (Olden and Møen, 2022). Based on these, one may be tempted to extrapolate these claims to more general setups. In what follows, we highlight that, unfortunately, this is not warranted and that proceeding as such can lead to non-negligible biases. The solution to these issues is relatively simple and involves adopting a "forward-engineering" approach to DDD setups (Baker et al., 2025), recognizing its particularities.

3.1 DDD setup with two periods, with covariates being important

In this section, we illustrate the challenges of leveraging standard regression-based and DiD tools to DDD setups using simple simulations in a setup where covariates are important for identification,

i.e., when Assumption DDD-CPT is satisfied only after accounting for covariates. We consider a design with four different time-invariant, unit-specific covariates, $X_i = (X_{i,1}, X_{i,2}, X_{i,3}, X_{i,4})'$, and two periods and two treatment-enabling-groups. The true $ATT(2,2)$ in our simulations is zero. To ease the exposition, we abstract from further details about the DGP and refer the reader to Section 5.1 and Supplemental Appendix B.1 for a more detailed discussion.

Based on the discussion on DDD without covariates above, it is natural to consider three alternative ways to incorporate covariates in the analysis. The first approach would be to "extrapolate" from (3.1), add the interactions of the time-invariant covariates with post-treatment dummies,

$$Y_{i,t} = \gamma_i + \gamma_{s,t} + \gamma_{q,t} + \tilde{\beta}_{\text{3wfe}} D_{i,t} + (X_i 1_{\{t=2\}})'\theta + u_{i,t}, \tag{3.3}$$

and interpret the OLS estimates of $\tilde{\beta}_{\text{3wfe}}$ as estimates of $ATT(2,2)$. The second natural way to proceed is similar, but it would leverage the Mundlak device and replace unit fixed effects in (3.1) with S-by-Q fixed effects, add covariates linearly,

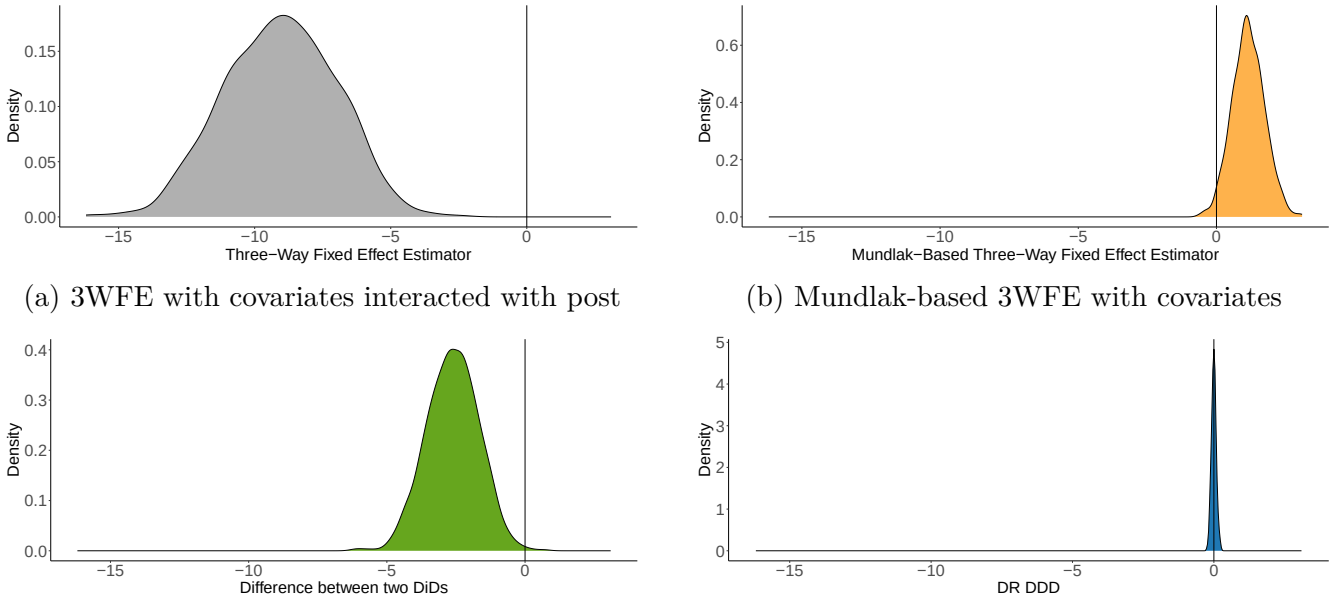
$$Y_{i,t} = \gamma_{s,q} + \gamma_{s,t} + \gamma_{q,t} + \check{\beta}_{\text{3wfe}} D_{i,t} + X_i'\theta + e_{i,t}, \tag{3.4}$$

and interpret the OLS estimates of $\check{\beta}_{\text{3wfe}}$ as estimates of $ATT(2,2)$. Both strategies leverage a presumption that we can extend the 3WFE regression (3.1) to allow for covariate-specific trends by linearly including X's into it. A third strategy that is also a priori intuitive presumes that we can write DDD estimates as the difference between two DiD estimates: one DiD using the subset with $S = 2$ and considering units treated if $Q = 1$, and another DiD using the subset with $S = \infty$ and considering units treated if $Q = 1$. Here, one could consider different estimation strategies, but we focus on the doubly robust (DR) DiD estimators proposed by Sant'Anna and Zhao (2020).

To check if such alternative strategies recover the $ATT(2,2)$, we draw $5,000$ units in each simulation draw, compute estimates using these three alternative estimators, and repeat this $1,000$ times—we defer all details of the data generating process to Section 5.1 and Supplemental Appendix B.1. Panels (a) and (b) from Figure 1 display the density of OLS estimates for the $D_{i,t}$ coefficient in (3.3) and (3.4), while Panel (c) displays the density of the DDD estimates based on the difference between two Sant'Anna and Zhao (2020) DR DiD estimators. These three panels make it clear that when covariates are necessary to justify the plausibility of the DDD research design, using any of these three procedures that are justified in DDD setups without covariates can lead to substantial biases and harm policy recommendations and evaluations. In other words, these results highlight that traditional 3WFE regression specifications are "too rigid" to be reliable for DDD analysis. They also highlight that one cannot simply claim that DDD is the difference between two DiD procedures.

A natural question that then arises is, what should we do instead? As we discuss in Section 4, one can form regression adjustment, inverse probability weighting, and doubly robust DDD estimators that do not suffer from the shortcoming highlighted in Panel (a) - (c) in Figure 1. Among these, we generally favor the DR DDD estimator as it is more robust against model misspecifications

Figure 1: Density of different DDD estimates for ATT(2,2): two-period setup with covariates



(a) 3WFE with covariates interacted with post

(b) Mundlak-based 3WFE with covariates

(c) Difference between two Doubly Robust DiDs

(d) Doubly Robust DDD

Notes: Simulation designs based on DGP 1 described in Section 5.1 and Supplemental Appendix B.1, with $n = 5,000$ and $1,000$ Monte Carlo repetitions. True $ATT(2,2)$ is zero and is indicated in the solid vertical line in all panels. Panel (a) displays the density of OLS estimates of $\tilde{\beta}_{\text{3wfe}}$ based on (3.3). Panel (b) displays the density of OLS estimates of $\hat{\beta}_{\text{3wfe}}$ based on (3.4). Panel (c) displays the density of the DDD estimates based on the difference between two doubly robust DiD estimators (Sant'Anna and Zhao, 2020). Panel (d) displays the density of the estimates based on our proposed doubly robust DDD estimator described in (3.5). All densities are computed across all simulation draws. Panels have the same x-axis range but different y-axis.

than the other alternatives. To form the DR DDD estimator for $ATT(2,2)$, we need to estimates for the outcome regression models $m^{S=g,Q=q}_{Y_2-Y_1}(X) \equiv \mathbb{E}\left[Y_2 - Y_1 | S = g, Q = q, X\right]$, and for the generalized propensity score model $p^{S=g,Q=q}(X) \equiv \mathbb{P}[S = g, Q = q|X]$. Let $\widehat{m}^{S=g,Q=q}_{Y_1-Y_0}(X)$ and $\widehat{p}^{S=g,Q=q}(X)$ be working models for these—e.g, a linear regression model and a multinomial logistic linear model. Based on these estimates, we propose the following DR DDD estimator for the $ATT(2,2)$:

$$\begin{aligned}\widehat{ATT}_{\text{dr}}(2,2) = \; & \mathbb{E}_n\left[\left(\widehat{w}^{S=2,Q=1}_{\text{trt}}(S,Q) - \widehat{w}^{S=2,Q=0}_{\text{comp}}(S,Q,X)\right)\left(Y_2 - Y_1 - \widehat{m}^{S=2,Q=0}_{Y_2-Y_1}(X)\right)\right] \\ & + \mathbb{E}_n\left[\left(\widehat{w}^{S=2,Q=1}_{\text{trt}}(S,Q) - \widehat{w}^{S=\infty,Q=1}_{\text{comp}}(S,Q,X)\right)\left(Y_2 - Y_1 - \widehat{m}^{S=\infty,Q=1}_{Y_2-Y_1}(X)\right)\right] \\ & - \mathbb{E}_n\left[\left(\widehat{w}^{S=2,Q=1}_{\text{trt}}(S,Q) - \widehat{w}^{S=\infty,Q=0}_{\text{comp}}(S,Q,X)\right)\left(Y_2 - Y_1 - \widehat{m}^{S=\infty,Q=0}_{Y_2-Y_1}(X)\right)\right],\end{aligned} \tag{3.5}$$

where $\mathbb{E}_n[A] = n^{-1}\sum_{i=1}^n A_i$ denotes the sample mean, and the estimated weights $\widehat{w}$ are given by

$$\widehat{w}^{S=2,Q=1}_{\text{trt}}(S,Q) \equiv \frac{1\{S=2,Q=1\}}{\mathbb{E}_n[1\{S=2,Q=1\}]}, \quad \widehat{w}^{S=g,Q=q}_{\text{comp}}(S,Q,X) \equiv \frac{\frac{1\{S=g,Q=q\}\cdot\widehat{p}^{S=2,Q=1}(X)}{\widehat{p}^{S=g,Q=q}(X)}}{\mathbb{E}_n\left[\frac{1\{S=g,Q=q\}\cdot\widehat{p}^{S=2,Q=1}(X)}{\widehat{p}^{S=g,Q=q}(X)}\right]}.$$

Interestingly, it is worth mentioning that although the DR DDD estimator in (3.5) cannot be expressed as the difference between two DR DiD estimators, it is a function of *three* DR DiD

estimators, each one using a particular subset of the untreated units as a comparison group.

For comparisons, we report in Panel (d) of Figure 1 the density of the estimates using the DR DDD estimates based on (3.5). Clearly, our proposed DR DDD estimator mitigated the biases associated with the other estimation strategies and led to substantially more precise estimates. All in all, the results in Figure 1 highlight that common DDD practices can lead to misleading conclusions. However, it is straightforward to bypass these limitations by adopting our DR DDD estimators.

3.2 DDD setups with variation in treatment timing

The practical challenges of estimating average treatment effects in DDD setups are not confined to the presence of covariates. Even in designs without covariates, the use of too-rigid 3WFE regression specifications like (3.1) can lead to misleading estimates when there is variation in treatment timing across groups (Strezhnev, 2023). In such cases, new identification and estimation concerns emerge that the recent DiD literature does not address. In particular, in this section, we highlight that unlike in staggered DiD procedures like Callaway and Sant'Anna (2021), pooling all not-yet-treated units and using them as a comparison group does not respect the triple-differences identification assumptions and, as such, can lead to biased estimates for the parameters of interest. We also discuss straightforward and computationally simple estimators that bypass these problems. Throughout this section, we assume that all identification assumptions discussed in Section 2.2 hold without covariates, i.e., by taking $X = 1$ almost surely.

To build intuition, we begin by noting that the way the DiD literature has addressed the shortcomings of using regression specifications akin to (3.1) to infer overall average treatment effects is to decompose the problem into a series of 2-period 2-group (2×2) DiDs; for an overview, see Roth et al. (2023) and Baker et al. (2025). A popular strategy involves using the units not yet treated by period t as a comparison group when estimating $ATT(g,t)$ (Callaway and Sant'Anna, 2021). It is thus intuitive and natural to build on Callaway and Sant'Anna (2021), Olden and Møen (2022), and (3.2), and attempt to estimate $ATT(g,t)$ in a DDD setup using

$$\begin{aligned}\widehat{ATT}_{\text{cs-nyt}}(g,t) = &\left[\left(\mathbb{E}_n\left[Y_t - Y_{g-1} | S = g, Q = 1\right]\right) - \left(\mathbb{E}_n\left[Y_t - Y_{g-1} | S = g, Q = 0\right]\right)\right] \\ &- \left[\left(\mathbb{E}_n\left[Y_t - Y_{g-1} | S > t, Q = 1\right]\right) - \left(\mathbb{E}_n\left[Y_t - Y_{g-1} | S > t, Q = 0\right]\right)\right] \end{aligned} \tag{3.6}$$

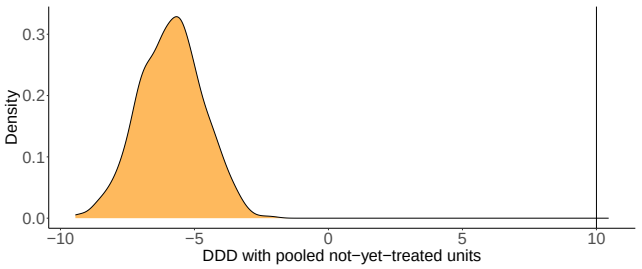
in any post-treatment periods $t \geqslant g$.[7] The question now is whether (3.6) indeed recovers $ATT(g,t)$'s under our identification assumptions.

To answer this practically relevant question, we conduct some Monte Carlo simulations for a setup with three time periods, $t = 1, 2, 3$, three treatment-enabling groups, $S \in \{2, 3, \infty\}$, and two

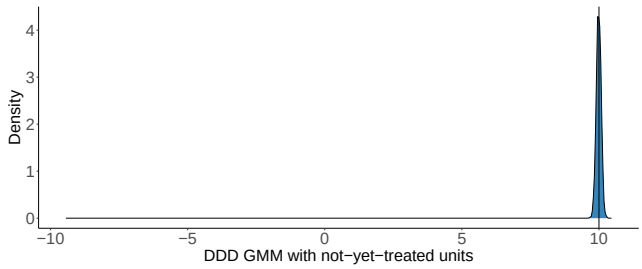
7 Since there are no covariates, we do not need to use three DiDs as we discussed in Section 3.1.

eligibility groups $Q = 1$ and $Q = 0$. We focus on $ATT(2, 2)$, i.e., the average treatment effect in period two of being treated in period two, among units treated in period two. The true $ATT(2, 2)$ in our simulations is 10. We considered a setup with $n = 5,000$ and conducted $1,000$ simulation draws. To ease the exposition, we abstract from further details about the DGP and refer the reader to Section 5.2 and Supplemental Appendix B.2 for a more detailed discussion. Panel (a) from Figure 2 displays the density of the DDD estimates for $ATT(2, 2)$ based on (3.6). This result makes it clear that, in general, (3.6) is not a valid estimator for the $ATT(2, 2)$ in DDD setups, as it is systematically biased. In fact, in our simulations, (3.6) always leads to a negative estimate while the true effect is positive. This bias arises because the DDD parallel trends assumption is more flexible than its DiD counterpart: it allows for treatment-enabling-groups- and partition-specific violations of DiD-type parallel trends. In particular, when the fraction of eligible units differs across treatment-enabling groups S, pooling not-yet-treated units may conflate trends across heterogeneous populations, violating the assumptions necessary to interpret differences as causal.

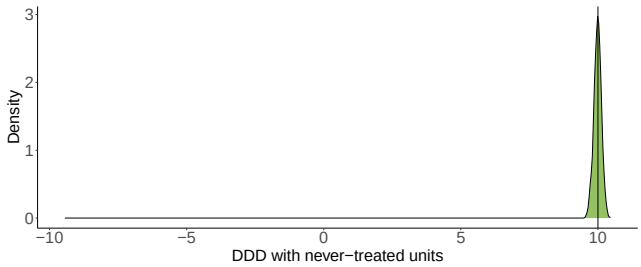
Figure 2: Density of different staggered DDD estimates for ATT(2,2), without covariates



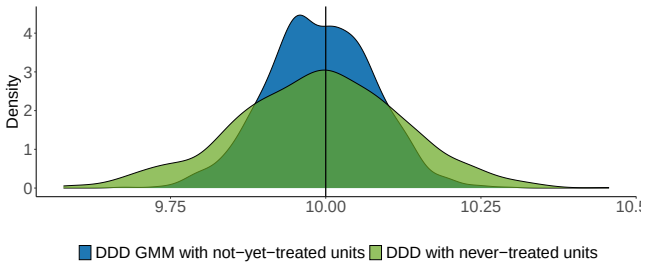
(a) DDD using pooled not-yet-treated units



(b) DDD GMM using all not-yet-treated units



(c) DDD using never-treated units



(d) DDD using never-treated and GMM with not-yet-treated units

Notes: Simulation designs based on the design described in Section 5.2 and Supplemental Appendix B.2, with $n = 5,000$ and $1,000$ Monte Carlo repetitions. The true $ATT(2, 2)$ is ten and is indicated in the solid vertical line in all panels. Panel (a) displays the density of DDD estimates that use the pooled not-yet-treated units as a comparison group as described in (3.6). Panel (b) displays the density of the estimates based on our proposed DDD GMM estimator that uses all not-yet-treated units as a comparison group described in (3.8). Panel (c) displays the density of the estimates based on our proposed DDD estimator that uses the never-treated units as a comparison group described in (3.7) with $g_c = \infty$. Panel (d) compares DDD estimates using never-treated units (yellow curve) with GMM-based DDD using not-yet-treated (green curve), on the same scale. All densities are computed across all simulation draws. Panels(a)-(c) have the same x-axis range but different y-axis.

