Journal of Econometrics 225 (2021) 200–230

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Difference-in-Differences with multiple time periods☆

Brantly Callaway [a], Pedro H.C. Sant'Anna [b,*]

[a] Department of Economics, University of Georgia, United States of America
[b] Department of Economics, Vanderbilt University, United States of America

A R T I C L E   I N F O

Article history:
Received 1 March 2019
Received in revised form 18 August 2020
Accepted 1 December 2020
Available online 17 December 2020

JEL classification:
C14
C21
C23
J23
J38

Keywords:
Difference-in-Differences
Dynamic treatment effects
Doubly robust
Event study
Variation in treatment timing
Treatment effect heterogeneity
Semi-parametric

A B S T R A C T

In this article, we consider identification, estimation, and inference procedures for treatment effect parameters using Difference-in-Differences (DiD) with (i) multiple time periods, (ii) variation in treatment timing, and (iii) when the "parallel trends assumption" holds potentially only after conditioning on observed covariates. We show that a family of causal effect parameters are identified in staggered DiD setups, even if differences in observed characteristics create non-parallel outcome dynamics between groups. Our identification results allow one to use outcome regression, inverse probability weighting, or doubly-robust estimands. We also propose different aggregation schemes that can be used to highlight treatment effect heterogeneity across different dimensions as well as to summarize the overall effect of participating in the treatment. We establish the asymptotic properties of the proposed estimators and prove the validity of a computationally convenient bootstrap procedure to conduct asymptotically valid simultaneous (instead of pointwise) inference. Finally, we illustrate the relevance of our proposed tools by analyzing the effect of the minimum wage on teen employment from 2001–2007. Open-source software is available for implementing the proposed methods.

1. Introduction

Difference-in-Differences (DiD) has become one of the most popular research designs used to evaluate causal effects of policy interventions. In its canonical format, there are two time periods and two groups: in the first period no one is treated, and in the second period some units are treated (the treated group), and some units are not (the comparison group). If, in the absence of treatment, the average outcomes for treated and comparison groups would have followed parallel paths over time (which is the so-called parallel trends assumption), one can estimate the average treatment effect for the treated subpopulation (ATT) by comparing the average change in outcomes experienced by the treated group to the average change in outcomes experienced by the comparison group. Methodological extensions of DiD methods often focus on this standard two periods, two groups setup; see, e.g., Heckman et al. (1997, 1998), Abadie (2005), Athey and Imbens (2006), Qin and Zhang (2008), Bonhomme and Sauder (2011), de Chaisemartin and D'Haultfœuille (2017), Botosaru and Gutierrez (2018), Callaway et al. (2018), and Sant'Anna and Zhao (2020).[1]

☆ A previous version of this paper has been circulated with the title "Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment". Code to implement the methods proposed in the paper is available in the R package `did` which is available on CRAN.

* Corresponding author.
E-mail addresses: brantly.callaway@uga.edu (B. Callaway), pedro.h.santanna@vanderbilt.edu (P.H.C. Sant'Anna).

1 See Section 6 of Athey and Imbens (2006) and Theorem S1 in de Chaisemartin and D'Haultfœuille (2017) for notable exceptions that cover multiple periods and multiple groups.

https://doi.org/10.1016/j.jeconom.2020.12.001
0304-4076/

Many DiD empirical applications, however, deviate from the canonical DiD setup and have more than two time periods and variation in treatment timing. In this article, we provide a unified framework for average treatment effects in DiD setups with multiple time periods, variation in treatment timing, and when the parallel trends assumption holds potentially only after conditioning on observed covariates. We concentrate our attention on DiD with staggered adoption, i.e., to DiD setups such that once units are treated, they remain treated in the following periods.

The core of our proposal relies on separating the DiD analysis into three separate steps: (*i*) identification of policy-relevant disaggregated causal parameters; (*ii*) aggregation of these parameters to form summary measures of the causal effects; and (*iii*) estimation and inference about these different target parameters. Our approach allows for estimation and inference on interpretable causal parameters allowing for arbitrary treatment effect heterogeneity and dynamic effects, thereby completely avoiding the issues of interpreting results of standard two-way fixed effects (TWFE) regressions as causal effects in DiD setups as pointed out by Borusyak and Jaravel (2017), de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2019), Sun and Abraham (2020), and Athey and Imbens (2018). In addition, it adds transparency and objectivity to the analysis (Rubin, 2007, 2008), and allows researchers to exploit a variety of estimation methods to answer different questions of interest.

The identification step of the analysis provides a blueprint for the other steps. In this paper, we pay particular attention to the disaggregated causal parameter that we call the *group-time average treatment effect*, i.e., the average treatment effect for group g at time t, where a "group" is defined by the time period when units are first treated. In the canonical DiD setup with two periods and two groups, these parameters reduce to the ATT which is typically the parameter of interest in that setup. An attractive feature of the group-time average treatment effect parameters is that they do not directly restrict heterogeneity with respect to observed covariates, the period in which units are first treated, or the evolution of treatment effects over time. As a consequence, these easy-to-interpret causal parameters can be directly used for learning about treatment effect heterogeneity, and/or to construct many other more aggregated causal parameters. We view this level of generality and flexibility as one of the main advantages of our proposal.

We provide sufficient conditions related to treatment anticipation behavior and conditional parallel trends under which these group-time average treatment effects are nonparametrically point-identified. A unique feature of our framework is that it shows how researchers can flexibly incorporate covariates into the staggered DiD setup with multiple groups and multiple periods. This is particularly important in applications in which differences in observed characteristics create non-parallel outcome dynamics between different groups – in this case, unconditional DiD strategies are generally not appropriate to recover sensible causal parameters of interest (Heckman et al., 1997, 1998; Abadie, 2005). We propose three different types of DiD estimands in staggered treatment adoption setups: one based on outcome regressions (Heckman et al., 1997, 1998), one based on inverse probability weighting (Abadie, 2005), and one based on doubly-robust methods (Sant'Anna and Zhao, 2020). We provide versions of these estimands both for the case with panel data and for the case with repeated cross sections data. To the best of our knowledge, this paper is the first to show how one can allow for covariate-specific trends across groups in DiD setups with variation in treatment timing. Our results also highlight that, in practice, one can rely on different types of parallel trends assumptions and allow some types of treatment anticipation behavior; our proposed estimands explicitly reflect these assumptions.

Our framework acknowledges that in some applications there may be many group-time average treatment effects and researchers may want to aggregate them into different summary causal effect measures. This characterizes the aggregation step of the analysis. We provide ways to aggregate the potentially large number of group-time average treatment effects into a variety of intuitive summary parameters and discuss specific aggregation schemes that can be used to highlight different sources of treatment effect heterogeneity across groups and time periods. In particular, we consider aggregation schemes that deliver a single overall treatment effect parameter with similarities to the ATT in the two period and two group case as well as partial aggregations that highlight heterogeneity along certain dimensions such as (*a*) how average treatment effects vary with length of exposure to the treatment (event-study-type estimands); (*b*) how average treatment effects vary across treatment groups; and (*c*) how cumulative average treatment effects evolve over calendar time. We also provide a formal discussion of the costs and benefits of balancing the sample in "event time" when analyzing dynamic treatment effects. Overall, our setup makes it clear that, in general, the "best" aggregation scheme is application-specific as it depends on the type of question one wants to answer.

Given that our identification results are constructive, we propose easy-to-use plug-in type (parametric) estimators for the causal parameters of interest. Although the outcome regression, inverse probability weighting and doubly-robust estimands are equivalent from the identification point of view, they suggest different types of DiD estimators one can use in practice. Here, we note that using doubly-robust estimators can be particularly attractive as they rely on less stringent modeling conditions than the outcome regression and the inverse probability weighting procedures.

In order to conduct asymptotically valid inference, we justify the use of a computationally convenient multiplier-type bootstrap procedure. This approach can be used to obtain simultaneous confidence bands for the group-time average treatment effects. Unlike commonly used pointwise confidence bands, our simultaneous confidence bands asymptotically cover the *entire path* of the group-time average treatment effects with fixed probability and take into account the dependency across different group-time average treatment effect estimators. Thus, our proposed confidence bands are arguably more suitable for visualizing the overall estimation uncertainty than more traditional pointwise confidence intervals.

We illustrate the practical relevance of our proposal by analyzing the effect of the minimum wage on teen employment. Here, we follow much empirical work on the effects of the minimum wage and exploit having access to panel data and variation in treatment timing across states (e.g., Card and Krueger, 1994; Neumark and Wascher, 2000, 2008; Dube et al., 2010, among many others) in order to estimate the effect of the minimum wage on employment. Interestingly, in our setup, using our approach leads to qualitatively different results than results from the TWFE estimator. This suggests that, at least in certain applications, using methods that are robust to treatment effect heterogeneity can lead to meaningful differences relative to more standard TWFE regressions.

Recent Related Literature: This paper is related to the recent and emerging literature on heterogeneous treatment effects in DiD and/or event studies with variation in treatment timing; see, e.g., de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2019), Imai et al. (2018), Borusyak and Jaravel (2017), Athey and Imbens (2018) and Sun and Abraham (2020). All these papers present, among other things, some negative results about the interpretation of parameters associated with standard TWFE linear regression specifications; see also Laporte and Windmeijer (2005), Wooldridge (2005a), Chernozhukov et al. (2013), and Gibbons et al. (2018) for earlier related results based on (one-way) fixed-effect estimators. Our proposed procedure completely bypasses the pitfalls highlighted in these papers as we clearly separate the identification, aggregation and estimation/inference steps of the analysis.

These aforementioned papers also propose alternative DiD estimators that do not suffer from the pitfalls associated with TWFE. Among these, perhaps the closest to our proposal are those of de Chaisemartin and D'Haultfœuille (2020), and Sun and Abraham (2020), though several major differences are worth stressing.

de Chaisemartin and D'Haultfœuille (2020) are focused on recovering an instantaneous treatment effect measure, while we pay particular attention to treatment effect dynamics. In fact, our framework allows one to form *families* of different aggregate parameters in a unified manner. Second, while we pay particular attention to the role played by pre-treatment covariates, de Chaisemartin and D'Haultfœuille (2020) mainly focus on unconditional DiD designs. On the other hand, the setup in de Chaisemartin and D'Haultfœuille (2020) is more general than ours as we consider staggered adoption designs and they allow for more general treatment selection. Nonetheless, we note that the unconditional versions of our parallel trends assumptions are weaker than the one in de Chaisemartin and D'Haultfœuille (2020), even if one were to specialize their setup to staggered adoption designs.

Sun and Abraham (2020) propose a parameter, cohort-specific average treatment effects, that translates our group-time average treatment effects from calendar time into event time. Sun and Abraham (2020) propose regression-based estimators of these parameters that have similar properties to our estimators in the specific case of staggered treatment adoption under an unconditional version of the parallel trends assumption. However, our approach is more general in several respects. First, we allow for parallel trends assumptions to hold after conditioning on covariates, and it is not clear how to adapt the regression based estimators in Sun and Abraham (2020) to this case. Second, we consider a wide variety of possible aggregations of group-time average treatment effects where Sun and Abraham (2020) focus particularly on the event study type of aggregation. Third, we make use of simultaneous inference procedures that explicitly account for potential multiple-testing problems; Sun and Abraham (2020) focus on pointwise inference. On the other hand, we do not have any results highlighting the pitfalls associated with using TWFE specifications with leads and lags of treatment indicators to conduct causal inference; these are unique to Sun and Abraham (2020).

We also note that Athey and Imbens (2018) consider a staggered treatment adoption setup similar to ours. However, the starting point of Athey and Imbens (2018) is an assumption that the treatment adoption date is fully randomized which is stronger than our parallel trends assumptions. We also note that Athey and Imbens (2018) abstract away from the important role played by covariates in the DiD analysis and do not consider aggregation schemes to summarize treatment effect heterogeneity like we do. On the other hand, we stress that the main focus of their paper is on providing design-based inference procedures for staggered DiD setups with random treatment dates. Their design-based inference procedures complement our sampling-based inference procedures.

Organization of the paper: The remainder of this article is organized as follows. Section 2 presents our main identification results. We discuss our different aggregation schemes in Section 3. Estimation and inference procedures for the treatment effects of interest are presented in Section 4. We revisit the effect of minimum wage on employment in Section 5. Section 6 concludes. Proofs as well as additional methodological results are reported in the Appendix. In the Supplementary Appendix, we present proofs for the results when only repeated cross-sections data is available, provide additional details about the empirical application, and present a small scale Monte Carlo simulation to illustrate the finite sample properties of our proposed estimators.

2. Identification

2.1. Setup

We first introduce the notation we use throughout the article. We consider the case with $\mathcal{T}$ periods and denote a particular time period by t where $t = 1, \ldots, \mathcal{T}$. In a canonical DiD setup, $\mathcal{T} = 2$ and no one is treated in period $t = 1$. Let $D_{i,t}$ be a binary variable equal to one if unit i is treated in period t and equal to zero otherwise. We make the following assumption about the treatment process:

Assumption 1 (*Irreversibility of Treatment*)**.** $D_1 = 0$ *almost surely* (a.s.). For $t = 2, \ldots, \mathcal{T}$,

$$D_{t-1} = 1 \text{ implies that } D_t = 1 \text{ a.s.}$$

Assumption 1 states that no one is treated at time $t = 1$, and that once a unit becomes treated, that unit will remain treated in the next period.[2] This assumption is also called staggered treatment adoption in the literature. We interpret this assumption as if units do not "forget" about the treatment experience.[3]

Define G as the time period when a unit first becomes treated. Under Assumption 1, for all units that eventually participate in the treatment, G defines which "group" they belong to. If a unit does not participate in any time period, we arbitrarily set $G = \infty$. We define G_g to be a binary variable that is equal to one if a unit is first treated in period g (i.e., $G_{i,g} = \mathbf{1}\{G_i = g\}$) and define C to be a binary variable that is equal to one for units that do not participate in the treatment in any time period (i.e., $C_i = \mathbf{1}\{G_i = \infty\} = 1 - D_{i,\mathcal{T}}$). Let $\bar{g} = \max_{i=1,\ldots,n} G_i$ be the maximum G in the dataset. Next, denote the generalized propensity score as $p_{g,s}(X) = P(G_g = 1|X, G_g + (1 - D_s)\left(1 - G_g\right) = 1)$. Note that $p_{g,s}(X)$ indicates the probability of being first treated at time g, conditional on pre-treatment covariates X and on either being a member of group g (in this case, $G_g = 1$) or a member of the "not-yet-treated" group by time s (in this case, $(1 - D_s)(1 - G_g) = 1$). Many of our results use a specialized version of this generalized propensity score, and, henceforth, we define $p_g(X) = p_{g,\mathcal{T}}(X) = P(G_g = 1|X, G_g + C = 1)$ which is the probability of being first treated in period g conditional on covariates and either being a member of group g or not participating in the treatment in any time period. Let $\mathcal{G} = \text{supp}(G)\backslash\{\bar{g}\} \subseteq \{2, 3, \ldots, \mathcal{T}\}$ denote the support of G excluding $\bar{g}$.[4] Likewise, let $\mathcal{X} = \text{supp}(X) \subseteq \mathbb{R}^k$ denote the support of the pre-treatment covariates. Finally, for a generic $\delta \geq 0$, let $\mathcal{G}_\delta = \mathcal{G} \cap \{2 + \delta, 3 + \delta, \ldots, \mathcal{T}\}$.

Next, we set up the potential outcomes framework. Here, we combine the dynamic potential outcomes framework of Robins (1986, 1987) with the multi-stage treatment adoption setup discussed by Heckman et al. (2016); see also Sianesi (2004). Let $Y_{i,t}(0)$ denote unit i's untreated potential outcome at time t if they remain untreated through time period $\mathcal{T}$; i.e., if they were not to participate in the treatment across all available time periods. For $g = 2, \ldots, \mathcal{T}$, let $Y_{i,t}(g)$ denote the potential outcome that unit i would experience at time t if they were to first become treated in time period g. Note that our potential outcomes notation accounts for potential dynamic treatment selection, though it also accommodates (pre-specified) treatment regimes (Murphy et al., 2001; Murphy, 2003). The observed and potential outcomes for each unit i are related through

$$Y_{i,t} = Y_{i,t}(0) + \sum_{g=2}^{\mathcal{T}} \left(Y_{i,t}(g) - Y_{i,t}(0)\right) \cdot G_{i,g} \tag{2.1}$$

In other words, we only observe one potential outcome path for each unit. For those that do not participate in the treatment in any time period, observed outcomes are untreated potential outcomes in all periods. For units that do participate in the treatment, observed outcomes are the unit-specific potential outcomes corresponding to the particular time period when that unit adopts the treatment.

We also impose the following random sampling assumption.

Assumption 2 (*Random Sampling*)**.** $\{Y_{i,1}, Y_{,i2}, \ldots, Y_{i,\mathcal{T}}, X_i, D_{i,1}, D_{i,2}, \ldots, D_{i,\mathcal{T}}\}_{i=1}^n$ is independent and identically distributed (*iid*).

Assumption 2 implies that we have access to panel data; our results extend essentially immediately to the case with repeated cross sections data and this case is developed in Appendix B. Here, we note that Assumption 2 allows us to view all potential outcomes as random. Furthermore, it does not impose restrictions between *potential outcomes* and treatment allocation, nor does it restrict the time series dependence of the observed random variables. On the other hand, Assumption 2 imposes that each unit i is randomly drawn from a large population of interest. For an alternative design-based inference approach, see Athey and Imbens (2018).

Henceforth, to keep the notation more concise, we will suppress the unit index i in our notation.

2.2. The group-time average treatment effect parameter

Given that different potential outcomes cannot be observed for the same unit at the same time, researchers often focus on identifying and estimating some average causal effects. For instance, in the canonical DiD setup with two time periods, the most popular treatment effect parameter of interest is the average treatment effect on the treated, denoted by[5]

$$ATT = \mathbb{E}[Y_2(2) - Y_2(0)|G_2 = 1].$$

[2] In applications, it can be the case that some units are already treated by the first time period. In our case, we would drop these units; this is analogous to the case with two time periods. The reason to drop these units is that untreated potential outcomes are never observed for this group which will imply that treatment effects are not identified for this group nor are they useful as a comparison group under a parallel trends assumption.

[3] See Han (2020), de Chaisemartin and D'Haultfœuille (2020) and Bojinov et al. (2020) for alternative setups where treatment can "turn off".

[4] When there is a "never treated" set of units with $G = \infty$, $\mathcal{G}$ only excludes this group. When such "never-treated" group is not available, we exclude the latest-treated group as there will be no valid untreated comparison group for them.

[5] Existence of expectations is assumed throughout.

In this paper, we consider a natural generalization of the *ATT* that is suitable to setups with multiple treatment groups and multiple time periods. More precisely, we use the average treatment effect for units who are members of a particular group g at a particular time period t, denoted by

$$ATT(g,t) = \mathbb{E}[Y_t(g) - Y_t(0)|G_g = 1],$$

as the main building block of our framework. We call this causal parameter the *group-time average treatment effect*.

Note that the $ATT(g,t)$ does not impose any restriction on treatment effect heterogeneity across groups or across time. Thus, focusing on the family of $ATT(g,t)$'s allow us to analyze how average treatment effects vary across different dimensions in a unified manner. For instance, by fixing a group g and varying time t, one is able to highlight how average treatment effects evolve over time for that specific group. By doing this for different groups, we can have a better understanding about how treatment effect dynamics vary across groups. In addition, as we discuss in Section 3, one can build on the $ATT(g,t)$'s to form more aggregated causal parameters that are constructed to answer specific questions like: (a) What was the average effect of participating in the treatment across all groups that participated in the treatment by time period $\mathcal{T}$? (b) Are average treatment effects heterogeneous across groups? (c) How do average treatment effects vary by length of exposure to the treatment? (d) How do cumulative average treatment effects evolve over calendar time? We view this level of generality and flexibility as one of the main advantages of our framework that first focuses on the family of $ATT(g,t)$'s.

2.3. Identifying assumptions

In order to identify the $ATT(g,t)$ and their functionals, we impose the following assumptions.

Assumption 3 (*Limited Treatment Anticipation*)**.** There is a known $\delta \geq 0$ such that

$$\mathbb{E}[Y_t(g)|X, G_g = 1] = \mathbb{E}[Y_t(0)|X, G_g = 1] \ a.s. \text{ for all } g \in \mathcal{G}, t \in \{1, \dots, \mathcal{T}\} \text{ such that } t < g - \delta.$$

Assumption 3 restricts anticipation of the treatment for all "eventually treated" groups. When $\delta = 0$, it imposes a "no-anticipation" assumption, see, e.g., Abbring and van den Berg (2003) and Sianesi (2004). This is likely to be the case when the treatment path is not a priori known and/or when units are not the ones who "choose" treatment status. However, Assumption 3 also allows for anticipation behavior, as long as we have a good understanding about the anticipation horizon δ. For instance, if units anticipate treatment by one period, Assumption 3 would hold with $\delta = 1$; see, e.g., Laporte and Windmeijer (2005) and Malani and Reif (2015) for the importance of accounting for potential anticipation behavior. Note that, under Assumption 3, $ATT(g,t) = 0$ for all pre-treatment periods such that $t < g - \delta$.

Next, we consider two alternative assumptions that impose restrictions on the evolution of untreated potential outcomes.

