Difference-in-Differences with Multiple Time Periods*

Brantly Callaway† Pedro H. C. Sant'Anna‡

March 1, 2019

Abstract

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 propose a simple two-step estimation strategy, 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. We also propose a semiparametric data-driven testing procedure to assess the credibility of the DID design in our context. 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.

JEL: C14, C21, C23, J23, J38.
Keywords: Difference-in-Differences, Event Study, Multiple Periods, Variation in Treatment Timing, Pre-Testing, Semi-Paramatric.

1 Introduction

Difference-in-Differences (DID) has become one of the most popular 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 individuals are treated (the treated group), and some individuals are not (the control group). If, in the absence of treatment, the average outcomes for treated and control groups would have followed parallel paths over time (which is the so-called parallel

*First complete version: March 23, 2018. 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". We thank Sebastian Calonico, Xiaohong Chen, Clement de Chaisemartin, Xavier D'Haultfoeuille, Bruno Ferman, Andrew Goodman-Bacon, Federico Gutierrez, Sukjin Han, Hugo Jales, Vishal Kamat, Tong Li, Catherine Maclean, Aureo de Paula, Donald Rubin, Bernhard Schmidpeter, Yuya Sasaki, Na'Ama Shenhav, Tymon Słoczyński, Sebastian Tello-Trillo, Jeffrey Wooldridge, Haiqing Xu and seminar participants at Yale University, Syracuse University, the University of Central Florida, the University of Mississippi, the University of Texas at Austin, the 2017 Southern Economics Association (SEA), the 2018 International Association for Applied Econometrics, the 2018 Latin American Workshop in Econometrics, the 2018 SEA Conference, the 2018 Latin American Meeting of the Econometric Society, and the 2018 Canadian Econometrics Study Group for valuable comments. Code to implement the methods proposed in the paper is available in the R package `did` which is available on CRAN.

†Department of Economics, Temple University. Email: brantly.callaway@temple.edu
‡Department of Economics, Vanderbilt University. Email: pedro.h.santanna@vanderbilt.edu

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 control group. Most methodological extensions of DID methods 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), Botosaru and Gutierrez (2017), de Chaisemartin and D'Haultfœuille (2017), and Callaway et al. (2018); 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.

Many DID empirical applications, however, deviate from the standard DID setup and have more than two time periods and variation in treatment timing. In this article, we consider identification, estimation, and inference procedures for average treatment effects in DID models 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 concentrate our attention on DID with staggered adoption, i.e., to DID setups such that once an individual/group is treated, he/she remains treated in the following periods. Importantly, we emphasize that our proposal does not rely on functional form restrictions about the potential outcomes and automatically accommodates general forms of treatment effect heterogeneity that can also vary with observed covariates and time.

We develop our approach in several steps. To fix ideas, consider the popular "two-way fixed effects" (TWFE) regression model

$$Y_{it} = \alpha_t + c_i + \beta D_{it} + \theta X_i + \epsilon_{it}, \tag{1.1}$$

where Y_{it} is the outcome of interest, α_t is a time fixed effect, c_i is an "individual/group" fixed effect,[1] D_{it} is a treatment indicator that is equal to one if an individual i is treated at time t and zero otherwise, X_i is a vector of observed characteristics, and ϵ_{it} is an error term. In some cases the β coefficient in (1.1) is of intrinsic interest, for example if treatment effects are homogeneous. On the other hand, in many situations β may not be closely related to the causal parameter of interest. For instance, when there is variation in treatment timing and treatment effects are dynamic, Borusyak and Jaravel (2017), de Chaisemartin and D'Haultfœuille (2018), Goodman-Bacon (2018), Abraham and Sun (2018) and Athey and Imbens (2018) point out that β represents a weighted average of these dynamic effects where some of these weights can be negative.[2] In such cases, β is not the relevant parameter for evaluating policy changes as it is possible, for example, for β to be negative even in the case where the effect of participating in the treatment is always positive.

In this paper, we propose a general framework that allows one to identify, estimate, and make inference about treatment effect parameters other than β in (1.1). The main building block of our analysis is the *group-time average treatment effects*, 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, they collapse to the ATT. Two attractive features of the group-time average treatment effect parameters are that (*i*) they are not determined by the estimation method one adopts (e.g., first difference or fixed

[1] Group fixed effects are defined at a different level of aggregation than the covariates X_i; otherwise, one cannot separately identity the effect of c_i and X_i on Y_{it}.

[2] See also Wooldridge (2005), Chernozhukov et al. (2013) and Słoczyński (2018) for results related to causal interpretation of β under other sources of treatment effect heterogeneity.

effect linear regression), and (*ii*) they do not directly restrict heterogeneity with respect to observed covariates, the time one is first treated (group), 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.

In some applications, our approach may deliver many group-time average treatment effects, and a second contribution of our paper is to consider different ways one can aggregate group-time average treatment effects into summary measures of the causal effects. Although the "best" way to aggregate these parameters is, in general, application specific, we consider several leading cases and aggregation schemes that are motivated by economic theory and the context of the analysis. In particular, we propose aggregation procedures depending on whether one is concerned with (*a*) selective treatment timing, i.e., allowing, for example, the possibility that individuals with the largest benefits from participating in a treatment choose to become treated earlier than those with a smaller benefit; (*b*) dynamic treatment effects – where the effect of a treatment can depend on the length of exposure to the treatment; or (*c*) calendar time effects – where the effect of treatment may depend on the time period. Overall, like the β coefficient in (1.1), our proposed aggregation procedures result in a single estimated "effect" of treatment.

A third contribution of this paper is to propose estimators and provide asymptotically valid inference procedures for the causal parameters of interest. In the same spirit as Abadie (2005), we consider inverse probability weighted estimators for the treatment effects. We extend Abadie (2005)'s estimators in two directions. First and most importantly, we account for variation in treatment timing. Second, our proposed estimators are of the Hájek (1971) type, whereas the Abadie (2005) estimator is of the Horvitz and Thompson (1952) type. In other words, our proposed weights are normalized to one, while the weights in Abadie (2005) are not. As discussed in Robins et al. (2007), Hájek-type estimators are sample bounded - i.e., the estimates are enforced to lie within the range of the observed outcomes - whereas Horvitz-Thompson estimators are not. In practice, this modification usually translates to estimators with improved finite sample properties; see, e.g., Busso et al. (2014).

In order to conduct asymptotically valid inference, we justify the use of a computationally convenient multiplier-type bootstrap procedure 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 probability $1-\alpha$, and take into account the dependency across different group-time average treatment effects estimators. Thus, our proposed confidence bands are arguably more suitable for visualizing the overall estimation uncertainty than more traditional pointwise confidence intervals.

Finally, it is worth mentioning that the reliability of the causal interpretation of all aforementioned results relies on the validity of a conditional parallel trends assumption. However, the conditional parallel trends assumption we adopt in this article is fundamentally untestable. On the other hand, we show that if one is willing to strengthen the conditional parallel trends assumption such that it holds not only in post-treatment but also in pre-treatment periods, this augmented conditional parallel trends assumption has testable implications when more than two time periods are available. A fourth contribution of this article is to take advantage of this observation and propose a falsification test based on it.

Our pre-test for the plausibility of the conditional parallel trends assumption is based on the integrated

moments approach, completely avoids having to select tuning parameters, and is fully data-driven. To the best of our knowledge, we are the first to note that we can use the conditional moment restrictions to pre-test for the reliability of the conditional parallel trends identifying assumption. We derive the asymptotic null distribution of our falsification test statistic, prove that it is consistent against fixed nonparametric alternatives, and show that critical values can be computed with the assistance of an easy to implement multiplier-type bootstrap. These results build on many papers in the goodness-of-fit literature – see, e.g., Bierens (1982), Bierens and Ploberger (1997), Stute (1997), and Escanciano (2006*b*, 2008); for a recent overview, see González-Manteiga and Crujeiras (2013). However, in contrast with most specification testing proposals, our null hypothesis involves multiple conditional moment restrictions instead of a single conditional moment restriction; see Escanciano (2008) for an exception.

We illustrate the appeal of our method by revisiting findings about the effect of the minimum wage on teen employment. Although classical economic models suggest that a wage floor should result in lower employment, there is a bulk of literature that finds no disemployment effects of the minimum wage – see, e.g., Card and Krueger (1994) and Dube et al. (2010), among many others. However, another strand of the literature argues that raising the minimum wage leads to lower employment – see, e.g., Neumark and Wascher (1992, 2000, 2008), Neumark et al. (2014), and Jardim et al. (2017).

We use data from 2001-2007, where the federal minimum wage was flat at $5.15 per hour. Using a period where the federal minimum wage is flat allows for a clear source of identification – state level changes in minimum wage policy. However, we also need to confront the issue that states changed their minimum wage policy at different points in time over this period – an issue not encountered in the case study approach to studying the employment effects of the minimum wage. In addition, for states that changed their minimum wage policy in later periods, we can pre-test the parallel trends assumption which serves as a check of the internal consistency of the models used to identify minimum wage effects.

We consider both an unconditional and conditional DID approach to estimating the effect of increasing the minimum wage on teen employment rates. For the unconditional DID case, we find that increases in the minimum wage tend to decrease teen employment, which is in line with most of the work on the minimum wage that uses a similar setup. As Dube et al. (2010) points out, such negative effects may be spurious given potential violations of the parallel trends assumption. Indeed, when we test for the reliability of the unconditional parallel trends assumption, we reject it at the usual significance levels. Next, we focus on conditional DID. First, we follow an approach suggested in Dube et al. (2010) and consider a TWFE regression model with region-year fixed effects. As in Dube et al. (2010), such an estimation strategy finds no adverse effect on teen employment. Nonetheless, one must bear in mind that, as discussed before, such a TWFE regression may not identify an easy to interpret causal parameter. To circumvent this issue, we use our conditional DID approach and find that increasing the minimum wage does tend to decrease teen employment, though the magnitude of the effect is slightly smaller than in the unconditional case. The contrast of the findings based on TWFE regression and our proposed method highlights the importance of taking treatment effect heterogeneity into account. On the other hand, when we apply our pre-test for the reliability of the conditional DID setup, we do find evidence against the conditional parallel trends assumption. Thus, one should interpret the findings with care.

Recent Related Literature: This paper is related to the recent and emerging literature on het-

erogeneous treatment effects in DID and/or event studies with variation in treatment timing; see, e.g., de Chaisemartin and D'Haultfœuille (2018), Goodman-Bacon (2018), Imai et al. (2018), Han (2018), Borusyak and Jaravel (2017), Athey and Imbens (2018) and Abraham and Sun (2018).[3] de Chaisemartin and D'Haultfœuille (2018), Goodman-Bacon (2018), Imai et al. (2018), and Han (2018) consider a general framework where individuals can dynamically select in and out of treatment over time, whereas Borusyak and Jaravel (2017), Abraham and Sun (2018), and Athey and Imbens (2018) focus on staggered adoption designs as we do in this paper.

de Chaisemartin and D'Haultfœuille (2018) show that, in a setting without covariates, the causal interpretation of β in (1.1) heavily relies on homogeneous treatment effect assumptions, and given that such assumptions are often implausible, they also propose alternative estimators that can recover a weighted average of treatment effects among the switchers. Their proposal differs from ours in many dimensions. For instance, while we pay particular attention to the role played by covariates, de Chaisemartin and D'Haultfœuille (2018) mainly focus on unconditional designs.[4] Second, whereas de Chaisemartin and D'Haultfœuille (2018) propose an estimator that recovers a particular weighted average of the many treatment effects, our setup allows one to recover the disaggregated causal parameter and form a *family* of different aggregate parameters in a unified manner. On the other hand, the setup in de Chaisemartin and D'Haultfœuille (2018) is more general than ours as we consider staggered adoption designs and they allow for more general treatment selection. Nonetheless, we note that our parallel trends assumption is strictly weaker than the one in de Chaisemartin and D'Haultfœuille (2018), even if one were to specialize their setup to staggered adoptions designs. Overall, given that our and de Chaisemartin and D'Haultfœuille (2018) setups do not nest each other, we emphasize that one should view this paper and de Chaisemartin and D'Haultfœuille (2018) as complements rather than substitutes.

Goodman-Bacon (2018) provides a detailed decomposition of β in TWFE models without covariates in terms of all possible two period−two group DID coefficients. This decomposition highlights that negative weighting appears when treatment effects vary over time essentially because already treated individuals sometimes act as control groups. He also proposes diagnostic checks to assess the relative importance of the different sources of variation in a given application. Finally, we note that in Goodman-Bacon (2018), the parameter of interest is directly tied to β in (1.1), which is much different from our proposal.

Imai et al. (2018) proposes matching-based estimators for the average treatment effect for treated individuals, k periods after they were treated. Han (2018) establishes nonparametric identification of different dynamic treatment effect parameters in settings where sequences of outcomes and treatment choices may influence one another in a dynamic manner. The frameworks of Imai et al. (2018) and Han (2018) are also much different from ours as neither rely on conditional parallel trends assumptions as we do, but instead adopt alternative sets of identifying assumptions.[5]

[3]Importantly, all the aforementioned papers are not yet published. Among these, the only two articles that predate the first complete version of our paper (posted in arXiv on March 23, 2018) are de Chaisemartin and D'Haultfœuille (2018) and Borusyak and Jaravel (2017).

[4]We note that de Chaisemartin and D'Haultfœuille (2018) allow for covariates in their model but in a rather restrictive form. Nonetheless, the same authors have considered covariates in a more flexible manner in previous work; see, e.g., Section 1.4 of the Web Appendix of de Chaisemartin and D'Haultfœuille (2017).