The key insight to address these problems is that we should be careful when choosing the comparison group to learn each $ATT(g, t)$. Such comparison groups must satisfy the DDD identification assumptions, which need to be verified group-by-group. Upon close inspection of Assumption

DDD-CPT, a natural solution is to avoid pooling across treatment-enabling groups and use one of them at a time. Doing so yields multiple valid comparisons for the same (g,t) group, generating an over-identified system. More precisely, for each available not-yet-enabled group $g_c > t$, we can use the following estimator for $ATT(g,t)$, $t \geqslant g$:

$$\widehat{ATT}_{g_c}(g,t) = \left[\left(\mathbb{E}_n\left[Y_t - Y_{g-1} | S = g, Q = 1\right]\right) - \left(\mathbb{E}_n\left[Y_t - Y_{g-1} | S = g, Q = 0\right]\right)\right] - \left[\left(\mathbb{E}_n\left[Y_t - Y_{g-1} | S = g_c, Q = 1\right]\right) - \left(\mathbb{E}_n\left[Y_t - Y_{g-1} | S = g_c, Q = 0\right]\right)\right]. \tag{3.7}$$

Note that when $g_c = \infty$, (3.7) uses the set of units that never enabled treatment $S = \infty$ as the comparison group. However, one is not restricted to this unique comparison group. In the context of our simulation, one can also use the units that enabled treatment in period three to learn about $ATT(2,2)$. This highlights that, in our context, the DDD model is over-identified. In this sense, instead of choosing which comparison group to use, we propose combining all available options and forming a more efficient (minimum variance) DDD estimator for the $ATT(g,t)$s. More concretely, we propose using

$$\widehat{ATT}_{\text{opt}}(g,t) = \frac{\mathbf{1}'\widehat{\Omega}^{-1}}{\mathbf{1}'\widehat{\Omega}_{g,t}^{-1}\mathbf{1}}\widehat{ATT}_{\text{dr}}(g,t), \tag{3.8}$$

where $\widehat{ATT}_{\text{dr}}(g,t)$ is the $k_{g,t} \times 1$-dimensional vector of all possible (non-collinear) estimators for $ATT(g,t)$ that uses a valid comparison group $g_c > t$, $\widehat{\Omega}_{g,t}$ is a consistent estimator of their variance-covariance matrix, and $\mathbf{1}$ is a ($k_{g,t} \times 1$-dimensional) vector of ones. We show that $\widehat{ATT}_{\text{opt}}(g,t)$ has a GMM interpretation based on re-centered influence functions in Remark 4.5.

Panels (b) and (c) of Figure 2 display the density of the DDD estimates for $ATT(2,2)$ based on our GMM-based DDD estimator (3.7) and our DDD estimator that only uses the never-treated as comparison group (3.7) with $g_c = \infty$, respectively. As it is easy to see, both estimators are correctly centered at the true $ATT(2,2)$. As all panels in Figure 2 have the same scale, it is challenging to compare our DDD estimates that combine all not-yet-treated units with our DDD estimates that use never-treated units as comparison groups. In Figure 2(d), we address this issue and display their densities based on the $1,000$ simulation draws. Overall, one can see that using all not-yet-treated units can yield substantial gains in precision. In fact, the results of our simulations indicate that confidence intervals based on $\widehat{ATT}_{g_c=\infty}(g,t)$ are around 50% wider than those based on $\widehat{ATT}_{\text{opt}}(g,t)$, underscoring the appeal of using our optimal DDD (GMM-based) estimator in terms of power.

Overall, the results in this section underscore the broader lesson that DDD designs with staggered adoption cannot be treated as simple extensions of DiD methods. The interaction between timing, eligibility, and heterogeneity in group composition introduces complexities that require more careful attention to identification and comparison group construction.

Although this section focused on settings without covariates, our methods naturally extend to setups where covariates are necessary for identification. These generalizations are discussed in Section 4, where we also explore extensions to event-study aggregations and treatment effect heterogeneity across groups and time.

4 The econometrics of DDD designs

In this section, we discuss the econometrics of DDD designs following the framework discussed in Section 2. We start by establishing nonparametric identification of the $ATT(g,t)$'s under the identification assumptions in Section 2.2. We then discuss estimation and inference procedures for $ATT(g,t)$'s and their event-study functional $ES(e)$ as defined in (2.3). Throughout this section, we focus on setups where covariates are important for identification, i.e., all the assumptions discussed in Section 2.2 are only plausible after you condition on covariates. Results for unconditional DDD setups follow as special cases by taking all covariates $X = 1$ for all units. We also focus on staggered treatment adoption DDD setups, as they nest DDD setups with a single treatment date.

4.1 Identification

In this section, we establish the nonparametric identification for the $ATT(g,t)$'s in all post-treatment periods $t \geqslant g$ under Assumptions S, SO, NA, and DDD-CPT. Furthermore, we show that one can use regression adjustment/outcome regression (RA), inverse probability weighting (IPW), or doubly robust estimands to recover the $ATT(g,t)$'s. We also show that one can potentially use different comparison groups, opening the door for combining these for potential efficiency gains.

Before formalizing our results, we need to introduce some additional notation. Let $m^{S=g,Q=q}_{Y_t - Y_{t'}}(X) \equiv \mathbb{E}\left[Y_t - Y_{t'} | S = g, Q = q, X\right]$ denote the population regression function of changes in outcomes from period t' to period t given covariates X among units that enabled treatment in period g ($S = g$) that belongs to eligibility group q ($Q = q$). Analogously, let $p^{S=g,Q=1}_{g',q'}(X) \equiv \mathbb{P}[S = g, Q = 1 | X, (S = g, Q = 1) \cup (S = g', Q = q')]$ denote the generalized propensity score. Note that $p^{S=g,Q=1}_{g',q'}(X)$ indicates the probability of a unit being observed in enabling group $S = g$ and being eligible for treatment ($Q = 1$), conditional on pre-treatment covariates X and on either being in the $S = g$ group and being eligible for treatment, or being in the $S = g'$ group with eligibility to treatment $Q = q'$.[8]

For any $g_c \in \mathcal{S}$ such that $g_c > \max\{g,t\}$, and any post-treatment period $t \geqslant g$, let the doubly

8 We use this notion of generalize propensity score as it allow us to focus on sequences of two-groups comparisons as in Lechner (2002) and Callaway and Sant'Anna (2021). One can understand these generalized propensity scores as $p^{S=g,Q=1}_{g,q}(X) = \frac{p^{S=g,Q=1}(X)}{p^{S=g,Q=1}(X) + p^{S=g,Q=q}(X)}$ with $p^{S=g,Q=q}(X) \equiv \mathbb{P}[S = g, Q = q | X]$. We favor $p^{S=g,Q=1}_{g,q}(X)$ as this is how we implement these when constructing our DDD estimators.

robust DDD estimand for the $ATT(g,t)$ be given by

$$\begin{aligned} ATT_{\mathrm{dr},g_c}(g,t) = \mathbb{E}\left[\left(w_{\mathrm{trt}}^{S=g,Q=1}(S,Q) - w_{g,0}^{S=g,Q=1}(S,Q,X)\right)\left(Y_t - Y_{g-1} - m_{Y_t-Y_{g-1}}^{S=g,Q=0}(X)\right)\right] \\ + \mathbb{E}\left[\left(w_{\mathrm{trt}}^{S=g,Q=1}(S,Q) - w_{g_c,1}^{S=g,Q=1}(S,Q,X)\right)\left(Y_t - Y_{g-1} - m_{Y_t-Y_{g-1}}^{S=g_c,Q=1}(X)\right)\right] \\ - \mathbb{E}\left[\left(w_{\mathrm{trt}}^{S=g,Q=1}(S,Q) - w_{g_c,0}^{S=g,Q=1}(S,Q,X)\right)\left(Y_t - Y_{g-1} - m_{Y_t-Y_{g-1}}^{S=g_c,Q=0}(X)\right)\right], \end{aligned} \tag{4.1}$$

where the weights w are given by

$$w_{\mathrm{trt}}^{S=g,Q=1}(S,Q) \equiv \frac{1\{S=g,Q=1\}}{\mathbb{E}[1\{S=g,Q=1\}]}, \quad w_{g',q'}^{S=g,Q=1}(S,Q,X) \equiv \frac{\frac{1\{S=g',Q=q'\}\cdot p_{g',q'}^{S=g,Q=1}(X)}{1-p_{g',q'}^{S=g,Q=1}(X)}}{\mathbb{E}\left[\frac{1\{S=g',Q=q'\}\cdot p_{g',q'}^{S=g,Q=1}(X)}{1-p_{g',q'}^{S=g,Q=q}(X)}\right]}. \tag{4.2}$$

Analogously, let the RA DDD estimand for the $ATT(g,t)$ be given by

$$ATT_{\mathrm{ra},g_c}(g,t) = \mathbb{E}\Big[w_{\mathrm{trt}}^{S=g,Q=1}(S,Q)\Big(Y_t - Y_{g-1} - m_{Y_t-Y_{g-1}}^{S=g,Q=0}(X) - m_{Y_t-Y_{g-1}}^{S=g_c,Q=1}(X) + m_{Y_t-Y_{g-1}}^{S=g_c,Q=0}(X)\Big)\Big] \tag{4.3}$$

and the IPW estimand be

$$\begin{aligned} ATT_{\mathrm{ipw},g_c}(g,t) = \mathbb{E}\left[\left(w_{\mathrm{trt}}^{S=g,Q=1}(S,Q) - w_{g,0}^{S=g,Q=1}(S,Q,X)\right)(Y_t - Y_{g-1})\right] \\ - \mathbb{E}\left[\left(w_{g_c,1}^{S=g,Q=1}(S,Q,X) - w_{g_c,0}^{S=g,Q=1}(S,Q,X)\right)(Y_t - Y_{g-1})\right]. \end{aligned} \tag{4.4}$$

Theorem 4.1. *Let Assumptions S, SO, NA, and DDD-CPT hold. Then, for all $g \in \mathcal{G}_{trt}$, $t \in \{2,\ldots,T\}$, and $g_c \in \mathcal{S}$ such that $t \geqslant g$ and $g_c > t$,*

$$ATT(g,t) = ATT_{dr,g_c}(g,t) = ATT_{ra,g_c}(g,t) = ATT_{ipw,g_c}(g,t). \tag{4.5}$$

Theorem 4.1 is the first main result of this paper. It establishes the nonparametric identification of all post-treatment $ATT(g,t)$'s in DDD setups. It extends the DiD identification results of Callaway and Sant'Anna (2021) to DDD setups. As such, it also extends the difference-in-differences identification results based on the RA approach of Heckman, Ichimura and Todd (1997), the IPW approach of Abadie (2005), and the DR approach of Sant'Anna and Zhao (2020) to DDD setups with multiple periods and variation in treatment time. Theorem 4.1 also highlights that one can use different parts of the data-generating process to identify the $ATT(g,t)$'s: the RA estimand only models the conditional expectation of evolution of outcomes among untreated units, the IPW approach only models the conditional probability of being observed in a given partition of the S-by-Q groups, whereas the DR approach exploits both components. A big advantage of the DR approach is that it is based on a Neyman-orthogonal moment condition (Belloni, Chernozhukov, Fernández-Val and Hansen, 2017), and, therefore, it is more robust against model misspecifications than the IPW and RA formulations. In fact, it is very easy to show that estimators based on $ATT_{\mathrm{dr},g_c}(g,t)$ enjoy a very attractive doubly-robust property (Sant'Anna and Zhao, 2020) that

allows for some forms of (global) model misspecifications.[9]

Another important result from Theorem 4.1 is that our DDD model is over-identified, as we can use multiple not-yet-treated enabling groups g_c as valid comparison groups. For instance, in a setup with $S \in \{2, 3, \infty\}$, we can set $g_c = 3$ or $g_c = \infty$ to identify $ATT(2, 2)$, and both will lead to the same target parameter. As a direct consequence of this result, any weighted sum of these estimands that use different g_c's will also lead to the $ATT(g, t)$, as long as the weights sum up to one. We formalize this result in the following corollary using the DR estimand, though this also works with the RA and IPW. Let $\mathcal{G}_c^{g,t} = \{g_c \in \mathcal{S} : g_c > \max\{g, t\}\}$.

Corollary 4.1. *Let Assumptions S, SO, NA, and DDD-CPT hold. Then, for all $g \in \mathcal{G}_{trt}$ and $t \in \{2, \ldots, T\}$ such that $t \geqslant g$, and any set of weights $w_{gc}^{g,t}$ that sum up to one over $\mathcal{G}_c^{g,t}$,*

$$ATT(g, t) = \sum_{g_c \in \mathcal{G}_c^{g,t}} w_{g_c}^{g,t} \, ATT_{dr,g_c}(g, t).$$

As Corollary 4.1 indicates that all weighted sums lead to the same $ATT(g, t)$, a natural way to choose these weights is to pick them such that we maximize precision in terms of minimizing the resulting asymptotic variance. In the next session, we will discuss this in greater detail, connecting these arguments to a formulation based on generalized methods of moments using re-centered influence functions.

Remark 4.1. As we discussed in Section 3, in two-period DDD setups without covariates, one can identify $ATT(2, 2)$ using the difference of two DiD estimands as in (3.2) (Olden and Møen, 2022). This equivalence breaks down when the DDD identification assumptions are only satisfied after you condition on covariates X—see Figure 1. The econometric reason for this failure of equivalence is that one needs to integrate the covariates using the covariate distribution among treated units, i.e., units with $S = 2$ and $Q = 1$. Proceeding as if $ATT(2, 2)$ were the difference of two DiD estimands would integrate X using the covariate distribution of untreated units ($S = \infty$ and $Q = 1$), leading to biases. The results in Theorem 4.1 address this problem by guaranteeing that one integrates out covariates using the correct reference distribution, which leads to a combination of three DiD estimands and not only two.

Remark 4.2. Although Theorem 4.1 and Corollary 4.1 allow one to use several different not-yet-treated cohorts g_c as the comparison group, it does not allow one to pool all not-yet-treated units and use that pooled set of units as the aggregate comparison group to identify $ATT(g, t)$ in DDD. This sharply contrasts DiD procedures such as those discussed in Callaway and Sant'Anna (2021)—see Figure 2a for an illustration of the bias that can arise by following this type of procedure. The econometric reasoning for such results is that Assumption DDD-CPT allows for both enabling-group- and eligibility-group-specific trends, and it does not impose that the proportion of units in

9 For an overview of doubly robust estimators in cross-sectional designs, see section 2 of Słoczyński and Wooldridge (2018), and Seaman and Vansteelandt (2018).

each eligibility group Q is the same across all enabling groups S. As such, Assumption DDD-CPT does not guarantee that, with probability one,

$$\begin{array}{rcl} \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S = g, Q = 1, X\right] & - & \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S = g, Q = 0, X\right] \\ & = & \\ \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S > t, Q = 1, X\right] & - & \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S > t, Q = 0, X\right], \end{array}$$

as it would be required to use the pooled, not-yet-treated units as a comparison group.

Remark 4.3. As Theorem 4.1 establishes nonparametric identification of the $ATT(g,t)$'s over all post-treatment periods and that $\mathbb{P}(G = g|G + e \in [1,T])$ is also nonparametrically identified, it follows that event-study parameters that aggregate across eligibility-groups, $ES(e)$ as defined in (2.3), is also nonparametrically identified. For instance, it follows that for any event-time $e \geqslant 0$,

$$ES(e) = \sum_{g \in \mathcal{G}_{\text{trt}}} \mathbb{P}(G = g|G + e \in [1,T]) ATT_{\text{dr},g_c}(g, g+e). \tag{4.6}$$

One can also replace $ATT_{\text{dr},g_c}(g, g+e)$ with their analogs in Corollary 4.1 or with the RA or IPW estimands in Theorem 4.1. One can also use Theorem 4.1 to establish the identification of many other aggregate summary causal parameters discussed in Section 3 of Callaway and Sant'Anna (2021).

Remark 4.4. One of the biggest appeals of DiD and DDD setups is the availability of pre-treatment periods that allow the assessment of the plausibility of PT assumptions, such as Assumption DDD-CPT. Under Assumption NA, a very popular way to assess the plausibility of PT is to construct event-study plots based on $ES(e)$ as in (2.3), consider both pre-treatment ($e < 0$) and post-treatment ($e \geqslant 0$) event times, and check whether pre-treatment event-study coefficients are all close to zero. It is straightforward to adapt this strategy in our DDD context by fixing the statistical estimand—for example, the $ATT_{\text{dr},g_c}(g,t)$ in (4.1)—consider pre-treatment periods $t < g$, and then aggregate them using cohort-size. More specifically, for any event-time $e < 0$,

$$ES(e) = \sum_{g \in \mathcal{G}_{\text{trt}}} \mathbb{P}(G = g|G + e \in [1,T]) ATT_{\text{dr},g_c}(g, g+e). \tag{4.7}$$

Note that when $e = -1$, $ES(e) = 0$ by construction, as we fix the baseline period at the last untreated period for group g, $g-1$. Based on these event-study aggregations, it is also possible to conduct sensitivity analysis for the plausibility of Assumption DDD-CPT using the results in Rambachan and Roth (2023).

4.2 Estimation and inference

In this section, we now propose simple-to-use plug-in estimators for the $ATT(g,t)$s and $ES(e)$s parameters, and discuss how one can conduct valid inference for these parameters. We focus on the doubly robust DDD estimator; the results for the RA and IPW DDD estimators are analogous.