Assumption 4 (*Conditional Parallel Trends Based on a "Never-Treated" Group*)**.** Let δ be as defined in Assumption 3. For each $g \in \mathcal{G}$ and $t \in \{2, \dots, \mathcal{T}\}$ such that $t \geq g - \delta$,

$$\mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = \mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, C = 1] \ a.s.$$

Assumption 5 (*Conditional Parallel Trends Based on "Not-Yet-Treated" Groups*)**.** Let δ be as defined in Assumption 3. For each $g \in \mathcal{G}$ and each $(s,t) \in \{2, \dots, \mathcal{T}\} \times \{2, \dots, \mathcal{T}\}$ such that $t \geq g - \delta$ and $t + \delta \leq s < \bar{g}$,

$$\mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = \mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, D_s = 0, G_g = 0] \ a.s.$$

Assumptions 4 and 5 are two different conditional parallel trends assumptions that generalize the two-period parallel trends assumption to the case where there are multiple time periods and multiple treatment groups; see, e.g., Heckman et al. (1997, 1998), Abadie (2005) and Sant'Anna and Zhao (2020). Both assumptions hold after conditioning on covariates X. This can be important in many applications in economics particularly in cases where there are covariate specific trends in outcomes over time and when the distribution of covariates is different across groups. For example, Heckman et al. (1997) motivate conditional parallel trends assumptions in the context of evaluating a job training program. For evaluating job training programs, the distribution of observed covariates such as age, employment history, and years of education is often quite different between individuals who participate in job training and those that do not. When the path of labor market outcomes (in the absence of participating in job training) depends on these covariates, a conditional parallel trends becomes more plausible than an unconditional parallel trends assumption. In fact, ignoring the presence of covariate-specific trends can result in important biases when evaluating causal effects of policy interventions using unconditional DiD methods.

Assumptions 4 and 5 differ from each other depending on the comparison group one is willing to use in a given application. More specifically, Assumption 4 states that, conditional on covariates, the average outcomes for the group first treated in period g and for the "never-treated" group would have followed parallel paths in the absence of treatment.

Assumption 5 imposes conditional parallel trends between group g and groups that are "not-yet-treated" by time $t+\delta$.[6] Importantly, both of these assumptions allow for covariate-specific trends and do not restrict the relationship between treatment timing and the potential outcomes, $Y_t(g)$'s. Thus, they are weaker than the randomization-based assumption made by Athey and Imbens (2018). We also note that the unconditional versions of Assumptions 4 and 5 are weaker than the parallel trends assumption imposed by de Chaisemartin and D'Haultfœuille (2020) and Sun and Abraham (2020) as they impose fewer restrictions on the evolution of $Y_t(0)$ in pre-treatment periods; see, e.g., Marcus and Sant'Anna (2020) for a comparison.

In our view, practitioners may favor Assumption 4 with respect to Assumption 5 when there is a sizeable group of units that do not participate in the treatment in any period, and, at the same time, these units are similar enough to the "eventually treated" units. When a "never-treated" group of units is not available or "too small", researchers may favor Assumption 5 as it allows one to use more groups as valid comparison units, which potentially leads to more informative inference procedures. However, it is important to stress that favoring Assumption 5 with respect to Assumption 4 also involves potential drawbacks. For instance, in the absence of treatment anticipation ($\delta = 0$), Assumption 4 does not restrict observed pre-treatment trends across groups, whereas Assumption 5 does; see, e.g., Marcus and Sant'Anna (2020). Not restricting pre-treatment trends may be particularly meaningful in applications where the economic environment during the "early-periods" was potentially different from the "later-periods". In these cases, the outcomes of different groups may evolve in a non-parallel manner during "early-periods", perhaps because the groups were exposed to different shocks, while trends become parallel in the "later-periods". We recommend taking these trade-offs into account when deciding which conditional parallel trends assumption is more appropriate for a given application.[7]

The final identifying assumption we impose is an overlap condition.

Assumption 6 (*Overlap*)**.** *For each $t \in \{2, \ldots, \mathcal{T}\}$, $g \in \mathcal{G}$, there exist some $\varepsilon > 0$ such that $P\left(G_g = 1\right) > \varepsilon$ and $p_{g,t}(X) < 1 - \varepsilon$ a.s.*

Assumption 6 extends the overlap assumption in Heckman et al. (1997, 1998), Abadie (2005), and Sant'Anna and Zhao (2020) to the multiple groups and multiple periods setup. It states that a positive fraction of the population starts treatment in period g, and that, for all g and t, the generalized propensity score is uniformly bounded away from one. Assumption 6 rules out "irregular identification", see, e.g., Khan and Tamer (2010).

Remark 1. Note that Assumptions 3 and 4 (Assumption 5) are intrinsically connected. For instance, when one imposes the "no-anticipation" condition (so that $\delta = 0$), Assumption 4 would then impose conditional parallel trends only for post-treatment periods $t \geq g$. If one allows for anticipation behavior (so that $\delta > 0$), Assumption 4 would then impose conditional parallel trends in some pre-treatment periods, too. In fact, the parallel trends assumptions become stronger as one increases δ. To the best of our knowledge, this trade-off between the strength of these assumptions has not been noticed before.

Remark 2. In some applications, practitioners may not be comfortable with using "never-treated" units as part of the comparison group because they behave very differently from the other "eventually treated" units. In these cases, practitioners could drop all "never-treated" units from the analysis and proceed with Assumption 5.

2.4. Nonparametric identification of the group-time average treatment effects

In this section, we show that the family of group-time average treatment effects are nonparametrically point-identified under the aforementioned assumptions. Furthermore, we show that one can use outcome regression (OR), inverse probability weighting (IPW), or doubly robust (DR) estimands to recover the $ATT(g,t)$'s. In addition, we also highlight the roles played by Assumption 3 and by Assumptions 4 and 5 when forming these different estimands.

Before formalizing all the results, we need to introduce some additional notation. Let $m^{nev}_{g,t,\delta}(X) = \mathbb{E}\left[Y_t - Y_{g-\delta-1} | X, C = 1\right]$ and $m^{ny}_{g,t,\delta}(X) = \mathbb{E}\left[Y_t - Y_{g-\delta-1} | X, D_{t+\delta} = 0, G_g = 0\right]$. These are population outcome regressions for the never-treated group and for the "not-yet-treated" by time $t+\delta$ group. Let

$$ATT^{nev}_{ipw}(g,t;\delta) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]} - \frac{\dfrac{p_g(X)\,C}{1-p_g(X)}}{\mathbb{E}\left[\dfrac{p_g(X)\,C}{1-p_g(X)}\right]}\right)\left(Y_t - Y_{g-\delta-1}\right)\right], \tag{2.2}$$

[6] Athey and Imbens (2006) and de Chaisemartin and D'Haultfœuille (2017) also consider using "not-yet-treated" units as comparison groups in related DiD procedures.

[7] It may be tempting to use statistical pre-tests to select between different versions of the parallel trends assumption. However, the results of Roth (2020) show that such a practice can lead to important distortions when conducting inference. Thus, we do not recommend following this path, but instead recommend taking the context of the application into account in order to choose the appropriate parallel trends assumption.

$$ATT_{or}^{nev}(g,t;\delta)=\mathbb{E}\left[\frac{G_g}{\mathbb{E}\left[G_g\right]}\left(Y_t-Y_{g-\delta-1}-m_{g,t,\delta}^{nev}\left(X\right)\right)\right], \tag{2.3}$$

$$ATT_{dr}^{nev}(g,t;\delta)=\mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]}-\frac{\dfrac{p_g\left(X\right)C}{1-p_g\left(X\right)}}{\mathbb{E}\left[\dfrac{p_g\left(X\right)C}{1-p_g\left(X\right)}\right]}\right)\left(Y_t-Y_{g-\delta-1}-m_{g,t,\delta}^{nev}\left(X\right)\right)\right]. \tag{2.4}$$

Analogously, let

$$ATT_{ipw}^{ny}(g,t;\delta)=\mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]}-\frac{\dfrac{p_{g,t+\delta}\left(X\right)\left(1-D_{t+\delta}\right)\left(1-G_g\right)}{1-p_{g,t+\delta}\left(X\right)}}{\mathbb{E}\left[\dfrac{p_{g,t+\delta}\left(X\right)\left(1-D_{t+\delta}\right)\left(1-G_g\right)}{1-p_{g,t+\delta}\left(X\right)}\right]}\right)\left(Y_t-Y_{g-\delta-1}\right)\right], \tag{2.5}$$

$$ATT_{or}^{ny}(g,t;\delta)=\mathbb{E}\left[\frac{G_g}{\mathbb{E}\left[G_g\right]}\left(Y_t-Y_{g-\delta-1}-m_{g,t,\delta}^{ny}\left(X\right)\right)\right], \tag{2.6}$$

$$ATT_{dr}^{ny}(g,t;\delta)=\mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]}-\frac{\dfrac{p_{g,t+\delta}\left(X\right)\left(1-D_{t+\delta}\right)\left(1-G_g\right)}{1-p_{g,t+\delta}\left(X\right)}}{\mathbb{E}\left[\dfrac{p_{g,t+\delta}\left(X\right)\left(1-D_{t+\delta}\right)\left(1-G_g\right)}{1-p_{g,t+\delta}\left(X\right)}\right]}\right)\left(Y_t-Y_{g-\delta-1}-m_{g,t,\delta}^{ny}\left(X\right)\right)\right]. \tag{2.7}$$

With some abuse of notation, we write $\bar{g}-\delta=\infty$ for any non-negative δ whenever $\bar{g}=\infty$.

Theorem 1. *Let Assumptions 1–3 and 6 hold.*

(*i*) *If Assumption 4 holds, then, for all g and t such that $g\in\mathcal{G}_\delta$, $t\in\{2,\dots\mathcal{T}-\delta\}$ and $t\geq g-\delta$,*

$$ATT(g,t)=ATT_{ipw}^{nev}(g,t;\delta)=ATT_{or}^{nev}(g,t;\delta)=ATT_{dr}^{nev}(g,t;\delta).$$

(*ii*) *If Assumption 5 holds, then, for all g and t such that $g\in\mathcal{G}_\delta$, $t\in\{2,\dots\mathcal{T}-\delta\}$ and $g-\delta\leq t<\bar{g}-\delta$,*

$$ATT(g,t)=ATT_{ipw}^{ny}(g,t;\delta)=ATT_{or}^{ny}(g,t;\delta)=ATT_{dr}^{ny}(g,t;\delta).$$

Theorem 1 is the first main result of this paper. It provides powerful identification results that extend the DiD identification results based on the outcome regression approach of Heckman et al. (1997, 1998), the IPW approach of Abadie (2005), and the DR approach of Sant'Anna and Zhao (2020) to the multiple-periods, multiple groups setup. In other words, Theorem 1 says that, from an identification point of view, one can recover the $ATT(g,t)$'s by exploiting different parts of the data generating process: the OR approach only relies on modeling the conditional expectation of the outcome evolution for the comparison groups, the IPW approach relies on modeling the conditional probability of being in group g, whereas the DR approach exploits both OR and IPW components.

In order to extend the results of Heckman et al. (1997, 1998), Abadie (2005), and Sant'Anna and Zhao (2020) to the multiple groups, multiple periods framework, we have to address two different challenges: one associated with an appropriate reference time period and one associated with an appropriate comparison group. Theorem 1 highlights how a solution to these challenges is directly connected to the limited anticipation and the conditional parallel trends assumptions. More specifically, Theorem 1 says that we can use the time period $t=g-\delta-1$ as an appropriate reference time period under Assumption 3 and either Assumption 4 or 5. This is the most recent time period when untreated potential outcomes are observed for units in group g. Interestingly, the more treatment anticipation is allowed (i.e., the higher δ is), the further back in time one needs to go.[8] Theorem 1 also suggests that the choice of comparison group is directly tied to the conditional parallel trends assumption one makes: under Assumption 4, one can use "never treated" units as a fixed comparison group for all "eventually treated" units; whereas, under Assumption 5, one can use the "not-yet-treated by time $t+\delta$" units as a valid comparison group for those who are first treated at time g. In this latter case, Theorem 1 also highlights that when all units eventually gets treated ($\bar{g}<\infty$), one is only able to identify the $ATT(g,t)$'s for time periods before the last treated group "effectively" starts their treatment, i.e., $t<\bar{g}-\delta$. In this case, one cannot identify the $ATT(g,t)$ for the last treated cohort, too.

Finally, we note that when pre-treatment covariates play no role in identification (i.e., Assumptions 3, 4, and 5 hold unconditionally on X), (2.2)–(2.4) collapse to

$$ATT_{unc}^{nev}(g,t;\delta)=E[Y_t-Y_{g-\delta-1}|G_g=1]-E[Y_t-Y_{g-\delta-1}|C=1], \tag{2.8}$$

[8] As mentioned in Remark 1, as one allows δ to increase, Assumptions 4 and 5 become more restrictive.

and (2.5)–(2.7) collapse to

$$ATT^{ny}_{unc}(g,t;\delta) = E[Y_t - Y_{g-\delta-1}|G_g = 1] - E[Y_t - Y_{g-\delta-1}|D_{t+\delta} = 0]. \tag{2.9}$$

These expressions for $ATT(g,t)$ clearly resemble the one for ATT in the canonical two-periods and two-groups case. As in that case, the average effect of participating in the treatment for units in group g is identified by taking the path of outcomes (i.e., the change in outcomes between the most recent period before they were affected by the treatment and the current period) actually experienced by that group and adjusting it by the path of outcomes experienced by a comparison group. Under the parallel trends assumption, this latter path is the path of outcomes that units in group g would have experienced if they had not participated in the treatment.

Remark 3. From (2.8) one can see that when Assumptions 3 and 4 hold unconditionally and there is no-anticipation, the $ATT(g,t)$ parameter can be obtained by first subsetting the data to only contain observations at time t and $g-1$, from units with either $G_g = 1$ or $C = 1$, and then, using only the observations of this subset, running the (population) linear regression

$$Y = \alpha_1^{g,t} + \alpha_2^{g,t}\cdot G_g + \alpha_3^{g,t}\cdot 1\{T = t\} + \beta^{g,t}\cdot\left(G_g \times 1\{T = t\}\right) + \epsilon^{g,t}. \tag{2.10}$$

It is then easy to verify that $\beta^{g,t} = ATT(g,t)$. Note that one would need to consider different partitions of the data to characterize different $ATT(g,t)$ in terms of regression parameters. Alternatively, one could use the interacted two-way fixed effects regression proposed by Sun and Abraham (2020).

Remark 4. When covariates are available, the $\tilde{\beta}^{g,t}$ coefficient of the population linear regression

$$Y = \tilde{\alpha}_1^{g,t} + \tilde{\alpha}_2^{g,t}\cdot G_g + \tilde{\alpha}_3^{g,t}\cdot 1\{T = t\} + \tilde{\beta}^{g,t}\cdot\left(G_g \times 1\{T = t\}\right) + \tilde{\gamma}\cdot X + \tilde{\epsilon}^{g,t}$$

that uses the same subset of data as in Remark 3 is, in general, not equal to $ATT(g,t)$ unless one is willing to assume (i) homogeneous (in X) treatment effects, i.e., $\mathbb{E}[Y_t(g) - Y_t(0)|G_g = 1, X] = \mathbb{E}[Y_t(g) - Y_t(0)|G_g = 1]$ a.s., and (ii) rule-out covariate-specific trends, i.e., for $\mathbb{E}[Y_t - Y_{t-1}|X, G] = \mathbb{E}[Y_t - Y_{t-1}|G]$ a.s. for all groups and time periods; see, e.g., Słoczyński (2018) for a related discussion. The characterizations of $ATT(g,t)$ discussed in Theorem 1 do not rely on these restrictive conditions.

Remark 5. Although the IPW, OR, and DR based estimands presented in Theorem 1 are identical from an identification standpoint, this is not the case when one wants to estimate and make inference about the $ATT(g,t)$. As we discuss in Section 4, DiD estimators based on the DR estimands (2.4) and (2.7) usually enjoy additional robustness against model-misspecifications when compared to the IPW and OR estimands.

Remark 6. Theorem 1 suggests that we can identify $ATT(g,t)$ only for groups in $\mathcal{G}_\delta \subseteq \mathcal{G}$ which can involve dropping some "early treated" groups due to anticipation effects. When $\delta = 0$, i.e. when there is no anticipation, $\mathcal{G}_\delta = \mathcal{G}$. Theorem 1 also suggests that we can identify $ATT(g,t)$ only until $t = \mathcal{T} - \delta$ because of potential treatment anticipation behavior. In applications where some units are known to never participate in the treatment (including periods after time period $\mathcal{T}$), however, we can identify $ATT(g,t)$ up to $t = \mathcal{T}$ by using these units as a valid comparison group for all time periods $t = \mathcal{T} - \delta + 1, \ldots, \mathcal{T}$, provided that an appropriate parallel trends assumption is satisfied.

Remark 7. From Theorem 1 it is clear that pre-treatment covariates play a prominent role in our analysis. Importantly, Assumptions 4 and 5 suggest that researchers should include pre-treatment covariates that are potentially associated with the outcome evolution of $Y(0)$ during post-treatment periods. We explicitly rule out incorporating post-treatment covariates as they can potentially be affected by the treatment; see, e.g., Wooldridge (2005b), for a related discussion under the unconfoundedness setup.

3. Summarizing group-time average treatment effects

The previous section shows that we can identify the $ATT(g,t)$'s by restricting treatment anticipation behavior and imposing a conditional parallel trends assumption. In many applications, the $ATT(g,t)$'s can be the ultimate causal parameters of interest. They can be used to highlight treatment effect heterogeneity across different groups g, at different points in time t, and across different lengths of treatment exposure, $e = t - g$. In other situations, however, researchers may want to combine these different $ATT(g,t)$'s to form more aggregated causal parameters. For instance, if the number of groups and time periods is relatively large, it may be challenging to interpret many group-time average treatment effects.

In this section, we consider different aggregation schemes for the $ATT(g,t)$'s that allow researchers to form a variety of summary measures of the causal effects of a given policy. Our aggregation schemes are of the form

$$\theta = \sum_{g\in\mathcal{G}}\sum_{t=2}^{\mathcal{T}} w(g,t)\cdot ATT(g,t), \tag{3.1}$$

where $w(g,t)$ are carefully-chosen (known or estimable) weighting functions specified by the researcher such that θ can be used to address a well-posed empirical/policy question. Difference choices of $w(g,t)$ allow researchers to highlight different types of treatment effect heterogeneity. We pay particular attention to aggregations that result in a single overall treatment effect summary parameter as well as to aggregations related to understanding dynamic effects as is commonly done in event-study analysis. Of course, many other aggregated parameters of the type (3.1) can be easily constructed following our framework. We illustrate this point by also summarizing heterogeneity with respect to group or by calendar time.

Before proceeding with the discussion on how to construct these different aggregated parameters, it is worth revisiting the two most popular treatment effect summary measures used by practitioners in DiD setups. These are based on the "static" and "dynamic" two-way fixed effects (TWFE) linear regression specifications

$$Y_{i,t} = \alpha_t + \alpha_g + \beta D_{i,t} + \epsilon_{i,t}, \tag{3.2}$$

$$Y_{i,t} = \alpha_t + \alpha_g + \sum_{e=-K}^{-2} \delta_e^{anticip} \cdot D_{i,t}^e + \sum_{e=0}^{L} \beta_e \cdot D_{i,t}^e + v_{i,t}, \tag{3.3}$$

respectively, where α_t is a time fixed effect, α_g is a group fixed effect, $\epsilon_{i,t}$ and $v_{i,t}$ are error terms, $D_{i,t}^e = 1\{t - G_i = e\}$ is an indicator for unit i being e periods away from initial treatment at time t, and K and L are positive constants. The parameter of interest in the static TWFE specification is β, which, in applications, is typically interpreted as an overall effect of participating in the treatment across groups and time periods. In the dynamic TWFE specification, practitioners usually focus on the β_e, $e \geq 0$, and these parameters are typically interpreted as measuring the effect of participating in the treatment at different lengths of exposure to the treatment.

Despite the popularity of these specifications, recent research has shown that one must be very careful in attaching a causal interpretation to these aggregated parameters. For instance, Borusyak and Jaravel (2017), Goodman-Bacon (2019), de Chaisemartin and D'Haultfœuille (2020), and Athey and Imbens (2018) have shown that, in general, β recovers a weighted average of some underlying treatment effect parameters but some of the weights on these parameters can be negative. This can potentially lead to particularly problematic cases such as the effect of the treatment being positive for all units, but the TWFE estimation resulting in estimates of β that are negative. Even in cases where the weights are not negative, the weights on underlying treatment effect parameters are entirely driven by the TWFE estimation strategy and are sensitive to the size of each group, the timing of treatment, and the total number of time periods (see Theorem 1 in Goodman-Bacon, 2019). The results in this section can be used in exactly the same setup to identify a single interpretable average treatment effect parameter and, thus, provide a way to circumvent the issues with the more common approach.

As discussed by Goodman-Bacon (2019), the "negative weight problem" associated with β arises when treatment effects evolve over time. Thus, one may wonder if such problems would still be present when considering more general, dynamic specifications such as (3.3). Sun and Abraham (2020) show that this is still the case as the $\beta_e' s$ associated with (3.3) do not recover easy-to-interpret causal parameters and still generally suffer from the same sorts of "negative weighting problems." In contrast to this, we provide a simple way to directly aggregate our group-time average treatment effects into average treatment effects across different lengths of exposure to the treatment.