[5]The identifying assumptions adopted by Imai et al. (2018), such as sequential ignorability or parallel trends conditional on covariates and past outcomes, do not nest nor are nested by our conditional parallel trends assumption; see, e.g., Section 6.5.4 of Imbens and Wooldridge (2009), Section 3.2.8 of Lechner (2010), Chabé-Ferret (2015, 2017), and Daw and

Although Borusyak and Jaravel (2017), Athey and Imbens (2018) and Abraham and Sun (2018) consider staggered adoption designs like we do, the focus of their analysis is very different from ours. For instance, Borusyak and Jaravel (2017) shows that the coefficients of TWFE models with treatment leads and lags are, in general, not point identified unless one introduces additional conditions, e.g., existence of valid comparison groups. They also show that β in "static" TWFE least squares models like (1.1) can be expressed as a weighted average of the dynamic effects, but these weights can be negative for long-run effects and also "overweight" short-run effects; to the best of our knowledge, they are the first to document this phenomenon in designs with staggered adoption. Although Borusyak and Jaravel (2017) briefly discuss potential solutions to these issues, they do not provide a formal analysis of alternative estimators like we do. In addition, they are mainly focused on parametric linear models whereas we use a semi/nonparametric approach.

Athey and Imbens (2018) mainly focus on providing design-based inference procedures for unconditional DID estimators under staggered adoption designs. However, the design-based inference procedures proposed by Athey and Imbens (2018) rely on a "random assignment" type assumption that is strictly stronger than our parallel trends assumption.

Abraham and Sun (2018) adapt de Chaisemartin and D'Haultfœuille (2018)'s results to the staggered design and show that the β coefficient in TWFE models does not have a clear causal interpretation under treatment effect heterogeneity. They also consider specifications with leads and lags of treatment indicators, formally show that these specifications are not suitable to "pre-test" for parallel trends, and propose an estimator that is able to recover an easy to understand weighted average of the average treatment effects. In contrast to our approach, Abraham and Sun (2018) mainly focus on unconditional DID designs, and our identification assumptions are strictly weaker than theirs (see Remark 1 and Appendix C in the Supplementary Appendix for further discussion). Our inference procedures are also quite different from theirs since we cover both the panel data and the repeated cross-section data case, explicitly account for potential multiple-testing problems when constructing confidence intervals, and propose pre-tests for the reliability of the parallel trends assumption. Finally, our general framework allows us to study additional parameters of interest that can take economic theory and/or the context of the analysis into account.

Organization of the paper: The remainder of this article is organized as follows. Section 2 presents our main identification results. We discuss estimation and inference procedures for the treatment effects of interest in Section 3. Section 4 describes our pre-tests for the credibility of the conditional parallel trends assumption. 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 Supplementary Appendix.

Hatfield (2018). Han (2018)'s identifying assumptions include two-way exclusion restrictions, weak separability and a strict monotonicity conditions, and sequential rank similarity. These also do not nest nor are nested by our conditional parallel trends assumption.

2 Identification

2.1 Framework

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 standard DID setup, $\mathcal{T} = 2$ and no one is treated in period 1. Let D_t be a binary variable equal to one if an individual is treated in period t and equal to zero otherwise. Also, define G_g to be a binary variable that is equal to one if an individual is first treated in period g, and define C as a binary variable that is equal to one for individuals in the control group – these are individuals who are never treated so the notation is not indexed by time. For each individual, exactly one of the G_g or C is equal to one. Denote the generalized propensity score as $p_g(X) = P(G_g = 1|X, G_g + C = 1)$. Note that $p_g(X)$ indicates the probability that an individual is treated conditional on having covariates X and conditional on being a member of group g or the control group. Finally, let $Y_t(1)$ and $Y_t(0)$ be the potential outcome at time t with and without treatment, respectively. The observed outcome in each period can be expressed as $Y_t = D_t Y_t(1) + (1 - D_t) Y_t(0)$.

Given that $Y_t(1)$ and $Y_t(0)$ cannot be observed for the same individual at the same time, researchers often focus on estimating some function of the potential outcomes. For instance, in the standard DID setup, the most popular treatment effect parameter is the average treatment effect on the treated, denoted by[6]

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

Unlike the two period and two group case, when there are more than two periods and variation in treatment timing, it is not obvious what the main causal parameter of interest should be. We focus on the average treatment effect for individuals who are members of a particular group g in a particular time period t, denoted by

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

We call this causal parameter the *group-time average treatment effect*. In particular, note that in the classical DID setup, $ATT(2,2)$ collapses to ATT.

In this article, we are interested in identifying and making inference about $ATT(g,t)$ and functionals of $ATT(g,t)$. At this stage, one may wonder about the advantages of first focusing on the family of $ATT(g,t)$ instead of directly focusing on more aggregate measures of treatment effects. In our view, the main advantage of first focusing on the family of $ATT(g,t)$ is to understand treatment effect heterogeneity across different dimensions in a unified manner. In addition, by first focusing on $ATT(g,t)$ one can later construct different summary treatment effect measures that can highlight different sources of heterogeneity. For instance, by identifying and estimating all possible $ATT(g,t)$, one would be able to answer questions like: (a) Are treatment effects heterogeneous by time of adoption? (b) Does the effect of the treatment increase over time? (c) Are short-run effects more pronounced than long-run effects? (d) Do treatment effect dynamics differ if people are first treated in recession years relative to expansion years? Note that all these questions are application dependent, and it is unlikely that a single summary measure of treatment effects can be used to answer them all. On the other hand, as we discuss in Section 2.3, one can build on the $ATT(g,t)$ to construct appropriate summary measures that take the context

[6] Existence of expectations is assumed throughout.

of the application into account. Given that the "best" way to construct summary measures is application/context dependent, we view this level of generality and flexibility as one of the main advantages of our framework that first focuses on the family of $ATT(g,t)$.

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

Assumption 1 (Sampling). *$\{Y_{i1}, Y_{i2}, \dots Y_{i\mathcal{T}}, X_i, D_{i1}, D_{i2}, \dots, D_{i\mathcal{T}}\}_{i=1}^n$ is independent and identically distributed (iid).*

Assumption 2 (Conditional Parallel Trends). *For all $t = 2, \dots, \mathcal{T}$, $g = 2, \dots, \mathcal{T}$ such that $g \leq t$,*

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

Assumption 3 (Irreversibility of Treatment). *For $t = 2, \dots, \mathcal{T}$,*

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

Assumption 4 (Overlap). *For all $g = 2, \dots, \mathcal{T}$, $P(G_g = 1) > 0$ and for some $\varepsilon > 0$, $p_g(X) < 1 - \varepsilon$ a.s..*

Assumption 1 implies that we are considering the case with panel data. The extension to the case with repeated cross sections is fully developed in Appendix B in the Supplementary Appendix though we note here that the arguments are very similar.

Assumption 2, which we refer to as the (conditional) parallel trends assumption throughout the paper, is the crucial identifying restriction for our DID model, and it generalizes the two-period DID assumption to the case where it holds in multiple periods for each group; see, e.g., Heckman et al. (1997, 1998), Blundell et al. (2004), and Abadie (2005). It states that, conditional on covariates, the average outcomes for the group first treated in period g and for the control group would have followed parallel paths in the absence of treatment. We require this assumption to hold for all groups g and all time periods t such that $g \leq t$; that is, it holds in all periods after group g is first treated. Note that Assumption 2 does not restrict the evolution of potential outcomes for periods $t < g$. This subtle distinction turns out to be important as we do not need to restrict "pre-trends" to nonparametrically identify the $ATT(g,t)$. Finally, it is important to emphasize that the parallel trends assumption holds only after conditioning on some covariates X, therefore allowing for X-specific time trends. All of our analysis continues to go through in the case where an unconditional parallel trends assumption holds by simply setting $X = 1$.

Assumption 3 states that once an individual becomes treated, that individual will also be treated in the next period. With regards to the minimum wage application, Assumption 3 says that once a state increases its minimum wage above the federal level, it does not decrease it back to the federal level during the analyzed period. Moreover, this assumption is consistent with most DID setups that exploit the enacting of a policy in some location while the policy is not enacted in another location.

Finally, Assumption 4 states that a positive fraction of the population start to be treated in period g, and that, for any possible value of the covariates X, there is some positive probability that an individual is not treated.[7] This is a standard covariate overlap condition; see, e.g., Heckman et al. (1997, 1998), Blundell et al. (2004), Abadie (2005).

[7] In our application on the minimum wage, we must take somewhat more care here as there are some periods where there are no states that increase their minimum wage. In this case, let $\mathcal{G}$ denote the set of first treatment times with $\mathcal{G} \subseteq \{1, \dots, \mathcal{T}\}$. Then, one can compute $ATT(g,t)$ for groups $g \in \mathcal{G}$ with $g \leq t$. This is a simple complication to deal with

Remark 1. In some applications, eventually all units are treated, implying that C is never equal to one. In such cases one can consider the "not yet treated" ($D_t = 0$) as a control group instead of the "never treated" ($C = 1$). That is, instead of relying on Assumption 2, one could instead assume that for all $t = 2, \ldots, \mathcal{T}$, $g = 2, \ldots, \mathcal{T}$ such that $g \leq t$,

$$\mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = \mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, D_t = 0] \ a.s.. \tag{2.1}$$

We present a detailed discussion of this case in Appendix C in the Supplementary Appendix.

When there is a "never treated" group, we note that the parallel trends assumption (2.1) and Assumption 2 are non-nested, though. On the other hand, the parallel trends assumption (2.1) is strictly weaker than the assumptions made by Abraham and Sun (2018), for example, as Abraham and Sun (2018) requires common trends for all groups g and (2.1) only requires parallel trends between those individuals treated at time g and the "supergroup" of those not yet treated.[8] In addition, we emphasize that our assumptions do not restrict the evolution of potential outcomes before treatment. See Appendix C in the Supplementary Appendix for additional details about this case.

2.2 Group-Time Average Treatment Effects

In this section, we introduce the nonparametric identification strategy for the group-time average treatment effect $ATT(g,t)$. Importantly, we allow for treatment effect heterogeneity and do not make functional form assumptions about the evolution of potential outcomes.

Theorem 1. *Under Assumptions 1 - 4 and for $2 \leq g \leq t \leq \mathcal{T}$, the group-time average treatment effect for group g in period t is nonparametrically identified, and given by*

$$ATT(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1})\right]. \tag{2.2}$$

Theorem 1 says that, under Assumptions 1 - 4, a simple weighted average of "long differences" of the outcome variable recovers the group-time average treatment effect. The weights depend on the generalized propensity score $p_g(X)$, and are normalized to one. The intuition for the weights is simple. One takes observations from the control group and group g, omitting other groups, and then weights up observations from the control group that have characteristics similar to those frequently found in group g and weights down observations from the control group that have characteristics that are rarely found in group g. Such a reweighting procedure guarantees that the covariates of group g and the control group are balanced. Interestingly, in the standard DID setup of two periods only, $\mathbb{E}[p_2(X)C/(1-p_2(X))] = \mathbb{E}[G_2]$, and the results of Theorem 1 reduce to Lemma 3.1 in Abadie (2005).

To shed light on the role of the "long difference", we give a sketch of how this argument works in the unconditional case, i.e., when $X = 1$. Recall that the key identification challenge is for $\mathbb{E}[Y_t(0)|G_g = 1]$

in practice, so we consider the notationally more convenient case where there are some individuals treated in all periods (possibly excluding period 1) in the main text of the paper.

[8] The same caveat applies to the assumptions made by de Chaisemartin and D'Haultfœuille (2018), though they do not restrict their attention to staggered treatment adoption as we do.

which is not observed when $g \leq t$. Under the parallel trends assumption,

$$\begin{aligned}\mathbb{E}[Y_t(0)|G_g = 1] &= \mathbb{E}[Y_t(0) - Y_{t-1}(0)|G_g = 1] + \mathbb{E}[Y_{t-1}(0)|G_g = 1] \\ &= \mathbb{E}[Y_t - Y_{t-1}|C = 1] + \mathbb{E}[Y_{t-1}(0)|G_g = 1]\end{aligned}$$

The first term is identified; it is the change in outcomes between $t-1$ and t experienced by the control group. If $g > t-1$, then the last term is identified. If not,

$$\mathbb{E}[Y_{t-1}(0)|G_g = 1] = \mathbb{E}[Y_{t-1} - Y_{t-2}|C = 1] + \mathbb{E}[Y_{t-2}(0)|G_g = 1]$$

which holds under the parallel trends assumption. If $g > t-2$, then every term above is identified. If not, one can proceed recursively in this same fashion until

$$\mathbb{E}[Y_t(0)|G_g = 1] = \mathbb{E}[Y_t - Y_{g-1}|C = 1] + \mathbb{E}[Y_{g-1}|G_g = 1],$$

implying the result for $ATT(g,t)$.

One final thing to consider in this section is the case when the parallel trends assumption holds without needing to condition on covariates. In this case, (2.2) simplifies to

$$ATT(g,t) = \mathbb{E}[Y_t - Y_{g-1}|G_g = 1] - \mathbb{E}[Y_t - Y_{g-1}|C = 1], \tag{2.3}$$

which is simpler than the weighted representation in (2.2) but also implies that all of our results will also cover the unconditional case which is commonly used in empirical work. We discuss an alternative regression based approach to obtaining $ATT(g,t)$ in Appendix D in the Supplementary Appendix.[9]

Remark 2. From (2.3) one can see that when the parallel trends assumption holds unconditionally, the $ATT(g,t)$ parameter can be obtained by first subsetting the data to only contain observations at time t and $g-1$, from groups with either $G_g = 1$ or $C = 1$, and then, using only the observations of this subset, running the (population) linear regression

$$Y = \alpha_1^{g,t} + \alpha_2^{g,t} \cdot G_g + \alpha_3^{g,t} \cdot 1\{T = t\} + \beta^{g,t} \cdot (G_g \times 1\{T = t\}) + \epsilon^{g,t}. \tag{2.4}$$

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. The weighted average representation in (2.2) avoids that.