First, notice that for any $g_c \in \mathcal{S}$ such that $g_c > \max\{g,t\}$, Theorem 4.1 suggests that we can estimate $ATT(g,t)$ by using the sample analogue of (4.1),

$$\begin{aligned}\widehat{ATT}_{\text{dr},g_c}(g,t) = \mathbb{E}_n\left[\left(\widehat{w}_{\text{trt}}^{S=g,Q=1}(S,Q) - \widehat{w}_{g,0}^{S=g,Q=1}(S,Q,X)\right)\left(Y_t - Y_{g-1} - \widehat{m}_{Y_t-Y_{g-1}}^{S=g,Q=0}(X)\right)\right] \\ + \mathbb{E}_n\left[\left(\widehat{w}_{\text{trt}}^{S=g,Q=1}(S,Q) - \widehat{w}_{g_c,1}^{S=g,Q=1}(S,Q,X)\right)\left(Y_t - Y_{g-1} - \widehat{m}_{Y_t-Y_{g-1}}^{S=g_c,Q=1}(X)\right)\right] \\ - \mathbb{E}_n\left[\left(\widehat{w}_{\text{trt}}^{S=g,Q=1}(S,Q) - \widehat{w}_{g_c,0}^{S=g,Q=1}(S,Q,X)\right)\left(Y_t - Y_{g-1} - \widehat{m}_{Y_t-Y_{g-1}}^{S=g_c,Q=0}(X)\right)\right],\end{aligned} \tag{4.8}$$

where the estimated weights $\widehat{w}$ are given by

$$\widehat{w}_{\text{trt}}^{S=g,Q=1}(S,Q) \equiv \frac{1\{S=g,Q=1\}}{\mathbb{E}_n[1\{S=g,Q=1\}]}, \quad \widehat{w}_{g',q'}^{S=g,Q=1}(S,Q,X) \equiv \frac{\dfrac{1\{S=g',Q=q'\}\cdot \widehat{p}_{g',q'}^{S=g,Q=1}(X)}{1-\widehat{p}_{g',q'}^{S=g,Q=1}(X)}}{\mathbb{E}_n\left[\dfrac{1\{S=g',Q=q'\}\cdot \widehat{p}_{g',q'}^{S=g,Q=1}(X)}{1-\widehat{p}_{g',q'}^{S=g,Q=q}(X)}\right]},$$

and $\widehat{m}_{Y_t-Y_{g-1}}^{S=g_c,Q=1}(X)$ and $\widehat{p}_{g',q'}^{S=g,Q=1}(X)$ are (potentially misspecified) working models for the outcome regression $m_{Y_t-Y_{g-1}}^{S=g_c,Q=1}(X)$ and the generalized propensity score $\widehat{p}_{g',q'}^{S=g,Q=1}(X)$. These estimators extend the DR DiD estimator of Callaway and Sant'Anna (2021) to the DDD setup, and remain consistent if *either* outcome regression or generalized propensity score models are correctly specified. It is also worth stressing that we do not need that *all* generalized propensity score working models or *all* outcome regression working models in (4.8) to be correctly specified to get a consistent DDD estimator for $ATT(g,t)$; it suffices that any of the working models within each of the 3 DR DiD components of (4.8) to be correctly specified, allowing a greater deal of estimation flexibility.[10]

As Corollary 4.1 highlights, one can also combine several $\widehat{ATT}_{\text{dr},g_c}(g,t)$ that leverage different comparison groups g_c, i.e., for any (consistently estimated) weights $\widehat{w}_{g_c}^{\text{g,t}}$ that sum up to one over $\mathcal{G}_c$,

$$\widehat{ATT}_{\text{dr},\widehat{w}}(g,t) = \sum_{g_c \in \mathcal{G}_c^{\text{g,t}}} \widehat{w}_{g_c}^{\text{g,t}}\ \widehat{ATT}_{\text{dr},g_c}(g,t) = \widehat{w}^{\text{g,t}\ \prime}\widehat{ATT}_{\text{dr}}(g,t), \tag{4.9}$$

where $\widehat{ATT}_{\text{dr}}(g,t)$ is the $k_{g,t} \times 1$ vector of $\widehat{ATT}_{\text{dr},g_c}(g,t)$ for all $g_c \in \mathcal{G}_c^{\text{g,t}}$, and $\widehat{w}^{\text{g,t}\ \prime}$ is a $k_{g,t} \times 1$ vector of (estimated) weights that sum up to one, i.e., for a generic vector of ones $\mathbf{1}$, $\mathbf{1}'w^{\text{g,t}} = 1$.

A natural question that arises is: how should one choose these weights $\widehat{w}_{g_c}^{\text{g,t}}$? We propose to choose the weights that lead to the asymptotically most precise (minimum variance) estimator for $ATT(g,t)$, that is, to pick weights that solve

$$\min_{w^{\text{g,t}}} w^{\text{g,t}\ \prime}\ \widehat{\Omega}_{g,t}\ w^{\text{g,t}} \qquad \text{subject to } \mathbf{1}'w^{\text{g,t}} = 1, \tag{4.10}$$

where $\widehat{\Omega}_{g,t}$ is a $k_{g,t} \times k_{g,t}$ consistent estimator for the variance-covariance matrix of $\widehat{ATT}_{\text{dr}}(g,t)$. Notice that the solution of (4.10) admits a closed-form solution, and the optimal weights are given

10 Some people may call this a multiply-robust estimator, as one has more than two opportunities to estimate the target parameter consistently. For simplicity, we retain the doubly robust terminology to avoid new acronyms.

by

$$\widehat{w}_{\text{opt}}^{\text{g,t}} = \frac{\widehat{\Omega}_{g,t}^{-1}\mathbf{1}}{\mathbf{1}'\widehat{\Omega}_{g,t}^{-1}\mathbf{1}}. \tag{4.11}$$

In turn, this implies that the linear combination of $\widehat{ATT}_{\text{dr},g_c}(g,t)$ that leads to the most precise estimator for $ATT(g,t)$ is given by

$$\widehat{ATT}_{\text{dr,opt}}(g,t) = \frac{\mathbf{1}'\widehat{\Omega}_{g,t}^{-1}}{\mathbf{1}'\widehat{\Omega}_{g,t}^{-1}\mathbf{1}}\widehat{ATT}_{\text{dr}}(g,t). \tag{4.12}$$

In many situations with multiple periods and variation in treatment time, researchers are interested in summarizing the $ATT(g,t)$'s into fewer parameters that highlight treatment effect heterogeneity with respect to the time elapsed since treatment take-up. That is, very often, researchers are interested in estimating event-study type parameters $ES(e)$ as defined in (2.3). A very natural estimator for $ES(e)$ is the plug-in estimator, where we replace $ATT(g,t)$ with $\widehat{ATT}_{\text{dr,opt}}(g,t)$ (or $\widehat{ATT}_{\text{dr},g_c}(g,t)$), and $\mathbb{P}(G=g|G+e\in[1,T])$ by its sample analogue, that is,

$$\widehat{ES}_{\text{dr,opt}}(e) = \sum_{g\in\mathcal{G}_{\text{trt}}} \mathbb{P}_n(G=g|G+e\in[1,T])\widehat{ATT}_{\text{dr,opt}}(g,g+e), \tag{4.13}$$

where $\mathbb{P}_n(G=g|G+e\in[1,T]) = \sum_{i=1}^n 1\{G_i=g\}1\{G_i+e\in[1,T]\}/\sum_{j=1}^n 1\{G_j+e\in[1,T]\}$. We can define $\widehat{ES}_{\text{dr},g_c}(e)$ analogously by replacing $\widehat{ATT}_{\text{dr,opt}}(g,g+e)$ with $\widehat{ATT}_{\text{dr},g_c}(g,g+e)$ on (4.13). Based on it, we can also estimate an overall summary parameter by averaging all post-treatment event times, i.e.,

$$\widehat{ES}_{\text{avg, opt}}\frac{1}{N_E}\sum_{e\in\mathcal{E}}\widehat{ES}_{\text{dr,opt}}(e). \tag{4.14}$$

Remark 4.5. It is also worth noticing that $\widehat{ATT}_{\text{dr,opt}}(g,t)$ in (4.12) can be interpreted as an optimal Generalized Method of Moments (GMM) estimator based on re-centered influence functions. To see this, let $\mathbb{IF}_{\text{dr},g_c}(g,t)$ denote the influence function of $\sqrt{n}\left(\widehat{ATT}_{\text{dr},g_c}(g,t) - ATT_{\text{dr},g_c}(g,t)\right)$. Let $\mathbb{RIF}_{\text{dr},g_c}(g,t) = \mathbb{IF}_{\text{dr},g_c}(g,t) + ATT_{\text{dr},g_c}(g,t)$ denote its re-centered influence function, and denote the $k_{g,t}\times 1$ vector of all $\mathbb{RIF}_{\text{dr},g_c}(g,t)$ for $g_c\in\mathcal{G}_c^{\text{g,t}}$ by $\mathbb{RIF}_{\text{dr}}(g,t)$. Since influence functions are mean zero, and that $ATT(g,t) = ATT_{\text{dr},g_c}(g,t)$ for any $g_c\in\mathcal{G}_c^{\text{g,t}}$, we have the vector of moment conditions $\mathbb{E}[\mathbb{RIF}_{\text{dr}}(g,t) - \theta^{g,t}] = 0$, with $\theta^{g,t} = ATT(g,t)$. From standard GMM results (Newey and McFadden, 1994), it follows that, under mild regularity conditions, the optimal (population) GMM estimator for $\theta^{g,t}$ is given by

$$\theta_{\text{opt}}^{g,t} = \frac{\mathbf{1}'\Omega_{g,t}^{-1}}{\mathbf{1}'\Omega_{g,t}^{-1}\mathbf{1}}\mathbb{E}[\mathbb{RIF}_{\text{dr}}(g,t)] = \frac{\mathbf{1}'\Omega_{g,t}^{-1}}{\mathbf{1}'\Omega_{g,t}^{-1}\mathbf{1}}ATT_{\text{dr}}(g,t),$$

where the last equality follows from $\mathbb{E}[\mathbb{IF}_{\text{dr}}(g,t)] = 0$. Thus, $\widehat{ATT}_{\text{dr,opt}}(g,t)$ in (4.12) is the sample-analogy of the efficient population GMM $\theta_{\text{opt}}^{g,t}$.

4.2.1 Asymptotic theory for ATT(g,t)'s

In what follows, we derive the large sample properties of our DR DDD estimators $\widehat{ATT}_{\text{dr},g_c}(g,t)$ and $\widehat{ATT}_{\text{dr,opt}}(g,t)$. All our results are derived for the large n, fixed T paradigm. For a generic Z, let $||Z|| = \sqrt{trace(Z'Z)}$ denote the Euclidean norm of Z and set $W_i = (Y_{i,t=1},\ldots,Y_{i,t=T},X_i',G_i,S_i,Q_i)'$; we will omit the index i to unclutter the notation. Let $g(\cdot)$ be a generic notation for the outcome regressions $m^{S=g',Q=q}_{Y_t-Y_{t'}}(X)$ and generalized propensity scores $p^{S=g,Q=1}_{g',q'}(X)$, and, with some abuse of notation, let $g(\cdot;\gamma)$ denote a parametric model for $g(\cdot)$ that is known up to the finite-dimensional parameters γ. For a generic $\kappa^{g,t}_{g_c} = (\gamma^{ps\,\prime}_{g,t,g_c},\gamma^{reg\,\prime}_{g,t,g_c})'$, with γ^{ps}_{g,t,g_c} and γ^{reg}_{g,t,g_c} being nuisance parameters for the generalized propensity score and outcome regressions, respectively, let

$$
\begin{aligned}
h^{g,t}_{g_c}(W;\kappa^{g,t}_{g_c}) = &\left(w^{S=g,Q=1}_{\text{trt}}(W) - w^{S=g,Q=1}_{g,0}(W;\gamma^{ps}_{g,t,g_c})\right)\left(Y_t - Y_{g-1} - m^{S=g,Q=0}_{Y_t-Y_{g-1}}(X;\gamma^{reg}_{g,t,g_c})\right)\\
&+\left(w^{S=g,Q=1}_{\text{trt}}(W) - w^{S=g,Q=1}_{g_c,1}(W;\gamma^{ps}_{g,t,g_c})\right)\left(Y_t - Y_{g-1} - m^{S=g_c,Q=1}_{Y_t-Y_{g-1}}(X;\gamma^{reg}_{g,t,g_c})\right)\\
&-\left(w^{S=g,Q=1}_{\text{trt}}(W) - w^{S=g,Q=1}_{g_c,0}(W;\gamma^{ps}_{g,t,g_c})\right)\left(Y_t - Y_{g-1} - m^{S=g_c,Q=0}_{Y_t-Y_{g-1}}(X;\gamma^{reg}_{g,t,g_c})\right),
\end{aligned}
$$

where the weights $w(W;\gamma^{ps}_{g,t,g_c})$ are defined similarly to those in (4.2), with the difference being that the true unknown generalized propensity score models are replaced by working parametric counterparts, $p^{S=g,Q=1}_{g',q'}(X;\gamma^{ps}_{g,t,g_c})$, and the true unknown outcome regression models $m^{S=g',Q=q}_{Y_t-Y_{g-1}}(X)$ are also replaced with parametric working models, $m^{S=g',Q=q}_{Y_t-Y_{g-1}}(X;\gamma^{reg}_{g,t,g_c})$. We denote the vector of pseudo-true parameters by $\kappa^{g,t}_{0,g_c}$ and let $\dot{h}^{g,t}_{g_c}(\kappa) = \partial h^{g,t}_{g_c}(W;\kappa)/\partial\kappa$.

To derive our results, we make the following relatively mild assumptions.

Assumption WM (Working Model Conditions)**.** (i) $g(x;\gamma)$ is a parametric model for $g(x)$, where $\gamma\in\Theta\subset\mathbb{R}^{d_k}$ is a compact set; (ii) the mapping $\theta\mapsto g(X;\theta)$ is a.s. continuous; (iii) the pseudo-true parameter $\theta_0\in\text{int}(\Theta)$ satisfies that for an appropriate criterion function $Q:\Theta\to\mathbb{R}$ and for any $\epsilon>0$, there exists some $\delta>0$ such that $\inf_{\theta\in\Theta:\|\theta-\theta_0\|\geqslant\epsilon} Q(\theta) - Q(\theta_0) > \delta$; (iv) there exists an open neighborhood $\Theta_0\subset\Theta$ containing θ_0 such that $g(X;\gamma)$ is a.s. continuously differentiable in a neighborhood of $\gamma_0\in\Theta_0$. In addition, (v) there exists some $\epsilon>0$ such that, for all $(g,g',q')\in\mathcal{G}_{\text{trt}}\times\mathcal{G}^{g,t}_c\times\{0,1\}$, we have that $0\leqslant p^{S=g,Q=1}_{g',q'}(X;\theta)\leqslant 1-\epsilon$ a.s. for all $\theta\in\text{int}(\Theta_{ps})$, where Θ_{ps} denotes the parameter space of γ for the generalized propensity score working model.

Assumption ALR ($\sqrt{n}$-Asymptotically Linear Representation)**.** Let $\hat{\theta}$ be a strongly consistent estimator of $\theta_0\mapsto g(x;\theta_0)$ and satisfy the following linear expansion

$$
\sqrt{n}\left(\widehat{\theta}-\theta_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} l\left(W_i;\theta_0\right) + o_p(1) \tag{4.15}
$$

where $l(\cdot;\cdot)$ is a function such that $\mathbb{E}\left[l\left(W_i;\theta_0\right)\right]=0$; $\mathbb{E}\left[l\left(W_i;\theta_0\right)\cdot l\left(W_i;\theta_0\right)'\right]<\infty$ and is positive definite; and $\lim_{\delta\to 0}\mathbb{E}\left[\sup_{\theta\in\Theta_0:\|\theta-\theta_0\|\leqslant\delta}\|l(W;\theta)-l(W;\theta_0)\|^2\right]=0$.

Assumption IC (Integrability Conditions)**.** For each $g \in \mathcal{G}_{\text{trt}}$, $t \in \{2, \ldots, T\}$, and $g' \in \mathcal{G}_c^{g,t}$, assume that $\mathbb{E}[\|h_{g_c}^{g,t}(W; \kappa_{0,g_c}^{g,t})\|^2] < \infty$ and $\mathbb{E}\left[\sup_{\kappa \in \Gamma_0} \left|\dot{h}_{g_c}^{g,t}(\kappa)\right|\right] < \infty$, where Γ_0 is a small neighborhood of the pseudo-true parameter $\kappa_{0,g_c}^{g,t}$.

Assumptions WM-IC are standard in the literature; see e.g., Abadie (2005); Wooldridge (2007); Sant'Anna and Zhao (2020); Callaway and Sant'Anna (2021). Assumptions WM and ALR impose a well-behaved parametric structure for the first-step estimators for the nuisance parameters. This assumption is made for statistical convenience and acknowledges that, in many DDD applications, the number of units in each group is small, making it difficult to adopt a nonparametric approach reliably. It is relatively straightforward to relax these conditions and allow for nonparametric or even Machine-Learning based estimators; see, e.g., Sant'Anna and Xu (2023) for detailed arguments on these in a DiD setup with compositional changes. Assumption IC imposes mild regularity constraints on the moments of the estimating equations, preventing ill-behaved variance properties and ensuring the stability of higher-order approximations.