3.1. Aggregations to highlight treatment effect heterogeneity

Next, we discuss several partial aggregations of the group-time average treatment effects in order to summarize different dimensions of treatment effect heterogeneity. Although there are additional possibilities, we focus our discussion below on how to answer three particular questions: (a) How does the effect of participating in the treatment vary with length of exposure to the treatment? (b) Do groups that are treated earlier have, on average, higher/lower average treatment effects than groups that are treated later? (c) What is the cumulative average treatment effect of the policy across all groups until some particular point in time? Throughout this section, to avoid notation clutter, we assume that units do not anticipate treatment, i.e., we consider the case where Assumption 3 holds with $\delta = 0$. We also assume that a "never treated" group is available.

3.1.1. How do average treatment effects vary with length of exposure to the treatment?

One of the most popular questions that arises in DiD setups with multiple time periods concerns treatment effect dynamics: How does the effect of participating in the treatment vary with length of exposure to the treatment? For instance, do average treatment effects increase/decrease with elapsed treatment time? Indeed, answering this type of question is often the main motivation for using the event study regression in (3.3), though, as we mentioned above, that sort of regression may not be suitable for such a task. In this section, we propose an aggregation scheme that is suitable to highlight treatment effect heterogeneity with respect to length of exposure to the treatment that does not suffer from the drawbacks associated with the event study regression in (3.3).

Let e denote event-time, i.e., $e = t - g$ denotes the time elapsed since treatment was adopted. Recall that G denotes the time period that a unit is first treated. Thus, a way to aggregate the $ATT(g,t)$'s to highlight treatment effect heterogeneity

Table 1
Weights on $ATT(g,t)$ for aggregated parameters.

Parameter	$w(g,t)$
$\theta_{es}(e)$	$w_e^{es}(g,t) = \mathbf{1}\{g+e \le \mathcal{T}\}\mathbf{1}\{t-g=e\}P(G=g \mid G+e \le \mathcal{T})$
$\theta_{es}^{bal}(e,e')$	$w_e^{es,bal}(g,t) = \mathbf{1}\{g+e' \le \mathcal{T}\}\mathbf{1}\{t-g=e\}P(G=g \mid G+e' \le \mathcal{T})$
$\theta_{sel}(\tilde{g})$	$w_{\tilde{g}}^{s}(g,t) = \mathbf{1}\{t \ge g\}\mathbf{1}\left\{g=\tilde{g}\right\} / \left(\mathcal{T}-g+1\right)$
$\theta_c(\tilde{t})$	$w_{\tilde{t}}^{c}(g,t) = \mathbf{1}\{t \ge g\}\mathbf{1}\left\{t=\tilde{t}\right\} P(G=g \mid G \le t)$
$\theta_c^{cumu}(\tilde{t})$	$w_{\tilde{t}}^{c,cumu}(g,t) = \mathbf{1}\{t \ge g\}\mathbf{1}\left\{t \le \tilde{t}\right\} P(G=g \mid G \le t)$
θ_W^O	$w_W^O(g,t) = \mathbf{1}\{t \ge g\}P(G=g \mid G \le \mathcal{T}) / \sum_{g\in\mathcal{G}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{t \ge g\}P(G=g \mid G \le \mathcal{T})$
θ_{sel}^O	$w_{sel}^O(g,t) = \mathbf{1}\{t \ge g\}P(G=g \mid G \le \mathcal{T}) / \left(\mathcal{T}-g+1\right)$

Notes: This table provides expressions for the weights on each $ATT(g,t)$ (as in Eq. (3.1)) for each parameter discussed in this section. In all cases except for $\theta_c^{cumu}(\tilde{t})$, the weights are all non-negative and sum to one. For $\theta_c^{cumu}(\tilde{t})$, the weights are all non-negative but sum up to $\tilde{t}-1$ (rather than one), but this is just a reflection of $\theta_c^{cumu}\left(\tilde{t}\right)$ being a cumulative treatment effect measure.

with respect to e is

$$\theta_{es}(e) = \sum_{g\in\mathcal{G}} \mathbf{1}\{g+e \le \mathcal{T}\}P(G=g \mid G+e \le \mathcal{T})ATT(g,g+e). \tag{3.4}$$

This is the average effect of participating in the treatment e time periods after the treatment was adopted across all groups that are ever observed to have participated in the treatment for exactly e time periods. Here, the "on impact" average effect of participating in the treatment occurs for $e=0$. $\theta_{es}(e)$ is the natural target for event study regressions that are common in applied work, though it completely avoids the pitfalls associated with the dynamic TWFE specification in (3.3).[9]

In event study regressions, it is common to plot β_e across different values of e and to interpret differences as being due to treatment effect dynamics. Similarly, one can plot $\theta_{es}(e)$ across different e's to better understand treatment effect dynamics. When doing so, it is important to be aware that these comparisons may include compositional changes that can complicate the interpretation of these parameters (note that the same complications arise for event study regressions as well). To see this, for $0 \le e_1 < e_2 \le \mathcal{T}-2$, consider the difference between $\theta_{es}(e_2)$ and $\theta_{es}(e_1)$ which is given by

$$\begin{aligned}
\theta_{es}(e_2) - \theta_{es}(e_1) = & \sum_{g\in\mathcal{G}} \mathbf{1}\{g+e_1 \le \mathcal{T}\}P(G=g \mid G+e_1 \le \mathcal{T})\underbrace{\left(ATT(g,g+e_2) - ATT(g,g+e_1)\right)}_{\text{dynamic effect for group } g} \\
& + \sum_{g\in\mathcal{G}} \mathbf{1}\{g+e_2 \le \mathcal{T}\}\underbrace{\left(P(G=g \mid G+e_2 \le \mathcal{T}) - P(G=g \mid G+e_1 \le \mathcal{T})\right)}_{\text{differences in weights}}ATT(g,g+e_2) \\
& - \sum_{g\in\mathcal{G}} \underbrace{\mathbf{1}\{\mathcal{T}-e_2 < g \le \mathcal{T}-e_1\}}_{\text{different composition of groups}}P(G=g \mid G+e_1 \le \mathcal{T})ATT(g,g+e_2).
\end{aligned} \tag{3.5}$$

From the above decomposition it becomes clear that comparing $\theta_{es}(e)$ at two different values of e provides a weighted average of the dynamic effect of participating in the treatment – the first component on the right-hand side of (3.5) – plus two extra undesirable terms. Both of these undesirable terms are due to different compositions of groups at different event times.[10] The first term arises because the weights at each length of exposure differ due to the changing composition of groups at each event time. The second term comes directly from different compositions of groups at each length of exposure. These two additional terms may prevent one from interpreting the differences in $\theta_{es}(e)$ across different values of e as being actual dynamic effects of participating in the treatment unless one is willing to impose that $ATT(g,g+e)$ does not vary with g for any $e \ge 0$; i.e., that dynamic effects are common across groups.[11] However, this sort of homogeneity condition may be deemed too strong in many applications.

A simple alternative causal parameter that can be used to highlight treatment effect dynamics with respect to length of exposure to the treatment and does not suffer from the issue of compositional changes highlighted in (3.5) arises from

[9] Many of the parameters in this section involve expressions that have similar components as the one for $\theta_{es}(e)$ in (3.4), and it is worth mentioning a few extra details for this case that are common to the other expressions below. The term involving the indicator function, $\mathbf{1}\{g+e \le \mathcal{T}\}$, limits consideration to identified group-time average treatment effects. The summation over groups with group specific weights, in this case given by $P(G=g \mid G+e \le \mathcal{T})$, calculates an average, weighted by group size, of $ATT(g,t)$'s that are involved in a particular aggregation. In addition, it is straightforward to show that $\theta_{es}(e)$ can be written in the form of θ in Eq. (3.1). Throughout this section, we have written each parameter of interest in its most intuitive form. Weights for each parameter in this section corresponding to the form of the weights in Eq. (3.1) are provided in Table 1.

[10] The composition changes mentioned here arise due to the staggered adoption of the treatment. For example, when $\mathcal{T}=3$, groups 2 and 3 both show up in the expression for $\theta_{es}(0)$, but only group 2 shows up in the expression for $\theta_{es}(1)$.

[11] If $ATT(g,g+e)$ does not vary with g for any $e \ge 0$, it is straightforward to show that the last two terms of (3.5) sum up to 0.

"balancing" the groups with respect to event time, i.e., to only aggregate the $ATT(g,t)$'s for a fixed set of groups that are exposed to the treatment for at least some particular number of time periods and thereby circumvent the issue of compositional changes across different values of e. In particular, for some event time e' with $0 \le e \le e' \le \mathcal{T}-2$, let

$$\theta_{es}^{bal}(e;e') = \sum_{g\in\mathcal{G}} \mathbf{1}\{g+e' \le \mathcal{T}\}ATT(g,g+e)P(G=g|G+e' \le \mathcal{T}). \tag{3.6}$$

Notice that the definition of $\theta_{es}^{bal}(e;e')$ is very similar to $\theta_{es}(e)$ except that it calculates the average group-time average treatment effect for units whose event time is equal to *e and who are observed to participate in the treatment for at least e′ periods*. In this case, since the composition of groups is the same across all values of e, the additional terms in (3.5) do not show up at all and differences in $\theta_{es}^{bal}(e;e')$ across different values of e cannot be due to differences in the composition of groups at different values of e. As an example, when one is interested in analyzing the evolution of treatment effects up to 5 periods after treatment was implemented, one can set $e'=5$ and, this way, the same groups of units will be used when computing $\theta_{es}^{bal}(0;5), \theta_{es}^{bal}(1;5), \ldots, \theta_{es}^{bal}(5;5)$.

The price one pays for "balancing" the groups with respect to event time is that fewer groups are used to compute these event-study-type estimands, which can lead to less informative inference. Thus, in practice, one should consider this "robustness" versus "efficiency" trade-off when choosing between θ_{es}^{bal} and θ_{es}.

Remark 8. We note that $\theta_{es}^{bal}(e;e')$ closely resembles the empirical practice of only reporting event-study-type coefficients for the event periods that do not suffer from compositional changes, see, e.g., McCrary (2007) and Bailey and Goodman-Bacon (2015). An important caveat is that our proposed event-study-type estimands $\theta_{es}^{bal}(e;e')$ are not based on dynamic TWFE specifications akin to (3.3), and therefore bypass the pitfalls associated with (3.3) highlighted by Sun and Abraham (2020).

3.1.2. How do average treatment effects vary across groups?

It is also straightforward to aggregate our group-time average treatment effects to understand heterogeneity in the effect of participating in the treatment across groups. Although understanding this sort of heterogeneity is relatively less common in applied work than trying to understand dynamic effects as discussed above, there are still a number of cases in economics where understanding this sort of heterogeneity may be of interest. For example, work on the effect of graduating during a recession on labor market outcomes (Oreopoulos et al., 2012) or the effect of job displacement across the business cycle (Farber, 2017) are related to heterogeneous effects across groups. More generally, these parameters are useful for understanding if the effect of participating in the treatment was larger for groups that are treated earlier relative to groups that are treated later. In addition, in the next section, these parameters will be the building block for our main measure of the overall effect of participating in the treatment. To consider heterogeneous effects across groups, we consider the following parameter

$$\theta_{sel}(\tilde{g}) = \frac{1}{\mathcal{T}-\tilde{g}+1}\sum_{t=\tilde{g}}^{\mathcal{T}} ATT(\tilde{g},t). \tag{3.7}$$

$\theta_{sel}(\tilde{g})$ is the average effect of participating in the treatment among units in group $\tilde{g}$, across all their post-treatment periods.

3.1.3. What is the cumulative average treatment effect of the policy across all groups until time $\tilde{t}$?

In some applications, researchers may want to construct an aggregated target parameter to highlight treatment effect heterogeneity with respect to calendar time. In economics, for example, researchers might wish to study heterogeneous treatment effects across the business cycle. The average effect of participating in the treatment in time period t (across groups that have adopted the treatment by period t) is given by

$$\theta_c(t) = \sum_{g\in\mathcal{G}} \mathbf{1}\{t \ge g\}P(G=g|G \le t)ATT(g,t) \tag{3.8}$$

An extension to this parameter is to think about the *cumulative* effect of participating in the treatment up to some particular time period. For instance, in active labor market applications, policy makers may want to know the cumulative average effect of a given training program on earnings from the year that the first group of people were trained until year $\tilde{t}$. This would provide a measure of the cumulative earnings gains induced by the training program. Alternatively, in health applications, researchers may want to measure how many COVID-19 cases have been averted by shelter-in-place orders up to day $\tilde{t}$. To consider the cumulative effect, consider the following parameter

$$\theta_c^{cumu}\left(\tilde{t}\right) = \sum_{t=2}^{\tilde{t}} \theta_c(t). \tag{3.9}$$

$\theta_c^{cumu}\left(\tilde{t}\right)$ can be interpreted as the cumulative average treatment effect among the units that have been treated by time $\tilde{t}$.

3.2. Aggregations into overall treatment effect parameters

Finally in this section, we consider some ideas for aggregating group time average treatment effects into an overall effect of participating in the treatment. One very simple idea is to just average all of the identified group-time average treatment effects together; i.e., to consider the parameter

$$\theta_W^O = \frac{1}{\kappa} \sum_{g \in \mathcal{G}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{t \geq g\} ATT(g,t) P(G=g|G \leq \mathcal{T}) \tag{3.10}$$

where $\kappa = \sum_{g \in \mathcal{G}} \sum_{t=2}^{\mathcal{T}} 1\{t \geq g\} P(G=g|G \leq \mathcal{T})$ (which ensures that the weights on $ATT(g,t)$ in the second term sum up to one). θ_W^O is a weighted average of each $ATT(g,t)$ putting more weight on $ATT(g,t)$'s with larger group sizes. Unlike β in the TWFE regression specification (3.2), this simple combination of $ATT(g,t)$'s immediately rules out troubling issues due to negative weights; as a particular example, when the effect of participating in the treatment is positive for all units, this aggregated parameter cannot be negative.

That being said, just requiring positive weights is a very minimal requirement of a reasonable overall treatment effect parameter. For example, one drawback of θ_W^O is that it systematically puts more weight on groups that participate in the treatment for longer. Instead, we suggest the following parameter as a general-purpose summary of the average effect of participating in the treatment

$$\theta_{sel}^O = \sum_{g \in \mathcal{G}} \theta_{sel}(g) P(G=g|G \leq \mathcal{T}) \tag{3.11}$$

where $\theta_{sel}(g)$ is the average effect of participating in the treatment for units in group g as defined in Eq. (3.7) above. θ_{sel}^O first computes the average effect for each group (across all time periods) and then averages these effects together across groups to summarize the overall average effect of participating in the treatment. Thus, θ_{sel}^O is the average effect of participating in the treatment experienced by all units that ever participated in the treatment. In this respect, its interpretation is the same as the ATT in the canonical DiD setup with two periods and two groups. This is an attractive property for a summary measure of the overall effect of participating in the treatment in the context of multiple time periods and variation in treatment timing.

Working by analogy, one can also define overall treatment effect parameters by averaging $\theta_{es}(e)$ across all event times or $\theta_c(t)$ across all time periods, i.e.,

$$\theta_{es}^O = \frac{1}{\mathcal{T}-1} \sum_{e=0}^{\mathcal{T}-2} \theta_{es}(e) \qquad \theta_c^O = \frac{1}{\mathcal{T}-1} \sum_{t=2}^{\mathcal{T}} \theta_c(t) \tag{3.12}$$

In our view, the appeal of these aggregations is likely to be somewhat more limited than that of θ_{sel}^O in most applications. For example, the interpretation of θ_{es}^O is complicated by the issue of the changing composition of groups across different values of e discussed above (similar arguments apply to θ_c^O as well).

As before, one can circumvent the issue of the changing composition of groups by balancing the sample with respect to event time. A (local) single summary parameter is given by

$$\theta_{es}^{O,bal}(e') = \frac{1}{e'+1} \sum_{e=0}^{e'} \theta_{es}^{bal}(e, e') \tag{3.13}$$

This is the average effect of participating in the treatment over the first e' periods of exposure to the treatment. This is also a reasonable alternative overall treatment effect parameter, but it should also be noted that it is local to groups that participated in the treatment for at least e' periods.

As a final comment, in general, none of the overall effect parameters considered in this section are equal to each other except in the special case where $ATT(g,t)$ is the same for all groups and all time periods. In that case, all of the aggregated parameters, including β from the TWFE regression, are equal to each other.

4. Estimation and inference

So far we have focused on the identification and aggregation stages of the analysis. In this section, we show how one can build on these results to form estimators for and conduct inference about the group-time average treatment effects and their summary measures described in Section 3. Given that the $ATT(g,t)$'s are the main building blocks of our analysis, we start with them.

First, it is important to notice that our identification results in Theorem 1 are constructive and suggest a simple and intuitive two-step estimation strategy to estimate the $ATT(g,t)$'s. In the first step, one estimates the nuisance functions for each group g and time period t – i.e., $p_g(x)$ and/or $m_{g,t,\delta}^{nev}(X)$ if one relies on Assumption 4, and $p_{g,t+\delta}(x)$ and/or $m_{g,t,\delta}^{ny}(X)$ if one relies on Assumption 5. In the second step, one plugs the fitted values of these estimated nuisance functions into the sample analogue of the considered $ATT(g,t)$ estimand to obtain estimates of the group-time average treatment effect.

A natural question that then arises is which type of approach one should use in practice: the outcome regression, inverse probability weighting, or the doubly-robust one. Although these three different approaches are equivalent from the *identification* perspective, this is not the case from the *estimation/inference* perspective. The OR approach requires researchers to *correctly* model the outcome evolution of the comparison group to estimate the group-time average treatment effects. This approach is explicitly connected with the conditional parallel trends assumption required in DiD analysis as this condition is usually expressed in terms of conditional expectations. The IPW approach, on the other hand, avoids explicitly modeling the outcome evolution of the comparison group and therefore does not rely on putative model restrictions directly tied to the parameter of interest. Instead, the IPW approach requires one to *correctly* model the conditional probability of unit i being in group g given their covariates X and that they are either in group g or in an appropriate comparison group. The DR approach combines both the OR and IPW approaches as it relies on modeling both the outcome evolution and the propensity score. However, it only requires one to correctly specify *either (but not necessarily both)* the outcome evolution for the comparison group or the propensity score model (Sant'Anna and Zhao, 2020). Thus, the DR approach enjoys additional robustness against model misspecifications when compared to the OR and IPW approaches. In addition, the DR approach potentially allows one to use a broader set of estimation methods such as those that involve penalization and some types of model selection, see, e.g. Belloni et al. (2017).

Given these attractive robustness features associated with the DR approach, in this section we consider estimators of the DR form; the discussion on how to proceed with the OR and IPW approaches is analogous and therefore omitted. We also focus on parametric estimators for the nuisance functions. We consider this case mainly for its practical appeal which is especially true in applications where the number of covariates is fairly large and the number of observations is only moderate.[12]

More concisely, let

$$\widehat{ATT}^{nev}_{dr}(g,t;\delta) = \mathbb{E}_n\left[\left(\widehat{w}^{treat}_g - \widehat{w}^{comp,nev}_g\right)\left(Y_t - Y_{g-\delta-1} - \widehat{m}^{nev}_{g,t,\delta}\left(X;\widehat{\beta}^{nev}_{g,t,\delta}\right)\right)\right], \tag{4.1}$$

$$\widehat{ATT}^{ny}_{dr}(g,t;\delta) = \mathbb{E}_n\left[\left(\widehat{w}^{treat}_g - \widehat{w}^{comp,ny}_g\right)\left(Y_t - Y_{g-\delta-1} - \widehat{m}^{ny}_{g,t,\delta}\left(X;\widehat{\beta}^{ny}_{g,t,\delta}\right)\right)\right], \tag{4.2}$$

where

$$\widehat{w}^{treat}_g = \frac{G_g}{\mathbb{E}_n\left[G_g\right]},\quad \widehat{w}^{comp,nev}_g = \frac{\dfrac{\widehat{p}_g\left(X;\widehat{\pi}_g\right)C}{1-\widehat{p}_g\left(X;\widehat{\pi}_g\right)}}{\mathbb{E}_n\left[\dfrac{\widehat{p}_g\left(X;\widehat{\pi}_g\right)C}{1-\widehat{p}_g\left(X;\widehat{\pi}_g\right)}\right]},\quad \widehat{w}^{comp,ny}_g = \frac{\dfrac{\widehat{p}_{g,t+\delta}\left(X;\widehat{\pi}_{g,t+\delta}\right)(1-D_{t+\delta})\left(1-G_g\right)}{1-\widehat{p}_{g,t+\delta}\left(X;\widehat{\pi}_{g,t+\delta}\right)}}{\mathbb{E}_n\left[\dfrac{\widehat{p}_{g,t+\delta}\left(X;\widehat{\pi}_{g,t+\delta}\right)(1-D_{t+\delta})\left(1-G_g\right)}{1-\widehat{p}_{g,t+\delta}\left(X;\widehat{\pi}_{g,t+\delta}\right)}\right]},$$

with $\widehat{p}_g\left(\cdot;\widehat{\pi}_g\right)$, $\widehat{p}_{g,t+\delta}(\cdot;\widehat{\pi}_{g,t+\delta})$, $\widehat{m}^{nev}_{g,t,\delta}(\cdot;\widehat{\beta}^{nev}_{g,t,\delta})$ and $\widehat{m}^{ny}_{g,t,\delta}(\cdot;\widehat{\beta}^{ny}_{g,t,\delta})$ being (parametric) estimators of $p_g(\cdot)$, $p_{g,t+\delta}(\cdot)$, $m^{nev}_{g,t,\delta}(\cdot)$ and $m^{ny}_{g,t,\delta}(\cdot)$, respectively, and for a generic Z, $\mathbb{E}_n[Z] = n^{-1}\sum_{i=1}^n Z_i$. $\widehat{ATT}^{nev}_{dr}(g,t;\delta)$ and $\widehat{ATT}^{nev}_{dr}(g,t;\delta)$ are our proposed DR DiD estimators for $ATT(g,t)$ when one invokes Assumption 4 and Assumption 5, respectively. These estimators extend the DR DiD estimators of Sant'Anna and Zhao (2020) from the two periods, two groups setup to the multiple groups, multiple periods setup while allowing for possible treatment anticipation. In addition, these estimators are of the Hájek (1971)-type and their associated weights are guaranteed to sum up to one in finite samples. As illustrated by Busso et al. (2014), this usually leads to improved finite sample properties.