In addition, note that when covariates are available, the $\tilde{\beta}^{g,t}$ coefficient of the population linear regression

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

that uses the same subset of data as before is, in general, not equal to $ATT(g,t)$ unless one is willing to assume that

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

a condition deemed too strong by most of the causal inference literature; see, e.g., Słoczyński (2018) for a

[9] Unlike the two period, two group case, there does not appear to be any advantage to trying to obtain $ATT(g,t)$ from a regression as it appears to require post-processing the regression output.

related discussion. The characterization of $ATT(g,t)$ in (2.2) does not rely on this restrictive condition.

2.3 Summarizing Group-time Average Treatment Effects

The previous section shows that the group-time average treatment effect $ATT(g,t)$ is identified for $g \leq t$. These are very useful parameters – they allow one to consider how the effect of treatment varies by group and time. However, in some applications there may be many of them, perhaps too many to easily interpret the effect of a given policy intervention. This section considers ways to aggregate group-time average treatment effects into fewer interpretable causal effect parameters. In applications, aggregating the group-time average treatment effects is also likely to increase statistical power, reducing estimation uncertainty.

The two simplest ways of combining $ATT(g,t)$ across g and t are

$$\frac{2}{\mathcal{T}(\mathcal{T}-1)} \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} ATT(g,t) \quad \text{and} \quad \frac{1}{\kappa} \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} ATT(g,t) P(G=g) \tag{2.5}$$

where $\kappa = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} 1\{g \leq t\} P(G=g)$ (which ensures that the weights on $ATT(g,t)$ in the second term sum to 1).[10] The first term in (2.5) is just the simple average of $ATT(g,t)$; the second is a weighted average of each $ATT(g,t)$ putting more weight on $ATT(g,t)$ with larger group sizes. Unlike β in the TWFE regression model, these simple combinations of $ATT(g,t)$ immediately rule out troubling cases due to negative weights; as a particular example, when the effect of participating in the treatment is positive for all individuals, these aggregated parameters cannot be negative. However, as we argue below, in many cases, it appears that researchers can choose more appropriate summary treatment effect measures that take into account economic theory and the context of the analysis. Interestingly, in the case with homogeneous treatment effects across groups and time, all the group-time average treatment effects are equal to each other and, therefore, all of the aggregated parameters that we consider in this section will be equal to each other.

In contrast to our approach in this section, the most common approach to estimating the effect of a binary treatment in a panel data setup is to interpret β in the following regression as the average treatment effect

$$Y_{it} = \alpha_t + c_i + \beta D_{it} + \theta X_i + \epsilon_{it},$$

where α_t is a time fixed effect and c_i is an individual/group fixed effect. Interestingly, Wooldridge (2005), Chernozhukov et al. (2013), Borusyak and Jaravel (2017), Goodman-Bacon (2018), Słoczyński (2018), de Chaisemartin and D'Haultfœuille (2018), Abraham and Sun (2018) and Athey and Imbens (2018) have shown that, in general, β does not represent an easy to interpret average treatment effect parameter. 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.

[10] Here we use the shorthand notation $P(G=g)$ to denote $P(G_g = 1 | G_1 + C = 0)$. Thus, $P(G=g)$ is the probability that an individual is first treated in period g conditional on not being in the control group or in the group first treated in period 1. Throughout this section, conditional probabilities such as $P(G=g|g \leq t)$ also implicitly condition on not being in the control group or in the group first treated in period 1.

In the following, we consider several common cases that are likely to occur in practice: (a) selective treatment timing, (b) dynamic treatment effects, and (c) calendar time effects. We provide some recommendations on constructing interpretable treatment effect parameters under each of these setups. It is worth mentioning that in each of these cases, $ATT(g,t)$ still provides the average causal effect of the treatment for group g in period t; the issue in this section is how to aggregate $ATT(g,t)$ into a smaller number of causal effect parameters.

Selective Treatment Timing In many cases, when to become treated is a choice variable. The parallel trends assumption does place some restrictions on how individuals select when to be treated. In particular, in order for the path of untreated potential outcomes to be the same for a particular group and the control group, the parallel trends assumption does not permit individuals to select into treatment in period t because they anticipate a negative "shock" to their untreated potential outcomes in that period. On the other hand, it does allow for some selection on the basis of time-invariant unobserved characteristics. In addition, it does not place restrictions on how treated potential outcomes are generated at all. Thus, our imposed DID assumptions fully allow for individuals to select into treatment on the basis of expected future values of treated potential outcomes.

While some forms of selective treatment timing are permitted under the parallel trends assumption and do not affect identification of group-time average treatment effects, they do have implications for the "best ways" to combine $ATT(g,t)$ into a single, easy to interpret treatment effect parameter. In particular, when there is selective treatment timing, the period when an individual is first treated may provide information about the size of the treatment effect. In such cases, we propose to summarize the causal effect of a policy by first aggregating $ATT(g,t)$ by group, and then combine group average treatment effects based on the size of each group.

More precisely, we first consider

$$\tilde{\theta}_S(g) = \frac{1}{\mathcal{T} - g + 1} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} ATT(g,t).$$

Note that $\tilde{\theta}_S(g)$ is the time-averaged treatment effect for individuals in group g, i.e., just a time-average of each available $ATT(g,t)$ for group g. Next, in order to further reduce the dimensionality of $\tilde{\theta}_S(g)$, one can average $\tilde{\theta}_S(g)$ across groups to get

$$\theta_S = \sum_{g=2}^{\mathcal{T}} \tilde{\theta}_S(g) P(G = g). \tag{2.6}$$

Note that θ_S appears to be quite similar to the second term in (2.5). The difference is in the weights. The second term in (2.5) puts more weight on groups that are exposed to treatment longer. The weights in (2.6) only depend on group size, not on the number of post-treatment periods available per group. For example, suppose there is positive selective treatment timing so that individuals who are treated earlier experience larger benefits from being treated than those who are treated later. In the presence of selective treatment timing, the approach in (2.5) would tend to overstate the effect of the treatment due to putting more weight on the groups that are treated the longest, which are precisely the ones that experience the

largest benefits of being treated. Thus, we argue that, in the presence of selective treatment timing, θ_S in (2.6) is a more natural causal parameter than the second term in (2.5).

Dynamic Treatment Effects In other cases, the effect of a policy intervention may depend on the length of exposure to it. To give some examples, Jacobson et al. (1993) argues that workers that are displaced from their jobs tend to have immediate large earnings effects that get smaller over time, and both the immediate effect and the dynamic effect are of interest. In the case of the minimum wage, Meer and West (2016) argue that increasing the minimum wage leads to lower job creation and thus that the effect of the minimum wage on employment is dynamic – one should expect larger effects in subsequent periods than in the initial period.

In the presence of dynamic treatment effects (but not selective treatment timing), we propose to summarize the effects of the policy by first aggregating $ATT(g,t)$ by the length of exposure to treatment (we denote this by e), and then (possibly) combining average effects based on length of exposure by averaging over different lengths of exposure. That is, we first consider the parameter

$$\tilde{\theta}_D(e) = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{t-g+1=e\} ATT(g,t) P(G=g|t-g+1=e),$$

which provides the average effect of treatment for individuals that have been treated for exactly e periods. For example, when $e=1$, it averages (based on group size) $ATT(g,t)$ for $g=t$ (groups that have been exposed to treatment for exactly one period). Averaging over all possible values of e results in the parameter

$$\theta_D = \frac{1}{\mathcal{T}-1} \sum_{e=1}^{\mathcal{T}-1} \tilde{\theta}_D(e). \tag{2.7}$$

The primary difference between θ_D, θ_S, and the second term in (2.5) is the weights. Relative to the other parameters, θ_D puts the most weight on $ATT(g,t)$ when g is much less than t, which corresponds to large values of e, because there are few groups available for large values of e. In the absence of selective treatment timing, these groups are informative about the dynamic effects of treatment for all groups. Hence, we argue that θ_D is appealing when treatment effects evolve over time.

Calendar Time Effects In other cases, calendar time may matter. For example, graduating during a recession may have a large effect on future earnings; see, e.g., Oreopoulos et al. (2012). The case with calendar time effects is similar to the case with dynamic treatment effects. Our proposed summary treatment effect parameter involves first computing an average treatment effect for all individuals that are treated in period t, and then averaging across all periods. Consider the parameter

$$\tilde{\theta}_C(t) = \sum_{g=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} ATT(g,t) P(G=g|g \leq t).$$

Here, $\tilde{\theta}_C(t)$ can be interpreted as the average treatment effect in period t for all groups that are treated by period t. With $\tilde{\theta}_C(t)$ at hand, one can compute

$$\theta_C = \frac{1}{\mathcal{T}-1}\sum_{t=2}^{\mathcal{T}} \tilde{\theta}_C(t),$$

which can be interpreted as the average treatment effect when calendar time matters. When calendar time matters, the most weight is put on groups that are treated in the earliest periods. This is because there are fewer groups available to estimate the average treatment effect in period t when t is small relative to the number of groups available to estimate the average treatment effect in period t when t is large.

Selective Treatment Timing and Dynamic Treatment Effects Finally, we consider the case where the timing of treatment is selected and there are dynamic treatment effects. This might very well be the most relevant case in studying the effect of increasing the minimum wage as (i) states are not likely to raise their minimum wage during a recession and (ii) the effect of the minimum wage takes some time to play out; see, e.g., Meer and West (2016).

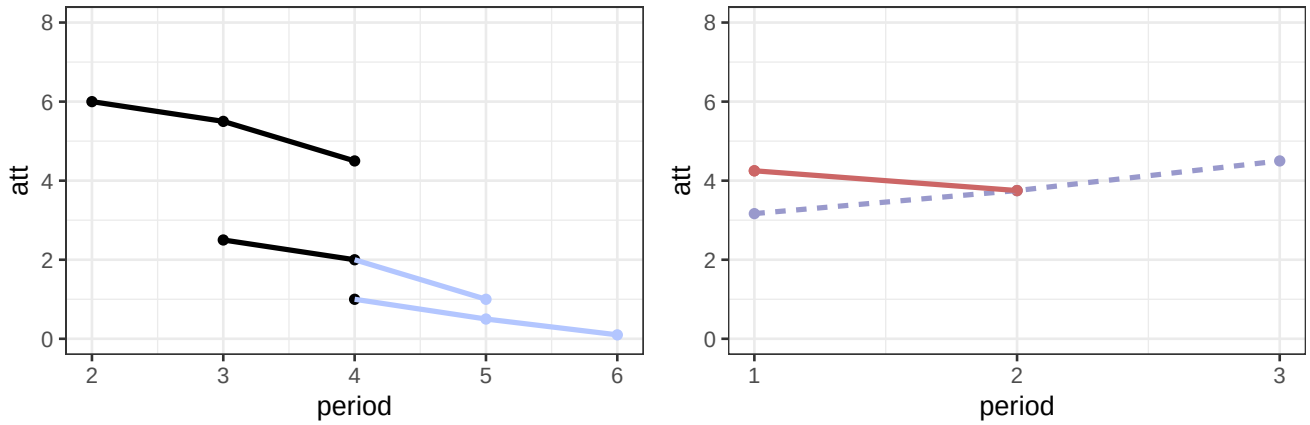
The fundamental problem with using the dynamic treatment effects approach when there is selective treatment timing is that the composition of the treated group changes when the length of exposure to treatment (e) changes. Without selective treatment timing, this does not matter because when an individual first becomes treated does not affect their outcomes. However, with selective treatment timing, changing the composition of the treatment group can have a big effect (See Figure 1 for an example where the dynamic treatment effect is declining with length of exposure to treatment for all groups but ignoring selective treatment timing leads to the opposite (wrong) conclusion – that the effect of treatment is increasing over time.).

To circumvent such an issue, we consider dynamic treatment effects only for $e \leq e'$ and for groups with at least e' periods of post-treatment data available. This setup removes the effect of selective treatment timing by keeping the same set of groups across all values of e. For example, one could consider the dynamic effect of treatment over three periods by averaging $ATT(g,t)$ for all the groups that have at least three periods of post-treatment observations while not utilizing $ATT(g,t)$ for groups that have less than three periods of post-treatment observations. Note that there is some trade-off here. Setting e' small results in many groups satisfying the requirement, but in only being able to study the effect of length of exposure to treatment for relatively few periods. Setting e' to be large decreases the number of available groups but allows one to consider the effect of length of exposure to treatment for relatively more periods.

Next, we describe how this proposed summary causal parameter is constructed. Let $\delta_{gt}(e,e') = 1\{t-g+1=e\}1\{T-g+1 \geq e'\}1\{e \leq e'\}$. Here, $\delta_{gt}(e,e')$ is equal to one in the period where group g has been treated for exactly e periods, if group g has at least e' post-treatment periods available, and if the length of exposure e is less than the post-treatment periods requirement e'.