In what follows, we omit W and X from the weights and outcome regressions to minimize notation, and for a generic $\kappa_{g_c}^{g,t}$, let

$$\psi_{g_c}^{g,t}(W; \kappa_{g_c}^{g,t}) = \psi_{S=g,Q=0}^{g,t}(W; \kappa_{g_c}^{g,t}) + \psi_{S=g_c,Q=1}^{g,t}(W; \kappa_{g_c}^{g,t}) - \psi_{S=g_c,Q=0}^{g,t}(W; \kappa_{g_c}^{g,t}), \tag{4.16}$$

where, for $(g', q') \in \{(g, 0), (g_c, 1), (g_c, 0)\}$, $\psi_{S=g',Q=q'}^{g,t}(W; \kappa_{g_c}^{g,t})$ is an influence function for one of the three DR DiD components of the DR DDD, and is given by

$$\psi_{S=g',Q=q'}^{g,t}(W; \kappa_{g_c}^{g,t}) = \psi_{S=g',Q=q'}^{g,t,1}(W; \kappa_{g_c}^{g,t}) - \psi_{S=g',Q=q'}^{g,t,0}(W; \kappa_{g_c}^{g,t}) - \psi_{S=g',Q=q'}^{g,t,est}(W; \kappa_{g_c}^{g,t}), \tag{4.17}$$

with

$$\begin{aligned}
\psi_{S=g',Q=q'}^{g,t,1}(W; \kappa_{g_c}^{g,t}) = {} & w_{\text{trt}}^{S=g,Q=1}\left(Y_t - Y_{g-1} - m_{Y_t-Y_{g-1}}^{S=a,Q=b}(\gamma_{g,t,g_c}^{reg})\right) \\
& - w_{\text{trt}}^{S=g,Q=1}\,\mathbb{E}\left[w_{\text{trt}}^{S=g,Q=1}\left(Y_t - Y_{g-1} - m_{Y_t-Y_{g-1}}^{S=a,Q=b}(\gamma_{g,t,g_c}^{reg})\right)\right] \\
\psi_{S=g',Q=q'}^{g,t,0}(W; \kappa_{g_c}^{g,t}) = {} & w_{g',q'}^{S=g,Q=1}(\gamma_{g,t,g_c}^{ps})\left(Y_t - Y_{g-1} - m_{Y_t-Y_{g-1}}^{S=a,Q=b}(\gamma_{g,t,g_c}^{reg})\right) \\
& - w_{g',q'}^{S=g,Q=1}(\gamma_{g,t,g_c}^{ps})\,\mathbb{E}\left[w_{g',q'}^{S=g,Q=1}(\gamma_{g,t,g_c}^{ps})\left(Y_t - Y_{g-1} - m_{Y_t-Y_{g-1}}^{S=a,Q=b}(\gamma_{g,t,g_c}^{reg})\right)\right]
\end{aligned}$$

and

$$\psi_{S=g',Q=q'}^{g,t,est}(W; \kappa_{g_c}^{g,t}) = l_{S=g',Q=q'}^{g,t,reg}(\gamma_{g,t,g_c}^{ref})' M_{S=g',Q=q'}^{g,t,1}(\kappa_{g_c}^{g,t}) + l_{S=g',Q=q'}^{g,t,ps}(\gamma_{g,t,g_c}^{ps})' M_{S=g',Q=q'}^{g,t,2}(\kappa_{g_c}^{g,t})$$

where $l_{S=g',Q=q'}^{g,t,reg}(\cdot)$ is the asymptotic linear representation of the outcome evolution for the group with $S = g'$ and $Q = q'$ as described in Assumption ALR, $l_{S=g',Q=q'}^{g,t,ps}(\cdot)$ is defined analogously for the generalized propensity score that uses group $S = a, Q = b$ as a comparison group, and

$$\begin{aligned}
M_{S=g',Q=q'}^{g,t,1}(\kappa_{g_c}^{g,t}) = {} & \mathbb{E}\left[\left(w_{\text{trt}}^{S=g,Q=1} - w_{g',q'}^{S=g,Q=1}(\gamma_{g,t,g_c}^{ps})\right)\dot{m}_{Y_t-Y_{g-1}}^{S=a,Q=b}(\gamma_{g,t,g_c}^{reg})\right], \\
M_{S=g',Q=q'}^{g,t,2}(\kappa_{g_c}^{g,t}) = {} & \mathbb{E}\left[\alpha_{g',q'}^{S=g,Q=1}(\gamma_{g,t,g_c}^{ps})\left(Y_t - Y_{g-1} - m_{Y_t-Y_{g-1}}^{S=a,Q=b}(\gamma_{g,t,g_c}^{reg})\right)\cdot \dot{p}_{g',q'}^{S=g,Q=q}(\gamma_{g,t,g_c}^{ps})\right] \\
& - \mathbb{E}\left[\alpha_{g',q'}^{S=g,Q=1}(\gamma_{g,t,g_c}^{ps})\left(\mathbb{E}\left[w_{g',q'}^{S=g,Q=1}(\gamma_{g,t,g_c}^{ps})\left(Y_t - Y_{g-1} - m_{Y_t-Y_{g-1}}^{S=a,Q=b}(\gamma_{g,t,g_c}^{reg})\right)\right]\right)\cdot \dot{p}_{g',q'}^{S=g,Q=q}(\gamma_{g,t,g_c}^{ps})\right]
\end{aligned}$$

with $\dot{m}^{S=a,Q=b}_{Y_t-Y_{g-1}}(\gamma^{reg}_{g,t,g_c}) = \partial m^{S=a,Q=b}_{Y_t-Y_{g-1}}(\gamma^{reg}_{g,t,g_c})/\partial\gamma^{reg}_{g,t,g_c}$, $\dot{p}^{S=g,Q=q}_{g',q'}(\gamma^{ps}_{g,t,g_c}) = \partial p^{S=g,Q=q}_{g',q'}(\gamma^{ps}_{g,t,g_c})/\gamma^{ps}_{g,t,g_c}$, and

$$\alpha^{S=g,Q=1}_{g',q'}(\gamma^{ps}_{g,t,g_c}) = \frac{1\{S=a,Q=b\}}{\left(1-p^{S=g,Q=1}_{g',q'}(X;\gamma^{ps}_{g,t,g_c})\right)^2} \Bigg/ \mathbb{E}\left[\frac{1\{S=a,Q=b\}\cdot p^{S=g,Q=1}_{g',q'}(X;\gamma^{ps}_{g,t,g_c})}{1-p^{S=g,Q=q}_{g',q'}(X;\gamma^{ps}_{g,t,g_c})}\right].$$

For each $g \in \mathcal{G}_{\text{trt}}$ and each $t \in \{2,3,\ldots,\}$, let $ATT_{\text{dr}}(g,t)$ denote the $k_{g,t}\times 1$ vector of $ATT_{\text{dr},g_c}(g,t)$ for all (non-collinear) $g_c \in \mathcal{G}^{\text{g,t}}_{\text{c}}$, and $\Omega_{g,t}$ be the asymptotic variance-covariance matrix of $\sqrt{n}\left(\widehat{ATT}_{\text{dr}}(g,t) - ATT_{\text{dr}}(g,t)\right)$, i.e., $\Omega_{g,t} = \mathbb{E}\left[\psi^{g,t}(W;\kappa^{g,t})\psi^{g,t}(W;\kappa^{g,t})'\right]$, with $\psi^{g,t}(W;\kappa^{g,t})$ being the $k_{g,t}\times 1$ vector that stacks all non-collinear $\psi^{g,t}_{g_c}(W;\kappa^{g,t}_{g_c})$ for $g_c \in \mathcal{G}^{\text{g,t}}_{\text{c}}$. Let $ATT_{\text{dr,opt}}(g,t) = (\mathbf{1}'\Omega^{-1}_{g,t}\mathbf{1})^{-1}\mathbf{1}'\Omega^{-1}_{g,t}\ ATT_{\text{dr}}(g,t)$, and for a generic set of weights that sum up to one, let $ATT_{\text{dr},w}(g,t) = \sum_{g_c\in\mathcal{G}^{\text{g,t}}_{\text{c}}} w^{\text{g,t}}_{g_c}\ ATT_{\text{dr},g_c}(g,t)$, and recall that $\widehat{ATT}_{\text{dr},\widehat{w}}(g,t)$ is its empirical analogue as defined in (4.9). Finally, let $\widehat{\Omega}_{g,t}$ be the empirical analogue of $\Omega_{g,t}$, where one replaces expectations by sample analogues and $\kappa^{g,t}$ with $\widehat{\kappa}^{g,t}$, and consider the following claim:

$$\begin{aligned}
&\text{For each } g\in\mathcal{G}_{\text{trt}}, t\in\{2,\ldots,T\} \text{ such that } t\geqslant g, \text{ and each } g_c\in\mathcal{G}^{\text{g,t}}_{\text{c}},\\
&\text{we have that, for each } (g',q')\in\{(g,0),(g_c,1),(g_c,0)\},\\
&\exists\gamma^{ps}_{0,g,t,g_c}\in\Theta^{ps}:\mathbb{P}(p^{S=g,Q=q}_{g',q'}(X;\gamma^{ps}_{0,g,t,g_c}) = p^{S=g,Q=q}_{g',q'}(X)) = 1 \text{ or}\\
&\exists\gamma^{reg}_{0,g,t,g_c}\in\Theta^{reg}:\mathbb{P}(m^{S=g,Q=q}_{g',q'}(X;\gamma^{reg}_{0,g,t,g_c}) = m^{S=g,Q=q}_{g',q'}(X)) = 1.
\end{aligned} \tag{4.18}$$

Claim (4.18) states that for each (g,t)-pair and each suitable comparison group g_c, either the working parametric model for the generalized propensity score is correctly specified, or the working outcome regression model for the comparison group is correctly specified for each of the three DiD components of our DDD estimator. Thus, eight possible working model combinations would lead to consistent DDD estimation of the $ATT(g,t)$ parameter.

The next theorem establishes the limiting distribution of $\widehat{ATT}_{\text{dr},g_c}(g,t)$ and $\widehat{ATT}_{\text{dr, opt}}(g,t)$.

Theorem 4.2 (Consistency and Asymptotic Normality)**.** *Let Assumptions S, SO, NA, DDD-CPT, WM, ALR, and IC hold. Then, for all $g\in\mathcal{G}_{trt}$, $t\in\{2,\ldots,T\}$, and $g_c\in\mathcal{G}^{g,t}_c$ such that $t\geqslant g$, provided that (4.18) is true,*

$$\sqrt{n}\left(\widehat{ATT}_{dr,g_c}(g,t) - ATT(g,t)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi^{g,t}_{g_c}(W_i;\kappa^{g,t}_{0,g_c}) + o_p(1) \xrightarrow{d} N(0,\Omega_{g,t,g_c}),$$

where $\Omega_{g,t,g_c} = \mathbb{E}\left[\psi^{g,t}_{g_c}(W_i;\kappa^{g,t}_{0,g_c})\psi^{g,t}_{g_c}(W_i;\kappa^{g,t}_{0,g_c})'\right]$. Furthermore,

$$\sqrt{n}\left(\widehat{ATT}_{dr,opt}(g,t) - ATT(g,t)\right) = \frac{\boldsymbol{1}'\Omega^{-1}_{g,t}}{\boldsymbol{1}'\Omega^{-1}_{g,t}\boldsymbol{1}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi^{g,t}(W_i;\kappa^{g,t}_0) + o_p(1) \xrightarrow{d} N(0,\Omega_{g,t,opt}),$$

where $\Omega_{g,t,opt} = \left(\boldsymbol{1}'\Omega^{-1}_{g,t}\boldsymbol{1}\right)^{-1} \leqslant \Omega_{g,t,g_c}$ for any $g_c\in\mathcal{G}^{g,t}_c$. In fact, for any set of weights w that sum up to one over the $\mathcal{G}^{g,t}_c$, $\Omega_{g,t,opt}\leqslant\Omega_{g,t,w}$, with $\Omega_{g,t,w}$ defined as the asymptotic variance of $\sqrt{n}\left(\widehat{ATT}_{dr,\widehat{w}}(g,t) - ATT_{dr,w}(g,t)\right)$.

Theorem 4.2 provides the influence function for estimating each $ATT(g,t)$, using different comparison groups g_c, as well as establishes the consistency and asymptotic normality of our DR DDD

estimator $\widehat{ATT}_{\text{dr},g_c}(g,t)$. Theorem 4.2 also highlights that combining different comparison groups as our DR DDD estimator $\widehat{ATT}_{\text{dr, opt}}(g,t)$ does is effective in terms of asymptotically improving precision. That is, Theorem 4.2 highlights that $\widehat{ATT}_{\text{dr, opt}}(g,t)$ is optimal in the sense that it asymptotically achieves the minimum variance across all weighted average estimators that combine multiple $\widehat{ATT}_{\text{dr},g_c}(g,t)$s. Importantly, Theorem 4.2 also highlights the doubly (or multiply) robust property of our DDD estimators: they recover the $ATT(g,t)$ provided that each of the three DR DiD estimators has a correctly specified outcome regression or generalized propensity score working model.

Remark 4.6. Although Theorem 4.2 provides pointwise inference results for each $ATT(g,t)$, it is straightforward to extend it to hold simultaneously across multiple $ATT(g,t)$'s. For instance, by letting $\widehat{ATT}_{\text{opt},t\geqslant g}$ and $ATT_{\text{opt},t\geqslant g}$ denote the vector of $\widehat{ATT}_{\text{dr,opt}}(g,t)$ and $ATT_{\text{dr,opt}}(g,t)$, respectively, for all $g \in \mathcal{G}_{\text{trt}}$, $t \in \{2,\ldots,T,\}$ such that $t \geqslant g$, it is straightforward to show that $\sqrt{n}\left(\widehat{ATT}_{\text{opt},t\geqslant g} - ATT_{\text{opt},t\geqslant g}\right) \xrightarrow{d} N(0,\Omega)$, with $\Omega = \mathbb{E}[\psi_{\text{opt}}^{t\geqslant g}(W_i;\kappa_0^{t\geqslant g})\psi_{\text{opt}}^{t\geqslant g}(W_i;\kappa_0^{t\geqslant g})']$, with $\psi_{\text{opt}}^{t\geqslant g}(W_i;\kappa_0^{t\geqslant g})$ the asymptotic linear representation of $\sqrt{n}\left(\widehat{ATT}_{\text{opt},t\geqslant g} - ATT_{\text{opt},t\geqslant g}\right)$. One can then construct simultaneous confidence bands using a simple-to-use multiplier bootstrap as discussed in Theorem 3 and Algorithm 1 of Callaway and Sant'Anna (2021). It is also straightforward to conduct cluster-robust inference; see Remark 10 of Callaway and Sant'Anna (2021). As these results are commonly accessible, we will not include them here to conserve space.

4.2.2 Asymptotic theory for event-study parameters

In this section, we derive large sample properties for our event-study estimator $\widehat{ES}_{\text{dr,opt}}(e)$ as defined in (4.13). Given that $\mathbb{P}_n(G=g|G+e\in[1,T])$ is an $\sqrt{n}$-consistent and asymptotically normal estimator of $\mathbb{P}(G=g|G+e\in[1,T])$, then for all $g\in\mathcal{G}_{\text{trt}}$, we have that

$$\sqrt{n}(\mathbb{P}_n(G=g|G+e\in[1,T]) - \mathbb{P}(G=g|G+e\in[1,T])) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi^{g,e}(W_i) + o_p(1), \tag{4.19}$$

with $\mathbb{E}[\xi^{g,e}(W)]=0$ and $\mathbb{E}[\xi^{g,e}(W)\xi^{g,e}(W)'] < \infty$ being positive definite, and

$$\xi^{g,e}(W) = \frac{1}{\mathbb{P}(G+e\in[1,T])}\cdot\left[1\{G=g,G+e\in[1,T]\} - \mathbb{P}(G=g|G+e\in[1,T])\cdot 1\{G+e\in[1,T]\}\right]$$

The following corollary can be used to conduct asymptotically valid (pointwise) inference for the event-study type parameter $ES(e)$.

Corollary 4.2. *Under the assumptions of Theorem 4.2, for each e such that $\mathbb{P}(1\leqslant G+e\leqslant T)$, as $n\to\infty$,*

$$\begin{aligned}\sqrt{n}(\widehat{ES}_{dr,opt}(e) - ES(e)) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_{opt}^{es,e}(W_i) + o_p(1)\\ &\xrightarrow{d} N(0,\mathbb{E}[l_{opt}^{es,e}(W)^2]),\end{aligned}$$

with $l_{opt}^{es,e}(W) = \sum_{g \in \mathcal{G}_{trt}} \Big(\mathbb{P}(G = g | G + e \in [1,T]) \cdot \dfrac{\boldsymbol{1}'\Omega_{g,t}^{-1}}{\boldsymbol{1}'\Omega_{g,t}^{-1}\boldsymbol{1}} \psi^{g,t}(W_i; \kappa_0^{g,t}) + \xi^{g,e}(W_i) \cdot ATT(g,t) \Big)$.

The results in Corollary 4.2 also apply to estimators of $ES(e)$ using $\widehat{ATT}_{\text{dr},g_c}(g, g+e)$ on (4.13). Corollary 4.2 focuses on pointwise inference procedures. Still, as discussed in Remark 4.6, it is straightforward to extend it to hold for all event-times e and conduct simultaneous-based inference. The asymptotic results for our overall summary parameter $\widehat{ES}_{\text{avg, opt}}$ as defined in (4.14) follow from the delta method and are omitted.

5 Monte Carlo Simulations

In this section, we evaluate the finite sample properties of our proposed DR DDD estimators via Monte Carlo simulations. We examine two scenarios: (i) when covariates play a crucial role in identification across two time periods, and (ii) when there are multiple time periods with variation in treatment timings. For the first scenario, we have panel data for two time periods, $t = 1, 2$, four covariates, two enabling-groups $S \in \{2, \infty\}$, and there are two eligibility groups: $Q = 1$ and $Q = 0$. In the setup with staggered adoption, we consider the simplest possible case with three time periods, $t = 1, 2, 3$, with $S \in\equiv \{2, 3, \infty\}$, and we abstract from covariates in the main text. We relegate simulation results with DDD staggered adoption with covariates to the Supplemental Appendix. In the main text, we compare the performance of different DDD estimators via graphs: one that presents the density of the point estimates across the 1,000 Monte Carlo repetitions, and one that presents the length of confidence intervals in each Monte Carlo draw. In the Supplemental Appendix, we also report the traditional summary statistics for the Monte Carlo involving average bias, root mean square error (RMSE), empirical 95% coverage probability, and the average length of a 95% confidence interval under 1,000 Monte Carlo repetitions. Light-gray confidence intervals mean that they do not contain the true parameter of interest, $ATT(2,2)$ in our simulations, and are appropriately colored when they contain it. We focus on results with $n = 5,000$ but report results for different sample sizes in the Supplemental Appendix B.