With the estimators for the $ATT(g,t)$'s in hand, one can use the analogy principle and combine these to estimate the summarized average treatment effect parameters discussed in Section 3.

Remark 9. In applications with limited covariate overlap (i.e., with propensity scores sufficiently close to one), IPW and DR estimators may lead to imprecise (irregular) inference procedures, see, e.g., Khan and Tamer (2010). In such cases, provided that one is comfortable with (parametric) extrapolation and is sufficiently confident that the outcome regression working models are correctly specified, relying on the OR estimation approach may lead to more informative inferences. Alternatively, one may choose to trim extreme propensity score estimates though proceeding in this manner would change the target parameter; i.e., we would not be recovering the $ATT(g,t)$'s; see, e.g., Crump et al. (2009) and Yang and Ding (2018) for related discussion in other contexts. In the rest of the paper, we abstract from these points.

4.1. Asymptotic theory for group-time average treatment effects

Next, we derive the asymptotic properties of our DR DiD estimators for the $ATT(g,t)$'s. To simplify exposition, we focus on the case with a never-treated comparison group as in (4.1); results that come from using the not-yet-treated group

[12] Alternatively, one could adopt a fully nonparametric approach. Let $f(x)$ be a generic notation for the nuisance functions. From Newey (1994), Chen et al. (2003), Ai and Chen (2003, 2007, 2012), and Chen et al. (2008), one can see that the use of nonparametric first-step estimators $\widehat{g}(x)$ of $g(x)$ is warranted provided that $\|\widehat{g}(x) - g(x)\|_{\mathcal{H}} = o_p\left(n^{-1/4}\right)$ for a pseudo-metric $\|\cdot\|_{\mathcal{H}}$, $\mathcal{H}$ being a vector space of functions. However, when the dimension of X is moderate or large, as is often the case in empirical applications, conditions ensuring that $\|\widehat{g}(x) - g(x)\|_{\mathcal{H}} = o_p\left(n^{-1/4}\right)$ can be rather stringent due to the so-called "curse of dimensionality".

as the comparison group as in (4.2) follow from symmetric arguments and are therefore omitted. We also note that the theoretical results in this section are justified within the large n, fixed $\mathcal{T}$ paradigm.

Let $\|Z\| = \sqrt{trace\,(Z'Z)}$ denote the Euclidean norm of Z and set $W = (Y_1, \ldots, Y_{\mathcal{T}}, X, D_1, \ldots, D_{\mathcal{T}})$. For a generic $\kappa_{g,t}^{nev} = \left(\pi_g', \beta_{g,t,\delta}^{nev}\right)'$, let

$$h_{g,t}^{dr,nev}\left(W; \kappa_{g,t}^{nev}, \delta\right) = \left(w_g^{treat}\left(W\right) - w_g^{comp,nev}\left(W; \pi_g\right)\right)\left(Y_t - Y_{g-\delta-1} - m_{g,t,\delta}^{nev}\left(X; \beta_{g,t,\delta}^{nev}\right)\right),$$

where the normalized weights $w_g^{treat}\left(W\right)$ and $w_g^{comp,nev}\left(W; \pi_g\right)$ are given by

$$w_g^{treat}\left(W\right) = \frac{G_g}{\mathbb{E}\left[G_g\right]}, \qquad w_g^{comp,nev}\left(W; \pi_g\right) = \frac{p_g\left(X; \pi_g\right)C}{1 - p_g\left(X; \pi_g\right)} \Bigg/ \mathbb{E}\left[\frac{p_g\left(X; \pi_g\right)C}{1 - p_g\left(X; \pi_g\right)}\right]. \tag{4.3}$$

Let $g\left(\cdot\right)$ be a generic notation for $p_g\left(\cdot\right)$ and $m_{g,t,\delta}^{nev}\left(\cdot\right)$. With some abuse of notation, let $g\left(\cdot; \gamma\right)$ be a generic notation for $p_g\left(\cdot; \pi_g\right)$ and $m_{g,t,\delta}^{nev}\left(\cdot; \beta_{g,t,\delta}^{nev}\right)$. The vector of pseudo-true parameters is given by $\kappa_{g,t}^{*,nev} = \left(\pi_g^{*\prime}, \beta_{g,t,\delta}^{*,nev\,\prime}\right)'$. Finally, let $\dot{h}_{g,t}^{dr,nev}\left(W; \kappa_{g,t}^{nev}\right) = \partial\, h_{g,t}^{dr,nev}\left(W; \kappa_{g,t}^{nev}\right) \big/ \partial \kappa_{g,t}^{nev}$.

Assumption 7. (i) $g\left(x; \gamma\right)$ is a parametric model for $g\left(x\right)$, where $\gamma \in \Theta \subset \mathbb{R}^k$, Θ being compact; (ii) $g\left(X; \gamma\right)$ is *a.s.* continuous at each $\gamma \in \Theta$; (iii) there exists a unique pseudo-true parameter $\gamma^* \in int\left(\Theta\right)$; (iv) $g\left(X; \gamma\right)$ is *a.s.* twice continuously differentiable in a neighborhood of γ^*, $\Theta^* \subset \Theta$; (v) the estimator $\widehat{\gamma}$ is strongly consistent for γ^* and satisfies the following linear expansion:

$$\sqrt{n}\left(\widehat{\gamma} - \gamma^*\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_{g,t}\left(W_i; \gamma^*\right) + o_p\left(1\right),$$

where $l_{g,t}\left(\cdot; \gamma\right)$ is a $k \times 1$ vector such that $\mathbb{E}\left[l_{g,t}\left(W; \gamma^*\right)\right] = 0$, $\mathbb{E}\left[l_{g,t}\left(W; \gamma^*\right) l_{g,t}\left(W; \gamma^*\right)'\right]$ exists and is positive definite and $\lim_{s \to 0} \mathbb{E}\left[\sup_{\gamma \in \Theta^*: \|\gamma - \gamma^*\| \le s}\left\|l_g\left(W; \gamma\right) - l_g\left(W; \gamma^*\right)\right\|^2\right] = 0$. In addition, (vi) for some $\varepsilon > 0$ and all $g \in \mathcal{G}$, $0 \le p_g\left(X; \pi_g\right) \le 1 - \varepsilon$ *a.s.*, for all $\pi \in int\left(\Theta^{ps}\right)$, where Θ^{ps} denotes the parameter space of π_g.

Assumption 8. For each $g \in \mathcal{G}$ and $t = \{2, \ldots, \mathcal{T} - \delta\}$, assume that $\mathbb{E}\left[\left\|h_{g,t}^{nev}\left(W; \kappa_{g,t}^{*,nev}, \delta\right)\right\|^2\right] < \infty$ and $\mathbb{E}\left[\sup_{\kappa \in \Gamma^*}\left|\dot{h}_{g,t}^{nev}\left(W; \kappa\right)\right|\right] < \infty$, where Γ^* is a small neighborhood of $\kappa_{g,t}^{*,nev}$.

Assumptions 7–8 are standard in the literature, see e.g. Abadie (2005), Wooldridge (2007), Bonhomme and Sauder (2011), Graham et al. (2012), and Sant'Anna and Zhao (2020). Assumption 7 requires that the first-step estimators are based on smooth parametric models and that the estimated parameters admit $\sqrt{n}$-asymptotically linear representations, whereas Assumption 8 imposes some weak integrability conditions. Under mild moment conditions, these requirements are fulfilled when one adopts linear/nonlinear outcome regressions or logit/probit models, for example, and estimates the unknown parameters by (nonlinear) least squares, quasi-maximum likelihood, or other alternative estimation methods, see e.g. Chapter 5 in van der Vaart (1998), Wooldridge (2007), Graham et al. (2012) and Sant'Anna and Zhao (2020). In other words, Assumptions 7–8 allow for flexible parametric specifications of the nuisance functions and accommodate different estimation methods.

In what follows, we write $w_g^{treat} = w_g^{treat}\left(W\right)$, $w_g^{comp}\left(\pi_g\right) = w_g^{comp,nev}\left(W; \pi_g\right)$, and $m_{g,t,\delta}^{nev}\left(\beta_{g,t,\delta}^{nev}\right) = m_{g,t,\delta}^{nev}\left(X; \beta_{g,t,\delta}^{nev}\right)$ to minimize notation. For a generic $\kappa_{g,t}^{nev} = \left(\pi_g', \beta_{g,t,\delta}^{nev\prime}\right)'$, define

$$\psi_{g,t,\delta}^{dr,nev}(W_i; \kappa_{g,t}^{nev}) = \psi_{g,t,\delta}^{treat,nev}(W_i; \beta_{g,t,\delta}^{nev}) - \psi_{g,t,\delta}^{comp,nev}(W_i; \pi_g, \beta_{g,t,\delta}^{nev}) - \psi_{g,t,\delta}^{est,nev}(W_i; \pi_g, \beta_{g,t,\delta}^{nev}), \tag{4.4}$$

with

$$\begin{aligned}
\psi_{g,t,\delta}^{treat,nev}(W; \beta_{g,t,\delta}^{nev}) &= w_g^{treat} \cdot \left(Y_t - Y_{g-\delta-1} - m_{g,t,\delta}^{nev}\left(\beta_{g,t,\delta}^{nev}\right)\right) \\
&\quad - w_g^{treat} \cdot \mathbb{E}\left[w_g^{treat} \cdot \left(Y_t - Y_{g-\delta-1} - m_{g,t,\delta}^{nev}\left(\beta_{g,t,\delta}^{nev}\right)\right)\right], \\
\psi_{g,t,\delta}^{comp,nev}(W; \pi_g, \beta_{g,t,\delta}^{nev}) &= w_g^{comp}\left(\pi_g\right) \cdot \left(Y_t - Y_{g-\delta-1} - m_{g,t,\delta}^{nev}\left(\beta_{g,t,\delta}^{nev}\right)\right) \\
&\quad - w_g^{comp}\left(\pi_g\right) \cdot \mathbb{E}\left[w_g^{comp}\left(\pi_g\right) \cdot \left(Y_t - Y_{g-\delta-1} - m_{g,t,\delta}^{nev}\left(\beta_{g,t,\delta}^{nev}\right)\right)\right],
\end{aligned}$$

and

$$\psi_{g,t}^{est,nev}(W; \pi_g, \beta_{g,t,\delta}^{nev}) = l_{g,t}^{or,nev}\left(\beta_{g,t,\delta}^{nev}\right)' \cdot M_{g,t,\delta}^{dr,nev,1} + l_g^{ps,nev}\left(\pi_g\right)' \cdot M_{g,t,\delta}^{dr,nev,2},$$

where $l_{g,t}^{or,nev}\left(\cdot\right)$ is the asymptotic linear representation of the estimator for the outcome evolution of the comparison groups as described in Assumption 7(iv), $l_g^{ps,nev}\left(\cdot\right)$ is defined analogously for the generalized propensity score, and

$$
\begin{aligned}
M_{g,t,\delta}^{dr,nev,1} &= \mathbb{E}\left[\left(w_g^{treat} - w_g^{comp}\left(\pi_g\right)\right) \cdot \dot{m}_{g,t,\delta}^{nev}\left(\beta_{g,t,\delta}^{nev}\right)\right], \\
M_{g,t,\delta}^{dr,nev,2} &= \mathbb{E}\left[\alpha_g^{ps,nev}\left(\pi_g\right) \cdot \left(Y_t - Y_{g-\delta-1} - m_{g,t,\delta}^{nev}\left(\beta_{g,t,\delta}^{nev}\right)\right) \cdot \dot{p}_g\left(\pi_g\right)\right] \\
&\quad - \mathbb{E}\left[\alpha_g^{ps,nev}\left(\pi_g\right) \cdot w_g^{comp}\left(\pi_g\right) \cdot \left(Y_t - Y_{g-\delta-1} - m_{g,t,\delta}^{nev}\left(\beta_{g,t,\delta}^{nev}\right)\right) \cdot \dot{p}_g\left(\pi_g\right)\right],
\end{aligned}
$$

with $\dot{m}_{g,t,\delta}^{nev}\left(\beta_{g,t,\delta}^{nev}\right) = \partial m_{g,t,\delta}^{nev}\left(X;\beta_{g,t,\delta}^{nev}\right)/\partial \beta_{g,t,\delta}^{nev}$, $\dot{p}_g\left(\pi_g\right) = \partial p_g\left(X;\pi_g\right)/\partial \pi_g$, and

$$
\alpha_g^{ps,nev}\left(\pi_g\right) = \frac{C}{\left(1 - p_g\left(X;\pi_g\right)\right)^2} \Bigg/ \mathbb{E}\left[\frac{p_g\left(X;\pi_g\right)C}{1 - p_g\left(X;\pi_g\right)}\right].
$$

Finally, let $ATT_{t\geq(g-\delta)}$ and $\widehat{ATT}_{t\geq(g-\delta)}^{dr,nev}$ denote the vector of $ATT(g,t)$ and $\widehat{ATT}_{dr}^{nev}(g,t;\delta)$, respectively, for all $g \in \mathcal{G}_\delta$, $t \in \{2,\ldots\mathcal{T}-\circ\}$ such that $t \geq g-\delta$. Analogously, let $\Psi_{t\geq(g-\delta)}^{dr,nev}$ denote the collection of $\psi_{g,t,\delta}^{dr,nev}$ across all $g \in \mathcal{G}_\delta$, $t \in \{2,\ldots\mathcal{T}-\delta\}$ such that $t \geq g-\delta$. Consider the following claim:

$$
\begin{aligned}
&\text{For each } g \in \mathcal{G}_\delta,\ t \in \{2,\ldots\mathcal{T}-\delta\} \text{ such that } t \geq g-\delta, \\
&\exists \pi_g^* \in \Theta^{ps} : P\left(p_g(X;\pi_g^*) = p_g\left(X\right)\right) = 1 \quad \text{or} \\
&\exists \beta_{g,t,\delta}^{*,nev} \in \Theta^{reg} : P\left(m_{g,t,\delta}^{nev}\left(X;\beta_{g,t,\delta}^{*,nev}\right) = m_{g,t,\delta}^{nev}\left(X\right)\right) = 1.
\end{aligned} \tag{4.5}
$$

Claim (4.5) says that 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.

The next theorem establishes the joint limiting distribution of $\widehat{ATT}_{t\geq(g-\delta)}^{dr,nev}$.

Theorem 2. *Under Assumptions 1–4, 6–8, for each g and t such that $g \in \mathcal{G}_\delta$, $t \in \{2,\ldots\mathcal{T}-\delta\}$ and $t \geq g-\delta$, provided that (4.5) is true,*

$$
\sqrt{n}(\widehat{ATT}_{dr}^{nev}(g,t;\delta) - ATT(g,t)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{g,t,\delta}^{dr,nev}(W_i;\kappa_{g,t}^{*,nev}) + o_p(1).
$$

Furthermore, as $n \to \infty$,

$$
\sqrt{n}(\widehat{ATT}_{t\geq(g-\delta)}^{dr,nev} - ATT_{t\geq(g-\delta)}) \xrightarrow{d} N(0,\Sigma)
$$

where $\Sigma = \mathbb{E}[\Psi_{t\geq(g-\delta)}^{dr,nev}(W)\Psi_{t\geq(g-\delta)}^{dr,nev}(W)']$.

Theorem 2 provides the influence function for estimating the vector of group-time average treatment effects, $ATT_{t\geq(g-\delta)}$, as well as its limiting distribution. Importantly, Theorem 2 emphasizes the DR property of $\widehat{ATT}_{dr}^{nev}(g,t;\delta)$: it recovers the $ATT(g,t)$ provided that either the propensity score working model or outcome regression working model for the "never treated" is correctly specified.

In order to conduct inference, one can show that the sample analogue of Σ is a consistent estimator for Σ, which leads directly to standard errors and pointwise confidence intervals. Instead of following this route, we propose to use a simple multiplier bootstrap procedure to conduct asymptotically valid inference. Our proposed bootstrap leverages the asymptotic linear representations derived in Theorem 2 and inherits important advantages. First, it is easy to implement and very fast to compute. Each bootstrap iteration simply amounts to "perturbing" the influence function by a random weight V, and it does not require re-estimating the propensity score in each bootstrap draw. Second, in each bootstrap iteration, there are always observations from each group. This can be a real problem with the traditional empirical bootstrap where there may be no observations from a particular group in some particular bootstrap iteration. Third, computation of simultaneously (in g and t) valid confidence bands is relatively straightforward. This is particularly important since researchers are likely to use confidence bands to visualize estimation uncertainty about $ATT(g,t)$. Unlike pointwise confidence bands, simultaneous confidences bands do not suffer from multiple-testing problems and are guaranteed to cover all $ATT(g,t)$'s with a probability at least $1-\alpha$. Finally, we note that our proposed bootstrap procedure can be readily modified to account for clustering, see Remark 10.

To proceed, let $\widehat{\Psi}_{t\geq(g-\delta)}^{dr,nev}(W)$ denote the sample-analogue of $\Psi_{t\geq(g-\delta)}^{dr,nev}(W)$, where population expectations are replaced by their empirical analogue, and the true nuisance functions and their derivatives are replaced by their estimators. Let $\{V_i\}_{i=1}^n$ be a sequence of *iid* random variables with zero mean, unit variance, and finite third moment, independent of the original sample $\{W_i\}_{i=1}^n$. A popular example involves *iid* Bernoulli variates $\{V_i\}$ with $P\left(V = 1-\kappa\right) = \kappa/\sqrt{5}$ and $P\left(V = \kappa\right) = 1-\kappa/\sqrt{5}$, where $\kappa = \left(\sqrt{5}+1\right)/2$, as suggested by Mammen (1993).

We define $\widehat{ATT}_{t\geq(g-\delta)}^{*,dr,nev}$, a bootstrap draw of $\widehat{ATT}_{t\geq(g-\delta)}^{dr,nev}$, via

$$
\widehat{ATT}_{t\geq(g-\delta)}^{*,dr,nev} = \widehat{ATT}_{t\geq(g-\delta)}^{dr,nev} + \mathbb{E}_n\left[V \cdot \widehat{\Psi}_{t\geq(g-\delta)}^{dr,nev}(W)\right]. \tag{4.6}
$$

The next theorem establishes the asymptotic validity of the multiplier bootstrap procedure proposed above.

Theorem 3. *Under the assumptions of Theorem 2*

$$\sqrt{n}\left(\widehat{ATT}^{*,dr,nev}_{t\geq(g-\delta)} - \widehat{ATT}^{dr,nev}_{t\geq(g-\delta)}\right) \xrightarrow[*]{d} N(0, \Sigma),$$

where Σ is as in Theorem 2, and $\xrightarrow[*]{d}$ denotes weak convergence (convergence in distribution) of the bootstrap law in probability, i.e., conditional on the original sample $\{W_i\}_{i=1}^n$. Additionally, for any continuous functional $\Gamma(\cdot)$,[13]

$$\Gamma\left(\sqrt{n}\left(\widehat{ATT}^{*,dr,nev}_{t\geq(g-\delta)} - \widehat{ATT}^{dr,nev}_{t\geq(g-\delta)}\right)\right) \xrightarrow[*]{d} \Gamma\left(N(0, \Sigma)\right).$$

We now describe a practical bootstrap algorithm to compute studentized confidence bands that cover $ATT(g, t)$ simultaneously over all $t \geq g - \delta$ with a pre-specified probability $1 - \alpha$ in large samples. This is similar to the bootstrap procedures used in Kline and Santos (2012), Belloni et al. (2017) and Chernozhukov et al. (2018) in different contexts.