Then, the average treatment effect for groups that have been treated for e periods and have at least

Figure 1: Example of Selective Treatment Timing and Dynamic Treatment Effects



Notes: In this example, there are three groups: G2 (first treated in period 2), G3 (first treated in period 3), and G4 (first treated in period 4). Suppose that the last period available in the sample is period 4; thus, the group-time average treatment effect is available in periods 2 through 4 – these are the dark lines in the left panel of the figure. The light lines in the left panel represent group-time average treatment effects that are not observed. Each group experiences a declining dynamic treatment effect, but there is also selective treatment timing. Groups that are treated earlier experience larger effects of the treatment. The right panel (dashed line) plots the dynamic treatment effect ignoring selective treatment timing and allowing the composition of the treated group to change. In particular, this means that group G4 is only included in the average for period 1, and group G3 only is included in the average for periods 1 and 2. In this case, selective treatment timing leads to exactly the wrong interpretation of the dynamic treatment effect – it appears as though the effect of the treatment is increasing. The solid line plots the dynamic treatment effect as suggested in Equation (2.8) that adjusts for selective treatment timing and for $e = 1, 2$ and $e' = 2$ and correctly determines the declining dynamic treatment effects.

e' post-treatment periods of data available is given by

$$\tilde{\theta}_{SD}(e, e') = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \delta_{gt}(e, e') ATT(g, t) P(G = g | \delta_{gt}(e, e') = 1) \tag{2.8}$$

which is defined for $e \leq e'$. Effectively, we put zero weight on $ATT(g, t)$ for groups that do not meet the minimum required number of periods in order to prevent the composition of groups from changing. Once $\tilde{\theta}_{SD}(e, e')$ is computed, one can further aggregate it to get

$$\theta_{SD}(e') = \frac{1}{\mathcal{T} - e'} \sum_{e=1}^{\mathcal{T}-e'} \tilde{\theta}_{SD}(e, e')$$

which should be interpreted as the average treatment effect for groups with at least e' periods of post-treatment data allowing for dynamic treatment effects and selective treatment timing. Such a causal parameter has the strengths of both θ_S and θ_D in (2.6) and (2.7), respectively.

3 Estimation and Inference

In this section, we study estimation and inference procedures for estimators corresponding to the estimands introduced in Section 2. Note that the nonparametric identification result in Theorem 1 suggests

a simple two-step strategy to estimate $ATT(g,t)$. In the first step, estimate the generalized propensity score $p_g(x) = P(G_g = 1|X = x, G_g + C = 1)$ for each group g, and compute the fitted values for the sample. In the second step, one plugs the fitted values into the sample analogue of $ATT(g,t)$ in (2.2) to obtain estimates of the group-time average treatment effect.

More concisely, we propose to estimate $ATT(g,t)$ by

$$\widehat{ATT}(g,t) = \mathbb{E}_n\left[\left(\frac{G_g}{\mathbb{E}_n[G_g]} - \frac{\frac{\hat{p}_g(X)C}{1-\hat{p}_g(X)}}{\mathbb{E}_n\left[\frac{\hat{p}_g(X)C}{1-\hat{p}_g(X)}\right]}\right)(Y_t - Y_{g-1})\right],$$

where $\hat{p}_g(\cdot)$ is an estimate of $p_g(\cdot)$, and for a generic Z, $\mathbb{E}_n[Z] = n^{-1}\sum_{i=1}^n Z_i$. As noted in Theorem 1, $ATT(g,t)$ is nonparametrically identified for $2 \leq g \leq t \leq \mathcal{T}$.

With $\widehat{ATT}(g,t)$ in hand, one can use the analogy principle and combine these to estimate the summarized average treatment effect parameters discussed in Section 2.3.

In what follows, we consider the case in which one imposes a parametric restriction on p_g and estimates it by maximum likelihood. This is perhaps the most popular approach adopted by practitioners. Nonetheless, under some additional regularity conditions, our results can be extended to allow nonparametric estimators for the $p_g(\cdot)$; see, e.g., Abadie (2005), Chen (2007), Chen et al. (2008), Donald and Hsu (2014) and Sant'Anna (2016, 2017). Finally, we note that when propensity score misspecification is a concern, one can use the data-driven specification tests proposed by Sant'Anna and Song (2019).

Assumption 5. *For all $g = 2,\ldots,\mathcal{T}$, (i) there exists a known function $\Lambda : \mathbb{R} \to [0,1]$ such that $p_g(X) = P(G_g = 1|X, G_g + C = 1) = \Lambda(X'\pi_g^0)$; (ii) $\pi_g^0 \in \text{int}(\Pi)$, where Π is a compact subset of $\mathbb{R}^k$; (iii) the support of X, $\mathcal{X}$, is a subset of a compact set S, and $\mathbb{E}[XX'|G_g + C = 1]$ is positive definite; (iv) let $\mathcal{U} = \{x'\pi : x \in \mathcal{X}, \pi \in \Pi\}$; $\forall\ u \in \mathcal{U}$, $\exists \varepsilon > 0$ such that $\Lambda(u) \in [\varepsilon, 1-\varepsilon]$, $\Lambda(u)$ is strictly increasing and twice continuously differentiable with first derivatives bounded away from zero and infinity, and bounded second derivative; (vi) $\mathbb{E}\left[Y_t^2\right] < \infty$ for all $t = 1,\ldots,\mathcal{T}$.*

Assumption 5 is standard in the literature (see, e.g., Section 9.2.2 in Amemiya (1985), Example 5.40 in van der Vaart (1998), or Assumption 4.2 in Abadie (2005)), and it allows for Logit and Probit models.

Under Assumption 5, π_g^0 can be estimated by maximum likelihood:

$$\hat{\pi}_g = \arg\max_{\pi} \sum_{i:G_{ig}+C_i=1} G_{ig}\ln\left(\Lambda\left(X_i'\pi\right)\right) + (1-G_{ig})\ln\left(1-\Lambda\left(X_i'\pi\right)\right).$$

Let $\mathcal{W} = (Y_1,\ldots,Y_{\mathcal{T}},X,G_1,\ldots,G_{\mathcal{T}},C)'$, $\hat{p}_g(X_i) = \Lambda(X_i'\hat{\pi}_g)$, $\dot{p}_g = \partial p_g(u)/\partial u$, $\dot{p}_g(X) = \dot{p}_g\left(X'\pi_g^0\right)$. Under Assumption 5, $\hat{\pi}_g$ is asymptotically linear, that is,

$$\sqrt{n}\left(\hat{\pi}_g - \pi_g^0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_g^{\pi}(\mathcal{W}_i) + o_p(1),$$

where

$$\xi_g^{\pi}(\mathcal{W}) = \mathbb{E}\left[\frac{(G_g+C)\dot{p}_g(X)^2}{p_g(X)(1-p_g(X))}XX'\right]^{-1} X\frac{(G_g+C)(G_g-p_g(X))\dot{p}_g(X)}{p_g(X)(1-p_g(X))}, \tag{3.1}$$

see Lemma A.2 in the Supplementary Appendix.

3.1 Asymptotic Theory for Group-Time Average Treatment Effects

Denote the normalized weights by

$$w_g^G = \frac{G_g}{\mathbb{E}\left[G_g\right]}, \qquad w_g^C = \frac{p_g\left(X\right)C}{1-p_g\left(X\right)} \Big/ \mathbb{E}\left[\frac{p_g\left(X\right)C}{1-p_g\left(X\right)}\right], \tag{3.2}$$

and define

$$\psi_{gt}(\mathcal{W}_i) = \psi_{gt}^G(\mathcal{W}_i) - \psi_{gt}^C(\mathcal{W}_i), \tag{3.3}$$

where

$$\begin{aligned}
\psi_{gt}^G(\mathcal{W}) &= w_g^G \ \left[(Y_t - Y_{g-1}) - \mathbb{E}\left[w_g^G\left(Y_t - Y_{g-1}\right)\right]\right], \\
\psi_{gt}^C(\mathcal{W}) &= w_g^C \ \left[(Y_t - Y_{g-1}) - \mathbb{E}\left[w_g^C\left(Y_t - Y_{g-1}\right)\right]\right] + M_{gt}'\,\xi_g^{\pi}\left(\mathcal{W}\right),
\end{aligned}$$

and

$$M_{gt} = \frac{\mathbb{E}\left[X\left(\frac{C}{1-p_g\left(X\right)}\right)^2 \dot{p}_g\left(X\right)\left[\left(Y_{it} - Y_{ig-1}\right) - \mathbb{E}\left[w_g^C\left(Y_t - Y_{g-1}\right)\right]\right]\right]}{\mathbb{E}\left[\frac{p_g\left(X\right)C}{1-p_g\left(X\right)}\right]}$$

which is a $k \times 1$ vector, with k the dimension of X, and $\xi_g^{\pi}\left(\mathcal{W}\right)$ is as defined in (3.1).

Finally, let $ATT_{g\leq t}$ and $\widehat{ATT}_{g\leq t}$ denote the vector of $ATT(g,t)$ and $\widehat{ATT}(g,t)$, respectively, for all $g = 2,\ldots,\mathcal{T}$ and $t = 2,\ldots,\mathcal{T}$ with $g \leq t$. Analogously, let $\Psi_{g\leq t}$ denote the collection of ψ_{gt} across all periods t and groups g such that $g \leq t$.

The next theorem establishes the joint limiting distribution of $\widehat{ATT}_{g\leq t}$.

Theorem 2. *Under Assumptions 1-5, for* $2 \leq g \leq t \leq \mathcal{T}$,

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

Furthermore,

$$\sqrt{n}(\widehat{ATT}_{g\leq t} - ATT_{g\leq t}) \xrightarrow{d} N(0,\Sigma)$$

where $\Sigma = \mathbb{E}[\Psi_{g\leq t}(\mathcal{W})\Psi_{g\leq t}(\mathcal{W})']$.

Theorem 2 provides the influence function for estimating the vector of group-time average treatment effects $ATT_{g\leq t}$, as well as its limiting distribution. In order to conduct inference, one can show that the sample analogue of Σ is a consistent estimator for Σ (see, e.g., Theorem 4.4 in Abadie (2005)) 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 $ATT(g,t)$ 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 3 below.

To proceed, let $\widehat{\Psi}_{g\leq t}(\mathcal{W})$ denote the sample-analogue of $\Psi_{g\leq t}(\mathcal{W})$, where population expectations are replaced by their empirical analogue, and the true generalized propensity score, p_g, and its derivatives, $\dot{p}_g$, are replaced by their MLE estimates, $\hat{p}_g$ and $\widehat{\dot{p}}_g$, respectively. Let $\{V_i\}_{i=1}^n$ be a sequence of *iid* random variables with zero mean, unit variance and bounded support, independent of the original sample $\{\mathcal{W}_i\}_{i=1}^n$. A popular example involves *iid* Bernoulli variates $\{V_i\}$ with $P(V=1-\kappa)=\kappa/\sqrt{5}$ and $P(V=\kappa)=1-\kappa/\sqrt{5}$, where $\kappa=\left(\sqrt{5}+1\right)/2$, as suggested by Mammen (1993).

We define $\widehat{ATT}^*_{g\leq t}$, a bootstrap draw of $\widehat{ATT}_{g\leq t}$, via

$$\widehat{ATT}^*_{g\leq t}=\widehat{ATT}_{g\leq t}+\mathbb{E}_n\left[V\cdot\widehat{\Psi}_{g\leq t}(\mathcal{W})\right]. \tag{3.4}$$

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

Theorem 3. *Under Assumptions 1-5,*

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

where $\Sigma=\mathbb{E}[\Psi_{g\leq t}(\mathcal{W})\Psi_{g\leq t}(\mathcal{W})']$ *as in Theorem 2, and* $\underset{*}{\overset{d}{\to}}$ *denotes weak convergence (convergence in distribution) of the bootstrap law in probability, i.e., conditional on the original sample* $\{\mathcal{W}_i\}_{i=1}^n$*. Additionally, for any continuous functional* $\Gamma(\cdot)$

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

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

Algorithm 1. 1) *Draw a realization of* $\{V_i\}_{i=1}^n$. 2) *Compute* $\widehat{ATT}^*_{g\leq t}$ *as in (3.4), 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,* $\widehat{\Sigma}^{1/2}(g,t)=\left(q_{0.75}(g,t)-q_{0.25}(g,t)\right)/\left(z_{0.75}-z_{0.25}\right)$*, where* $q_p(g,t)$ *is the* p*th sample quantile of the* $\hat{R}^*(g,t)$ *in the* B *draws, and* z_p *is the* p*th quantile of the standard normal distribution.* 5) *For each bootstrap draw, compute t-test*$_{g\leq t}=\max_{(g,t)}\left|\hat{R}^*(g,t)\right|\widehat{\Sigma}(g,t)^{-1/2}$. 5) *Construct* $\widehat{c}_{1-\alpha}$ *as the empirical*

$(1-a)$*-quantile of the* B *bootstrap draws of* $t\text{-test}_{g\leq t}$. 6) *Construct the bootstrapped simultaneous confidence band for* $ATT(g,t)$, $g \leq t$, *as* $\widehat{C}(g,t) = [\widehat{ATT}(g,t) \pm \widehat{c}_{1-\alpha}\widehat{\Sigma}(g,t)^{-1/2}/\sqrt{n}]$.

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 3, for any* $0 < \alpha < 1$, *as* $n \to \infty$,

$$P\left(ATT(g,t) \in \widehat{C}(g,t) : g \leq t\right) \to 1-\alpha,$$

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

Remark 3. 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. As suggested in Abadie et al. (2017), if one takes the traditional view in the panel data case that the treatment and control groups are fixed, and one obtains random samples from these subpopulations, then clustering is not recommended. However, if one wants to allow for reassignment of the control and treatment groups across samples, then one should cluster.[11]

In the case that one wishes to account for clustering, we note that this is 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 need 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 (2016, 2018). If the number of clusters is "small," however, the application of the aforementioned bootstrap procedure is not warranted.[12]

Remark 4. 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).