5.1 Simulations for DDD with two periods and covariates

We describe the data-generating process (DGP) for the 2-period DDD setup. For a generic four-dimensional vector O, the conditional probability of each unit belonging to a subgroup $(g, q) \in \{2, \infty\} \times \{0, 1\}$ is

$$\mathbb{P}[S = g, Q = q | O] \equiv p^{S=g,Q=q}(O) = \frac{\exp(f_{S=g,Q=q}^{ps}(O))}{\sum_{(g,q) \in \mathcal{S}_{\text{des-1}} \times \{0,1\}} \exp(f_{S=g,Q=q}^{ps}(O))}, \tag{5.1}$$

where $f_{S=g,Q=q}^{ps}(O))$ is a linear index with heterogeneous coefficients across sub-groups; we defined these in the Supplemental Appendix B.1 to save space. Of course, each unit belongs to a single

subgroup, and we assigned these subgroups as follows:

$$(S,Q) := \begin{cases} (\infty,0), & \text{if } U \leqslant p^{S=\infty,Q=0}(O), \\ (\infty,1), & \text{if } p^{S=\infty,Q=0}(O) < U \leqslant \sum_{j=0}^{1} p^{S=\infty,Q=j}(O), \\ (2,0), & \text{if } \sum_{j=0}^{1} p^{S=\infty,Q=j}(O) < U \leqslant \sum_{j=0}^{1} p^{S=\infty,Q=j}(O) + p^{S=2,Q=0}(O), \\ (2,1), & \text{if } \sum_{j=0}^{1} p^{S=\infty,Q=j}(O) + p^{S=2,Q=0}(O) < U, \end{cases} \tag{5.2}$$

with U being a uniform random variable in $[0,1]$, independent of all other variables.

The potential outcomes are defined as

$$\begin{aligned} Y_{i,1}(\infty) &= f^{reg}(O_i,S_i) + \nu_i(O_i,S_i,Q_i) + \varepsilon_{i,1}(\infty) \\ Y_{i,2}(\infty) &= 2f^{reg}(O_i,S_i) + \nu_i(O_i,S_i,Q_i) + \varepsilon_{i,2}(\infty) \\ Y_{i,2}(2) &= 2f^{reg}(O_i,S_i) + \nu_i(O_i,S_i,Q_i) + \varepsilon_{i,2}(2), \end{aligned} \tag{5.3}$$

where $f^{reg}(O_i,S_i)$ is a linear regression specification with heterogeneous coefficients across the enabling groups S, $\nu_i(O_i,S_i,Q_i)$ is a time-invariant unobserved heterogeneity correlated with covariates and sub-groups, and $\varepsilon_{i,1}(\infty), \varepsilon_{i,2}(\infty)$ and $\varepsilon_{i,2}(2)$ are independent standard normal random variables; we provide a precise definition of $f^{reg}(O_i,S_i)$ and $\nu_i(O_i,S_i,Q_i)$ in the Supplemental Appendix B.1. Note that our designs' $ATT(2,2)$ equals zero, though there is treatment effect heterogeneity across units. We observe untreated outcomes for all units in period $t=1$; in period $t=2$, we observed $Y_{i,2}(2)$ if unit i belongs to group $S=2$, $Q=1$, and observe $Y_{i,=2}(\infty)$ otherwise.

Building on Kang and Schafer (2007) and Sant'Anna and Zhao (2020), we allow propensity score and/or outcome regression models to be misspecified. We consider four different types of DGP: DGP 1, where all models are correctly specified; DGP 2, where outcome models are correctly specified but the propensity score model is misspecified; DGP 3, where the propensity score is correctly specified but outcome regressions are not; and DGP 4, where all models are misspecified. The source of misspecification in these nuisance models is related to whether they depend on X or Z, where X is a nonlinear transformation of all the Z's. In our simulations, the observed data is $W_i = \{Y_{i,1}, Y_{i,2}, S_i, Q_i, X_i\}_{i=1}^n$, so using Z as linear covariates in these nuisance models lead to working model misspecification; we relegate to the Supplemental Appendix B.1 the definition of X's and Z's.

We compare the performance of four different estimators for $ATT(2,2)$, just like in Section 3.1: our DR DDD estimator as defined in (3.5) (we label it as DRDDD), 3WFE OLS estimator for $\tilde{\beta}_{\text{3wfe}}$ based on (3.3) (we label it as 3WFE), 3WFE OLS estimator of $\check{\beta}_{\text{3wfe}}$ based on (3.4) (we label it as M-3WFE), and the difference of two Sant'Anna and Zhao (2020)' DR DiD estimators (we label it as DRDID-DIF). We summarize the results of our simulations in Figure 3, where we consider a sample size $n = 5,000$ and conducted 1,000 Monte Carlo repetitions. The left panels display the density of the point estimates across all Monte Carlo draws, and the right panels display the 95% confidence intervals for each Monte Carlo draw. See Table OA-1 in the Supplemental Appendix

for additional results.

The results in Figure 3 are self-explanatory and firmly support our theoretical results. When outcome regression or propensity score models (but not necessarily both) are correctly specified, our DR DDD estimators are appropriately centered (so they are unbiased), their confidence intervals are the narrowest across all other estimators, and they still have appropriate coverage across the first three DGPs (94.4%, 94.5%, and 94.6%, respectively). For instance, when all working models are correctly specified, the average length of the 95% confidence interval of the M-3WFE estimator is 7 times longer than our DR DDD estimator; this difference is much larger for the other considered estimators. In fact, the performance of all other estimators in all our considered DGPs is poor, as they have non-negligible bias, high RMSE, and poor coverage properties. In DGP 4, when all working models are misspecified, we note that our estimator is biased, directly affecting the confidence intervals' coverage probabilities. None of the considered DDD estimators perform well when all working models are misspecified (DGP 4), highlighting that all estimators indeed depend on modeling assumptions.

5.2 Simulations for DDD with variation in treatment timing

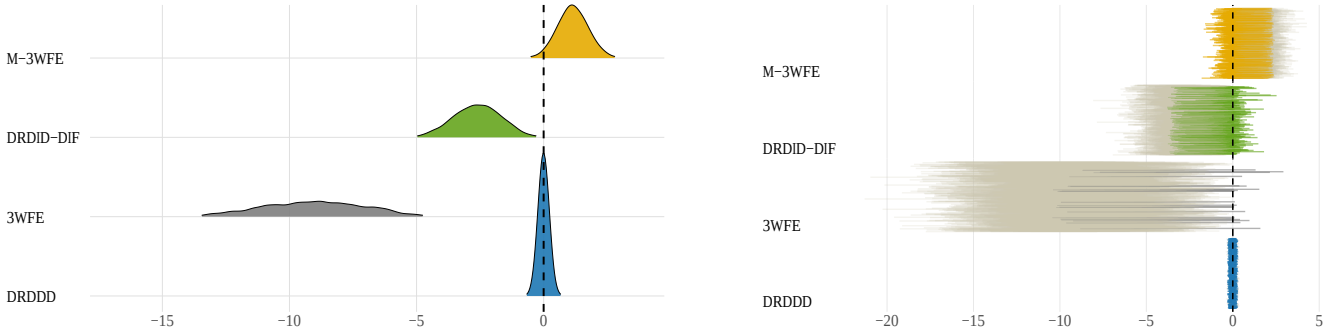
We now discuss the staggered DDD setup with three time periods, $t = 1, 2, 3$, three enabling groups, $S \in \{2, 3, \infty\}$, and two eligibility groups $Q \in \{0, 1\}$. We abstract from covariates and defer a discussion about them to the Supplemental Appendix.

Each unit i, we have that $p^{S=2,Q=0} = 0.20$, $p^{S=2,Q=1} = 0.15$, $p^{S=3,Q=0} = 0.30$, $p^{S=3,Q=1} = 0.20$, $p^{S=\infty,Q=0} = 0.05$, and $p^{S=\infty,Q=1} = 0.10$. We then randomly assign the realized value of (S, Q) based on the above distribution. The potential outcomes are generated as

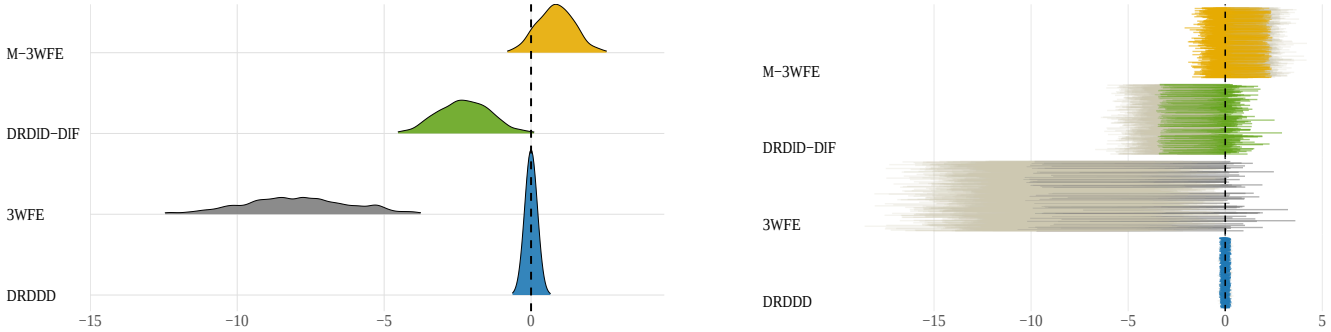
$$\begin{aligned}
Y_{i,1}(\infty) &= (1 + Q_i)\alpha + \nu_i(S_i, Q_i) + \varepsilon_{i,1}(\infty) \\
Y_{i,2}(\infty) &= (2 + Q_i)\alpha + 1.1\nu_i(S_i, Q_i) + \varepsilon_{i,2}(\infty) \\
Y_{i,3}(\infty) &= (3 + Q_i)\alpha + 1.2\nu_i(S_i, Q_i) + \varepsilon_{i,3}(\infty) \\
Y_{i,2}(2) &= (2 + Q_i)\alpha + 1.1\nu_i(S_i, Q_i) + ATT(2,2)Q_i + \varepsilon_{i,2}(2) \\
Y_{i,3}(2) &= (3 + Q_i)\alpha + 1.2\nu_i(S_i, Q_i) + ATT(2,3)Q_i + \varepsilon_{i,3}(2) \\
Y_{i,3}(3) &= (3 + Q_i)\alpha + 1.2\nu_i(S_i, Q_i) + ATT(3,3)Q_i + \varepsilon_{i,3}(3),
\end{aligned} \tag{5.4}$$

where we set $\alpha = 278.5$, $ATT(2,2) = 10$, $ATT(2,3) = 20$, $ATT(3,3) = 25$, all $\varepsilon_{i,t}(\cdot)$ are independent standard normal, and $\nu_i(S_i, Q_i)$ is an unobserved heterogeneity term that we formally define in the Supplemental Appendix B.2. The observed data is represented as $W_i = \{Y_{i,1}, Y_{i,2}, Y_{i,3}, S_i, Q_i\}_{i=1}^n$, where the $Y_{i,1} = Y_{i,1}(\infty)$, $Y_{i,2} = Y_{i,2}(2)$ for units with $S_i = 2, Q_i = 1$, and $Y_{i,2} = Y_{i,2}(\infty)$ otherwise, and $Y_{i,3} = Y_{i,3}(3)$ for units with $S_i = 3, Q_i = 1$, $Y_{i,3} = Y_{i,3}(2)$ for units with $S_i = 2, Q_i = 1$, and $Y_{i,3} = Y_{i,3}(\infty)$ for all other units. In what follows, we focus our attention on estimators for $ATT(2,2)$, though we compare estimators for $ATT(2,3)$ and $ATT(3,3)$ in the

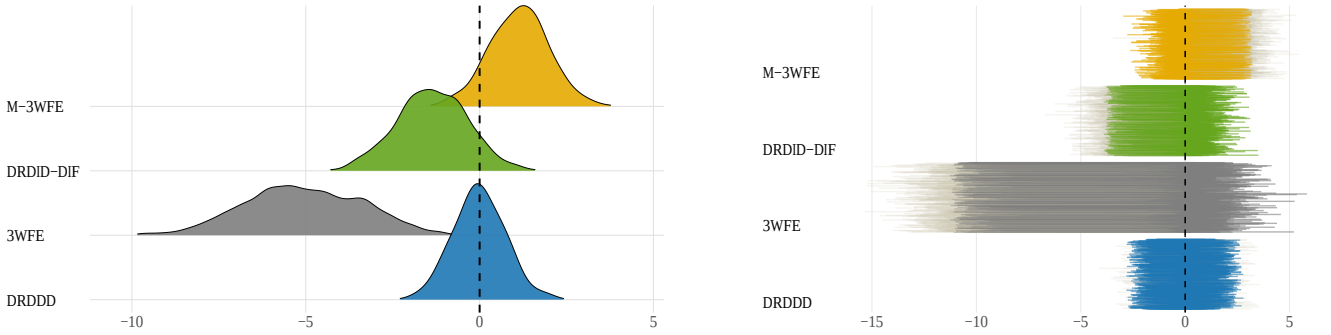
Figure 3: Monte Carlo Simulation Results for DDD: two-period setup with covariates



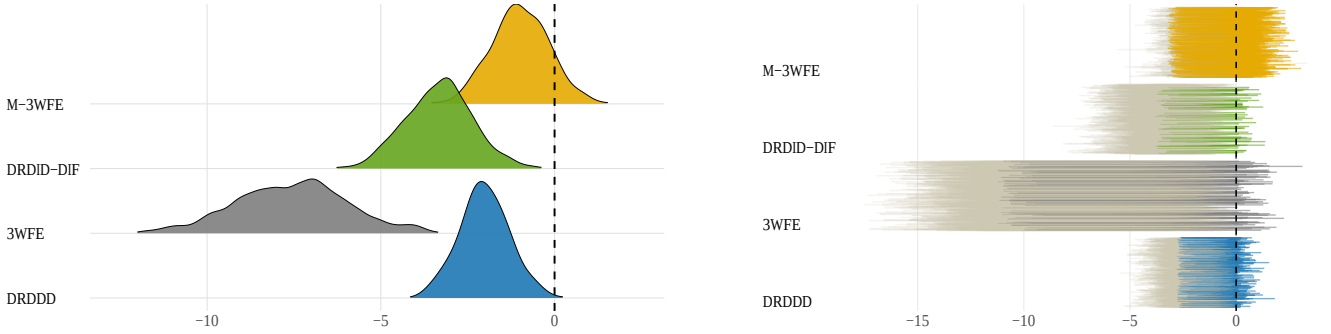
(a) DGP 1: All working models are correctly specified



(b) DGP 2: Outcome working models are correctly specified



(c) DGP 3: Propensity-score working models are correctly correct



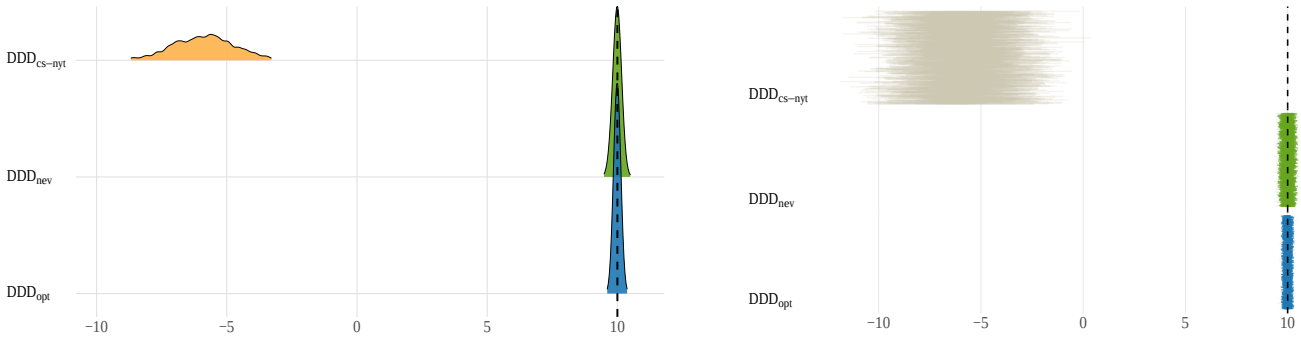
(d) DGP 4: All working models are misspecified

Notes: Simulation designs as discussed in text, with $n = 5,000$ and $1,000$ Monte Carlo repetitions. True $ATT(2,2)$ is zero and is indicated in the solid vertical line in all panels. 3WFE corresponds to the OLS estimates of $\tilde{\beta}_{3\text{wfe}}$ based on (3.3). M-3WFE corresponds to the OLS estimates of $\check{\beta}_{3\text{wfe}}$ based on (3.4). DRDID-DIF corresponds to the difference between two doubly robust DiD estimators (Sant'Anna and Zhao, 2020). DRDDD corresponds to our proposed doubly robust DDD estimator described in (3.5). All densities (left) and confidence intervals (right) are computed across all simulation draws. Light grey areas in the right plots indicate confidence intervals that exclude the true $ATT(2,2)$, where increased prominence suggests lower empirical coverage.

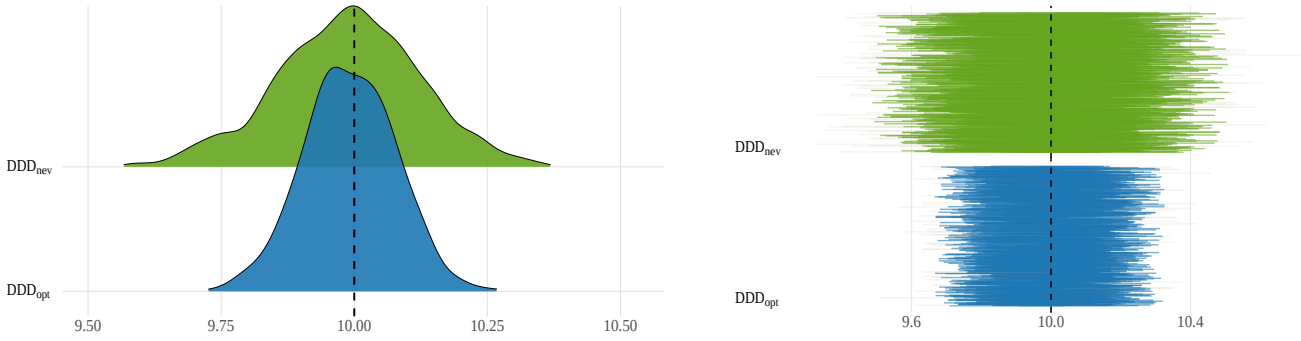
Supplemental Appendix B.2. We consider $n = 5,000$ here and refer to Supplemental Appendix B.2 for information on other sample sizes.