Algorithm 1. (1) Draw a realization of $\{V_i\}_{i=1}^n$. (2) Compute $\widehat{ATT}^{*,dr,nev}_{t\geq(g-\delta)}$ as in (4.6), denote its (g, t)-element as $\widehat{ATT}^{*}(g, t)$, and form a bootstrap draw of its limiting distribution as

$$\hat{R}^*(g, t) = \sqrt{n}\left(\widehat{ATT}^{*}(g, t) - \widehat{ATT}(g, t)\right).$$

(3) Repeat steps 1–2 B times. (4) Compute a bootstrap estimator of the main diagonal of $\Sigma^{1/2}$ such as the bootstrap interquartile range normalized by the interquartile range of the standard normal distribution, $\widehat{\Sigma}^{1/2}(g, t) = (q_{0.75}(g, t) - q_{0.25}(g, t))/(z_{0.75} - z_{0.25})$, where $q_p(g, t)$ is the *pth* sample quantile of the $\hat{R}^*(g, t)$ in the B draws, and z_p is the *pth* quantile of the standard normal distribution. (5) For each bootstrap draw, compute $t\text{-}test_{t\geq(g-\delta)} = \max_{(g,t)}\left|\hat{R}^*(g, t)\right|\widehat{\Sigma}(g, t)^{-1/2}$. (5) Construct $\widehat{c}_{1-\alpha}$ as the empirical $(1 - a)$-quantile of the B bootstrap draws of $t\text{-}test_{t\geq(g-\delta)}$. (6) Construct the bootstrapped simultaneous confidence band for $ATT(g, t)$, $t \geq (g - \delta)$, as $\widehat{C}(g, t) = [\widehat{ATT}^{nev}_{dr}(g, t; \delta) \pm \widehat{c}_{1-\alpha}\widehat{\Sigma}(g, t)^{-1/2}/\sqrt{n}]$.

The next corollary to Theorem 3 states that the simultaneous confidence band for $ATT(g, t)$ described in Algorithm 1 has correct asymptotic coverage.

Corollary 1. *Under the assumptions of Theorem 2, for any $0 < \alpha < 1$, as $n \to \infty$,*

$$P\left(ATT(g, t) \in \widehat{C}(g, t) \quad \forall t \in \{2, \ldots, \mathcal{T}\}, g \in \mathcal{G}_\delta : t \geq g - \delta\right) \to 1 - \alpha,$$

where $\widehat{C}(g, t)$ is as defined in Algorithm 1.

Remark 10. In DiD applications, it is common to use "cluster-robust" inference procedures; see, e.g., Wooldridge (2003) and Bertrand et al. (2004). However, we note that the choice of whether to cluster or not is usually not obvious, and depends on the kind of uncertainty one is trying to reflect; see, e.g., Abadie et al. (2017) for a discussion in a cross-sectional setup.[14] In the case that one wishes to account for clustering to reflect "cluster-based" sampling uncertainty, we note that this can be done in a straightforward manner using a small modification of the multiplier bootstrap described above, provided that the number of cluster is "large." More precisely, instead of drawing observation-specific V's, one simply needs to draw cluster-specific V's; see, e.g., Sherman and Le Cessie (2007), Kline and Santos (2012), Cheng et al. (2013), and MacKinnon and Webb (2018, 2020). If the number of clusters is "small," however, the application of the aforementioned bootstrap procedure is not warranted.[15]

Remark 11. In Algorithm 1 we have required an estimator for the main diagonal of Σ. However, we note that if one takes $\widehat{\Sigma}(g, t) = 1$ for all (g, t), the result in Corollary 1 continues to hold. However, the resulting "constant width" simultaneous confidence band may be of larger length; see, e.g., Montiel Olea and Plagborg-Møller (2018) and Freyberger and Rai (2018).

Remark 12. The above results focus on making inference about $ATT(g, t)$'s in (effective) post-treatment periods $t \geq g - \delta$. Although the limited anticipation condition in Assumption 3 implies that $ATT(g, t) = 0$ for all $t < g - \delta$ regardless of the group g, it is common practice to also estimate these pre-treatment parameters and use them to assess the credibility of the underlying identifying assumptions. Note that our DiD estimands (2.2)–(2.7) can be easily adjusted to include these by simply replacing the "long differences" $(Y_t - Y_{g-\delta-1})$ with the "short differences" $(Y_t - Y_{t-1})$ for all $t < g - \delta$. All our results continue to hold when one augments $\widehat{ATT}^{dr,nev}_{t\geq(g-\delta)}$ to also include these estimates for the $ATT(g, t)$'s in the pre-treatment periods $t < g - \delta$.

13 Since the number of periods $\mathcal{T}$ is fixed, $\Gamma(\cdot)$ should be interpreted as a continuous functional between Euclidean spaces.

14 The formal results in Abadie et al. (2017) focus on the cross section case and rely on additional functional form restrictions that we do not impose in this paper. Fully extending the results of Abadie et al. (2017) to the semiparametric panel data case is beyond the scope of our paper.

15 In such cases, provided that one is comfortable imposing additional functional form assumptions, one could use alternative procedures such as Conley and Taber (2011) and Ferman and Pinto (2019). Extending these proposals to our setup is beyond the scope of this paper though.

4.2. Asymptotic theory for summary parameters

Assume, for simplicity, that Assumption 3 holds with $\delta = 0$. In this section, we discuss how one can estimate and make inference about the summary measures of the causal effects discussed in Section 3. More concisely, we consider parameters of the form of θ as defined in (3.1), which covers all of the aggregated parameters discussed in Section 3.

Given the discussion in Section 4.1, a natural way to estimate θ is to use the plug-in type estimators

$$\hat{\theta} = \sum_{g \in \mathcal{G}} \sum_{t=2}^{\mathcal{T}} \widehat{w}(g,t) \widehat{ATT}_{dr}^{nev}(g,t;0),$$

where $\widehat{w}(g,t)$ are estimators for $w(g,t)$ such that for all $g \in \mathcal{G}$ and $t = 2, \ldots, \mathcal{T}$,

$$\sqrt{n}\left(\widehat{w}(g,t) - w(g,t)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{g,t}^{w}(\mathcal{W}_i) + o_p(1),$$

with $\mathbb{E}\left[\xi_{gt}^{w}(\mathcal{W})\right] = 0$ and $\mathbb{E}\left[\xi_{gt}^{w}(\mathcal{W})\xi_{gt}^{w}(\mathcal{W})'\right]$ finite and positive definite. Estimators based on the sample analogue of the weights discussed in Section 3 satisfy this condition.

Let

$$l^{w}(W_i) = \sum_{g \in \mathcal{G}} \sum_{t=2}^{\mathcal{T}} \left(w(g,t) \cdot \psi_{g,t,0}^{dr,nev}(W_i; \kappa_{g,t}^{*,nev}) + \xi_{g,t}^{w}(W_i) \cdot ATT(g,t) \right),$$

where $\psi_{g,t,\delta}^{dr,nev}$ are as defined in (4.4).

The following result follows immediately from Theorem 2, and can be used to conduct asymptotically valid inference for the summary causal parameters θ.

Corollary 2. *Under the assumptions of Theorem 2,*

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l^{w}(W_i) + o_p(1) \\ &\xrightarrow{d} N\left(0, \mathbb{E}\left[l^{w}(W)^2\right]\right)\end{aligned}$$

Corollary 2 implies that one can construct standard errors and confidence intervals for summary treatment effect parameters based on a consistent estimator of $\mathbb{E}\left[l^{w}(W)^2\right]$ or by using a bootstrap procedure like the one in Algorithm 1. The main advantage of using the bootstrap procedure akin to Algorithm 1 is that inference procedures would be robust against multiple-testing problems. This is particularly attractive when considering $\theta_{es}(e)$, $\theta_{es}^{bal}(e;e')$, $\theta_{sel}(\tilde{g})$, and $\theta_c(t)$, as practitioners would probably analyze how these parameters differ across event-times e, groups $\tilde{g}$, and calendar-time t.

Remark 13. As discussed in Remark 10, the validity of the "cluster-robust" multiplier bootstrap procedure relies on the number of clusters being "large." In some applications such a condition may be more plausible when analyzing the aggregated parameter θ than when analyzing the $ATT(g,t)$ themselves.

5. The effect of minimum wage policy on teen employment

In this section, we illustrate the empirical relevance of our proposed methods. To do this, we apply our methods to study the effect of the minimum wage on teen employment. The main goal of this section is to compare results arising from using a TWFE specification (as is most common in applications) to results coming from our proposed method. We think that this comparison is important in order to get a sense of whether the theoretical limitations of TWFE discussed in recent work end up translating into meaningful differences in applications. Moreover, one might expect that understanding the effect of a minimum wage change on employment is a challenging case for TWFE as the effect of the minimum wage may be dynamic (Meer and West, 2016) and the timing of minimum wage changes varies across states. Unlike TWFE, the approach that we have proposed in the current paper is robust to these challenges.

By far the most common approach to trying to understand the effect of the minimum wage on employment is to exploit variation in the timing of minimum wage increases across states. Our identification strategy follows this approach. In particular, we consider a time period from 2001–2007 where the federal minimum wage was flat at \$5.15 per hour. We focus on county level teen employment in states whose minimum wage was equal to the federal minimum wage at the beginning of the period. Some of these states increased their minimum wage over this period – these become treated groups. In particular, we define groups by the time period when a state first increased its minimum wage. Others did not increase their minimum wage – these are the untreated group. This setup allows us to have more data than local case study approaches. On the other hand, it also allows us to have cleaner identification (state-level minimum wage policy changes) than in studies with more periods; the latter setup is more complicated than ours particularly because

Table 2
Summary statistics for main dataset.
Source: Quarterly Workforce Indicators and 2000 County Data Book.

	Treated counties	Untreated counties	Diff.	P-val on Diff.
Midwest	0.59	0.34	0.25	0.00
South	0.27	0.59	−0.32	0.00
West	0.14	0.07	0.07	0.00
Population (1000s)	94.32	53.43	40.89	0.00
White	0.89	0.83	0.06	0.00
HS Graduates	0.59	0.55	0.04	0.00
Poverty Rate	0.13	0.16	−0.03	0.00
Median Inc. (1000s)	33.91	31.89	2.02	0.00

Notes: Summary statistics for counties located in states that raised their minimum wage between Q2 of 2003 and Q1 of 2007 (treated) and states whose minimum wage was effectively set at the federal minimum wage for the entire period (untreated). The sample consists of 2284 counties.

of the variation in the federal minimum wage over time. It also allows us to check for internal consistency of identifying assumptions — namely whether or not the identifying assumptions hold in periods before particular states raised their minimum wages.

We use county level data on teen employment and other county characteristics. County level teen employment comes from the Quarterly Workforce Indicators (QWI), as in Dube et al. (2016); see Dube et al. (2016) for a detailed discussion of this dataset. Other pre-treatment county characteristics come from the 2000 County Data Book. These include county population in 2000, the fraction of the population that is white, educational characteristics from 1990, median income in 1997, and the fraction of the population below the poverty level in 1997. After dropping ten states due to their minimum wage being higher than the federal minimum wage in 2000, seven other states for lack of data on teen employment, and four other states in the Northern census region, our final sample includes county-level data from 29 states. We provide additional details on constructing the data in the Supplementary Appendix.

Summary statistics for county characteristics are provided in Table 2. There are some notable differences in county characteristics between counties in states that increased their minimum wage and in states that did not increase their minimum wage. Treated counties are much less likely to be in the South. They also have much higher population (on average 94,000 compared to 53,000 for untreated counties). The proportion of white residents is higher in treated counties (on average, 89% compared to 83% for untreated counties). There are smaller differences in the fraction with high school degrees and the poverty rate though the differences are both statistically significant. Treated counties have a somewhat higher fraction of high school graduates and a somewhat lower poverty rate.

5.1. Results

In the following we discuss different sets of results using different identification strategies. In particular, we consider the cases in which one would assume that the parallel trends assumption would hold unconditionally, and when it holds only after controlling on observed characteristics X. In the main text, we consider the case where never-treated counties are the comparison group and where we do not allow for any anticipation effects (i.e., $\delta = 0$). We provide results using the not-yet-treated counties as the comparison group and allowing for one year anticipation in the Supplementary Appendix; results from those cases are quite similar to the ones presented here.

The first set of results comes from using the unconditional parallel trends assumption to estimate the effect of raising the minimum wage on teen employment. The results for group-time average treatment effects are reported in Panel (a) of Fig. 1 along with a simultaneous 95% confidence band. All inference procedures use clustered bootstrapped standard errors at the county level, and account for the autocorrelation of the data. The plot contains pre-treatment estimates that can be used to "pre-test" the parallel trends assumption as well as treatment effect estimates in post-treatment periods.

The group-time average treatment effect estimates provide support for the view that increasing the minimum wage led to a reduction in teen employment. For 5 out of 7 group-time average treatment effects, there is a clear statistically significant negative effect on employment. The other two are marginally insignificant (and negative). The group-time average treatment effects range from 2.3% lower teen employment to 13.6% lower teen employment. The simple average (weighted only by group size) is 5.2% lower teen employment, and the average effect of a minimum wage increase across all groups that increased their minimum wage (corresponding to an estimate of θ_{sel}^{O} above) is 3.9% lower teen employment (see Panel (a) of Table 3). A two-way fixed effects model with a post treatment dummy variable also provides similar results, indicating 3.7% lower teen employment due to increasing the minimum wage. In light of the literature on the minimum wage these results are not surprising as they correspond to the types of regressions that tend to find that increasing the minimum wage decreases employment; see the discussion in Dube et al. (2010).

As in Meer and West (2016), there also appears to be a dynamic effect of increasing the minimum wage. For Illinois (the only state in the group that first raised its minimum wage in 2004), teen employment is estimated to be 3.4% lower on average in 2004 than it would have been if the minimum wage had not been increased. In 2005, teen employment

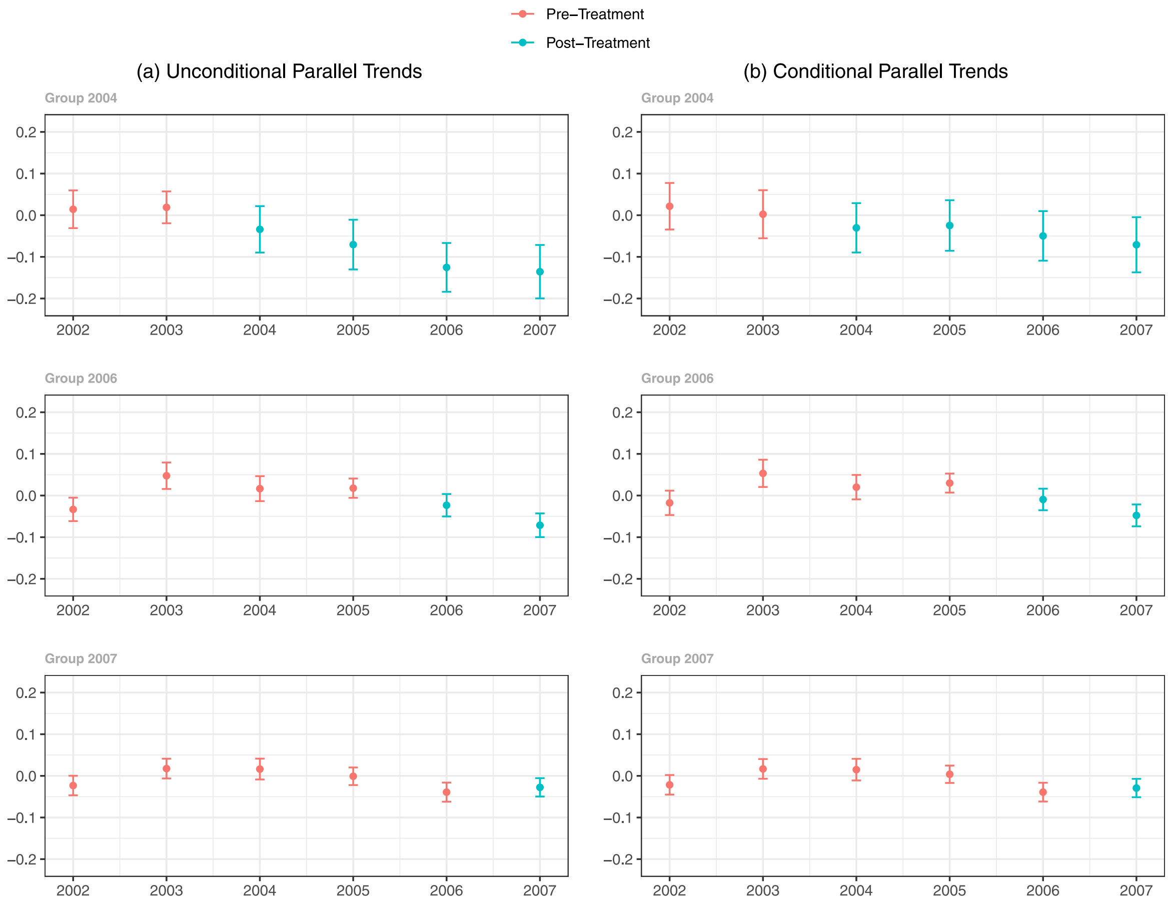

Fig. 1. Minimum wage group-time average treatment effects. *Notes:* The effect of the minimum wage on teen employment estimated under the unconditional parallel trends assumption (Panel (a)) and the conditional parallel trends assumption (Panel (b)). Red lines give point estimates and simultaneous 95% confidence bands for pre-treatment periods allowing for clustering at the county level. Under the null hypothesis of the parallel trends assumption holding in all periods, these should be equal to 0. Blue lines provide point estimates and simultaneous 95% confidence bands for the treatment effect of increasing the minimum wage allowing for clustering at the county level. The top row includes states that increased their minimum wage in 2004, the middle row includes states that increased their minimum wage in 2006, and the bottom row includes states that increased their minimum wage in 2007. The estimates in Panel (b) use the doubly robust estimator discussed in the text.

is estimated to be 7.1% lower; in 2006, 12.5% lower; and in 2007, 13.6% lower. For states first treated in 2006, there is a small effect in 2006: 2.3% lower teen employment; however, it is larger in 2007: 7.1% lower teen employment.

Panel (a) of Table 3 reports aggregated treatment effect measures. First, we consider how the effect of increasing the minimum changes by the amount of time that the policy has been in place. These parameters paint largely the same picture as the group-time average treatment effects. The effect of increasing the minimum wage on teen employment appears to be negative and increasing in magnitude the longer states are exposed to the higher minimum wage. In particular, in the first year that a state increases its minimum wage, teen employment is estimated to decrease by 2.7%, in the second year it is estimated to decrease by 7.1%, in the third year by 12.5%, and in the fourth year by 13.6%. Notice that the last two dynamic treatment effect estimates are exactly the same as the estimates coming from Illinois alone because Illinois is the only state that is treated for at least two years. These results are robust to keeping the composition of groups constant by "balancing" the groups across different lengths of exposure to the treatment (see the row in Table 3 labeled 'Event Study w/ Balanced Groups'). When we restrict the sample to only include groups that had a minimum wage increase for at least one full year (i.e., we keep groups 2004 and 2006 but not 2007), we estimate that the effect of increasing the minimum wage on impact is 2.7% lower teen employment and 7.1% lower teen employment one year after the increase.[16]

[16] Notice that these estimates are exactly the same as in the first two periods for the dynamic treatment effect estimates that do not hold the composition of groups constant across different lengths of exposure. The reason that they are the same for initial exposure is coincidental as the

Table 3
Minimum wage aggregated treatment effect estimates.

(a) Unconditional parallel trends					
	Partially aggregated				Single parameters
TWFE					−0.037
					(0.006)
Simple weighted average					−0.052
					(0.006)
Group-specific effects	g=2004	g=2006	g=2007		
	−0.091	−0.047	−0.028		−0.039
	(0.019)	(0.008)	(0.007)		(0.007)
Event study	e=0	e=1	e=2	e=3	
	−0.027	−0.071	−0.125	−0.136	−0.090
	(0.006)	(0.009)	(0.021)	(0.023)	(0.013)
Calendar time effects	t=2004	t=2005	t=2006	t=2007	
	−0.034	−0.071	−0.055	−0.050	−0.052
	(0.019)	(0.02)	(0.009)	(0.006)	(0.013)
Event study	e=0	e=1			
w/ Balanced groups	−0.027	−0.071			−0.049
	(0.009)	(0.009)			(0.008)
(b) Conditional parallel trends					
	Partially aggregated				Single parameters
TWFE					−0.008
					(0.006)
Simple weighted average					−0.033
					(0.007)
Group-specific effects	g=2004	g=2006	g=2007		
	−0.044	−0.029	−0.029		−0.031
	(0.020)	(0.008)	(0.008)		(0.007)
Event study	e=0	e=1	e=2	e=3	
	−0.024	−0.041	−0.050	−0.071	−0.046
	(0.006)	(0.009)	(0.022)	(0.026)	(0.013)
Calendar time effects	t=2004	t=2005	t=2006	t=2007	
	−0.030	−0.025	−0.030	−0.049	−0.033
	(0.022)	(0.021)	(0.009)	(0.007)	(0.012)
Event study	e=0	e=1			
w/ Balanced groups	−0.016	−0.041			−0.028
	(0.010)	(0.009)			(0.008)

Notes: The table reports aggregated treatment effect parameters under the unconditional and conditional parallel trends assumptions and with clustering at the county level. The row 'TWFE' reports the coefficient on a post-treatment dummy variable from a two-way fixed effects regression. The row 'Simple Weighted Average' reports the weighted average (by group size) of all available group-time average treatment effects as in Eq. (3.10). The row 'Group-Specific Effects' summarizes average treatment effects by the timing of the minimum wage increase; here, g indexes the year that a county is first treated. The row 'Event Study' reports average treatment effects by the length of exposure to the minimum wage increase; here, e indexes the length of exposure to the treatment. The row 'Calendar Time Effects' reports average treatment effects by year; here, t indexes the year. The row 'Event Study w/ Balanced Groups' reports average treatment effects by length of exposure using a fixed set of groups at all lengths of exposure; here, e indexes the length of exposure and the sample consists of counties that have at least two years of exposure to minimum wage increases. The column 'Single Parameters' represents a further aggregation of each type of parameter, as discussed in the text. The estimates in Panel (b) use the doubly robust estimator discussed in the text.