3.2 Asymptotic Theory for Summary Parameters

Let θ generically represent one of the parameters from Section 2.3, including the ones indexed by some variable (for example, $\tilde{\theta}_S(g)$ or $\tilde{\theta}_{SD}(e,e')$). Notice that all of the parameters in Section 2.3 can be expressed as weighted averages of $ATT(g,t)$. Write this generically as

$$\theta = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} w_{gt} ATT(g,t)$$

[11] The formal results in Abadie et al. (2017) focus on the cross section case and rely on additional functional restrictions that we do not impose in this paper. Therefore, the aforementioned "recommendation" should be interpreted with care. Fully extending the results of Abadie et al. (2017) to the semiparametric panel data case is well beyond the scope of our paper.

[12] 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 (2018). Extending these proposals to our setup is beyond the scope of this paper though.

where w_{gt} are some potentially random weights. θ can be estimated by

$$\hat{\theta} = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \widehat{w}_{gt} \widehat{ATT}(g,t),$$

where $\widehat{w}_{gt}$ are estimators for w_{gt} such that for all $g, t = 2, \ldots, \mathcal{T}$,

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

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 2.3 satisfy this condition.

Let

$$l^{w}\left(\mathcal{W}_i\right) = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} w_{gt} \cdot \psi_{gt}(\mathcal{W}_i) + \xi_{gt}^{w}(\mathcal{W}_i) \cdot ATT(g,t),$$

where $\psi_{gt}(\mathcal{W})$ are as defined in (3.3).

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 Assumptions 1-5,*

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

Corollary 2 implies that one can construct standard errors and confidence intervals for summary treatment effect parameters based on a consistent estimator of $\mathbb{E}\left[l^{w}\left(\mathcal{W}\right)^2\right]$ or by using a bootstrap procedure like the one in Algorithm 1.

Remark 5. As discussed in Remark 3, 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.

4 Pre-testing the Conditional Parallel Trends Assumption

So far, we have discussed how one can nonparametrically identify and conduct asymptotically valid inference about causal treatment effect parameters using conditional DID models with multiple periods and variation in treatment timing. The credibility of our results crucially relies on the conditional parallel trends assumption stated in Assumption 2. This assumption is fundamentally untestable. However, when one imposes a stronger version of the conditional parallel trends assumption, that is, that Assumption 2 holds *for all* periods t, and not only for the periods $g \leq t$, one can assess the reliability of the parallel trends assumption. Relative to Assumption 2, the additional time periods are ones where $g > t$ which are pre-treatment time periods. In this section, we describe how one can construct such a test in our context. Interestingly, our proposed testing procedure exploits more information than simply testing

whether $ATT(g,t)$ are equal to zero for all $2 \leq t < g$, and therefore is able to detect a broader set of violations of the stronger conditional parallel trends condition.

Before proceeding, we state the "augmented" conditional parallel trends assumption that allows us to "pre-test" for the conditional parallel trends assumption stated in Assumption 2.

Assumption 6 (Augmented Conditional Parallel Trends)**.** *For all* $t = 2, \ldots, \mathcal{T}$*,* $g = 2, \ldots, \mathcal{T}$*,*

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

In order to understand how such an assumption leads to testable implications, note that, under Assumption 6, for $2 \leq t < g \leq \mathcal{T}$, $\mathbb{E}[Y_t(0)|X, G_g = 1]$ can be expressed as

$$\begin{aligned}\mathbb{E}[Y_t(0)|X, G_g = 1] &= \mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, C = 1] + \mathbb{E}[Y_{t-1}(0)|X, G_g = 1] \\ &= \mathbb{E}[Y_t - Y_{t-1}|X, C = 1] + \mathbb{E}[Y_{t-1}|X, G_g = 1], \end{aligned} \tag{4.1}$$

where the second equality follows since for individuals in group g when $g > t$, $Y_{t-1}(0)$ is observed since treatment has not occurred yet. Using exactly the same logic, $Y_t(0)$ is also the observed outcome for individuals in group g when $g > t$. Thus, the construction of our test is based on comparing $\mathbb{E}[Y_t(0)|X, G_g = 1]$ in (4.1) to $\mathbb{E}[Y_t|X, G_g = 1]$ for all periods such $2 \leq t < g$: under Assumption 6 these conditional expectations should be equal.

Formally, the null hypothesis we seek to test is

$$H_0 : \mathbb{E}[Y_t - Y_{t-1}|X, G_g = 1] - \mathbb{E}[Y_t - Y_{t-1}|X, C = 1] = 0 \text{ } a.s. \text{ for all } 2 \leq t < g \leq \mathcal{T}. \tag{4.2}$$

One option to assess H_0 is to nonparametrically estimate each conditional expectation in (4.2), and compare how close their difference is to zero. Such a procedure would deviate from the estimation and inference procedures described in Section 3 and would involve choosing smoothing parameters such as bandwidths, assuming additional smoothness conditions of these expectations, potentially ruling out discrete covariates X, and would also suffer from the "curse of dimensionality" when the dimension of X is moderate.

Alternatively, one can test an *implication* of H_0 by using the results of Theorem 1, and compare how close to zero are the estimates of $ATT(g,t)$ for all $2 \leq t < g \leq \mathcal{T}$. Although intuitive, such a procedure does not exploit all the restrictions imposed by H_0. For instance, deviations from H_0 in opposite directions for different values of X could offset each other, implying that one may fail to reject the plausibility of the conditional parallel trends assumption, even when H_0 is violated in some directions. See Remark 6 at the end of this section for more details about this case.

We adopt an alternative approach that avoids all the aforementioned drawbacks: it is compatible with the framework adopted in Section 3, it does not involve choosing bandwidths, does not impose additional smoothness conditions, does not suffer from the "curse of dimensionality," and exploits all the testable restrictions implied by the augmented conditional parallel trends assumption. Our proposal builds on the integrated conditional moments (ICM) approach commonly used in the goodness-of-fit literature; see, e.g., Bierens (1982), Bierens and Ploberger (1997), Stute (1997), Stinchcombe and White (1998), and Escanciano (2006*a*,*b*, 2008). To the best of our knowledge, we are the first to propose to use ICM to assess

the plausibility of the parallel trends assumption, even when there is no treatment timing variation.

Let w_g^G and w_g^C be defined as in (3.2). After some algebra, under Assumptions 1-5, we can rewrite H_0 as

$$H_0 : \mathbb{E}\left[\left(w_g^G - w_g^C\right)\left(Y_t - Y_{t-1}\right)|X\right] = 0 \ a.s. \text{ for all } 2 \leq t < g \leq \mathcal{T}, \tag{4.3}$$

see Lemma A.4 in the Appendix. In fact, by exploiting Lemma 1 in Escanciano (2006*b*), we can further characterize (4.3) as

$$H_0 : \mathbb{E}\left[\left(w_g^G - w_g^C\right)\gamma(X,u)\left(Y_t - Y_{t-1}\right)\right] = 0 \ \forall u \in \Xi \text{ for all } 2 \leq t < g \leq \mathcal{T}, \tag{4.4}$$

where Ξ is a properly chosen space, and the parametric family $\{\gamma(\cdot,u) : u \in \Xi\}$ is a family of weighting functions such that the equivalence between (4.3) and (4.4) holds. The most popular weighting functions include $\gamma(X,u) = \exp(iX'u)$ as in Bierens (1982) and $\gamma(X,u) = 1\{X \leq u\}$ as in Stute (1997). In the following, to ease the notation, we concentrate our attention on the indicator functions, $\gamma(X,u) = 1\{X \leq u\}$, with $\Xi = \mathcal{X}$, the support of the covariates X.

The advantage of the representation in (4.4) is that it resembles the expression for $ATT(g,t)$ in (2.2), and therefore we can use a similar estimation procedure that avoids the use of smoothing parameters. To see this, let

$$J(u,g,t,p_g) = \mathbb{E}\left[\left(w_g^G - w_g^C\right)1(X \leq u)\left(Y_t - Y_{t-1}\right)\right],$$

and, for each u in the support of X, we can estimate $J(u,g,t,p_g)$ by

$$\widehat{J}(u,g,t,\hat{p}_g) = \mathbb{E}_n\left[\left(\frac{G_g}{\mathbb{E}_n[G_g]} - \frac{\frac{\hat{p}_g(X)C}{1-\hat{p}_g(X)}}{\mathbb{E}_n\left[\frac{\hat{p}_g(X)C}{1-\hat{p}_g(X)}\right]}\right)1(X \leq u)\left(Y_t - Y_{t-1}\right)\right],$$

where $\hat{p}_g$ is a first-step estimator of p_g.

With $\widehat{J}(u,g,t,\hat{p}_g)$ in hand, one should reject H_0 when it is not "too close" to zero across different values of u, g, and t, $2 \leq t < g \leq \mathcal{T}$. In order to evaluate the distance from $\widehat{J}(u,g,t,\hat{p}_g)$ to zero, we consider the Cramér-von Mises norm,

$$CvM_n = \int_{\mathcal{X}} \left|\sqrt{n}\widehat{J}_{g>t}(u)\right|_M^2 F_{n,X}(du)$$

where $J_{g>t}(u)$ and $\widehat{J}_{g>t}(u)$ denote the vector of $J(u,g,t,p_g)$ and $\hat{J}(u,g,t,\hat{p}_g)$, respectively, for all $g = 2,\ldots,\mathcal{T}$ and $t = 2,\ldots,\mathcal{T}$, such that $2 \leq t < g \leq \mathcal{T}$, $|A|_M$ denotes the weighted seminorm $\sqrt{A'MA}$ for a positive semidefinite matrix M and a real vector A, and $F_{n,X}$ is the empirical CDF of X. To simplify exposition and leverage intuition, we fix M to be a $(\mathcal{T}-1)^2 \times (\mathcal{T}-1)^2$ diagonal matrix such that its (g,t)-th diagonal element is given by $1\{g > t\}$. As a result, we can write CvM_n as

$$CvM_n = \sum_{g=2}^{\mathcal{T}}\sum_{t=2}^{\mathcal{T}} 1\{g > t\}\int_{\mathcal{X}}\left|\sqrt{n}\widehat{J}(u,g,t,\hat{p}_g)\right|^2 F_{n,X}(du). \tag{4.5}$$

This choice of test statistic is similar to the one used by Escanciano (2008) in a different context. However, one can choose some other M or other norms as well.

The key step to derive the asymptotic properties of CvM_n is to study the process $\sqrt{n}\widehat{J}(u,g,t,\hat{p}_g)$. Here, note that in contrast to $\widehat{ATT}(g,t)$, $\widehat{J}(u,g,t,p_g)$ is infinite dimensional (since it involves a continuum of u), and therefore we need to use uniform (instead of pointwise) arguments. Furthermore, we must account for the uncertainty inherited by using the estimated generalized propensity scores $\hat{p}_g$ instead of the unknown true p_g. To accomplish this, we build on the existing literature on empirical processes with a first step estimation of the propensity score; see, e.g., Donald and Hsu (2014) and Sant'Anna (2017) for applications in the causal inference context. As before, we focus on the case where the p_g is estimated parametrically.

Define

$$\psi_{ugt}^{test}(\mathcal{W}_i) = \psi_{ugt}^{G,test}(\mathcal{W}_i) - \psi_{ugt}^{C,test}(\mathcal{W}_i), \tag{4.6}$$

where

$$\begin{aligned}
\psi_{ugt}^{G,test}(\mathcal{W}) &= w_g^G \left[(Y_t - Y_{t-1})\,1(X \leq u) - \mathbb{E}\left[w_g^G 1(X \leq u)\,(Y_t - Y_{t-1})\right]\right], \\
\psi_{ugt}^{C,test}(\mathcal{W}) &= w_g^C \left[(Y_t - Y_{t-1})\,1(X \leq u) - \mathbb{E}\left[w_g^C 1(X \leq u)\,(Y_t - Y_{t-1})\right]\right] + M_{ugt}^{test\,\prime}\,\xi_g^{\pi}(\mathcal{W}),
\end{aligned}$$

with $\xi_g^{\pi}(\mathcal{W})$ as defined in (3.1), and

$$M_{ugt}^{test} = \frac{\mathbb{E}\left[X\left(\frac{C}{1-p_g(X)}\right)^2 \dot{p}_g(X)\left[1(X \leq u)\,(Y_t - Y_{t-1}) - \mathbb{E}\left[w_g^C 1(X \leq u)\,(Y_t - Y_{t-1})\right]\right]\right]}{\mathbb{E}\left[\frac{p_g(X)\,C}{1-p_g(X)}\right]}.$$

Let $\Psi_{g>t}^{test}(\mathcal{W}_i;u)$ denote the vector of $\psi_{ugt}^{test}(\mathcal{W}_i)$ across all periods t and groups g such that $2 \leq t < g \leq \mathcal{T}$.

The next theorem establishes the weak convergence of the process $\sqrt{n}\widehat{J}_{g>t}(u)$ under H_0, characterizes the limiting null distribution of CvM_n, and shows that our proposed test is consistent. From these results, we can conclude that our proposed test controls size, and if Assumption 6 does not hold, our test procedure rejects H_0 with probability approaching one as n goes to infinity. Hence, our tests can indeed be used to assess the reliability of our main identification assumption.