We compare the performance of three staggered DDD estimators for the $ATT(2,2)$ as in Section 3.2. More precisely, we consider our optimally GMM-weighted estimator $\widehat{ATT}_{\text{opt}}(g,t)$ as defined in (3.8), which is a special case of the (4.12) without covariates, i.e., with $X_i = 1$ for all units; we refer to this estimator as DDD_{opt}. We also consider our DDD estimator that uses never-treated units as the comparison group as defined in (3.7) with $g_c = \infty$; we refer to this estimator as DDD_{nev}.[11]. Finally, we consider a DDD estimator that pools all not-yet-treated units a la Callaway and Sant'Anna (2021), $\widehat{ATT}_{\text{cs-nyt}}(g,t)$, as defined in (3.6); we refer to this estimator as $DDD_{\text{cs-nyt}}$. Our theoretical results indicated that DDD_{opt} and DDD_{nev} should both be valid DDD estimators under our identification assumptions. At the same time, we have no statistical guarantee about the performance of $DDD_{\text{cs-nyt}}$. Our theoretical results also suggest that DDD_{opt} should be more precisely estimated than DDD_{nev}, as it uses more information.

Figure 4: Monte Carlo Simulation Results for DDD: staggered setup without covariates



(a) Comparison across the three DDD estimators



(b) Comparing the performance of DDD_{nev} and DDD_{opt}

Notes: Simulation designs are detailed in the text, using $n = 5,000$ and 1,000 Monte Carlo repetitions. The true $ATT(2,2)$ is 10, marked by a solid vertical line in all panels. DDD_{nev} denotes our DDD estimator with $g_c = \infty$ from Equation (4.1). DDD_{opt} is our proposed DDD estimator from Equation (4.12). DDD_{cs-nyt} is the estimator pooling all not-yet-treated units as defined in (3.6). In the top-left panel, we plot the densities of the DDD_{nev}, DDD_{opt}, and DDD_{cs-nyt} estimates across all simulation draws. The bottom-left panel zooms into the densities for DDD_{nev} and DDD_{opt}. The top-right panel shows the confidence intervals for these estimators, while the bottom-right panel focuses on DDD_{nev} and DDD_{opt} only. Light grey areas in the right plots indicate confidence intervals that exclude the true $ATT(2,2)$, where increased prominence suggests lower empirical coverage.

Figure 4 summarizes our simulation results; see also Table OA-2 in the Supplemental Appendix.

11 This estimator is a special case of (4.8) when covariates are trivial

Panel (a) of Figure 4 compares the performance of the three estimators. As it is evident, $DDD_{\text{cs-nyt}}$ is severely biased in our design, and its confidence interval never covered the true $ATT(2,2)$ in the Monte Carlo draws. This result further highlights that proceeding as if DDD is just the difference of two DiD procedures can lead to misleading conclusions. At the same time, the simulations highlight that DDD_{nev} and DDD_{opt} are unbiased and have good coverage properties, 93.2% and 95%, respectively. In Panel (b), we drop $DDD_{\text{cs-nyt}}$ and focus on comparing the performance of our two proposed DDD estimators. As the plot makes it clear, the gains in efficiency of using all not-yet-treated comparison groups as in DDD_{opt} are notable. The average length of the confidence intervals of DDD_{nev} is more than 50% higher than that of DDD_{opt}. Thus, in practice, we recommend favoring DDD_{opt}, especially when the sample size of the never-enabling comparison group is low.

6 Concluding remarks

This paper studied DDD estimators, paying close attention to situations where covariates are important for identification and to setups with staggered treatment adoption. Our findings challenge the conventional wisdom that DDD can be understood as the difference between two DiDs. We showed that when DDD-type parallel trends hold after conditioning on covariates, DDD estimators cannot generally be expressed as such, even in cases with only two time periods. In addition, when treatment adoption is staggered, pooling all not-yet-treated units a la Callaway and Sant'Anna (2021) is not generally valid, and proceeding as such can lead to misleading conclusions even when covariates are not crucial for the DDD identification arguments. These results highlight the need for more careful consideration when applying DDD strategies.

To address these challenges, we proposed DR DDD estimators that can appropriately handle covariates and can also be used in DDD setups with staggered treatment adoption. Importantly, we proposed a DR DDD estimator that leverages information across different comparison groups, and our simulation results highlighted that the gains in precision can be substantial compared to alternatives. As such, we recommend practitioners to favor our proposed DDD estimator $\widehat{ATT}_{\text{dr,opt}}(g,t)$ in applications. We are finishing an `R` package that will automate all these DDD estimators, hopefully making it easier to adopt.

References

Abadie, Alberto, "Semiparametric Difference-in-Difference Estimators," *The Review of Economic Studies*, 2005, *72*, 1–19.

Abbring, Jaap H. and Gerard J. van den Berg, "The nonparametric identification of treatment effects in duration models," *Econometrica*, 2003, *71* (5), 1491–1517.

Antwi, Yaa Akosa, Asako S Moriya, and Kosali Simon, "Effects of Federal Policy to Insure Young Adults: Evidence from the 2010 Affordable Care Act's Dependent-Coverage Mandate," *American Economic Journal: Economic Policy*, November 2013, *5* (4), 1–28.

Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna, "Difference-in-Differences Designs: A Practitioner's Guide," *arXiv:2503.13323*, 2025.

Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen, "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 2017, *85* (1), 233–298.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event Study Designs: Robust and Efficient Estimation," *Review of Economic Studies*, 2024, *91* (6), 3253–3285.

Callaway, Brantly and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of econometrics*, 2021, *225* (2), 200–230.

Caron, Laura, "Triple Difference Designs with Heterogeneous Treatment Effects," *arXiv:2502.19620*, 2025.

Chaisemartin, Clément De and Xavier d'Haultfoeuille, "Two-way fixed effects estimators with heterogeneous treatment effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

Chiu, Albert, Xingchen Lan, Ziyi Liu, and Yiqing Xu, "Causal Panel Analysis under Parallel Trends: Lessons from A Large Reanalysis Study," *American Political Science Review*, 2025, *Forthcoming.*

Garthwaite, Craig, Tal Gross, and Matthew J. Notowidigdo, "Public Health Insurance, Labor Supply, and Employment Lock *," *The Quarterly Journal of Economics*, 03 2014, *129* (2), 653–696.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277.

Gruber, Jonathan, "The Incidence of Mandated Maternity Benefits," *American Economic Review*, June 1994, *84* (3), 622–641.

Hansen, Casper Worm and Asger Mose Wingender, "National and Global Impacts of Genetically Modified Crops," *American Economic Review: Insights*, June 2023, *5* (2), 224–240.

Heckman, James J., Hidehiko Ichimura, and Petra Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 1997, *64* (4), 605–654.

Kang, Joseph D. Y. and Joseph L. Schafer, "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 2007, *22* (4), 523 – 539.

Khan, Shakeeb and Elie Tamer, "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 2010, *78* (6), 2021–2042.

Lechner, Michael, "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," *The Review of Economics and Statistics*, 2002, *84* (2), 205–220.

Malani, Anup and Julian Reif, "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform," *Journal of Public Economics*, 2015, *124*, 1–17.

Newey, Whitney K. and Daniel McFadden, "Large sample estimation and hypothesis testing," in "Handbook of Econometrics," Vol. 4, Amsterdam: North-Holland: Elsevier, 1994, chapter 36, pp. 2111–2245.

Olden, Andreas and Jarle Møen, "The triple difference estimator," *The Econometrics Journal*, 2022, *25* (3), 531–553.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," *The Review of Economic Studies*, 02 2023, *90* (5), 2555–2591.

Robins, James, "A New Approach To Causal Inference in Mortality Studies With a Sustained Exposure Period - Application To Control of the Healthy Worker Survivor Effect," *Mathematical Modelling*, 1986, *7*, 1393–1512.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe, "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," *Journal of Econometrics*, 2023, *235* (2), 2218–2244.

Sant'Anna, Pedro H.C. and Jun Zhao, "Doubly Robust Difference-in-Differences Estimators," *Journal of Econometrics*, 2020, *219* (1), 101–122.

_ **and Qi Xu**, "Difference-in-Differences with Compositional Changes," 2023. Working Paper.

Seaman, Shaun R and Stijn Vansteelandt, "Introduction to Double Robust Methods for Incomplete Data," *Statistical Science*, 2018, *33* (2), 184–197.

Słoczyński, Tymon and Jeffrey M Wooldridge, "A General Double Robustness Result for Estimating Average Treatment Effects," *Econometric Theory*, 2018, *34* (1), 112–133.

Strezhnev, Anton, "Decomposing Triple-Differences Regression under Staggered Adoption," 2023.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Słoczyński, Tymon, "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *The Review of Economics and Statistics*, 2022, *104* (3), 501—509.

_ , "When Should We (Not) Interpret Linear IV Estimands as LATE?," *arXiv:2011.06695*, 2024.

Walker, W. Reed, "The Transitional Costs of Sectoral Reallocation: Evidence From the Clean Air Act and the Workforce*," *The Quarterly Journal of Economics*, 2013, *128* (4), 1787–1835.

Wooldridge, Jeffrey M., "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 2007, *141* (2), 1281–1301.

Wooldridge, Jeffrey M, "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators," 2021. Working Paper.

Better Understanding Triple Differences Estimators: Supplemental Appendix

Marcelo Ortiz-Villavicencio Pedro H.C. Sant'Anna

May 15, 2025

This Supplementary Appendix includes: (a) proofs for the results presented in the main paper; and (b) details about the data-generating process (DGP) used in the Monte Carlo simulations to illustrate the finite sample properties of our DR DDD method.

A Proofs of Main Results

We begin by proving auxiliary lemmas that will be used later to establish the main results of the paper. Initially, consider the conditional $ATT(g,t)$, and for simplicity, we omit the unit indexing i.

$$CATT_X(g,t) \equiv \mathbb{E}[Y_t(g) - Y_t(\infty)|X, S = g, Q = 1]$$

Lemma A.1. *Let Assumptions S, SO, NA, and DDD-CPT hold. Then, for all $g \in \mathcal{G}_{trt}$, $t \in \{2,\dots,T\}$, and $g_c \in \mathcal{S}$ such that $t \geqslant g$ and $g_c > t$,*

$$\begin{aligned} CATT_X(g,t) = &(\mathbb{E}[Y_t - Y_{g-1}|X, S = g, Q = 1] - \mathbb{E}[Y_t - Y_{g-1}|X, S = g, Q = 0]) \\ &- (\mathbb{E}[Y_t - Y_{g-1}|X, S = g_c, Q = 1] - \mathbb{E}[Y_t - Y_{g-1}|X, S = g_c, Q = 0]) \ \textit{almost surely.} \end{aligned}$$

Proof. All equalities below are understood to hold almost surely (a.s.), conditioning on X throughout.

$$\begin{aligned} CATT_X(g,t) &= \mathbb{E}[Y_t(g) - Y_{g-1}(\infty)|X, S = g, Q = 1] - \mathbb{E}[Y_t(\infty) - Y_{g-1}(\infty)|X, S = g, Q = 1] \\ &= \mathbb{E}[Y_t(g) - Y_{g-1}(\infty)|X, S = g, Q = 1] - \mathbb{E}[Y_t(\infty) - Y_{g-1}(\infty)|X, S = g, Q = 0] \\ &\quad - \mathbb{E}[Y_t(\infty) - Y_{g-1}(\infty)|X, S = g', Q = 1] + \mathbb{E}[Y_t(\infty) - Y_{g-1}(\infty)|X, S = g', Q = 0] \\ &= \mathbb{E}[Y_t - Y_{g-1}|X, S = g, Q = 1] - \mathbb{E}[Y_t - Y_{g-1}|X, S = g, Q = 0] \\ &\quad - \mathbb{E}[Y_t - Y_{g-1}|X, S = g_c, Q = 1] + \mathbb{E}[Y_t - Y_{g-1}|X, S = g_c, Q = 0], \end{aligned}$$

where the first equality is derived by algebraically adding and subtracting the term $\mathbb{E}[Y_{g-1}(\infty)|X, S = g, Q = 1]$. The second equality is obtained based on Assumption DDD-CPT, and the final equality arises from the definition in (2.1) along with Assumption NA. □

Proof of Theorem 4.1

Proof. We start by showing that $ATT(g,t) = ATT_{ra,g_c}(g,t)$. Given Lemma A.1,

$$\begin{aligned}
ATT(g,t) &= \mathbb{E}[CATT_x(g,t)|S=g,Q=1] \\
&= \mathbb{E}\left[\mathbb{E}[Y_t - Y_{g-1}|X,S=g,Q=1]|S=g,Q=1\right] - \mathbb{E}\Bigg[\mathbb{E}[Y_t - Y_{g-1}|X,S=g,Q=0] \\
&\quad + \mathbb{E}[Y_t - Y_{g-1}|X,S=g_c,Q=1] - \mathbb{E}[Y_t - Y_{g-1}|X,S=g_c,Q=0]\Bigg|S=g,Q=1\Bigg] \\
&= \mathbb{E}[Y_t - Y_{g-1}|S=g,Q=1] - \mathbb{E}\Bigg[m^{S=g,Q=0}_{Y_t-Y_{g-1}}(X) \\
&\quad + m^{S=g_c,Q=1}_{Y_t-Y_{g-1}}(X) - m^{S=g_c,Q=0}_{Y_t-Y_{g-1}}(X)\Bigg|S=g,Q=1\Bigg] \\
&= \mathbb{E}\left[Y_t - Y_{g-1} - m^{S=g,Q=0}_{Y_t-Y_{g-1}}(X) - m^{S=g_c,Q=1}_{Y_t-Y_{g-1}}(X) + m^{S=g_c,Q=0}_{Y_t-Y_{g-1}}(X)|S=g,Q=1\right] \\
&= \mathbb{E}\Bigg[w^{S=g,Q=1}_{\text{trt}}(S,Q)\Big(Y_t - Y_{g-1} - m^{S=g,Q=0}_{Y_t-Y_{g-1}}(X) - m^{S=g_c,Q=1}_{Y_t-Y_{g-1}}(X) + m^{S=g_c,Q=0}_{Y_t-Y_{g-1}}(X)\Big)\Bigg]
\end{aligned}$$

Hence, we have established that $ATT(g,t) = ATT_{ra,g_c}(g,t)$. Our next objective is to demonstrate the equality $ATT(g,t) = ATT_{\text{ipw},g_c}(g,t)$. Specifically, we aim to prove that

$$\mathbb{E}[w^{S=g,Q=1}_{g',q'}(S,Q,X)\cdot(Y_t - Y_{g-1})] = \frac{\mathbb{E}[1\{S=g,Q=1\}\cdot\mathbb{E}[Y_t - Y_{g-1}|X,S=g',Q=q']]}{\mathbb{E}[1\{S=g,Q=1\}]}.$$

By LIE and conditional probabilities, we observe that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1\{S=g',Q=q'\}\cdot p^{S=q,Q=1}_{g',q'}(X)}{1-p^{S=q,Q=1}_{g',q'}(X)}\cdot(Y_t - Y_{g-1})\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[1\{S=g,Q=1\}|X]\cdot 1\{S=g',Q=q'\}}{\mathbb{E}[1\{S=g',Q=q'\}|X]}\cdot(Y_t - Y_{g-1})\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[1\{S=g,Q=1\}|X]}{\mathbb{E}[1\{S=g',Q=q'\}|X]}\mathbb{E}\left[1\{S=g',Q=q'\}\cdot(Y_t - Y_{g-1})|X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}[1\{S=g,Q=1\}|X]\cdot\mathbb{E}\left[(Y_t - Y_{g-1})|X,S=g',Q=q'\right]\right] \\
&= \mathbb{E}\left[1\{S=g,Q=1\}\cdot\mathbb{E}\left[(Y_t - Y_{g-1})|X,S=g',Q=q'\right]\right] \qquad \text{(A.1)}
\end{aligned}$$

Given that

$$\begin{aligned}
\mathbb{E}\left[\frac{1\{S=g',Q=q'\}\cdot p^{S=q,Q=1}_{g',q'}(X)}{1-p^{S=q,Q=1}_{g',q'}(X)}\right] &= \mathbb{E}\left[\frac{\mathbb{E}[1\{S=g,Q=1\}|X]\cdot 1\{S=g',Q=q'\}}{\mathbb{E}[1\{S=g',Q=q'\}|X]}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[1\{S=g,Q=1\}|X]\cdot\mathbb{E}[1\{S=g',Q=q'\}|X]}{\mathbb{E}[1\{S=g',Q=q'\}|X]}\right] \\
&= \mathbb{E}\left[\mathbb{E}[1\{S=g,Q=1\}|X]\right] \\
&= \mathbb{E}[1\{S=g,Q=1\}]. \qquad \text{(A.2)}
\end{aligned}$$

Combining (A.1) and (A.2)

$$\begin{aligned}\mathbb{E}[w^{S=g,Q=1}_{g',q'}(S,Q,X)\cdot(Y_t-Y_{g-1})] &= \frac{\mathbb{E}\left[1\{S=g,Q=1\}\cdot\mathbb{E}\left[(Y_t-Y_{g-1})|X,S=g',Q=q'\right]\right]}{\mathbb{E}[1\{S=g,Q=1\}]}\\ &= \mathbb{E}\left[\mathbb{E}[Y_t-Y_{g-1}|X,S=g',Q=q']|S=g,Q=1\right]\end{aligned}$$

Thus, analogous reasoning as in RA demonstrates that $ATT(g,t) = ATT_{\text{ipw},g_c}(g,t)$. Finally, we need to show that $ATT(g,t) = ATT_{\text{dr},g_c}(g,t)$.