Our summary parameters aggregated by group and by calendar time are also consistent with the idea that increasing the minimum wage had a negative effect on county level teen employment relative to what would have happened in the absence of the minimum wage increase.

The second set of results comes from using the conditional parallel trends assumption; that is, we assume only that counties with *the same characteristics* would follow the same trend in teen employment in the absence of treatment. The county characteristics that we use are region of the country, county population, county median income, the fraction of the population that is white, the fraction of the population with a high school education, and the county's poverty rate. We use the doubly robust estimation procedure discussed above. Thus, estimation requires a first step estimation of the generalized propensity score and outcome regression discussed above. For each generalized propensity score, we estimate

results holding group composition constant do not include the group first treated in 2007 (the estimated effect of the minimum wage in 2007 for the group of states first treated in 2007 is 2.76% lower teen employment which just happens to correspond to the estimated effect for the balanced groups). On the other hand, for the second period, they correspond by construction because both estimates only include the groups first treated in 2004 and 2006.

a logit model that includes each county characteristic along with quadratic terms for population and median income.[17] For the outcome regressions, we use the same specification for the covariates.

Before presenting these results, we note that our doubly robust estimation procedure is not computationally demanding. Our estimates of group-time average treatment effects in this section (across all groups and time periods and including our multiplier bootstrap with 1000 iterations) run in 3.0 seconds on a laptop with a 2.80-GHz Intel i5 processor with 8GB of RAM and without using any parallel processing.

For comparison's sake, we first estimate the coefficient on a post-treatment dummy variable in a model with unit fixed effects and region-year fixed effects. This is very similar to one of the sorts of models that Dube et al. (2010) find to eliminate the correlation between the minimum wage and employment. Like Dube et al. (2010), using this specification, we find that the estimated coefficient is small and not statistically different from 0. However, one must have in mind that the approach we proposed in this article is different from the two-way fixed effects regression. In particular, we explicitly identify group-time average treatment effects for different groups and different times, allowing for arbitrary treatment effect heterogeneity as long as the conditional parallel trends assumption is satisfied. Thus, our causal parameters have a clear interpretation. As pointed out by Wooldridge (2005a), Chernozhukov et al. (2013), de Chaisemartin and D'Haultfœuille (2020), Borusyak and Jaravel (2017), Goodman-Bacon (2019) and Słoczyński (2018), the same may not be true for two-way fixed effects regressions in the presence of treatment effect heterogeneity.[18]

The results using our approach are available in Panel (b) in Fig. 1 and Panel (b) in Table 3. Interestingly, we find quite different results using our approach than are suggested by the two-way fixed effects regression approach. In particular, we continue to find evidence that increasing the minimum wage tended to reduce teen employment. The estimated group-time average treatment effects range from 0.9% lower teen employment (not statistically different from 0) in 2006 for the group of states first treated in 2006 to 7.1% lower teen employment in 2007 for states first treated in 2004. Now, 3 of 7 group-time average treatment effects are statistically significant. The average effect of increasing the minimum wage on teen employment across all groups that increased their minimum wage is a 3.1% reduction in teen employment. This estimate is much different from the TWFE estimate. In addition, the pattern of dynamic treatment effects where the magnitude of the effect of increasing the minimum wage tends to increase with length of exposure is the same as in the unconditional case.

Overall, our results suggest that increasing the minimum wage decreased teen employment relative to what it would have been without the policy change. However, there are some important limitations of our application. First, some of the estimates of pseudo group-time average treatment effects in pre-treatment periods in Fig. 1 are significantly different from zero which provides some suggestive evidence against the parallel trends assumption.

Second, as discussed in the Supplementary Appendix, there is some heterogeneity in the size of the minimum wage increase itself across states which could complicate the interpretation of our results. Together, these suggest that our results should be interpreted with some caution. That being said, we think that the key takeaway from the application is that, (implicitly) holding the main identifying assumptions constant, in a prominent application in economics that has many very common features (treatment effect heterogeneity, dynamic effects, and staggered treatment adoption) the choice of estimation method can potentially lead to qualitatively different conclusions.

6. Conclusion

This paper has considered Difference-in-Differences methods in the case where there are more than two time periods and units can become treated at different points in time – a commonly encountered setup in empirical work in economics. In this setup, we have proposed group-time average treatment effects, $ATT(g,t)$, that are the average treatment effect in period t for the group of units first treated in period g. Unlike the more common approach of including a post-treatment dummy variable in a two-way fixed effects regression, $ATT(g,t)$ corresponds to a well defined treatment effect parameter. We also showed that once $ATT(g,t)$ has been obtained for different values of g and t, they can be aggregated into other parameters to more concisely summarize heterogeneity with respect to some particular dimension of interest (such as length of exposure to the treatment) or, alternatively, into a single overall treatment effect parameter. In addition, our approach is suitable (i) for cases where the parallel trends assumption holds only after conditioning on covariates, (ii) using different comparison groups such as the never-treated or not-yet-treated, and (iii) when units can anticipate participating in the treatment and may adjust their behavior before the treatment is implemented. We view such flexibility as an important component of our proposed methodology.

We also provided nonparametric identification results leading to outcome regression, inverse probability weighting, and doubly robust estimands. Given that our nonparametric identification results are constructive, we proposed to estimate $ATT(g,t)$ using its sample analogue. We established consistency and asymptotic normality of the proposed

17 Using the propensity score specification tests proposed by Sant'Anna and Song (2019), we fail to reject the null hypothesis that these models are correctly specified at the usual significance levels.

18 Our approach is also different from that of Dube et al. (2010) in several other ways that are worth mentioning. We focus on teen employment; Dube et al. (2010) consider employment in the restaurant industry. Their most similar specification to the one mentioned above includes census division-time fixed effects rather than region-time fixed effects though the results are similar. Finally, our period of analysis is different from theirs; in particular, there are no federal minimum wage changes over the periods we analyze.

estimators, and proved the validity of a powerful, but easy to implement, multiplier bootstrap procedure to construct simultaneous confidence bands for $ATT(g,t)$. The computational costs of our approach are generally low, and code for implementing our approach is available in the R `did` package.

Finally, we applied our approach to study the effect of minimum wage increases on teen employment. We found some evidence that increasing the minimum wage led to reductions in teen employment. More interestingly though, in some cases we found notable differences between the results coming from our approach relative to the more common two-way fixed effects approach. These differences suggest that using an approach that is robust to treatment effect heterogeneity and dynamics should be strongly considered by applied researchers.

Acknowledgments

We thank the Editor, Elie Tamer, the Associate Editor, two anonymous referees, Stéphane Bonhomme, Carol Caetano, Sebastian Calonico, Xiaohong Chen, Clément de Chaisemartin, Xavier D'Haultfoeuille, Bruno Ferman, John Gardner, Andrew Goodman-Bacon, Federico Gutierrez, Sukjin Han, Hugo Jales, Andrew Johnston, Vishal Kamat, Qi Li, Tong Li, Jason Lindo, Catherine Maclean, Matt Masten, Magne Mogstad, Tom Mroz, Aureo de Paula, Jonathan Roth, Donald Rubin, Bernhard Schmidpeter, Yuya Sasaki, Na'Ama Shenhav, Tymon Słoczyński, Sebastian Tello-Trillo, Alex Torgovitsky, Jeffrey Wooldridge, Haiqing Xu and several seminar and conference audiences for comments and suggestions.

Appendix A. Proofs of main results

We provide the proofs of our results in this appendix. Before proceeding, we first state and prove several auxiliary lemmas that help us to prove our main theorems.

Let

$$ATT_X(g,t) = \mathbb{E}[Y_t(g) - Y_t(0)|X, G_g = 1].$$

Lemma A.1. *Let Assumptions 1, 2, 3, 4, and 6 hold. Then, for all g and t such that $g \in \mathcal{G}_\delta$, $t \in \{2, \ldots \mathcal{T} - \delta\}$ and $t \geq g - \delta$,*

$$ATT_X(g,t) = \mathbb{E}[Y_t - Y_{g-\delta-1}|X, G_g = 1] - \mathbb{E}[Y_t - Y_{g-\delta-1}|X, C = 1] \text{ a.s.}$$

Proof of Lemma A.1. In what follows, take all equalities to hold almost surely (a.s.). Then, we have that

$$\begin{aligned}
ATT_X(g,t) &= \mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \mathbb{E}[Y_t(0) - Y_{g-\delta-1}(0)|X, G_g = 1] \\
&= \mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \sum_{\ell=0}^{t-g-\delta} \mathbb{E}[\Delta Y_{t-\ell}(0)|X, G_g = 1] \\
&= \mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \sum_{\ell=0}^{t-g-\delta} \mathbb{E}[\Delta Y_{t-\ell}(0)|X, C = 1] \\
&= \mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \mathbb{E}[Y_t(0) - Y_{g-\delta-1}(0)|X, C = 1] \\
&= \mathbb{E}[Y_t - Y_{g-\delta-1}|X, G_g = 1] - \mathbb{E}[Y_t - Y_{g-\delta-1}|X, C = 1]
\end{aligned}$$

where the first equality follows from adding and subtracting $\mathbb{E}[Y_{g-\delta-1}(0)|X, G_g = 1]$, the second equality from simple algebra, the third equality by Assumption 4, the fourth equality by simple algebra, and the last equality from (2.1) and Assumption 3. □

Lemma A.2. *Let Assumptions 1–3, 5 and 6 hold. Then, for all g and t such that $g \in \mathcal{G}_\delta$, $t \in \{2, \ldots \mathcal{T} - \delta\}$ with $g - \delta \leq t < \bar{g}$*

$$ATT_X(g,t) = \mathbb{E}[Y_t - Y_{g-\delta-1}|X, G_g = 1] - \mathbb{E}[Y_t - Y_{g-\delta-1}|X, D_{t+\delta} = 0, G_g = 0] \text{ a.s.}$$

Proof of Lemma A.2. The proof follows similar steps as the proof of Lemma A.1. Taking all equalities to hold almost surely (a.s.), we have that

$$\begin{aligned}
ATT_X(g,t) &= \mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \mathbb{E}[Y_t(0) - Y_{g-\delta-1}(0)|X, G_g = 1] \\
&= \mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \sum_{\ell=0}^{t-g-\delta} \mathbb{E}[\Delta Y_{t-\ell}(0)|X, G_g = 1] \\
&= \mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \sum_{\ell=0}^{t-g-\delta} \mathbb{E}[\Delta Y_{t-\ell}(0)|X, D_{t+\delta} = 0, G_g = 0] \\
&= \mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \mathbb{E}[Y_t(0) - Y_{g-\delta-1}(0)|X, D_{t+\delta} = 0, G_g = 0] \\
&= \mathbb{E}[Y_t - Y_{g-\delta-1}|X, G_g = 1] - \mathbb{E}[Y_t - Y_{g-\delta-1}|X, D_{t+\delta} = 0, G_g = 0]
\end{aligned}$$

where the first equality follows from adding and subtracting $\mathbb{E}[Y_{g-\delta-1}(0)|X, G_g = 1]$, the second equality from simple algebra, the third equality by Assumption 5 with $s = t + \delta$, the fourth equality by simple algebra, and the last equality from (2.1) and Assumption 3. □

Now, we are ready to proceed with the proofs of our main theorems.

Proof of Theorem 1.

Part 1: Identification when Assumption 4 is invoked.

In this case, given the result in Lemma A.1,

$$\begin{aligned} ATT(g,t) &= \mathbb{E}[ATT_X(g,t)|G_g = 1] \\ &= \mathbb{E}\left[\left(\mathbb{E}[Y_t - Y_{g-\delta-1}|X, G_g = 1] - \mathbb{E}[Y_t - Y_{g-\delta-1}|X, C = 1]\right)\Big| G_g = 1\right] \\ &= \mathbb{E}[Y_t - Y_{g-\delta-1}|G_g = 1] - \mathbb{E}\left[\underbrace{\mathbb{E}[Y_t - Y_{g-\delta-1}|X, C = 1]}_{=m^{nev}_{g,t,\delta}(X)}|G_g = 1\right] \\ &= \mathbb{E}\left[\frac{G_g}{\mathbb{E}\left[G_g\right]}\left(Y_t - Y_{g-\delta-1} - m^{nev}_{g,t,\delta}(X)\right)\right]. \end{aligned}$$

Hence, we have that $ATT(g,t) = ATT^{nev}_{or}(g,t;\delta)$.

Next, to show that $ATT(g,t) = ATT^{nev}_{ipw}(g,t;\delta)$, it suffices to show that

$$\frac{\mathbb{E}\left[\frac{p_g(X)C}{(1-p_g(X))}(Y_t - Y_{g-\delta-1})\right]}{\mathbb{E}\left[\frac{p_g(X)C}{(1-p_g(X))}\right]} = \frac{\mathbb{E}\left[G_g \cdot \mathbb{E}[Y_t - Y_{g-\delta-1}|X, C = 1]\right]}{\mathbb{E}\left[G_g\right]}. \tag{A.1}$$

Towards this end, by noticing that

$$p_g(X) = \frac{\mathbb{E}\left[G_g|X\right]}{\mathbb{E}\left[G_g + C|X\right]},\quad 1 - p_g(X) = \frac{\mathbb{E}\left[C|X\right]}{\mathbb{E}\left[G_g + C|X\right]}, \tag{A.2}$$

it follows that

$$\begin{aligned} \mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right] &= \mathbb{E}\left[\frac{\mathbb{E}\left[G_g|X\right]C}{\mathbb{E}\left[C|X\right]}\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}\left[G_g|X\right]\mathbb{E}\left[C|X\right]}{\mathbb{E}\left[C|X\right]}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[G_g|X\right]\right] \\ &= \mathbb{E}\left[G_g\right]. \end{aligned} \tag{A.3}$$

Next, by exploiting (A.2) and applying the law of iterated expectations, we have that

$$\begin{aligned} \mathbb{E}\left[\frac{p_g(X)C}{(1-p_g(X))}(Y_t - Y_{g-\delta-1})\right] &= \mathbb{E}\left[\frac{\mathbb{E}\left[G_g|X\right]C}{\mathbb{E}\left[C|X\right]}(Y_t - Y_{g-\delta-1})\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}\left[G_g|X\right]}{\mathbb{E}\left[C|X\right]}\mathbb{E}\left[C\cdot(Y_t - Y_{g-\delta-1})\big|X\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[G_g|X\right]\cdot\mathbb{E}\left[(Y_t - Y_{g-\delta-1})\big|X, C = 1\right]\right] \\ &= \mathbb{E}\left[G_g\cdot\mathbb{E}\left[(Y_t - Y_{g-\delta-1})\big|X, C = 1\right]\right]. \end{aligned}$$

Thus, combined this with (A.3), we establish (A.1), implying that $ATT(g,t) = ATT^{nev}_{ipw}(g,t;\delta)$.

Finally, notice that

$$\begin{aligned}
ATT_{dr}^{nev}(g,t;\delta) &= \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]} - \frac{\dfrac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\dfrac{p_g(X)C}{1-p_g(X)}\right]}\right)\left(Y_t - Y_{g-\delta-1} - m_{g,t,\delta}^{nev}(X)\right)\right] \\
&= \underbrace{\mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]} - \frac{\dfrac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\dfrac{p_g(X)C}{1-p_g(X)}\right]}\right)\left(Y_t - Y_{g-\delta-1}\right)\right]}_{\equiv ATT_{ipw}^{nev}(g,t;\delta)} \\
&\quad - \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]} - \frac{\dfrac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\dfrac{p_g(X)C}{1-p_g(X)}\right]}\right) m_{g,t,\delta}^{nev}(X)\right] \\
&= ATT(g,t) - \frac{1}{\mathbb{E}\left[G_g\right]}\mathbb{E}\left[\left(G_g - \frac{\mathbb{E}\left[G_g|X\right]C}{\mathbb{E}\left[C|X\right]}\right) m_{g,t,\delta}^{nev}(X)\right] \\
&= ATT(g,t) - \frac{1}{\mathbb{E}\left[G_g\right]}\mathbb{E}\left[\left(\mathbb{E}\left[G_g|X\right] - \mathbb{E}\left[G_g|X\right]\right)\cdot m_{g,t,\delta}^{nev}(X)\right] \\
&= ATT(g,t).
\end{aligned}$$

where the third equality follows from $ATT_{ipw}^{nev}(g,t;\delta) = ATT(g,t)$, (A.2) and (A.3), and the fourth equality from the law of iterated expectations.

Part 2: Identification when Assumption 5 is invoked.
In this case, given the result in Lemma A.2,

$$\begin{aligned}
ATT(g,t) &= \mathbb{E}[ATT_X(g,t)|G_g = 1] \\
&= \mathbb{E}\left[\left(\mathbb{E}[Y_t - Y_{g-\delta-1}|X, G_g = 1] - \mathbb{E}[Y_t - Y_{g-\delta-1}|X, D_{t+\delta} = 0, G_g = 0]\right)\Big| G_g = 1\right] \\
&= \mathbb{E}[Y_t - Y_{g-\delta-1}|G_g = 1] - \mathbb{E}\left[\underbrace{\mathbb{E}[Y_t - Y_{g-\delta-1}|X, D_{t+\delta} = 0, G_g = 0]}_{=m_{g,t,\delta}^{ny}(X)}|G_g = 1\right] \\
&= \mathbb{E}\left[\frac{G_g}{\mathbb{E}\left[G_g\right]}\left(Y_t - Y_{g-1-a} - m_{g,t,\delta}^{ny}(X)\right)\right].
\end{aligned}$$

Hence, we have that $ATT(g,t) = ATT_{or}^{ny}(g,t;\delta)$.
Next, to show that $ATT(g,t) = ATT_{ipw}^{ny}(g,t;\delta)$, it suffices to show that

$$\frac{\mathbb{E}\left[\dfrac{p_{g,t+\delta}(X)(1-D_{t+\delta})\left(1-G_g\right)}{1-p_{g,t+\delta}(X)}(Y_t - Y_{g-\delta-1})\right]}{\mathbb{E}\left[\dfrac{p_{g,t+\delta}(X)(1-D_{t+\delta})\left(1-G_g\right)}{1-p_{g,t+\delta}(X)}\right]} = \frac{\mathbb{E}\left[G_g\cdot\mathbb{E}[Y_t - Y_{g-\delta-1}|X, D_{t+\delta} = 0, G_g = 0]\right]}{\mathbb{E}\left[G_g\right]}. \tag{A.4}$$

Towards this end, recall that $p_{g,t+\delta}(X) = P(G_g = 1|X, G_g + (1-D_{t+\delta})\left(1-G_g\right) = 1)$ and also notice that

$$p_{g,t+\delta}(X) = \frac{\mathbb{E}\left[G_g|X\right]}{\mathbb{E}\left[G_g + (1-D_{t+\delta})\left(1-G_g\right)|X\right]},\quad 1-p_g(X) = \frac{\mathbb{E}\left[(1-D_{t+\delta})\left(1-G_g\right)|X\right]}{\mathbb{E}\left[G_g + (1-D_{t+\delta})\left(1-G_g\right)|X\right]}, \tag{A.5}$$

it follows that by the law of iterated expectations,

$$\begin{aligned}
\mathbb{E}\left[\frac{p_{g,t+\delta}(X)(1-D_{t+\delta})\left(1-G_g\right)}{1-p_{g,t+\delta}(X)}\right] &= \mathbb{E}\left[\frac{\mathbb{E}\left[G_g|X\right](1-D_{t+\delta})\left(1-G_g\right)}{\mathbb{E}\left[(1-D_{t+\delta})\left(1-G_g\right)|X\right]}\right] \\
&= \mathbb{E}\left[G_g\right].
\end{aligned} \tag{A.6}$$

Next, by exploiting (A.5) and applying the law of iterated expectations, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\frac{p_{g,t+\delta}(X)\left(1-D_{t+\delta}\right)\left(1-G_g\right)}{1-p_{g,t+\delta}(X)}\left(Y_t-Y_{g-\delta-1}\right)\right] \\
&=\mathbb{E}\left[\frac{\mathbb{E}\left[G_g|X\right]\left(1-D_{t+\delta}\right)\left(1-G_g\right)}{\mathbb{E}\left[\left(1-D_{t+\delta}\right)\left(1-G_g\right)|X\right]}\left(Y_t-Y_{g-\delta-1}\right)\right] \\
&=\mathbb{E}\left[\frac{\mathbb{E}\left[G_g|X\right]}{\mathbb{E}\left[\left(1-D_{t+\delta}\right)\left(1-G_g\right)|X\right]}\mathbb{E}\left[\left(1-D_{t+\delta}\right)\left(1-G_g\right)\cdot\left(Y_t-Y_{g-\delta-1}\right)\middle|X\right]\right] \\
&=\mathbb{E}\left[\mathbb{E}\left[G_g|X\right]\cdot\mathbb{E}\left[\left(Y_t-Y_{g-\delta-1}\right)\middle|X,D_{t+\delta}=0,G_g=0\right]\right] \\
&=\mathbb{E}\left[G_g\cdot\mathbb{E}\left[\left(Y_t-Y_{g-\delta-1}\right)\middle|X,D_{t+\delta}=0,G_g=0\right]\right].
\end{aligned}
$$

Thus, combined this with (A.6), we establish (A.4), implying that $ATT(g,t)=ATT^{ny}_{ipw}(g,t;\delta)$.