Theorem 4. *Suppose Assumptions 1-5 hold. Then,*

1. *If Assumption 6 holds, i.e., under the null hypothesis (4.4), as $n \to \infty$,*

$$\sqrt{n}\widehat{J}_{g>t}(u) \Rightarrow \mathbb{G}(u) \text{ in } l^{\infty}(\mathcal{X}),$$

where $\Rightarrow$ denote weak convergence in the sense of J. Hoffmann-Jϕrgensen (see, e.g., Definition 1.3.3 in van der Vaart and Wellner (1996)), $\mathcal{X}$ is the support of X, and $\mathbb{G}$ is a zero-mean Gaussian process with covariance function

$$V(u_1,u_2) = \mathbb{E}[\Psi_{g>t}^{test}(\mathcal{W};u_1)\,\Psi_{g>t}^{test}(\mathcal{W};u_2)'].$$

In particular, as $n \to \infty$.

$$CvM_n \xrightarrow{d} \int_{\mathcal{X}} |\mathbb{G}(u)|_M^2 \, F_X(du)$$

2. If Assumption 6 does not hold, i.e., under the negation of the null hypothesis (4.4)

$$\lim_{n\to\infty} P\left(CvM_n > c_\alpha^{CvM}\right) = 1,$$

where $c_\alpha^{CvM} = \inf\{c \in [0,\infty) : \lim_{n\to\infty} P(CvM_n > c) = \alpha\}$.

From Theorem 4, we see that the asymptotic distribution of CvM_n depends on the underlying data generating process (DGP) and standardization is complicated. To overcome this problem, we propose to compute critical values with the assistance of the multiplier bootstrap akin to the one discussed in Theorem 3.

To proceed, let $\widehat{\Psi}_{g>t}^{test}(\cdot;u)$ denote the sample-analogue of $\Psi_{g>t}^{test}(\cdot;u)$, where population expectations are replaced by their empirical analogues, and the true generalized propensity score, p_g, and its derivatives, $\dot{p}_g$, are replaced by their MLE estimates, $\hat{p}_g$ and $\widehat{\dot{p}}_g$, respectively. Let

$$\widehat{J}_{g>t}^{*}(u) = \mathbb{E}_n\left[V \cdot \widehat{\Psi}_{g>t}^{test}(\mathcal{W};u)\right], \tag{4.7}$$

where $\{V_i\}_{i=1}^n$ is defined as in Section 3. The next algorithm provides a step-by-step procedure to approximate c_α, the critical value of our test CvM_n.

Algorithm 2. 1) *Draw a realization of* $\{V_i\}_{i=1}^n$. 2) *For each* $u \in \mathcal{X}$, *compute* $\widehat{J}_{g>t}^{*}(u)$ *as in (4.7).* 3) *Compute* $CvM_n^* = \int_{\mathcal{X}} \left|\sqrt{n}\widehat{J}_{g>t}^{*}(u)\right|_M^2 F_{n,X}(du)$. 4) *Repeat steps 1-3 B times.* 5) *Construct* $\widehat{c}_{1-\alpha}^{CvM}$ *as the empirical* $(1-a)$*-quantile of the B bootstrap draws of* CvM_n^*.

The next theorem establishes the asymptotic validity of the multiplier bootstrap described in Algorithm 2.

Theorem 5. *Suppose Assumptions 1-5 hold. Then, under the null hypothesis (4.4) and under fixed alternatives (i.e., the negation of (4.4)),*

$$\sqrt{n}\widehat{J}_{g>t}^{*}(u) \underset{*}{\Rightarrow} \mathbb{G}(u) \text{ in } l^\infty(\mathcal{X}),$$

where $\mathbb{G}(u)$ *in* $l^\infty(\mathcal{X})$ *is the same Gaussian process of Theorem 4 and* $\underset{*}{\Rightarrow}$ *indicates weak convergence in probability under the bootstrap law, see Giné and Zinn (1990). In particular,*

$$CvM_n^* \underset{*}{\xrightarrow{d}} \int_{\mathcal{X}} |\mathbb{G}(u)|_M^2 F_X(du).$$

Remark 6. As described above, our proposed test CvM_n fully exploits the null hypothesis (4.4), and can detect a broad set of violations against the conditional parallel trends assumption. However, sometimes researchers are also interested in visualizing deviations from the conditional parallel trends assumption, but our proposed Cramér-von Mises test does not directly provide that. In such cases, we note that one can test an implication of the augmented conditional parallel trends assumption, at the cost of losing power against some directions. Namely, under the augmented conditional parallel trends assumptions, $ATT(g,t)$ should be equal to 0 in periods before individuals become treated, that is, when $g > t$. This test is simple to implement in practice though it is distinct from the tests commonly employed in DID with multiple periods and multiple groups (see, e.g., Autor et al. (2007) and Angrist and Pischke

(2009)) which we briefly discuss in Appendix D in the Supplementary Appendix. In fact, as formally shown by Abraham and Sun (2018), traditional regression-based tests for (unconditional) pre-trends may be unreliable in settings with treatment effect heterogeneity. Our proposal does not suffer from this drawback.

Let $ATT_{g>t}$ denote the "ATT" in periods before an individual in group g is treated (and also satisfying $2 \leq g$). Using exactly the same arguments as in Section 3, one can establish the limiting distribution of an estimator of $ATT_{g>t}$ (we omit the details for brevity). And one can implement a test of the augmented parallel trends assumption using a Wald-type test. We also found it helpful in the application to obtain the joint limiting distribution of estimators of $ATT_{g\leq t}$ and $ATT_{g>t}$ (once again using the same arguments as in Section 3) and then reporting uniform confidence bands that cover both pre-tests and estimates of $ATT(g,t)$ across all $g = 2,\ldots,\mathcal{T}$ and $t = 2,\ldots,\mathcal{T}$. From these uniform confidence bands, one can immediately infer whether or not the implication of the augmented parallel trends assumption is violated.

5 The Effect of Minimum Wage Policy on Teen Employment

In this section, we illustrate the empirical relevance of our proposed methods by studying the effect of the minimum wage on teen employment.

From 1999-2007, the federal minimum wage was flat at \$5.15 per hour. In July 2007, the federal minimum wage was raised from \$5.15 to \$5.85. 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. Others did not – 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 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 as well as minimum wage levels by state 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 county characteristics come from the 2000 County Data Book. These include county population in 2000, the fraction of the population that is black, educational characteristics from 1990, median income in 1997, and the fraction of the population below the poverty level in 1997.

For forty-one states, the federal minimum wage was binding in quarter 2 of 1999. We omit two states that raised their minimum wage between then and the first quarter of 2004. We drop several other states for lack of data. We also drop states in the Northern census region because all but two of them had minimum wages higher than the federal minimum wage at the beginning of the period and census region is an important control in the minimum wage literature. We use quarterly employment in the first quarter of each year from 2001 to 2007 for employment among teenagers. Alternatively, we could use more periods of data, but this would come at the cost of losing several states due to lack of data. Also, we choose first quarter employment because it is further away from the federal minimum wage increase

in Q3 of 2007. Our final sample includes county level teen employment for 29 states matched with county characteristics.

Our strategy is to divide the observations based on the timing of when a state increased its minimum wage above the federal minimum wage. States that did not raise their minimum wage during this period form the untreated group. We also have groups of states that increased their minimum wage during 2004, 2006, and 2007.[13] Before 2004, Illinois did not have a state minimum wage. In Q1 of 2004, Illinois set a state minimum wage of $5.50 which was 35 cents higher than the federal minimum wage. In Q1 of 2005, Illinois increased its minimum wage to $6.50 where it stayed for the remainder of the period that we consider. No other states changed their minimum wage policy by the first quarter of 2005. In the second quarter of 2005, Florida and Wisconsin set a state minimum wage above the federal minimum wage. In Q3 of 2005, Minnesota also set a state minimum wage. Florida and Wisconsin each gradually increased their minimum wages over time, while Minnesota's was flat over the rest of the period. These three states constitute the treated group for 2006. West Virginia increased its minimum wage in Q3 of 2006; Michigan and Nevada increased their minimum wages in Q4 of 2006; Colorado, Maryland, Missouri, Montana, North Carolina, and Ohio increased their state minimum wages in Q1 of 2007. These states form the 2007 treated group. Among these there is some heterogeneity in the size of the minimum wage increase. For example, North Carolina only increased its minimum wage to $6.15 though each state increased its minimum wage to strictly more than the new federal minimum wage of $5.85 per hour in Q3 of 2007. At the other extreme, Michigan increased its minimum wage to $6.95 and then to $7.15 by Q2 of 2007.[14]

Summary statistics for county characteristics are provided in Table 1. As discussed above, treated counties are much less likely to be in the South. They also have much lower population (on average 53,000 compared to 94,000 for treated counties). The proportion of black residents is much higher in treated counties (on average, 10% compared to 6% 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 smaller fraction of high school graduates and a somewhat higher poverty rate.

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 for observed characteristics X.

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 Figure 2 along with a uniform 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 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 increases on

[13] To be precise, we use only employment data from the first quarter of each year. A state is considered to raise its minimum wage in year y if it raised its minimum wage in Q2, Q3, or Q4 of year $y-1$ or in Q1 of year y.

[14] See Appendix E in the Supplementary Appendix for additional details about the adoption time and the spatial distribution of the state-level minimum wage policy changes in our sample.

Table 1: Summary Statistics for Main Dataset

	Treated States	Untreated States	Diff	P-val on Difference
Midwest	0.59	0.34	0.259	0.00
South	0.27	0.59	-0.326	0.00
West	0.14	0.07	0.067	0.00
Black	0.06	0.10	-0.042	0.00
HS Graduates	0.59	0.55	0.327	0.00
Population (1000s)	94.32	53.43	40.896	0.00
Poverty Rate	0.13	0.16	-0.259	0.00
Median Income (1000s)	33.91	31.89	2.024	0.00

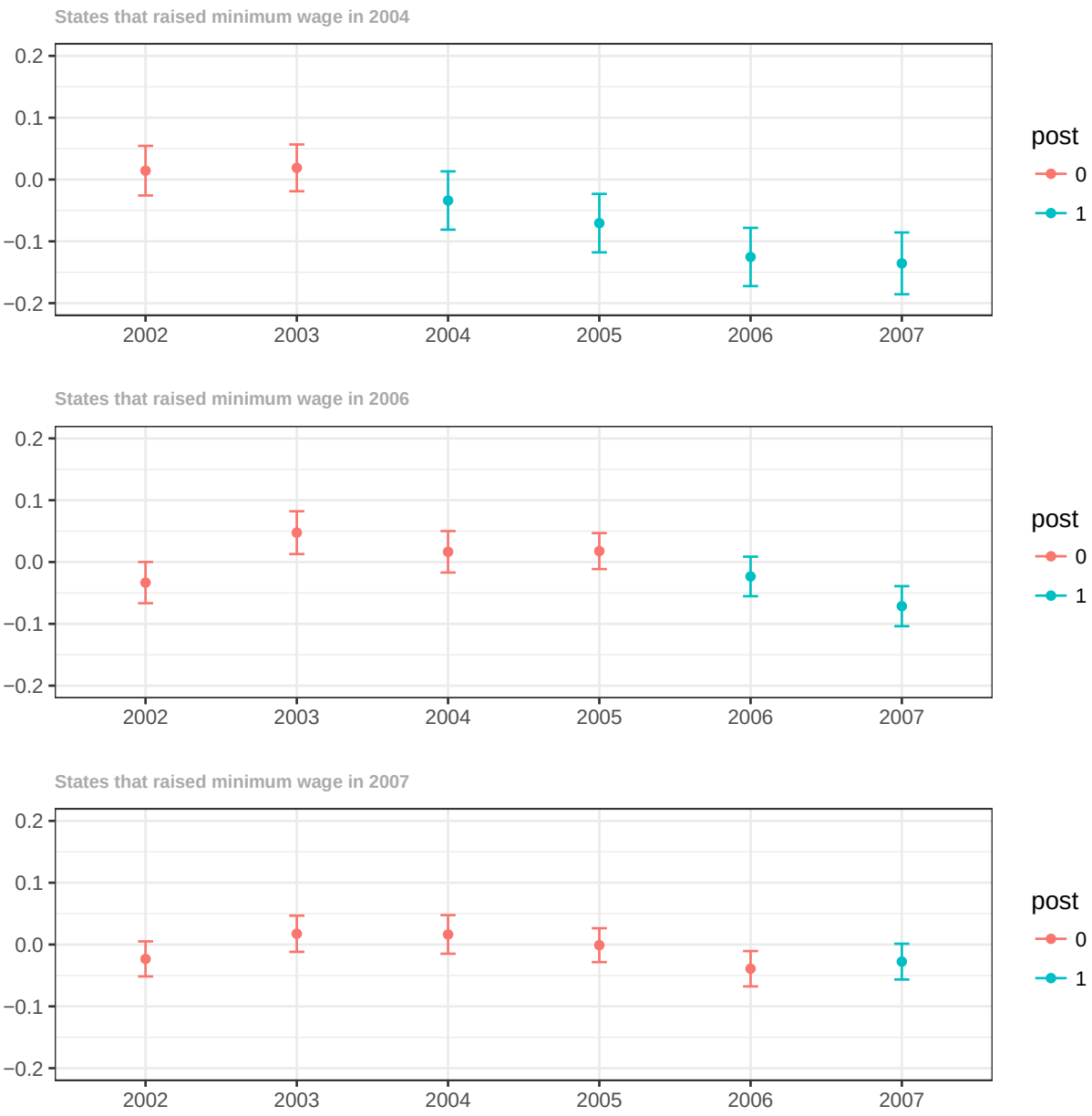
Notes: Summary statistics for counties located in states that raised their minimum wage between Q2 of 2003 and Q1 of 2007 (treated) and states whose minimum wage was effectively set at the federal minimum wage for the entire period (untreated). The sample consists of 2284 counties. *Sources:* Quarterly Workforce Indicators and 2000 County Data Book

the minimum wage lead to a reduction in teen employment. For 4 out of 7 group-time average treatment effects, there is a clear statistically significant negative effect on employment. The other three 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 (see Table 2). 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 3.4% lower on average in 2004 than it would have been if the minimum wage had not been increased. In 2005, teen employment 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.