$$\begin{aligned}
&ATT_{\text{dr},g_c}(g,t)\\
&= \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g,0}(S,Q,X)\right)\left(Y_t - Y_{g-1} - m^{S=g,Q=0}_{Y_t-Y_{g-1}}(X)\right)\right]\\
&\quad + \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g_c,1}(S,Q,X)\right)\left(Y_t - Y_{g-1} - m^{S=g_c,Q=1}_{Y_t-Y_{g-1}}(X)\right)\right]\\
&\quad - \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g_c,0}(S,Q,X)\right)\left(Y_t - Y_{g-1} - m^{S=g_c,Q=0}_{Y_t-Y_{g-1}}(X)\right)\right]\\
&= \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g,0}(S,Q,X)\right)(Y_t - Y_{g-1})\right]\\
&\quad + \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g_c,1}(S,Q,X)\right)(Y_t - Y_{g-1})\right]\\
&\quad - \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g_c,0}(S,Q,X)\right)(Y_t - Y_{g-1})\right]\\
&\quad - \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g,0}(S,Q,X)\right)m^{S=g,Q=0}_{Y_t-Y_{g-1}}(X)\right]\\
&\quad - \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g,0}(S,Q,X)\right)m^{S=g_c,Q=1}_{Y_t-Y_{g-1}}(X)\right]\\
&\quad + \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g,0}(S,Q,X)\right)m^{S=g_c,Q=0}_{Y_t-Y_{g-1}}(X)\right]\\
&= \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g,0}(S,Q,X)\right)(Y_t - Y_{g-1})\right]\\
&\quad - \mathbb{E}\left[\left(w^{S=g,Q=1}_{g_c,1}(S,Q,X) - w^{S=g,Q=1}_{g_c,0}(S,Q,X)\right)(Y_t - Y_{g-1})\right]\\
&\quad - \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g,0}(S,Q,X)\right)m^{S=g,Q=0}_{Y_t-Y_{g-1}}(X)\right]\\
&\quad - \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g,0}(S,Q,X)\right)m^{S=g_c,Q=1}_{Y_t-Y_{g-1}}(X)\right]\\
&\quad + \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g,0}(S,Q,X)\right)m^{S=g_c,Q=0}_{Y_t-Y_{g-1}}(X)\right]\\
&= ATT(g,t)\\
&\quad - \frac{1}{\mathbb{E}[1\{S=g,Q=1\}]}\mathbb{E}\left[\left(1\{S=g,Q=1\} - \frac{1\{S=g,Q=0\}\mathbb{E}[1\{S=g,Q=1\}|X]}{\mathbb{E}[1\{S=g,Q=0\}|X]}\right)m^{S=g,Q=0}_{Y_t-Y_{g-1}}(X)\right]\\
&\quad - \frac{1}{\mathbb{E}[1\{S=g,Q=1\}]}\mathbb{E}\left[\left(1\{S=g,Q=1\} - \frac{1\{S=g_c,Q=1\}\mathbb{E}[1\{S=g,Q=1\}|X]}{\mathbb{E}[1\{S=g_c,Q=1\}|X]}\right)m^{S=g_c,Q=1}_{Y_t-Y_{g-1}}(X)\right]\\
&\quad + \frac{1}{\mathbb{E}[1\{S=g,Q=1\}]}\mathbb{E}\left[\left(1\{S=g,Q=1\} - \frac{1\{S=g_c,Q=0\}\mathbb{E}[1\{S=g,Q=1\}|X]}{\mathbb{E}[1\{S=g_c,Q=0\}|X]}\right)m^{S=g_c,Q=0}_{Y_t-Y_{g-1}}(X)\right]\\
&= ATT(g,t) - \frac{1}{\mathbb{E}[1\{S=g,Q=1\}]}\mathbb{E}\left[(\mathbb{E}[1\{S=g,Q=1\}|X] - \mathbb{E}[1\{S=g,Q=1\}|X])\cdot m^{S=g,Q=0}_{Y_t-Y_{g-1}}(X)\right]\\
&\quad - \frac{1}{\mathbb{E}[1\{S=g,Q=1\}]}\mathbb{E}\left[(\mathbb{E}[1\{S=g,Q=1\}|X] - \mathbb{E}[1\{S=g,Q=1\}|X])\cdot m^{S=g_c,Q=1}_{Y_t-Y_{g-1}}(X)\right]\\
&\quad + \frac{1}{\mathbb{E}[1\{S=g,Q=1\}]}\mathbb{E}\left[(\mathbb{E}[1\{S=g,Q=1\}|X] - \mathbb{E}[1\{S=g,Q=1\}|X])\cdot m^{S=g_c,Q=0}_{Y_t-Y_{g-1}}(X)\right]\\
&= ATT(g,t),
\end{aligned}$$

The second and third equalities result from straightforward algebraic manipulations. The fourth equality derives from $ATT(g,t) = ATT_{\text{ipw},g_c}(g,t)$ and references equations (A.1) and (A.2). The fifth equality is a consequence of the LIE. This concludes the proof. □

Proof of Corollary 4.1

Proof. Based on results from Theorem 4.1, it is established that for any $g \in \mathcal{G}_{trt}$ and $t \in \{2,\dots,T\}$ where $t \geqslant g$, and for $g_c \in \mathcal{S}$ such that $g_c > t$, the following holds: $ATT(g,t) = ATT_{\text{dr},g_c}(g,t)$. Multiplying both sides of the previous expression by an arbitrary weight $w_{g_c}^{g,t}$ and summing over all admissible comparison groups gives

$$\begin{aligned}
\sum_{g_c \in \mathcal{G}_c^{g,t}} w_{g_c}^{g,t} ATT(g,t) &= \sum_{g_c \in \mathcal{G}_c^{g,t}} w_{g_c}^{g,t} ATT_{\text{dr},g_c}(g,t) \\
ATT(g,t) \underbrace{\sum_{g_c \in \mathcal{G}_c^{g,t}} w_{g_c}^{g,t}}_{=1} &= \sum_{g_c \in \mathcal{G}_c^{g,t}} w_{g_c}^{g,t} ATT_{\text{dr},g_c}(g,t) \\
ATT(g,t) &= \sum_{g_c \in \mathcal{G}_c^{g,t}} w_{g_c}^{g,t} ATT_{\text{dr},g_c}(g,t).
\end{aligned}$$

The left-hand side of the second equality simplifies because $ATT(g,t)$ does not depend on g_c; allowing us to take it outside the summation. Since the weights sum to one, the desired result is obtained. □

Proof of Theorem 4.2

Proof. For the case of $\widehat{ATT}_{\text{dr},g_c}(g,t)$, Theorem 4.1 establishes that $ATT(g,t)$ is point-identified for all $g \in \mathcal{G}_{trt}$, $t \in \{2,\dots,T\}$, and $g_c \in \mathcal{G}_c^{g,t}$ such that $t \geqslant g$. Additionally, we observe that $\widehat{ATT}_{\text{dr},g_c}(g,t)$ is comprised of a function involving three DR DiDs. Therefore, the asymptotic linear representation of $\sqrt{n}\left(\widehat{ATT}_{\text{dr,opt}}(g,t) - ATT(g,t)\right)$ follows from Theorem A.1(a) in Sant'Anna and Zhao (2020). This is due to the fact that $\psi_{g_c}^{g,t}(W_i; \kappa_{0,g_c}^{g,t})$ is a weighted sum of the influence functions for each DR DiD, with weights reflecting the number of units in each of them. The asymptotic normality follows from the Lindeberg-Lévy central limit theorem.

According to Corollary 4.1, the estimator $\widehat{ATT}_{\text{dr,opt}}(g,t)$ implies that $ATT(g,t)$ is over-identified. This means that different $g_c \in \mathcal{G}_c^{g,t}$ can be used, and any weighted sum of these estimands will identify our parameter of interest. For each comparison group g_c, the estimator $\widehat{ATT}_{\text{dr,opt}}(g,t)$ is determined by a function involving three DR DiD terms. Consequently, its asymptotic linear representation, as established in Theorem A.1(a) from Sant'Anna and Zhao (2020), is modified by the factor $\mathbf{1}'\Omega_{g,t}^{-1}/\mathbf{1}'\Omega_{g,t}^{-1}\mathbf{1}$. This adjustment reflects the incorporation of the $k_{g,t} \times 1$ vector comprising all $\mathbb{RIF}_{\text{dr},g_c}(g,t)$ for every $g_c \in \mathcal{G}_c^{g,t}$, given that $E[\mathbb{IF}_{dr}(g,t)] =$

0. As before, asymptotic normality follows from an implementation of the Lindeberg-Lévy central limit theorem.

To demonstrate that $\Omega_{g,t,\text{opt}} = \left(\mathbf{1}'\Omega_{g,t}^{-1}\mathbf{1}\right)^{-1} \leqslant \Omega_{g,t,g_c}$ for any $g_c \in \mathcal{G}_c^{g,t}$, we can reformulate our minimum variance problem into a GMM problem. This involves a vector of moment conditions represented as $\mathbb{E}[\mathbb{RIF}_{dr}(g,t) - \theta^{g,t}] = 0$. Under the conventional regularity conditions that underpin GMM theory (Newey and McFadden, 1994), the asymptotic variance associated with any set of weights w is expressed as:

$$\Omega_{g,t,w} = (\mathbf{1}'W\mathbf{1})^{-1}\mathbf{1}'W\Omega_{g,t}W\mathbf{1}(\mathbf{1}'W\mathbf{1})^{-1}$$

where W denotes any positive definite weight matrix. According to Efficient GMM theory, choosing $W = \Omega_{g,t}^{-1}$ minimizes the asymptotic variance, leading to $\Omega_{g,t,\text{opt}} = \left(\mathbf{1}'\Omega_{g,t}^{-1}\mathbf{1}\right)^{-1}$. More generally, for any $W > 0$, we have:

$$(\mathbf{1}'W\mathbf{1})^{-1}\mathbf{1}'W\Omega_{g,t}W\mathbf{1}(\mathbf{1}'W\mathbf{1})^{-1} - \left(\mathbf{1}'\Omega_{g,t}^{-1}\mathbf{1}\right)^{-1} > 0$$

Consequently, it follows that $\Omega_{g,t,\text{opt}} \leqslant \Omega_{g,t,w}$. □

Proof of Corollary 4.2

Proof. The population event-study parameter is defined by $ES(e) = \sum_{g\in\mathcal{G}_{trt}} \mathbb{P}(G = g|G + e \in [1,T]) \cdot ATT(g, g+e)$ and its sample analogue is provided by Equation (4.13). By adding and subtracting the term $\mathbb{P}(G = g|G + e \in [1,T]) \cdot \widehat{ATT}_{dr,opt}(g,t)$, followed by multiplying the bias term by $\sqrt{n}$, we get the following expression

$$\begin{aligned}
&\sqrt{n}\left(\widehat{ES}_{\text{dr,opt}}(e) - ES(e)\right) \\
&= \sqrt{n}\sum_{g\in\mathcal{G}_{trt}} \mathbb{P}(G = g|G + e \in [1,T]) \cdot \left(\widehat{ATT}_{dr,opt}(g,t) - ATT(g,t)\right) \\
&\quad + \sqrt{n}\sum_{g\in\mathcal{G}_{trt}} (\mathbb{P}_n(G = g|G + e \in [1,T]) - \mathbb{P}(G = g|G + e \in [1,T])) \cdot ATT(g,t) + o_p(1).
\end{aligned}$$

Given the asymptotic linear representation of both $\sqrt{n}(\widehat{ATT}_{dr,opt}(g,t) - ATT(g,t))$ and $\sqrt{n}(\mathbb{P}_n(G = g|G + e \in [1,T]) - \mathbb{P}(G = g|G + e \in [1,T]))$ as stated in Theorem 4.2 and Equation (4.19), respectively, we can substitute these expressions into the representation above and arrange the summation over $i = 1, \ldots, n$ to obtain

$$\begin{aligned}
&\sqrt{n}\left(\widehat{ES}_{\text{dr,opt}}(e) - ES(e)\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \underbrace{\Big[\sum_{g\in\mathcal{G}_{\text{trt}}} \left(\mathbb{P}(G = g|G + e \in [1,T]) \cdot \frac{\mathbf{1}'\Omega_{g,t}^{-1}}{\mathbf{1}'\Omega_{g,t}^{-1}\mathbf{1}}\psi^{g,t}(W_i;\kappa_0^{g,t}) + \xi^{g,e}(W_i) \cdot ATT(g,t)\right)\Big]}_{l_{\text{opt}}^{es,e}(W_i)} + o_p(1).
\end{aligned}$$

As before, asymptotic normality is established through the Lindeberg-Lévy central limit

theorem. This concludes the proof. □

B Additional Details about Monte Carlo Simulations

This section provides additional details about the Monte Carlo designs discussed in the main text.

B.1 More details on two-period DDD setup with covariates

As discussed in Section 5.1, we considered a two-period DDD setup with covariates where, for a generic four-dimensional vector O, the conditional probability of each unit belonging to a subgroup $(g,q) \in \{2,\infty\} \times \{0,1\}$ is

$$\mathbb{P}[S=g,Q=q|O] = p^{S=g,Q=q}(O) = \frac{\exp(f^{ps}_{S=g,Q=q}(O))}{\sum_{(g,q)\in\mathcal{S}_{\text{des-1}}\times\{0,1\}}\exp(f^{ps}_{S=g,Q=q}(O))}, \tag{B.1}$$

such that, for each $(g,q) \in \mathcal{S}_{\text{des-1}} \times \{0,1\}$, we define the linear process $f^{ps}_{S=g_c,Q=q}(O) = cO'\gamma_{g_c,q}$, where c is a scalar that guarantees reasonable overlap, and $\gamma_{g_c,q}$ is a 4×1 vector of coefficients for each variable in O given by

$$\gamma_{\infty,0} = \begin{pmatrix} -1 \\ 0.5 \\ -0.25 \\ -0.1 \end{pmatrix}; \gamma_{\infty,1} = \begin{pmatrix} -0.5 \\ 2 \\ 0.5 \\ -0.2 \end{pmatrix}; \gamma_{2,0} = \begin{pmatrix} 3 \\ -1.5 \\ 0.75 \\ -0.3 \end{pmatrix}; \text{and } c = \begin{cases} 0.2, & \text{if } S=\infty, Q=0 \\ 0.2, & \text{if } S=\infty, Q=1 \\ 0.05, & \text{if } S=2, Q=0 \end{cases}$$

In this design, the probabilities for each subgroup are denoted as $\{p^{S=2,Q=0}(O), p^{S=2,Q=1}(O), p^{S=\infty,Q=0}(O), p^{S=\infty,Q=1}(O)\}$, and their sum is equal to one. Define the random variable $U \sim \text{Uniform}[0,1]$. The assignment process to each group is defined as in (5.2).

The potential outcomes are defined as in (5.3). We define $f^{reg}(O,S)$ as

$$f^{reg}(O,S) = 1\{S_i=2\}\cdot f^{reg}_{S=2}(O) + 1\{S_i=\infty\}\cdot f^{reg}_{S=\infty}(O),$$

with $f^{reg}_{S=g}(O) = \alpha + O'\beta_g$, with $\alpha = 2010$, $\beta_2 = (27.4, 13.7, 13.7, 13.7)'$ and $\beta_\infty = 0.5\beta_2$. We define the unobserved heterogeneity term $\nu_i(O_i, S_i, Q_i)$ as

$$\nu_i(O_i,S_i,Q_i) \overset{d}{\sim} N(1\{S_i=2\}Q_i f^{reg}_{S=2}(O_i) + 1\{S_i=\infty\}Q_i f^{reg}_{S=\infty}(O_i), 1).$$

Finally, as discussed in our main text, we allow misspecification of propensity score and/or outcome regression models. In all four DGPs, the observed data is $W_i = \{Y_{i,1}, Y_{i,2}, S_i, Q_i, X_i\}_{i=1}^n$, where the covariates X_i are generated as follows. Let $Z_i = (Z_{i,1}, Z_{i,2}, Z_{i,3}, Z_{i,4})' \overset{d}{\sim} N(0, I_4)$. The observed vector of covariates $X_i = (X_{i,1}, X_{i,2}, X_{i,3}, X_{i,4})'$ where for every $k = 1, \cdots, 4$,

$X_k = (\tilde{X}_k - \mathbb{E}[\tilde{X}_k]) \Big/ \sqrt{\text{Var}(\tilde{X}_k)}$ such that

$$\begin{aligned}
\tilde{X}_1 &= \exp(0.5Z_1), \\
\tilde{X}_2 &= 10 + Z_2 / (1 + \exp(Z_1)), \\
\tilde{X}_3 &= (0.6 + Z_1 Z_3/25)^3, \\
\tilde{X}_4 &= (20 + Z_1 + Z_4)^2.
\end{aligned}$$

These transformations build on Kang and Schafer (2007) and Sant'Anna and Zhao (2020). We consider four different DGPs depending on whether the propensity score and/or the outcome regressions are misspecified. The specifics are outlined below:

- *DGP 1*: All nuisance functions depend on X, namely $f^{reg}_{S=g}(X)$ and $f^{ps}_{S=g,Q=q}(X)$. All working models are correctly specified in this DGP, as X is observed in the data.
- *DGP 2*: Regression models depend on X, $f^{reg}_{S=g}(X)$, and the propensity score depends on Z $f^{ps}_{S=g,Q=q}(Z)$. The working model for propensity score is misspecified, whereas the working models for the outcomes are correctly specified.
- *DGP 3*: Regression models depend on Z, $f^{reg}_{S=g}(Z)$, and propensity score depends on X, $f^{ps}_{S=g,Q=q}(X)$. The working model for propensity score is correctly specifies, whereas the working models for the outcomes are misspecified.
- *DGP 4*: All nuisances functions depend on Z, namely $f^{reg}_{S=g}(Z)$ and $f^{ps}_{S=g,Q=q}(Z)$. All working models are misspecified.

We graphically illustrate the results in Figure 3 when $n = 5,000$. The following table presents the traditional summary statistics for the Monte Carlo involving average bias, root mean square error (RMSE), empirical 95% coverage probability, and the average length of a 95% confidence interval under 1,000 Monte Carlo repetitions. We report these results for $n = 1,000$, $n = 5,000$, and $n = 10,000$. These results are self-explanatory.