Finally, notice that

$$
\begin{aligned}
ATT^{ny}_{dr}(g,t;\delta)&=\mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]}-\frac{\dfrac{p_{g,t+\delta}(X)\left(1-D_{t+\delta}\right)\left(1-G_g\right)}{1-p_{g,t+\delta}(X)}}{\mathbb{E}\left[\dfrac{p_{g,t+\delta}(X)\left(1-D_{t+\delta}\right)\left(1-G_g\right)}{1-p_{g,t+\delta}(X)}\right]}\right)\left(Y_t-Y_{g-\delta-1}-m^{ny}_{g,t,\delta}(X)\right)\right] \\
&=ATT^{ny}_{ipw}(g,t;\delta)-\frac{1}{\mathbb{E}\left[G_g\right]}\mathbb{E}\left[\left(G_g-\frac{\mathbb{E}\left[G_g|X\right]\left(1-D_{t+\delta}\right)\left(1-G_g\right)}{\mathbb{E}\left[\left(1-D_{t+\delta}\right)\left(1-G_g\right)|X\right]}\right)m^{ny}_{g,t,\delta}(X)\right] \\
&=ATT(g,t)-\frac{1}{\mathbb{E}\left[G_g\right]}\mathbb{E}\left[\left(\mathbb{E}\left[G_g|X\right]-\mathbb{E}\left[G_g|X\right]\right)\cdot m^{nev}_{g,t,\delta}(X)\right] \\
&=ATT(g,t).
\end{aligned}
$$

where the second equality follows from (A.5) and (A.6), and the third equality from $ATT^{nev}_{ipw}(g,t;\delta)=ATT(g,t)$ and the law of iterated expectations. □

Proof of Theorem 2. From Theorem 1 it follows that $ATT(g,t)$'s are point-identified for all groups g and time periods t such that $g\in\mathcal{G}_\delta$, $t\in\{2,\ldots\mathcal{T}-\delta\}$ and $t\geq g-\delta$. For each (g,t) pair, the asymptotic linear representation of $\sqrt{n}(\widehat{ATT}^{nev}_{dr}(g,t;\delta)-ATT(g,t))$ follows from Theorem A.1(a) of Sant'Anna and Zhao (2020), whereas

$$\sqrt{n}(\widehat{ATT}^{dr,nev}_{t\geq(g-\delta)}-ATT_{t\geq(g-\delta)})\xrightarrow{d}N(0,\mathbb{E}[\Psi^{dr,nev}_{t\geq(g-\delta)}(W)\Psi^{dr,nev}_{t\geq(g-\delta)}(W)'])$$

follows from the Lindeberg–Lévy central limit theorem. □

Proof of Theorem 3. Note that, by the conditional multiplier central limit theorem, see Lemma 2.9.5 in van der Vaart and Wellner (1996), as $n\to\infty$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i\cdot\Psi^{dr,nev}_{t\geq(g-\delta)}(W_i)\underset{*}{\xrightarrow{d}}N(0,\Sigma),\tag{A.7}$$

where $\Sigma=\mathbb{E}[\Psi^{dr,nev}_{t\geq(g-\delta)}(W)\Psi^{dr,nev}_{t\geq(g-\delta)}(W)']$. Thus, to conclude the proof that

$$\sqrt{n}\left(\widehat{ATT}^{*,dr,nev}_{t\geq(g-\delta)}-\widehat{ATT}^{dr,nev}_{t\geq(g-\delta)}\right)\underset{*}{\xrightarrow{d}}N(0,\Sigma),$$

it suffices to show that, for all g and t such that $g\in\mathcal{G}_\delta$, $t\in\{2,\ldots\mathcal{T}-\delta\}$ and $t\geq g-\delta$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i\cdot\left[\widehat{\psi}^{dr,nev}_{g,t,\delta}(W_i;\widehat{\kappa}^{nev}_{g,t})-\psi^{dr,nev}_{g,t,\delta}(W_i;\kappa^{*,nev}_{g,t})\right]=o_{p^*}(1),\tag{A.8}$$

where $\kappa^{*,nev}_{g,t}=\left(\pi^{*\prime}_g,\beta^{*,nev\,\prime}_{g,t,\delta}\right)'$ is the vector of pseudo-true finite-dimensional parameters.

Towards this end, recall that

$$\widehat{\psi}^{dr,nev}_{g,t,\delta}(W_i;\widehat{\kappa}^{nev}_{g,t})=\widehat{\psi}^{treat,nev}_{g,t,\delta}(W_i;\widehat{\beta}^{nev}_{g,t,\delta})-\widehat{\psi}^{comp,nev}_{g,t,\delta}(W_i;\widehat{\pi}_g,\widehat{\beta}^{nev}_{g,t,\delta})-\widehat{\psi}^{est,nev}_{g,t,\delta}(W_i;\widehat{\pi}_g,\widehat{\beta}^{nev}_{g,t,\delta})$$

where

$$\begin{aligned}
\widehat{\psi}^{treat,nev}_{g,t,\delta}(W;\widehat{\beta}^{nev}_{g,t,\delta}) &= \widehat{w}^{treat}_{g}\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(\widehat{\beta}^{nev}_{g,t,\delta}\right)\right)\\
&\quad-\widehat{w}^{treat}_{g}\cdot\mathbb{E}_n\left[\widehat{w}^{treat}_{g}\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(\widehat{\beta}^{nev}_{g,t,\delta}\right)\right)\right],\\
\widehat{\psi}^{comp,nev}_{g,t,\delta}(W;\widehat{\pi}_g,\widehat{\beta}^{nev}_{g,t,\delta}) &= \widehat{w}^{comp,nev}_{g}\left(\widehat{\pi}_g\right)\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(\widehat{\beta}^{nev}_{g,t,\delta}\right)\right)\\
&\quad-\widehat{w}^{comp}_{g}\left(\widehat{\pi}_g\right)\cdot\mathbb{E}_n\left[w^{comp}_{g}\left(\widehat{\pi}_g\right)\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(\widehat{\beta}^{nev}_{g,t,\delta}\right)\right)\right],\\
\widehat{\psi}^{est,nev}_{g,t}(W;\widehat{\pi}_g,\widehat{\beta}^{nev}_{g,t,\delta}) &= l^{or,nev}_{g,t}\left(\widehat{\beta}^{nev}_{g,t,\delta}\right)'\cdot\widehat{M}^{dr,nev,1}_{g,t,\delta}+l^{ps,nev}_{g}\left(\widehat{\pi}_g\right)'\cdot\widehat{M}^{dr,nev,2}_{g,t,\delta},
\end{aligned}$$

with

$$\widehat{w}^{treat}_{g}=\frac{G_g}{\mathbb{E}_n\left[G_g\right]},\qquad \widehat{w}^{comp,nev}_{g}\left(\widehat{\pi}_g\right)=\frac{\dfrac{\widehat{p}_g\left(X;\widehat{\pi}_g\right)C}{1-\widehat{p}_g\left(X;\widehat{\pi}_g\right)}}{\mathbb{E}_n\left[\dfrac{\widehat{p}_g\left(X;\widehat{\pi}_g\right)C}{1-\widehat{p}_g\left(X;\widehat{\pi}_g\right)}\right]},$$

and

$$\begin{aligned}
\widehat{M}^{dr,nev,1}_{g,t,\delta} &= \mathbb{E}_n\left[\left(\widehat{w}^{treat}_{g}-\widehat{w}^{comp,nev}_{g}\left(\widehat{\pi}_g\right)\right)\cdot\dot{m}^{nev}_{g,t,\delta}\left(\widehat{\beta}^{nev}_{g,t,\delta}\right)\right],\\
\widehat{M}^{dr,nev,2}_{g,t,\delta} &= \mathbb{E}_n\left[\widehat{\alpha}^{ps,nev}_{g}\left(\widehat{\pi}_g\right)\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(\widehat{\beta}^{nev}_{g,t,\delta}\right)\right)\cdot\dot{p}_g\left(\widehat{\pi}_g\right)\right]\\
&\quad-\mathbb{E}_n\left[\widehat{\alpha}^{ps,nev}_{g}\left(\widehat{\pi}_g\right)\cdot\widehat{w}^{comp,nev}_{g}\left(\widehat{\pi}_g\right)\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(\widehat{\beta}^{nev}_{g,t,\delta}\right)\right)\cdot\dot{p}_g\left(\widehat{\pi}_g\right)\right],\\
\widehat{\alpha}^{ps,nev}_{g}\left(\widehat{\pi}_g\right) &= \frac{C}{\left(1-p_g\left(X;\widehat{\pi}_g\right)\right)^2}\Bigg/\mathbb{E}_n\left[\frac{p_g\left(X;\widehat{\pi}_g\right)C}{1-p_g\left(X;\widehat{\pi}_g\right)}\right].
\end{aligned}$$

We first show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i\cdot\left(\widehat{\psi}^{treat,nev}_{g,t,\delta}(W_i;\widehat{\beta}^{nev}_{g,t,\delta})-\psi^{treat,nev}_{g,t,\delta}(W_i;\beta^{*,nev}_{g,t,\delta})\right)=o_{p^*}(1)\,. \tag{A.9}$$

Using the mean-value theorem, we write

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i\cdot\widehat{\psi}^{treat,nev}_{g,t,\delta}(W_i;\widehat{\beta}^{nev}_{g,t,\delta})\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i\cdot\widehat{w}^{treat}_{g,i}\cdot\left(Y_{i,t}-Y_{i,g-\delta-1}-m^{nev}_{g,t,\delta}\left(W_i;\beta^{*,nev}_{g,t,\delta}\right)\right)\\
&-\sqrt{n}\left(\widehat{\beta}^{nev}_{g,t,\delta}-\beta^{*,nev}_{g,t,\delta}\right)'\frac{1}{n}\sum_{i=1}^{n}V_i\cdot\widehat{w}^{treat}_{g,i}\cdot\dot{m}^{nev}_{g,t,\delta}\left(W_i;\bar{\beta}^{nev}_{g,t,\delta}\right)\\
&-\mathbb{E}_n\left[\widehat{w}^{treat}_{g}\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(\widehat{\beta}^{nev}_{g,t,\delta}\right)\right)\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i\cdot\widehat{w}^{treat}_{g,i}\\
&=\widehat{I}^{1,treat}_{g,t,\delta}-\widehat{I}^{2,treat}_{g,t,\delta}-\widehat{I}^{3,treat}_{g,t,\delta},
\end{aligned}$$

where $\bar{\beta}^{nev}_{g,t,\delta}$ is an intermediate point that satisfies $\left|\bar{\beta}^{nev}_{g,t,\delta}-\beta^{*,nev}_{g,t,\delta}\right|\leq\left|\widehat{\beta}^{nev}_{g,t,\delta}-\beta^{*,nev}_{g,t,\delta}\right|$ a.s. From the strong law of large numbers and the fact that V is mean zero, independent of W, it follows that

$$\begin{aligned}
\widehat{I}^{1,treat}_{g,t,\delta} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i\cdot w^{treat}_{g,i}\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(W_i;\beta^{*,nev}_{g,t,\delta}\right)\right)+o_{p^*}(1)\\
\widehat{I}^{3,treat}_{g,t,\delta} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i\cdot w^{treat}_{g,i}\cdot\mathbb{E}\left[w^{treat}_{g}\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(\beta^{*,nev}_{g,t,\delta}\right)\right)\right]+o_{p^*}(1)
\end{aligned}$$

Similarly, from Assumptions 7 and 8, and the strong law of large numbers, we conclude that $\widehat{I}^{2,treat}_{g,t,\delta}=o_{p^*}(1)$. Now (A.9) follows from combining these results.

Next, we show

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i\cdot\left(\widehat{\psi}^{comp,nev}_{g,t,\delta}(W_i;\widehat{\pi}_g,\widehat{\beta}^{nev}_{g,t,\delta})-\psi^{comp,nev}_{g,t,\delta}(W_i;\pi^*_g,\beta^{*,nev}_{g,t,\delta})\right)=o_{p^*}(1)\,. \tag{A.10}$$

Again, by the mean value theorem, we write

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i\cdot\widehat{\psi}^{comp,nev}_{g,t,\delta}(W_i;\widehat{\pi}_g,\widehat{\beta}^{nev}_{g,t,\delta})\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i\cdot\widehat{w}^{comp,nev}_{g}\left(W_i;\widehat{\pi}_g\right)\cdot\left(Y_{i,t}-Y_{i,g-\delta-1}-m^{nev}_{g,t,\delta}\left(W_i;\beta^{*,nev}_{g,t,\delta}\right)\right)\\
&-\sqrt{n}\left(\widehat{\beta}^{nev}_{g,t,\delta}-\beta^{*,nev}_{g,t,\delta}\right)'\frac{1}{n}\sum_{i=1}^{n} V_i\cdot\widehat{w}^{comp,nev}_{g}\left(W_i;\widehat{\pi}_g\right)\cdot\dot{m}^{nev}_{g,t,\delta}\left(W_i;\bar{\beta}^{nev}_{g,t,\delta}\right)\\
&-\mathbb{E}_n\left[\widehat{w}^{comp,nev}_{g}\left(\widehat{\pi}_g\right)\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(\widehat{\beta}^{nev}_{g,t,\delta}\right)\right)\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i\cdot\widehat{w}^{comp,nev}_{g}\left(W_i;\widehat{\pi}_g\right)\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i\cdot\frac{\dfrac{p_g\left(X_i;\pi^*_g\right)C_i}{1-p_g\left(X;\pi^*_g\right)}}{\mathbb{E}_n\left[\dfrac{\widehat{p}_g\left(X;\widehat{\pi}_g\right)C}{1-\widehat{p}_g\left(X;\widehat{\pi}_g\right)}\right]}\cdot\left(Y_{i,t}-Y_{i,g-\delta-1}-m^{nev}_{g,t,\delta}\left(W_i;\beta^{*,nev}_{g,t,\delta}\right)\right)\\
&+\sqrt{n}\left(\widehat{\pi}_g-\pi^*_g\right)'\frac{1}{n}\sum_{i=1}^{n} V_i\cdot\frac{\dfrac{C_i}{\left(1-p_g\left(X_i;\bar{\pi}_g\right)\right)^2}}{\mathbb{E}_n\left[\dfrac{\widehat{p}_g\left(X;\widehat{\pi}_g\right)C}{1-\widehat{p}_g\left(X;\widehat{\pi}_g\right)}\right]}\cdot\left(Y_{i,t}-Y_{i,g-\delta-1}-m^{nev}_{g,t,\delta}\left(W_i;\beta^{*,nev}_{g,t,\delta}\right)\right)\cdot\dot{p}_g\left(X_i;\bar{\pi}_g\right)\\
&-\sqrt{n}\left(\widehat{\beta}^{nev}_{g,t,\delta}-\beta^{*,nev}_{g,t,\delta}\right)'\frac{1}{n}\sum_{i=1}^{n} V_i\cdot\widehat{w}^{comp,nev}_{g}\left(W_i;\widehat{\pi}_g\right)\cdot\dot{m}^{nev}_{g,t,\delta}\left(W_i;\bar{\beta}^{nev}_{g,t,\delta}\right)\\
&-\widehat{M}^{comp}_{g,t,\delta}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i\cdot\frac{\dfrac{p_g\left(X_i;\pi^*_g\right)C_i}{1-p_g\left(X_i;\pi^*_g\right)}}{\mathbb{E}_n\left[\dfrac{\widehat{p}_g\left(X;\widehat{\pi}_g\right)C}{1-\widehat{p}_g\left(X;\widehat{\pi}_g\right)}\right]}\\
&-\widehat{M}^{comp}_{g,t,\delta}\sqrt{n}\left(\widehat{\pi}_g-\pi^*_g\right)'\frac{1}{n}\sum_{i=1}^{n} V_i\cdot\frac{\dfrac{C_i}{\left(1-p_g\left(X_i;\bar{\pi}_g\right)\right)^2}}{\mathbb{E}_n\left[\dfrac{\widehat{p}_g\left(X;\widehat{\pi}_g\right)C}{1-\widehat{p}_g\left(X;\widehat{\pi}_g\right)}\right]}\cdot\dot{p}_g\left(X_i;\bar{\pi}_g\right)\\
&=\widehat{I}^{1,comp}_{g,t,\delta}+\widehat{I}^{2,comp}_{g,t,\delta}-\widehat{I}^{3,comp}_{g,t,\delta}-\widehat{I}^{4,comp}_{g,t,\delta}-\widehat{I}^{5,comp}_{g,t,\delta},
\end{aligned}
$$

where $\bar{\beta}^{nev}_{g,t,\delta}$ and $\bar{\pi}_g$ are intermediate points that satisfy $\left|\bar{\beta}^{nev}_{g,t,\delta}-\beta^{*,nev}_{g,t,\delta}\right|\leq\left|\widehat{\beta}^{nev}_{g,t,\delta}-\beta^{*,nev}_{g,t,\delta}\right|$ a.s. and $\left|\bar{\pi}_g-\pi^*_g\right|\leq\left|\widehat{\pi}_g-\pi^*_g\right|$ a.s., respectively, and

$$
\widehat{M}^{comp}_{g,t,\delta}=\mathbb{E}_n\left[\widehat{w}^{comp,nev}_{g}\left(\widehat{\pi}_g\right)\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(\widehat{\beta}^{nev}_{g,t,\delta}\right)\right)\right].
$$

From the strong law of large numbers and the fact that V is mean zero, has variance one, and is independent of W, it follows that

$$
\begin{aligned}
\widehat{I}^{1,comp}_{g,t,\delta}&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i\cdot w^{comp,nev}_{g}\left(W_i;\pi^*_g\right)\cdot\left(Y_{i,t}-Y_{i,g-\delta-1}-m^{nev}_{g,t,\delta}\left(W_i;\beta^{*,nev}_{g,t,\delta}\right)\right)+o_{p^*}(1),\\
\widehat{I}^{3,treat}_{g,t,\delta}&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i\cdot w^{comp,nev}_{g}\left(W_i;\pi^*_g\right)\cdot\mathbb{E}\left[w^{comp}_{g}\cdot\left(Y_t-Y_{g-\delta-1}-m^{nev}_{g,t,\delta}\left(\beta^{*,nev}_{g,t,\delta}\right)\right)\right]+o_{p^*}(1).
\end{aligned}
$$

Similarly, from Assumptions 7 and 8, and the strong law of large numbers, we conclude that

$$
\widehat{I}^{2,comp}_{g,t,\delta}=\widehat{I}^{4,comp}_{g,t,\delta}=\widehat{I}^{5,comp}_{g,t,\delta}=o_{p^*}(1).
$$

Now (A.10) follows from combining these results.

Next, we show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i \cdot \left(\widehat{\psi}^{est,nev}_{g,t}(W_i;\widehat{\pi}_g,\widehat{\beta}^{nev}_{g,t,\delta}) - \psi^{est,nev}_{g,t}(W_i;\pi^*_g,\beta^{*,nev}_{g,t,\delta})\right) = o_{p^*}(1)\,. \tag{A.11}$$

From the strong law of large numbers and Assumptions 7 and 8,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i \cdot \widehat{\psi}^{est,nev}_{g,t}(W_i;\widehat{\pi}_g,\widehat{\beta}^{nev}_{g,t,\delta})\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i \cdot \left(l^{or,nev}_{g,t}\left(W_i;\widehat{\beta}^{nev}_{g,t,\delta}\right)' \cdot M^{dr,nev,1}_{g,t,\delta} + l^{ps,nev}_{g}\left(W_i;\widehat{\pi}_g\right)' \cdot M^{dr,nev,2}_{g,t,\delta}\right) + o_{p^*}(1)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i \cdot l^{est}_{g,t}\left(W_i;\widehat{\kappa}^{nev}_{g,t}\right) + o_{p^*}(1)\,,
\end{aligned}$$

where, for a generic $\kappa^{nev}_{g,t} = \left(\pi'_g, \beta^{nev'}_{g,t,\delta}\right)'$,

$$l^{est}_{g,t}\left(W;\kappa^{nev}_{g,t}\right) = l^{or,nev}_{g,t}\left(W;\beta^{nev}_{g,t,\delta}\right)' \cdot M^{dr,nev,1}_{g,t,\delta} + l^{ps,nev}_{g}\left(W_i;\pi_g\right)' \cdot M^{dr,nev,2}_{g,t,\delta}$$

Thus, from Lemma 4.3 in Newey and McFadden (1994) and Assumption 7, it follows that

$$\begin{aligned}
&Var^*\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i \cdot \left(l^{est}_{g,t}\left(W_i;\widehat{\kappa}^{nev}_{g,t,\delta}\right) - l^{est}_{g,t}\left(W_i;\kappa^{*,nev}_{g,t,\delta}\right)\right)\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(l^{est}_{g,t}\left(W_i;\widehat{\kappa}^{nev}_{g,t,\delta}\right) - l^{est}_{g,t}\left(W_i;\kappa^{*,nev}_{g,t,\delta}\right)\right)^2\\
&= o_p(1)\,,
\end{aligned}$$

which, in turn, implies (A.11).