Table 2 reports aggregated treatment effect measures. Allowing for dynamic treatment effects is perhaps the most useful for our study. 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 getting stronger 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 more than two years. These results are robust to keeping the treated group constant to make sure that selective treatment timing does not bias the results (see the row in Table 2 labeled 'Selectivity and Dynamics'). When we restrict the sample to only include groups with at least two years of exposure to treatment

Figure 2: Minimum Wage Results under Unconditional Parallel Trends



Notes: The effect of the minimum wage on teen employment estimated under the unconditional parallel trends assumption. Red lines give point estimates and uniform 95% confidence bands for pre-treatment periods allowing for clustering at the county level. Under the null hypothesis of the unconditional parallel trends assumption holding in all periods, these should be equal to 0. Blue lines provide point estimates and uniform 95% confidence bands for the treatment effect of increasing the minimum wage allowing for clustering at the county level. The top panel includes states that increased their minimum wage in 2004, the middle panel includes states that increased their minimum wage in 2006, and the bottom panel includes states that increased their minimum wage in 2007. No states raised their minimum wages in other years prior to 2007.

(and only considering the first two periods of exposure which keeps the groups constant across length of exposure), we estimate that the effect of minimum wage increases in the first period of exposure is 2.7% lower teen employment and 7.1% lower teen employment in the second period.[15]

Table 2: Aggregate Treatment Effect Parameters under Unconditional Parallel Trends

	Partially Aggregated				**Single Parameters**
Standard DID					-0.037
					(0.006)
Simple Weighted Average					-0.052
					(0.006)
Selective Treatment Timing	g=2004	g=2006	g=2007		
	-0.091	-0.047	-0.028		-0.039
	(0.019)	(0.008)	(0.007)		(0.007)
Dynamic Treatment Effects	e=1	e=2	e=3	e=4	
	-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)
Selectivity and Dynamics	e=1	e=2			
	-0.027	-0.071			-0.049
	(0.009)	(0.009)			(0.008)

Notes: The table reports aggregated treatment effect parameters under the unconditional parallel trends assumption and with clustering at the county level. The row 'Standard DID' reports the coefficient on a post-treatment dummy variable from a two-way fixed effects regression. The row 'Single Weighted Average' reports the weighted average (by group size) of all available group-time average treatment effects as in Equation (2.5). The row 'Selective Treatment Timing' allows for period that a county is first treated to affect its group-time average treatment effect; here, g indexes the year that a county is first treated. The row 'Dynamic Treatment Effects' allows for the effect of the minimum wage to depend on length of exposure; here, e indexes the length of exposure to the treatment. The row 'Calendar Time Effects' allows the effect of the minimum wage to change across years; here, t indexes the year. The row 'Selectivity and Dynamics' allows for the effect of the minimum wage to depend on length of exposure while making sure that the composition of the treatment group does not change with e; 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.

Allowing for calendar time effects or selective treatment timing also is consistent with the idea that states that increased their minimum wage experienced negative effects on teen employment relative to what they would have experienced if they had not increased their minimum wage.

We consider testing the unconditional parallel trends assumption. First, since the confidence bands in Figure 2 are uniform, one can immediately infer that the unconditional parallel trends assumption should be rejected based on the implication of the unconditional parallel trends assumption that the "ATT" in periods before treatment should be equal to 0. Likewise, our proposed test also rejects the unconditional parallel trends assumption (p-value: 0.000). The estimated uniform confidence bands in Figure 2 also provide some insight into how to think about our pre-tests. For the group first treated in 2004, the parallel trends assumption is not rejected in any period. For the group first treated in 2006,

[15]Notice that these estimates are exactly the same as in the first two periods for the dynamic treatment effect estimates that do not condition on the group remaining constant. The reason that they are the same for the first period is coincidental; 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 in the latter case. For the second period, they correspond by construction.

it is rejected in 2003; for the group first treated in 2007, it is rejected in 2006. Interestingly, with the exception of 2006 for the group first treated in 2007, in each of the cases where it is rejected, the placebo estimates are positive.

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. Estimation requires a first step estimation of the generalized propensity score. For each generalized propensity score, we estimate a logit model that includes each county characteristic along with quadratic terms for population and median income.[16] In particular, the conditional results allow for differential trends in teen employment across different regions as well as in the other county characteristics mentioned above. In what follows, all inference procedures use clustered bootstrapped standard errors at the county level.

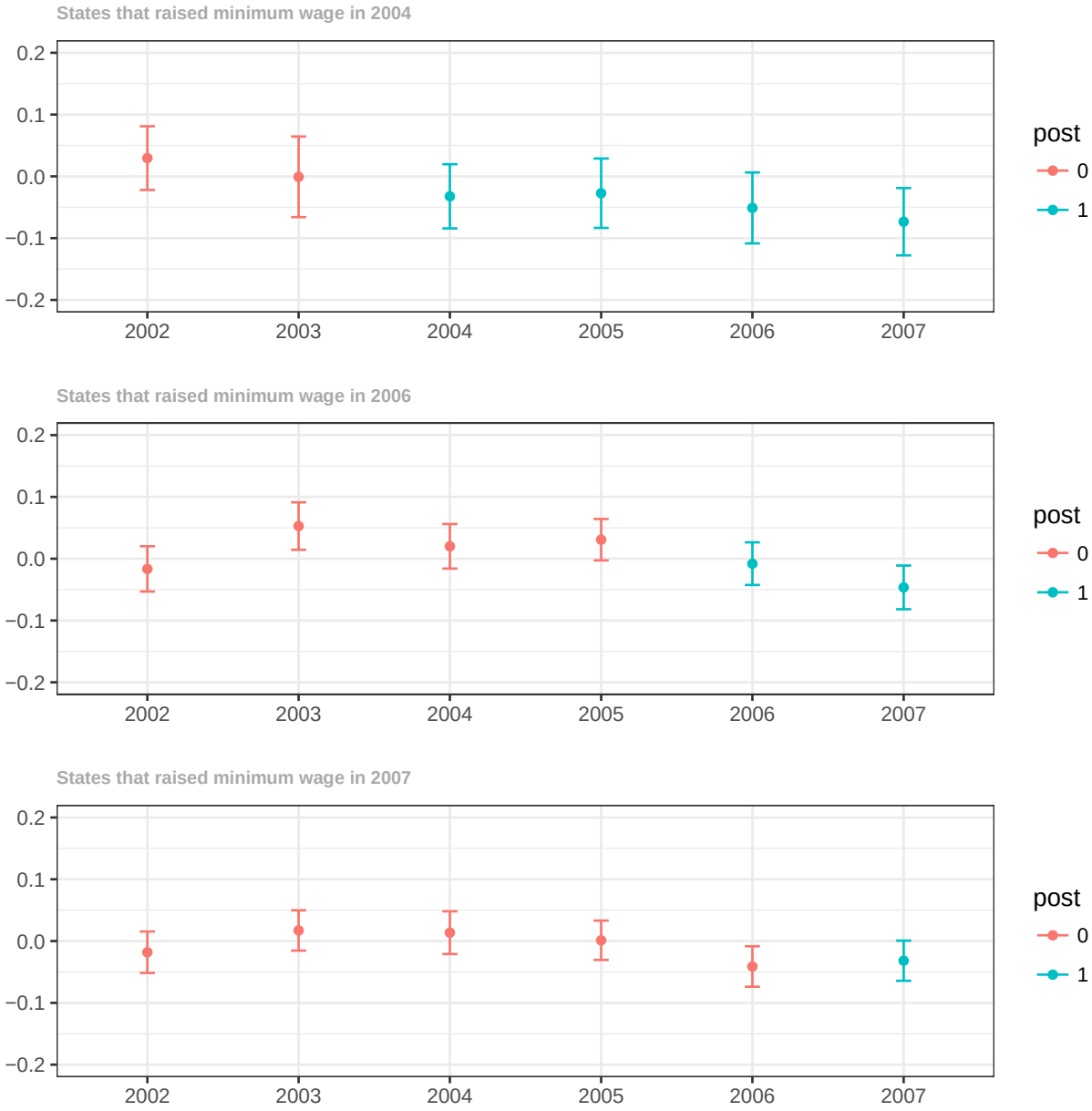
For comparison's sake, we first estimate the coefficient on a post-treatment dummy variable in a model with individual fixed effects and region-year fixed effects. This is very similar to one of the sorts of models that Dube et al. (2010) finds to eliminate the correlation between the minimum wage and 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 (2005), Chernozhukov et al. (2013), de Chaisemartin and D'Haultfœuille (2018), Borusyak and Jaravel (2017), Goodman-Bacon (2018) and Słoczyński (2018), the same may not be true for two-way fixed effect regressions in the presence of treatment effect heterogeneity.[17]

The results using our approach are available in Figure 3 and Table 3. Interestingly, we find quite different results using our approach than are suggested by the two-way fixed effect 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 0.8% lower teen employment (not statistically different from 0) in 2006 for the group of states first treated in 2006 to 7.3% lower teen employment in 2007 for states first treated in 2004. Now only 2 of 7 group-time average treatment effects are statistically significant. The pattern of dynamic treatment effects where the effect of minimum wage increases tends to increase with length of exposure is the same as in the unconditional case. Similarly, using our aggregated treatment effect parameters, allowing for dynamic treatment effects, we estimate that increasing the minimum wage led on average to 4.8% lower teen employment. Allowing for dynamic

[16] 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.

[17] 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) considers 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.

Figure 3: Minimum Wage Results under Conditional Parallel Trends



Notes: The effect of the minimum wage on teen employment estimated under the conditional parallel trends assumption. Red lines give point estimates and uniform 95% confidence bands for pre-treatment periods allowing for clustering at the county level. Under the null hypothesis of the conditional parallel trends assumption holding in all periods, these should be equal to 0. Blue lines provide point estimates and uniform 95% confidence bands for the treatment effect of increasing the minimum wage allowing for clustering at the county level. The top panel includes states that increased their minimum wage in 2004, the middle panel includes states that increased their minimum wage in 2006, and the bottom panel includes states that increased their minimum wage in 2007. No states raised their minimum wages in other years prior to 2007.

treatment effects and selective treatment timing, we estimate that increasing the minimum wage lowers teen employment by 2.8%.

Table 3: Aggregate Treatment Effect Parameters under Conditional Parallel Trends

	Partially Aggregated				**Single Parameters**
Standard DID					-0.008
					(0.006)
Simple Weighted Average					-0.034
					(0.008)
Selective Treatment Timing	g=2004	g=2006	g=2007		
	-0.046	-0.027	-0.032		-0.032
	(0.020)	(0.008)	(0.008)		(0.007)
Dynamic Treatment Effects	e=1	e=2	e=3	e=4	
	-0.026	-0.041	-0.051	-0.073	-0.048
	(0.006)	(0.010)	(0.025)	(0.024)	(0.014)
Calendar Time Effects	t=2004	t=2005	t=2006	t=2007	
	-0.032	-0.027	-0.021	-0.040	-0.030
	(0.019)	(0.024)	(0.011)	(0.007)	(0.013)
Selectivity and Dynamics	e=1	e=2			
	-0.016	-0.041			-0.028
	(0.009)	(0.010)			(0.008)

Notes: The table reports aggregated treatment effect parameters under the conditional parallel assumption and with clustering at the county level. The row 'Standard DID' reports the coefficient on a post-treatment dummy variable from a fixed effects regression with individual fixed effects and region-year fixed effects. The row 'Single Weighted Average' reports the weighted average (by group size) of all available group-time average treatment effects as in Equation (2.5). The row 'Selective Treatment Timing' allows for period that a county is first treated to affect its group-time average treatment effect; here, g indexes the year that a county is first treated. The row 'Dynamic Treatment Effects' allows for the effect of the minimum wage to depend on length of exposure; here, e indexes the length of exposure to the treatment. The row 'Calendar Time Effects' allows the effect of the minimum wage to change across years; here, t indexes the year. The row 'Selectivity and Dynamics' allows for the effect of the minimum wage to depend on length of exposure while making sure that the composition of the treatment group does not change with e; 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 evidence of the negative effect of minimum wage increases is somewhat mitigated by the fact that we reject the conditional parallel trends assumption in pre-treatment periods. This is immediately evident from Figure 3 because we can reject that the "ATT" is equal to zero in 2 out of 11 pre-treatment periods. Using the consistent Cramér-von Mises tests discussed in Section 4, we also reject the conditional parallel trends assumption (p-value: 0.000). In addition, we conducted our test of the augmented conditional parallel trends assumption separately for states first treated in 2004 because the pre-treatment "ATT" is not statistically significant in any period for this group. Here, we reject the augmented conditional parallel trends assumption. This is an interesting result because the "visual" test often conducted in empirical work would incorrectly lead the researcher to believe that the conditional parallel trends assumption is valid for states first treated in 2004.