Table OA-1: Monte Carlo results for $ATT(2,2)$ in DGP 1 – DGP 4, with two periods and covariates

	DGP 1				DGP 2				DGP 3				DGP 4			
Estimator	Bias	RMSE	Cov. 95	ALCI	Bias	RMSE	Cov. 95	ALCI	Bias	RMSE	Cov. 95	ALCI	Bias	RMSE	Cov. 95	ALCI
							$n = 1000$									
DRDDD	-0.003	0.188	0.953	0.732	-0.008	0.185	0.944	0.729	0.032	1.606	0.952	6.310	-2.072	2.629	0.747	6.304
3WFE	-9.298	10.301	0.674	22.431	-8.038	9.090	0.780	22.426	-4.985	6.160	0.977	24.794	-7.755	8.574	0.894	24.714
DRDID-DIF	-2.681	3.361	0.762	8.100	-2.302	3.016	0.798	7.862	-1.280	2.515	0.911	8.514	-3.402	4.054	0.628	8.332
M-3WFE	1.192	1.809	0.855	5.275	0.865	1.621	0.903	5.331	1.188	2.219	0.898	7.177	-1.076	2.185	0.900	7.161
							$n = 5000$									
DRDDD	-0.002	0.083	0.944	0.324	0.000	0.084	0.951	0.323	-0.014	0.746	0.939	2.794	-2.019	2.141	0.190	2.792
3WFE	-9.059	9.291	0.022	10.055	-7.929	8.156	0.069	10.065	-5.049	5.324	0.601	11.125	-7.563	7.733	0.101	11.097
DRDID-DIF	-2.616	2.782	0.203	3.635	-2.270	2.442	0.283	3.520	-1.372	1.706	0.698	3.832	-3.326	3.463	0.064	3.755
M-3WFE	1.154	1.293	0.532	2.358	0.847	1.050	0.708	2.387	1.137	1.419	0.710	3.206	-0.985	1.284	0.777	3.197
							$n = 10000$									
DRDDD	0.001	0.058	0.954	0.229	0.002	0.060	0.942	0.228	-0.008	0.495	0.944	1.973	-2.013	2.078	0.024	1.973
3WFE	-9.248	9.349	0.000	7.106	-7.970	8.091	0.001	7.119	-5.108	5.231	0.146	7.870	-7.597	7.677	0.000	7.852
DRDID-DIF	-2.686	2.759	0.010	2.568	-2.295	2.381	0.036	2.489	-1.405	1.556	0.480	2.716	-3.339	3.404	0.001	2.660
M-3WFE	1.153	1.226	0.223	1.667	0.829	0.935	0.520	1.688	1.138	1.272	0.499	2.267	-0.957	1.127	0.619	2.260

Notes: This table summarizes the Monte Carlo experiments for four distinct DGPs as discussed in Section 5.1. The DRDDD row includes our proposed doubly robust estimators with $g_c = \infty$. The 3WFE and M-3WFE rows display OLS estimates for 3WFE models in equations (3.3) and (3.4). The DRDID-DIF row shows the difference between two Doubly Robust DiDs. Columns represent average bias, RMSE, coverage probability at 95% (Cov. 95), and average confidence interval length (ALCI) for each estimator. The 95% confidence intervals use point-wise asymptotic critical values. Results span sample sizes $n = \{1{,}000, 5{,}000, 10{,}000\}$ over 1,000 simulations, with the true $ATT(2,2)$ being zero.

B.2 More details on staggered DDD setups

We now provide more details about the simulation designs for staggered DDD as discussed in Section 5.2. The potential outcomes are as defined in (5.4) with the unobserved heterogeneity term $\nu_i(S_i, Q_i)$ being defined as

$$\nu_i(S_i, Q_i) \overset{d}{\sim} N(1\{S_i = 2\}(2 + Q_i)\alpha + 1\{S_i = 2\}(2 + Q_i)\alpha + 1\{S_i = \infty\}Q_i\alpha, 1), \tag{B.2}$$

where $\alpha = 278.5$. All the other relevant information is described in Section 5.2.

As we summarize the simulation results in Section 5.2 via graphs, below we present a more traditional summary of our simulation results using a table. We stress that, in period 3, all the considered DDD estimators coincide as there is only one possible comparison group at that time period, the never-enabling units. This explains the results for $ATT(2,3)$ and $ATT(3,3)$ in Table OA-2. The $ATT(2,2)$ results are self-explanatory and highlight that (a) proceeding as if DDD is just the difference between two DiD procedures can lead to misleading conclusions, (b) using our proposed DDD estimators bypasses these limitations, and (c) the gains in precision of using our optimally weighted estimator can be large.

Table OA-2: Monte Carlo results for Staggered DDD without covariates

	$ATT(2,2)$				$ATT(2,3)$				$ATT(3,3)$			
Estimator	Bias	RMSE	Cov. 95	ALCI	Bias	RMSE	Cov. 95	ALCI	Bias	RMSE	Cov. 95	ALCI
						$n = 1000$						
DDD_{opt}	-0.003	0.192	0.956	0.748	-0.004	0.298	0.947	1.140	-0.007	0.286	0.941	1.083
DDD_{nev}	-0.001	0.287	0.947	1.131	-0.004	0.298	0.947	1.140	-0.007	0.286	0.941	1.083
DDD_{cs-nyt}	-15.815	16.034	0.000	10.833	-0.004	0.298	0.947	1.140	-0.007	0.286	0.941	1.083
						$n = 5000$						
DDD_{opt}	-0.009	0.086	0.950	0.335	-0.007	0.134	0.938	0.511	0.001	0.127	0.935	0.487
DDD_{nev}	-0.012	0.135	0.932	0.507	-0.007	0.134	0.938	0.511	0.001	0.127	0.935	0.487
DDD_{cs-nyt}	-15.875	15.919	0.000	4.845	-0.007	0.134	0.938	0.511	0.001	0.127	0.935	0.487
						$n = 10000$						
DDD_{opt}	0.001	0.062	0.941	0.237	0.001	0.097	0.940	0.361	-0.001	0.090	0.941	0.344
DDD_{nev}	0.000	0.094	0.941	0.359	0.001	0.097	0.940	0.361	-0.001	0.090	0.941	0.344
DDD_{cs-nyt}	-15.897	15.918	0.000	3.428	0.001	0.097	0.940	0.361	-0.001	0.090	0.941	0.344

Notes: This table summarizes the Monte Carlo experiments for the DGP as discussed in Section 5.2. DDD_{nev} denotes our DDD estimator with $g_c = \infty$ from Equation (4.1). DDD_{opt} is our proposed DDD estimator from Equation (4.12). DDD_{cs-nyt} is the estimator pooling all not-yet-treated units as defined in (3.6). Columns represent average bias, RMSE, coverage probability at 95% (Cov. 95), and average confidence interval length (ALCI) for each estimator. The 95% confidence intervals use point-wise asymptotic critical values. Results span sample sizes $n = \{1{,}000, 5{,}000, 10{,}000\}$ over 1,000 simulations, with the true $ATT(2,2) = 10$, $ATT(2,3) = 20$, and $ATT(3,3) = 25$.

B.3 Staggered DDD setups with covariates

In this section, we expand on the DGP mentioned in Section 5.2 by incorporating covariates into the analysis. Similar to before, there are three time periods $t = 1, 2, 3$, three enabling groups $S \in \{2, 3, \infty\}$, and two eligibility groups $Q \in \{0, 1\}$.

Without loss of generality, we define $f^{ps}_{S=g,Q=q}(W) = c_q W^\top \gamma_g, \forall (g, q) \in \mathcal{S}_{des-2} \times \{0, 1\}$, where c_q is a scalar controlling the overlap condition on the propensity scores, and γ_g is a 4×1 vector of coefficients for each variable in W. For an arbitrary four-dimensional vector O, the conditional probability of each unit belonging to a subgroup $(g, q) \in \mathcal{S}_{des-2} \times \{0, 1\}$ is

$$\mathbb{P}[G = g, Q = q \mid O] = p^{S=g,Q=q}(O) = \frac{exp(f^{ps}_{S=g,Q=q}(O))}{\sum_{(g,q)\in\mathcal{S}_{des-2}\times\{0,1\}} exp(f^{ps}_{S=g,Q=q}(O))} \tag{B.3}$$

Then, we set $c_q = 0.4$ if $Q = 0$, $c_q = -0.4$ if $Q = 1$ and,

$$\gamma_2 = \begin{pmatrix} -1 \\ 0.5 \\ -0.25 \\ -0.1 \end{pmatrix}; \gamma_3 = \begin{pmatrix} -0.5 \\ 1 \\ -0.1 \\ -0.25 \end{pmatrix}; \gamma_\infty = \begin{pmatrix} -0.25 \\ 0.1 \\ -1 \\ -0.1 \end{pmatrix}$$

In this design, the probabilities for each subgroup are denoted as $\big\{p^{S=2,Q=0}(V), p^{S=2,Q=1}(V), p^{S=3,Q=0}(V), p^{S=3,Q=1}(V), p^{S=\infty,Q=0}(V), p^{S=\infty,Q=1}(V)\big\}$, which sum to one. Define the random variable $U \sim \text{Uniform}[0, 1]$. The assignment process to each group is determined as follows:

$$(S, Q) := \begin{cases} (\infty, 0), & \text{if } U \leqslant p^{S=\infty,Q=0}(O), \\ (\infty, 1), & \text{if } U \leqslant \sum_{j=0}^{1} p^{S=\infty,Q=j}(O), \\ (2, 0), & \text{if } U \leqslant \sum_{j=0}^{1} p^{S=\infty,Q=j}(O) + p^{S=2,Q=0}(V), \\ (2, 1), & \text{if } U \leqslant \sum_{j=0}^{1} p^{S=\infty,Q=j}(O) + \sum_{j=0}^{1} p^{S=2,Q=j}(O), \\ (3, 0), & \text{if } U \leqslant \sum_{j=0}^{1} p^{S=\infty,Q=j}(O) + \sum_{j=0}^{1} p^{S=2,Q=j}(O) + p^{S=3,Q=0}(O), \\ (3, 1), & \text{if } \sum_{j=0}^{1} p^{S=\infty,Q=j}(O) + \sum_{j=0}^{1} p^{S=2,Q=j}(O) + p^{S=3,Q=0}(O) < U. \end{cases} \tag{B.4}$$

Next, we define the outcome regression component of our DGP. In this model, covariates O enter the working model linearly. Let $f^{reg}(O_i) = \alpha + O'\beta$, where α is a scalar and β is a 4×1 vector of coefficients for each variable in O. We set $\alpha = 210$ and $\beta = (27.4, 13.7, 13.7, 13.7)'$. The untreated potential outcomes (which are observed for all units) at period $t = 1$ are defined as follows:

$$Y_{i,1}(\infty) = (1 + Q_i) \cdot f^{reg}(O_i) + \nu_i(O_i, S_i, Q_i) + \varepsilon_{i,1}(\infty)$$

where $\nu_i(O_i, S_i, Q_i)$ denotes a time-invariant unobserved heterogeneity with $\nu_i(O_i, S_i, Q_i) \sim$

$N(M_i \cdot f^{reg}(O_i) + Q_i \cdot f^{reg}(O_i), 1)$ where $M_i = S_i$ if $S_i \in \mathcal{G}_{trt}$, zero otherwise, and $\varepsilon_{i,1}(\infty) \sim N(0,1)$. Subsequently, we generate the potential outcomes at period $t = 2$ for each $g \in \mathcal{G}$:

$$\begin{aligned}
Y_{i,2}(\infty) &= (2 + Q_i) f^{reg}(O_i) + 2\nu_i(O_i, S_i, Q_i) + \varepsilon_{i,2}(\infty) \\
Y_{i,2}(2) &= (2 + Q_i) f^{reg}(O_i) + 2\nu_i(O_i, S_i, Q_i) + Q_i \cdot ATT(2,2) + \varepsilon_{i,2}(2) \\
Y_{i,2}(3) &= (2 + Q_i) f^{reg}(O_i) + 2\nu_i(O_i, S_i, Q_i) + \varepsilon_{i,2}(3)
\end{aligned}$$

where $\varepsilon_{i,2}(\cdot) \sim N(0,1)$. The realized outcomes at $t = 2$ are given by

$$Y_{i,2} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,2}(g)$$

Then, we generate the potential outcomes in the period $t = 3$ for each $g \in \mathcal{G}$:

$$\begin{aligned}
Y_{i,3}(\infty) &= (3 + Q_i) f^{reg}(O_i) + 3\nu_i(O_i, S_i, Q_i) + \varepsilon_{i,3}(\infty) \\
Y_{i,3}(2) &= (3 + Q_i) f^{reg}(O_i) + 3\nu_i(O_i, S_i, Q_i) + Q_i \cdot ATT(2,3) + \varepsilon_{i,3}(2) \\
Y_{i,3}(3) &= (3 + Q_i) f^{reg}(O_i) + 3\nu_i(O_i, S_i, Q_i) + Q_i \cdot ATT(3,3) + \varepsilon_{i,3}(3)
\end{aligned}$$

where $\varepsilon_{i,3}(\cdot) \sim N(0,1)$. The realized outcomes at $t = 3$ are given by

$$Y_{i,3} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,3}(g)$$

Finally, we set $ATT(2,2) = 10$, $ATT(2,3) = 20$ and $ATT(3,3) = 25$. Given that we account for potential misspecification of nuisance functions, we can proceed as outlined in Section B.1. The observed data is $W_i = \{Y_{i,1}, Y_{i,2}, Y_{i,3}, S_i, Q_i, X_i\}_{i=1}^n$, where covariates X_i and Z_i are generated in the same manner as described in Section B.1. Depending on whether the propensity score and/or outcome regression are misspecified, these specifications will result in four distinct DGPs. Table OA-3 presents summary statistics from 1,000 Monte Carlo experiments, which include average bias, root mean squared error (RMSE), coverage probability at 95%, and average confidence interval length. We provide these statistics for the $ATT(2,2)$ under the design mentioned in the current section with sample sizes $n = 1,000$, $n = 5,000$, and $n = 10,000$. Results for $ATT(2,3)$ and $ATT(3,3)$ are omitted to save space, as they produce similar patterns. These results clearly illustrate the effectiveness of our proposed estimator and align with findings from the previous sections.

Table OA-3: Monte Carlo results for $ATT(2,2)$ in DGP 1 – DGP 4, with multiple periods and covariates

	DGP 1				DGP 2				DGP 3				DGP 4			
Estimator	Bias	RMSE	Cov. 95	ALCI	Bias	RMSE	Cov. 95	ALCI	Bias	RMSE	Cov. 95	ALCI	Bias	RMSE	Cov. 95	ALCI
							$n = 1000$									
$DRDDD_{opt}$	0.009	0.236	0.943	0.901	-0.002	0.230	0.945	0.893	-0.175	3.300	0.943	12.396	-1.499	3.765	0.917	12.722
$DRDDD_{nev}$	0.012	0.298	0.939	1.137	-0.002	0.288	0.937	1.121	-0.367	3.984	0.937	14.726	0.947	3.956	0.937	14.403
$DRDDD_{cs-nyt}$	-8.877	9.294	0.079	10.298	-8.138	8.546	0.123	10.280	-8.962	9.891	0.404	16.173	-11.244	12.075	0.232	16.642
DRCS-DIF	-3.765	4.668	0.704	10.348	-4.377	5.109	0.608	10.116	-3.919	5.355	0.789	14.263	-6.866	7.767	0.520	14.161
							$n = 5000$									
$DRDDD_{opt}$	0.006	0.106	0.947	0.407	-0.002	0.105	0.935	0.399	0.083	1.450	0.953	5.629	-1.170	1.956	0.856	5.954
$DRDDD_{nev}$	0.005	0.131	0.949	0.507	-0.005	0.131	0.944	0.497	0.012	1.746	0.948	6.639	1.128	2.032	0.900	6.568
$DRDDD_{cs-nyt}$	-8.949	9.028	0.000	4.557	-8.063	8.140	0.000	4.547	-8.636	8.821	0.005	7.192	-11.009	11.173	0.002	7.576
DRCS-DIF	-3.793	3.983	0.108	4.566	-4.339	4.483	0.024	4.479	-3.844	4.165	0.333	6.356	-6.720	6.924	0.013	6.280
							$n = 10000$									
$DRDDD_{opt}$	-0.001	0.073	0.955	0.289	-0.004	0.073	0.948	0.283	0.047	1.021	0.951	4.004	-1.088	1.544	0.820	4.249
$DRDDD_{nev}$	-0.004	0.091	0.956	0.359	-0.004	0.092	0.937	0.351	0.036	1.201	0.950	4.711	1.159	1.658	0.846	4.666
$DRDDD_{cs-nyt}$	-8.928	8.965	0.000	3.226	-8.111	8.151	0.000	3.211	-8.746	8.842	0.000	5.090	-11.108	11.188	0.000	5.373
DRCS-DIF	-3.770	3.861	0.004	3.221	-4.347	4.415	0.000	3.170	-3.896	4.051	0.063	4.511	-6.852	6.941	0.000	4.419

Notes: This table presents the Monte Carlo experiments for the DGPs as detailed in the text. DDD_{nev} denotes our DDD estimator with $g_c = \infty$ from Equation (4.1). DDD_{opt} is our proposed DDD estimator from Equation (4.12). DDD_{cs-nyt} is the estimator pooling all not-yet-treated units as defined in (3.6). *DRCS-DIF* uses the differences between two DiDs following Callaway and Sant'Anna (2021) as described in Remark 4.1. Each column shows average bias, RMSE, coverage probability at 95% (Cov. 95), and average confidence interval length (ALCI) for each estimator, using point-wise asymptotic critical values for the confidence intervals. The results cover sample sizes $n = \{1{,}000, 5{,}000, 10{,}000\}$ over 1,000 simulations, with the true $ATT(2,2) = 10$.

References

Callaway, Brantly and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of econometrics*, 2021, *225* (2), 200–230.

Kang, Joseph D. Y. and Joseph L. Schafer, "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 2007, *22* (4), 523 – 539.

Newey, Whitney K. and Daniel McFadden, "Large sample estimation and hypothesis testing," in "Handbook of Econometrics," Vol. 4, Amsterdam: North-Holland: Elsevier, 1994, chapter 36, pp. 2111–2245.

Sant'Anna, Pedro H.C. and Jun Zhao, "Doubly Robust Difference-in-Differences Estimators," *Journal of Econometrics*, 2020, *219* (1), 101–122.