Taking (A.9), (A.10), and (A.11) together, we then establish (A.8). Thus, by (A.7), we have

$$\sqrt{n}\left(\widehat{ATT}^{*,dr,nev}_{t\geq(g-\delta)} - \widehat{ATT}^{dr,nev}_{t\geq(g-\delta)}\right) \underset{*}{\overset{d}{\to}} N(0,\Sigma).$$

Finally, by the continuous mapping theorem, see e.g. Theorem 10.8 in Kosorok (2008), for any continuous functional $\Gamma(\cdot)$

$$\Gamma\left(\sqrt{n}\left(\widehat{ATT}^{*,dr,nev}_{t\geq(g-\delta)} - \widehat{ATT}^{dr,nev}_{t\geq(g-\delta)}\right)\right) \underset{*}{\overset{d}{\to}} \Gamma\left(N(0,\Sigma)\right),$$

concluding our proof. □

Appendix B. Additional results for repeated cross sections

In this section we extend our identification results to the case where one has access to repeated cross sections data instead of panel data. Here we assume that for each unit in the pooled sample, we observe $(Y, G_2, \ldots, G_{\mathcal{T}}, C, T, X)$ where $T \in \{1, \ldots, \mathcal{T}\}$ denotes the time period when that unit is observed. Let $T_t = 1$ if an observation is observed at time t, and zero otherwise.

We assume that random samples are available for each time period.

Assumption B.1. Conditional of $T = t$, the data are independent and identically distributed from the distribution of $(Y_t, G_2, \ldots, G_{\mathcal{T}}, C, X)$, for all $t = 1, \ldots, \mathcal{T}$., with $(G_2, \ldots, G_{\mathcal{T}}, C, X)$ being invariant to T.

Assumption B.1 implies that our sample consists of random draws from the mixture distribution

$$F_M(y, g_2, \ldots, g_{\mathcal{T}}, c, t, x) = \sum_{t=1}^{\mathcal{T}} \lambda_t \cdot F_{Y,G_2,\ldots,G_{\mathcal{T}},C,X|T}(y, g_2, \ldots, g_{\mathcal{T}}, c, x|t),$$

where $\lambda_t = P(T_t = 1)$. It also rules-out compositional changes across time. This assumption is related to the sampling assumption imposed by Abadie (2005) and Sant'Anna and Zhao (2020) in the two periods, two groups DiD setup. Notice that, once one conditions on the time period, then expectations under the mixture distribution correspond to population expectations. Also, because X, G_g, and C are observed for all units, by the stationarity condition one can use draws from the mixture distribution to estimate the generalized propensity score. With some abuse of notation, we then use $p_{g,s}(X)$

as a short notation for $\mathbb{E}_M\left[G_g|X, G_g+(1-D_s)\left(1-G_g\right)=1\right]$, where $\mathbb{E}_M[\cdot]$ denotes expectations with respect to $F_M(\cdot)$. Also, we use $p_g(X)=p_{g,\mathcal{T}}(X)=\mathbb{E}_M\left[G_g|X, G_g+C=1\right]$.

Before formalizing all the results, we need to introduce some additional notation. Let $m_{c,t}^{rc,nev}(x)\equiv\mathbb{E}_M[Y|X=x, C=1, T=t]$, $m_{g,t}^{rc,treat}(x)\equiv\mathbb{E}_M[Y|X=x, G_g=1, T=t]$ and $m_{s,t}^{rc,ny}(x)\equiv\mathbb{E}_M[Y|X=x, D_s=0, G_g=0, T=t]$. Consider the weights

$$
\begin{aligned}
w^{treat}(a,b) &= T_b\cdot G_a/\,\mathbb{E}_M\left[T_b\cdot G_a\right],\\
w_{nev}^{comp}(a,b) &= \frac{T_b\cdot p_a(X)\,C}{1-p_a(X)}\Bigg/\,\mathbb{E}_M\left[\frac{T_b\cdot p_a(X)\,C}{1-p_a(X)}\right],\\
w_{ny}^{comp}(a,b,s) &= \frac{T_b\cdot p_{a,s}(X)\,(1-D_b)\,(1-G_a)}{1-p_{a,s}(X)}\Bigg/\,\mathbb{E}_M\left[\frac{T_b\cdot p_{a,s}(X)\,(1-D_b)\,(1-G_a)}{1-p_{a,s}(X)}\right].
\end{aligned}
$$

Finally, consider the outcome regression (OR) estimands,

$$
\begin{aligned}
ATT_{or,rc}^{nev}(g,t;\delta) &= \mathbb{E}_M\left[\frac{G_g}{\mathbb{E}_M\left[G_g\right]}\left(\left(m_{g,t}^{rc,treat}(X)-m_{g,g-\delta-1}^{rc,treat}(X)\right)-\left(m_{c,t}^{rc,nev}(X)-m_{c,g-\delta-1}^{rc,nev}(X)\right)\right)\right],\\
ATT_{or,rc}^{ny}(g,t;\delta) &= \mathbb{E}_M\left[\frac{G_g}{\mathbb{E}_M\left[G_g\right]}\left(\left(m_{g,t}^{rc,treat}(X)-m_{g,g-\delta-1}^{rc,treat}(X)\right)-\left(m_{t+\delta,t}^{rc,ny}(X)-m_{t+\delta,g-\delta-1}^{rc,ny}(X)\right)\right)\right],
\end{aligned}
$$

the inverse probability weighted (IPW) estimands

$$
\begin{aligned}
ATT_{ipw,rc}^{nev}(g,t;\delta) &= \mathbb{E}_M\left[\left(w^{treat}(g,t)-w^{treat}(g,g-\delta-1)\right)\cdot Y\right]\\
&\quad-\mathbb{E}_M\left[\left(w_{nev}^{comp}(g,t)-w_{nev}^{comp}(g,g-\delta-1)\right)\cdot Y\right],\\
ATT_{ipw,rc}^{ny}(g,t;\delta) &= \mathbb{E}_M\left[\left(w^{treat}(g,t)-w^{treat}(g,g-\delta-1)\right)\cdot Y\right]\\
&\quad-\mathbb{E}_M\left[\left(w_{ny}^{comp}(g,t,t+\delta)-w_{ny}^{comp}(g,g-\delta-1,t+\delta)\right)\cdot Y\right],
\end{aligned}
$$

and the doubly-robust (DR) estimands

$$
\begin{aligned}
ATT_{dr,rc}^{nev}(g,t;\delta) &= \mathbb{E}_M\left[\frac{G_g}{\mathbb{E}_M\left[G_g\right]}\left(\left(m_{g,t}^{rc,treat}(X)-m_{g,g-\delta-1}^{rc,treat}(X)\right)-\left(m_{c,t}^{rc,nev}(X)-m_{c,g-\delta-1}^{rc,nev}(X)\right)\right)\right]\\
&\quad+\mathbb{E}_M\left[w^{treat}(g,t)\left(Y-m_{g,t}^{rc,treat}(X)\right)-w^{treat}(g,g-\delta-1)\left(Y-m_{g,g-\delta-1}^{rc,treat}(X)\right)\right]\\
&\quad-\mathbb{E}_M\left[w_{nev}^{comp}(g,t)\left(Y-m_{c,t}^{rc,nev}(X)\right)-w_{nev}^{comp}(g,g-\delta-1)\left(Y-m_{c,g-\delta-1}^{rc,nev}(X)\right)\right],\\
ATT_{dr,rc}^{ny}(g,t;\delta) &= \mathbb{E}_M\left[\frac{G_g}{\mathbb{E}_M\left[G_g\right]}\left(\left(m_{g,t}^{rc,treat}(X)-m_{g,g-\delta-1}^{rc,treat}(X)\right)-\left(m_{t+\delta,t}^{rc,ny}(X)-m_{t+\delta,g-\delta-1}^{rc,ny}(X)\right)\right)\right]\\
&\quad+\mathbb{E}_M\left[w^{treat}(g,t)\left(Y-m_{g,t}^{rc,treat}(X)\right)-w^{treat}(g,g-\delta-1)\left(Y-m_{g,g-\delta-1}^{rc,treat}(X)\right)\right]\\
&\quad-\mathbb{E}_M\left[w_{ny}^{comp}(g,t,t+\delta)\left(Y-m_{t+\delta,t}^{rc,ny}(X)\right)-w_{ny}^{comp}(g,g-\delta-1,t+\delta)\left(Y-m_{t+\delta,g-\delta-1}^{rc,ny}(X)\right)\right].
\end{aligned}
$$

The OR, IPW and DR estimands respectively generalize Heckman et al. (1997), Abadie (2005) and Sant'Anna and Zhao (2020) estimands for the two groups, two periods DiD setup to the staggered DiD setup with multiple periods and multiple groups.

Theorem B.1. *Let Assumptions 1, 3, 6, and B.1 hold.*

(i) *If Assumption 4 in the main text holds, then, for all g and t such that $g\in\mathcal{G}_\delta$, $t\in\{2,\ldots\mathcal{T}-\delta\}$ and $t\geq g-\delta$,*

$$ATT(g,t)=ATT_{ipw,rc}^{nev}(g,t;\delta)=ATT_{or,rc}^{nev}(g,t;\delta)=ATT_{dr,rc}^{nev}(g,t;\delta).$$

(ii) *If Assumption 5 in the main text holds, then, for all g and t such that $g\in\mathcal{G}_\delta$, $t\in\{2,\ldots\mathcal{T}-\delta\}$ and $t\geq g-\delta$,*

$$ATT(g,t)=ATT_{ipw,rc}^{ny}(g,t;\delta)=ATT_{or,rc}^{ny}(g,t;\delta)=ATT_{dr,rc}^{ny}(g,t;\delta).$$

We defer the proof of Theorem B.1 to the Supplementary Appendix. The identification results in Theorem B.1 suggest a simple two-step estimation procedure for the $ATT(g,t)$ with repeated cross-section data that is analogous to the panel data case discussed in Section 4. The asymptotic properties of such two-step estimators follow from analogous arguments; the details are omitted for brevity. Likewise, we can aggregate these estimators to provide summary measures of the causal effects like those discussed in Section 3.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.12.001.

References

Abadie, A., 2005. Semiparametric difference-in-difference estimators. Rev. Econom. Stud. 72, 1–19.
Abadie, A., Athey, S., Imbens, G., Wooldridge, J., 2017. When should you adjust standard errors for clustering? Working Paper. pp. 1–33.
Abbring, J.H., van den Berg, G.J., 2003. The nonparametric identification of treatment effects in duration models. Econometrica 71 (5), 1491–1517.
Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containin unknown functions. Econometrica 71 (6), 1795–1843.
Ai, C., Chen, X., 2007. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. J. Econometrics 141 (1), 5–43.
Ai, C., Chen, X., 2012. The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. J. Econometrics 170 (2), 442–457.
Athey, S., Imbens, G.W., 2006. Identification and inference in nonlinear difference in differences models. Econometrica 74 (2), 431–497.
Athey, S., Imbens, G.W., 2018. Design-based analysis in difference-in-differences settings with staggered adoption. Working Paper.
Bailey, M.J., Goodman-Bacon, A., 2015. The war on poverty's experiment in public medicine: Community health centers and the mortality of older Americans. Amer. Econ. Rev. 105 (3), 1067–1104.
Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C., 2017. Program evaluation and causal inference with high-dimensional data. Econometrica 85 (1), 233–298.
Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates? Q. J. Econ. 119 (1), 249–275.
Bojinov, I., Rambachan, A., Shephard, N., 2020. Panel experiments and dynamic causal effects: A finite population perspective. Working Paper.
Bonhomme, S., Sauder, U., 2011. Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. Rev. Econ. Stat. 93 (May), 479–494.
Borusyak, K., Jaravel, X., 2017. Revisiting event study designs. Working Paper. pp. 1–33.
Botosaru, I., Gutierrez, F.H., 2018. Difference-in-differences when the treatment status is observed in only one period. J. Appl. Econometrics 33 (1), 73–90.
Busso, M., Dinardo, J., McCrary, J., 2014. New evidence on the finite sample properties of propensity score reweighting and matching estimators. Rev. Econ. Stat. 96 (5), 885–895.
Callaway, B., Li, T., Oka, T., 2018. Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. J. Econometrics 206 (2), 395–413.
Card, D., Krueger, A.B., 1994. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. Amer. Econ. Rev. 84 (4), 772–793.
de Chaisemartin, C., D'Haultfœuille, X., 2017. Fuzzy differences-in-differences. Rev. Econom. Stud. 85, 999–1028.
de Chaisemartin, C., D'Haultfœuille, X., 2020. Two-way fixed effects estimators with heterogeneous treatment effects. Amer. Econ. Rev. 110 (9), 2964–2996.
Chen, X., Hong, H., Tarozzi, A., 2008. Semiparametric efficiency in GMM models with auxiliary data. Ann. Statist. 36 (2), 808–843.
Chen, X., Linton, O., Van Keilegom, I., 2003. Estimation of semiparametric models when the criterion function is not smooth. Econometrica 71 (5), 1591–1608.
Cheng, G., Yu, Z., Huang, J.Z., 2013. The cluster bootstrap consistency in generalized estimating equations. J. Multivariate Anal. 115, 33–47.
Chernozhukov, V., Fernández-Val, I., Hahn, J., Newey, W., 2013. Average and quantile effects in nonseparable panel models. Econometrica 81 (2), 535–580.
Chernozhukov, V., Fernández-Val, I., Luo, Y., 2018. The sorted effects method: Discovering heterogeneous effects beyond their averages. Econometrica 86 (6), 1911–1938.
Conley, T., Taber, C., 2011. Inference with "difference in differences" with a small number of policy changes. Rev. Econ. Stat. 93 (1), 113–125.
Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A., 2009. Dealing with limited overlap in estimation of average treatment effects. Biometrika 96 (1), 187–199.
Dube, A., Lester, T.W., Reich, M., 2010. Minimum wage effects across state borders: Estimates using contiguous counties. Rev. Econ. Stat. 92 (4), 945–964.
Dube, A., Lester, T.W., Reich, M., 2016. Minimum wage shocks, employment flows, and labor market frictions. J. Lab. Econ. 34 (3), 663–704.
Farber, H.S., 2017. Employment, hours, and earnings consequences of job loss: US evidence from the Displaced Workers Survey. J. Lab. Econ. 35 (S1), S235–S272.
Ferman, B., Pinto, C., 2019. Inference in differences-in-differences with few treated groups and heteroskedasticity. Rev. Econ. Stat. 101 (3), 452–467.
Freyberger, J., Rai, Y., 2018. Uniform confidence bands: Characterization and optimality. J. Econometrics 204 (1), 119–130.
Gibbons, C.E., Suárez Serrato, J.C., Urbancic, M.B., 2018. Broken or fixed effects? J. Econometr. Methods 8 (1).
Goodman-Bacon, A., 2019. Difference-in-differences with variation in treatment timing. NBER Working Paper n. 25018. Working Paper.
Graham, B., Pinto, C., Egel, D., 2012. Inverse probability tilting for moment condition models with missing data. Rev. Econom. Stud. 79 (3), 1053–1079.
Hájek, J., 1971. Discussion of 'An essay on the logical foundations of survey sampling, Part I', by D. Basu. In: Godambe, V.P., Sprott, D.A. (Eds.), Foundations of Statistical Inference. Holt, Rinehart, and Winston, Toronto.
Han, S., 2020. Identification in nonparametric models for dynamic treatment effects. J. Econometrics (forthcoming).
Heckman, J.J., Humphries, J.E., Veramendi, G., 2016. Dynamic treatment effects. J. Econometrics 191 (2), 276–292.
Heckman, J.J., Ichimura, H., Smith, J., Todd, P., 1998. Characterizing selection bias using experimental data. Econometrica 66 (5), 1017–1098.
Heckman, J.J., Ichimura, H., Todd, P., 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. Rev. Econom. Stud. 64 (4), 605–654.
Imai, K., Kim, I.S., Wang, E., 2018. Matching methods for causal inference with time-series cross-section data. Working Paper.
Khan, S., Tamer, E., 2010. Irregular identification, support conditions, and inverse weight estimation. Econometrica 78 (6), 2021–2042.
Kline, P., Santos, A., 2012. A score based approach to wild bootstrap inference. J. Econometr. Methods 1 (1), 1–40.
Kosorok, M.R., 2008. Introduction to Empirical Processes and Semiparametric Inference. Springer, New York, NY.
Laporte, A., Windmeijer, F., 2005. Estimation of panel data models with binary indicators when treatment effects are not constant over time. Econom. Lett. 88 (3), 389–396.
MacKinnon, J.G., Webb, M.D., 2018. The wild bootstrap for few (treated) clusters. Econom. J. 21 (2), 114–135.
MacKinnon, J.G., Webb, M.D., 2020. Randomization inference for difference-in-differences with few treated clusters. J. Econometrics 218 (2), 435–450.
Malani, A., Reif, J., 2015. Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform. J. Publ. Econ. 124, 1–17.

Mammen, E., 1993. Bootstrap and wild bootstrap for high dimensional linear models. Ann. Statist. 21 (1), 255–285.
Marcus, M., Sant'Anna, P.H.C., 2020. The role of parallel trends in event study settings : An application to environmental economics. J. Assoc. Environ. Resour. Econom. (forthcoming).
McCrary, J., 2007. The effect of court-ordered hiring quotas on the composition and quality of police. Amer. Econ. Rev. 97 (1), 318–353.
Meer, J., West, J., 2016. Effects of the minimum wage on employment dynamics. J. Hum. Resour. 51 (2), 500–522.
Montiel Olea, J.L., Plagborg-Møller, M., 2018. Simultaneous confidence bands: Theory, implementation, and an application to SVARs. J. Appl. Econometrics 1–64.
Murphy, S.A., 2003. Optimal dynamic treatment regimes. J. R. Stat. Soc. Ser. B Stat. Methodol. 65 (2), 331–366.
Murphy, S.A., Van der Laan, M.J., Robins, J.M., Bierman, K.L., Coie, J.D., Greenberg, M.T., Lochman, J.E., McMahon, R.J., Pinderhughes, E., 2001. Marginal mean models for dynamic regimes. J. Amer. Statist. Assoc. 96 (456), 1410–1423.
Neumark, D., Wascher, W., 2000. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Comment. Amer. Econ. Rev. 90 (5), 1362–1396.
Neumark, D., Wascher, W.L., 2008. Minimum Wages. The MIT Press, Cambridge, MA.
Newey, W.K., 1994. The asymptotic variance of semiparametric estimators. Econometrica 62 (6), 1349–1382.
Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Handbook of Econometrics, Vol. 4. Elsevier, North-Holland, Amsterdam, pp. 2111–2245.
Oreopoulos, P., von Wachter, T., Heisz, A., 2012. The short- and long-term career effects of graduating in a recession. Am. Econ. J.: Appl. Econ. 4 (1), 1–29.
Qin, J., Zhang, B., 2008. Empirical-likelihood-based difference-in-differences estimators. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (8), 329–349.
Robins, J.M., 1986. A new approach to causal inference in mortality studies with a sustained exposure period - Application to control of the healthy worker survivor effect. Math. Modelling 7, 1393–1512.
Robins, J.M., 1987. Addendum to 'A new approach to causal inference in mortality studies with a sustained exposure period - Application to control of the healthy worker survivor effect'. Comput. Math. Appl. 14 (9–12), 923–945.
Roth, J., 2020. Pre-test with caution: Event-study estimates after testing for parallel trends. Working Paper. pp. 1–84.
Rubin, D.B., 2007. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. Stat. Med. 26 (1), 20–36.
Rubin, D.B., 2008. For objective causal inference, design trumps analysis. Ann. Appl. Stat. 2 (3), 808–840.
Sant'Anna, P.H.C., Song, X., 2019. Specification tests for the propensity score. J. Econometrics 210, 379–404.
Sant'Anna, P.H., Zhao, J., 2020. Doubly robust difference-in-differences estimators. J. Econometrics 219 (1), 101–122.
Sherman, M., Le Cessie, S., 2007. A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. Comm. Statist. Simulation Comput. 26 (3), 901–925.
Sianesi, B., 2004. An evaluation of the Swedish system of active labor market programs in the 1990s. Rev. Econ. Stat. 86 (1), 133–155.
Słoczyński, T., 2018. A general weighted average representation of the ordinary and two-stage least squares estimands. Working Paper.
Sun, L., Abraham, S., 2020. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Working Paper.
van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press, Cambridge.
van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes. Springer, New York.
Wooldridge, J.M., 2003. Cluster-sample methods in applied econometrics. Am. Econ. Rev. P P 93 (2), 133–138.
Wooldridge, J.M., 2005a. Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. Rev. Econ. Stat. 87 (2), 385–390.
Wooldridge, J.M., 2005b. Violating ignorability of treatment by controlling for too many factors. Econometr. Theory 21 (5), 1026–1028.
Wooldridge, J.M., 2007. Inverse probability weighted estimation for general missing data problems. J. Econometrics 141 (2), 1281–1301.
Yang, S., Ding, P., 2018. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. Biometrika 105 (2), 487–493.