Overall, our results suggests that the minimum wage decreased teen employment in states that increased their minimum wage relative to what it would have been had those states not increased their minimum wage. Nonetheless, our proposed tests indicate that the parallel trends assumption should be rejected in pre-treatment periods, implying that the DID research design may lead to non-reliable conclusions. Perhaps not surprisingly, given the amount of disagreement in the minimum wage literature, our results should be interpreted with care and are ultimately inconclusive.

6 Conclusion

This paper has considered Difference-in-Differences methods in the case where there are more than two periods and individuals can become treated at different points in time – a commonly encountered setup in empirical work in economics. In this setup, we have suggested computing group-time average treatment effects, $ATT(g,t)$, that are the average treatment effect in period t for the group of individuals first treated in period g. Unlike the more common approach of running a regression with a post-treatment dummy variable, $ATT(g,t)$ corresponds to a well defined treatment effect parameter. And once $ATT(g,t)$ has been obtained for different values of g and t, they can be aggregated into a single parameter, though the exact implementation depends on the particular case. We view such a flexibility as a plus of our proposed methodology.

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 estimators, and proved the validity of a powerful, but easy to implement, multiplier bootstrap procedure to construct simultaneous confidence bands for $ATT(g,t)$. Importantly, we have also proposed a new pre-test for the reliability of the conditional parallel trends assumption.

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 and found strikingly different results from the more common approach of interpreting the coefficient on a post-treatment dummy variable as the effect of the minimum wage on employment. However, using the pre-tests developed in the current paper, we found evidence against both the unconditional and conditional parallel trends assumption.

Our results can be extended to other situations of practical interest. For instance, one can combine our proposal with Callaway and Li (2018) in order to consider group-time quantile treatment effects. In light of our empirical application, we note that it is worth considering DID procedures that relax the conditional parallel trends assumption. A possibility in this direction is to use conditional moment inequalities. More precisely, one could assume that, for all $t = 2, \ldots, \mathcal{T}$, $g = 2, \ldots, \mathcal{T}$, such that $g \leq t$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \,|X, G_g = 1\right] \geq \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \,|X, C = 1\right] \ a.s.. \tag{6.1}$$

Note that (6.1) implies that

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

Thus, this one-sided relaxation of the conditional parallel trends assumption suggests that, under (6.1), $\widehat{ATT}(g,t)$ would be an estimator for the upper bound of the $ATT(g,t)$. By combining our pre-test procedure with Andrews and Shi (2013), one would then be able to assess the reliability of (6.1). These extensions are beyond the scope of this article and are left for future research.

References

Abadie, A. (2005), "Semiparametric difference-in-difference estimators," *Review of Economic Studies*, 72, 1–19.

Abadie, A., Athey, S., Imbens, G., and Wooldridge, J. (2017), "When should you adjust standard errors for clustering?," *Working Paper*.

Abraham, S., and Sun, L. (2018), "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Working Paper*.

Amemiya, T. (1985), *Advanced Econometrics*, Cambridge: Harvard University Press.

Andrews, D. W. K., and Shi, X. (2013), "Inference based on conditional moment inequalities," *Econometrica*, 81(2), 609–666.

Angrist, J. D., and Pischke, J.-S. (2009), *Mostly Harmless Econometrics: An Empiricist ' s Companion*, Princeton, NJ: Princeton University Press.

Athey, S., and Imbens, G. W. (2006), "Identification and inference in nonlinear difference in differences models," *Econometrica*, 74(2), 431–497.

Athey, S., and Imbens, G. W. (2018), "Design-based analysis in difference-in-differences settings with staggered adoption," *Working Paper*.

Autor, D. H., Kerr, W. R., and Kugler, A. D. (2007), "Do employment protections reduce productivity? Evidence from U.S. states," *The Economic Journal*, 117, 189–217.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017), "Program evaluation and causal inference With high-dimensional data," *Econometrica*, 85(1), 233–298.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004), "How much should we trust differences-in-differences estimates?," *The Quarterly Journal of Economics*, 119(1), 249–275.

Bierens, H. J. (1982), "Consistent model specification tests," *Journal of Econometrics*, 20(1982), 105–134.

Bierens, H. J., and Ploberger, W. (1997), "Asymptotic theory of integrated conditional moment tests," *Econometrica*, 65(5), 1129–1151.

Blundell, R., Dias, M. C., Meghir, C., and van Reenen, J. (2004), "Evaluating the employment impact of a mandatory job search program," *Journal of the European Economic Association*, 2(4), 569–606.

Bonhomme, S., and Sauder, U. (2011), "Recovering distributions in difference-in-differences models: a comparison of selective and comprehensive schooling," *Review of Economics and Statistics*, 93(May), 479–494.

Borusyak, K., and Jaravel, X. (2017), "Revisiting event study designs," *Working Paper*.

Botosaru, I., and Gutierrez, F. H. (2017), "Difference-in-differences when the treatment status is observed in only one period," *Journal of Applied Econometrics*, (March 2017), 73–90.

Busso, M., Dinardo, J., and McCrary, J. (2014), "New evidence on the finite sample properties of propensity score reweighting and matching estimators," *The Review of Economics and Statistics*, 96(5), 885–895.

Callaway, B., and Li, T. (2018), "Quantile treatment effects in difference in differences models with panel data," *Working Paper*.

Callaway, B., Li, T., and Oka, T. (2018), "Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods," *Journal of Econometrics*, 206(2), 395–413.

Card, D., and Krueger, A. B. (1994), "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania," *American Economic Review*, 84(4), 772–793.

Chabé-Ferret, S. (2015), "Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes," *Journal of Econometrics*, 185(1), 110–123.

Chabé-Ferret, S. (2017), "Should we combine difference in differences with conditioning on pre-treatment outcomes?," *Working Paper*.

Chen, X. (2007), "Large sample sieve estimation of semi-nonparametric models," in *Handbook of Econometrics*, eds. J. J. Heckman, and E. E. Leamer, Vol. 6B, Amsterdam: Elsevier, chapter 76, pp. 5549–5632.

Chen, X., Hong, H., and Tarozzi, A. (2008), "Semiparametric efficiency in GMM models with auxiliary data," *The Annals of Statistics*, 36(2), 808–843.

Cheng, G., Yu, Z., and Huang, J. Z. (2013), "The cluster bootstrap consistency in generalized estimating equations," *Journal of Multivariate Analysis*, 115, 33–47.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013), "Average and quantile effects in nonseparable panel models," *Econometrica*, 81(2), 535–580.

Chernozhukov, V., Fernández-Val, I., and Luo, Y. (2017), "The sorted effects method: discovering heterogeneous effects beyond their averages," *Working Paper*.

Conley, T., and Taber, C. (2011), "Inference with "difference in differences" with a small number of policy changes," *Review of Economics and Statistics*, 93(1), 113–125.

Daw, J. R., and Hatfield, L. A. (2018), "Matching and regression to the mean in difference-in-differences analysis," *Health Services Research*, 53(6), 4138–4156.

de Chaisemartin, C., and D'Haultfœuille, X. (2017), "Fuzzy differences-in-differences," *The Review of Economic Studies*, (February), 1–30.

de Chaisemartin, C., and D'Haultfœuille, X. (2018), "Two-way fixed effects estimators with heterogeneous treatment effects," *Working Paper*.

Donald, S. G., and Hsu, Y.-C. (2014), "Estimation and inference for distribution functions and quantile functions in treatment effect models," *Journal of Econometrics*, 178(3), 383–397.

Dube, A., Lester, T. W., and Reich, M. (2010), "Minimum wage effects across state borders: Estimates using contiguous counties," *Review of Economics and Statistics*, 92(4), 945–964.

Dube, A., Lester, T. W., and Reich, M. (2016), "Minimum wage shocks, employment flows, and labor market frictions," *Journal of Labor Economics*, 34(3), 663–704.

Escanciano, J. C. (2006*a*), "A consistent diagnostic test for regression models using projections," *Econometric Theory*, 22, 1030–1051.

Escanciano, J. C. (2006*b*), "Goodness-of-fit tests for linear and nonlinear time series models," *Journal of the American Statistical Association*, 101(474), 531–541.

Escanciano, J. C. (2008), "Joint and marginal specification tests for conditional mean and variance models," *Journal of Econometrics*, 143(1), 74–87.

Ferman, B., and Pinto, C. (2018), "Inference in differences-in-differences with few treated groups and heteroskedasticity," *The Review of Economics and Statistics*, p. rest_a_00759.

Freyberger, J., and Rai, Y. (2018), "Uniform confidence bands: characterization and optimality," *Journal of Econometrics*, Forthcoming.

Giné, E., and Zinn, J. (1990), "Bootstrapping general empirical measures," *The Annals of Probability*, 18(2), 851–869.

González-Manteiga, W., and Crujeiras, R. M. (2013), "An updated review of Goodness-of-Fit tests for regression models," *Test*, 22(3), 361–411.

Goodman-Bacon, A. (2018), "Difference-in-differences with variation in treatment timing," *Working Paper*.

Hájek, J. (1971), "Discussion of 'An essay on the logical foundations of survey sampling, Part I', by D. Basu," in *Foundations of Statistical Inference*, eds. V. P. Godambe, and D. A. Sprott, Toronto: Holt, Rinehart, and Winston.

Han, S. (2018), "Identification in nonparametric models for dynamic treatment effects," *Working Paper*.

Heckman, J. J., Ichimura, H., Smith, J., and Todd, P. (1998), "Characterizing selection bias using experimental data," *Econometrica*, 66(5), 1017–1098.

Heckman, J. J., Ichimura, H., and Todd, P. (1997), "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme," *The Review of Economic Studies*, 64(4), 605–654.

Horvitz, D. G., and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47(260), 663–685.

Imai, K., Kim, I. S., and Wang, E. (2018), "Matching methods for causal inference with time-series cross-section data," *Working Paper*.

Imbens, G. W., and Wooldridge, J. M. (2009), "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature*, 47(1), 5–86.

Jacobson, L. S., Lalonde, R. J., and Sullivan, D. G. (1993), "Earnings losses of displaced workers," *American Economic Review*, 83(4), 685–709.

Jardim, E., Long, M., Plotnick, R., van Inwegen, E., Vigdor, J., and Wething, H. (2017), "Minimum wage increases, wages, and low-wage employment: Evidence from Seattle," *Working Paper*.

Kline, P., and Santos, A. (2012), "A score based approach to wild bootstrap inference," *Journal of Econometric Methods*, 1(1), 1–40.

Lechner, M. (2010), "The estimation of causal effects by difference-in-difference methods," *Foundations and Trends in Econometrics*, 4(3), 165–224.

MacKinnon, J. G., and Webb, M. D. (2016), "Randomization inference for difference-in-differences with few treated clusters," *Working Paper*.

MacKinnon, J. G., and Webb, M. D. (2018), "The wild bootstrap for few (treated) clusters," *The Econometrics Journal*, 21(2), 114–135.

Mammen, E. (1993), "Bootstrap and wild bootstrap for high dimensional linear models," *The Annals of Statistics*, 21(1), 255–285.

Meer, J., and West, J. (2016), "Effects of the minimum wage on employment dynamics," *Journal of Human Resources*, 51(2), 500–522.

Montiel Olea, J. L., and Plagborg-Møller, M. (2018), "Simultaneous confidence bands: Theory, implementation, and an application to SVARs," *Journal of Applied Econometrics*, pp. 1–64.

Neumark, D., Salas, J. M., and Wascher, W. (2014), "Revisiting the minimum wage-employment debate: Throwing out the baby with the bathwater?," *Industrial and Labor Relations Review*, 67(SUPPL), 608–648.

Neumark, D., and Wascher, W. (1992), "Evidence on employment effects of minimum wages and subminimum wage provisions from panel data on state minimum wage laws," *Industrial and Labor Relations Review*, 46(1), 55–81.

Neumark, D., and Wascher, W. (2000), "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Comment," *American Economic Review*, 90(5), 1362–1396.

Neumark, D., and Wascher, W. L. (2008), *Minimum Wages*, Cambridge, MA: The MIT Press.

Oreopoulos, P., von Wachter, T., and Heisz, A. (2012), "The short- and long-term career effects of graduating in a recession," *American Economic Journal: Applied Economics*, 4(1), 1–29.

Qin, J., and Zhang, B. (2008), "Empirical-likelihood-based difference-in-differences estimators," *Journal of the Royal Statistical Society. Series B (Methodological)*, 75(8), 329–349.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007), "Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable," *Statistical Science*, 22(4), 544–559.

Sant'Anna, P. H. C. (2016), "Program evaluation with right-censored data," *Working Paper*.

Sant'Anna, P. H. C. (2017), "Nonparametric tests for treatment effect heterogeneity with duration outcomes," *Working Paper*.

Sant'Anna, P. H. C., and Song, X. (2019), "Specification tests for the propensity score," *Journal of Econometrics*, (Forthcoming).

Sherman, M., and Le Cessie, S. (2007), "A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models," *Communications in Statistics - Simulation and Computation*, 26(3), 901–925.

Słoczyński, T. (2018), "A general weighted average representation of the ordinary and two-stage least squares estimands," *Working Paper*.

Stinchcombe, M. B., and White, H. (1998), "Consistent specification testing with nuisance parameters present only under the alternative," *Econometric theory*, 14, 295–325.

Stute, W. (1997), "Nonparametric model checks for regression," *The Annals of Statistics*, 25(2), 613–641.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press.

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer.

Wooldridge, J. M. (2003), "Cluster-sample methods in applied econometrics," *American Economic Review P&P*, 93(2), 133–138.

Wooldridge, J. M. (2005), "Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models," *Review of Economics and Statistics*, 87(2), 385–390.