

What’s Trending in Difference-in-Differences?

A Synthesis of the Recent Econometrics Literature

Jonathan Roth* Pedro H. C. Sant’Anna[†] Alyssa Bilinski[‡] John Poe[§]

January 3, 2022

Abstract

This paper synthesizes recent advances in the econometrics of difference-in-differences (DiD) and provides concrete recommendations for practitioners. We begin by articulating a simple set of “canonical” assumptions under which the econometrics of DiD are well-understood. We then argue that recent advances in DiD methods can be broadly classified as relaxing some components of the canonical DiD setup, with a focus on *(i)* multiple periods and variation in treatment timing, *(ii)* potential violations of parallel trends, or *(iii)* alternative frameworks for inference. Our discussion highlights the different ways that the DiD literature has advanced beyond the canonical model, and helps to clarify when each of the papers will be relevant for empirical work. We conclude by discussing some promising areas for future research.

1 Introduction

Differences-in-differences (DiD) is one of the most popular methods in the social sciences for estimating causal effects in non-experimental settings. The last few years have seen a dizzying array of new methodological papers on DiD and related designs, making it challenging for practitioners to keep up with rapidly evolving best practices. Furthermore, the recent literature has addressed a variety of different components of difference-in-differences analyses, which has made it difficult even for experts in the field to understand how all of the new developments fit together. In this paper, we attempt to synthesize some of the

*Brown University. jonathanroth@brown.edu

[†]Microsoft and Vanderbilt University. pedro.h.santanna@vanderbilt.edu

[‡]Brown University. alyssa_bilinski@brown.edu

[§]University of Michigan. john@johndavidpoe.com

recent advances on DiD and related designs and to provide concrete recommendations for practitioners.

Our starting point in Section 2 is the “canonical” difference-in-differences model, where two time periods are available, there is a treated population of units that receives a treatment of interest beginning in the second period, and a comparison population that does not receive the treatment in either period. The key identifying assumption is that the average outcome among the treated and comparison populations would have followed “parallel trends” in the absence of treatment. We also assume that the treatment has no causal effect before its implementation (no anticipation). Together, these assumptions allow us to identify the average treatment effect on the treated (ATT). If we observe a large number of independent clusters from the treated and comparison populations, the ATT can be consistently estimated using a two-way fixed effects (TWFE) regression specification, and clustered standard errors provide asymptotically valid inference.

In practice, DiD applications typically do not meet all of the requirements of the canonical DiD setup. The recent wave of DiD papers have each typically focused on relaxing one or two of the key assumptions in the canonical framework while preserving the others. We taxonomize the recent DiD literature by characterizing which of the key assumptions in the canonical model are relaxed. We focus on recent advances that (i) allow for multiple periods and variation in treatment timing (Section 3); (ii) consider potential violations of parallel trends (Section 4); or (iii) depart from the assumption of observing a sample of many independent clusters sampled from a super-population (Section 5). Section 6 briefly summarizes some other areas of innovation. In what follows, we briefly describe each of these strands of literature.

Multiple periods and variation in treatment timing: One strand of the DiD literature has focused on settings where there are more than two time periods and units are treated at different point in times. Multiple authors have noted that the coefficients from standard TWFE models may not represent a straightforward weighted average of unit-level treatment effects when treatment effects are allowed to be heterogeneous across time or units. In short, TWFE regressions make both “clean” comparisons between treated and not-yet-treated units as well as “forbidden” comparisons between units who are both already-treated. When treatment effects are heterogeneous, these “forbidden” comparisons potentially lead to severe drawbacks such as TWFE coefficients having the opposite sign of all individual-level treatment effects due to “negative weighting” problems. Even if all of the weights are positive, the weights “chosen” by TWFE regressions may not correspond with the most policy-relevant parameter.

We discuss a variety of straightforward-to-implement strategies that have been proposed

to bypass the limitations associated with TWFE regressions and identify/estimate causal parameters of interests under rich sources of treatment effect heterogeneity. These procedures rely on generalizations of the parallel trends assumption to the multi-period setting. A common theme is that these new estimators isolate “clean” comparisons between treated and not-yet-treated groups, and then aggregate them using user-specified weights to estimate a target parameter of economic interest. We discuss differences between some of the recent proposals — such as the exact comparison group used and the generalization of the parallel trends assumption needed for validity — and provide concrete recommendations for practitioners. We also briefly discuss extensions to more complicated settings such as when treatments turn on-and-off over time or are non-binary.

Non-parallel trends: A second strand of the DiD literature focuses on the possibility that the parallel trends assumption may be violated. One set of papers considers the setting where parallel trends holds only conditional on observed covariates, and proposes new estimators that are valid under a conditional parallel trends assumption. However, even if one conditions on covariates, there are often concerns that the necessary parallel trends assumption may be violated due to time-varying confounding factors. It is therefore common practice to test for pre-treatment differences in trends (“pre-trends”) as a test of the plausibility of the (conditional) parallel trends assumption. While intuitive, researchers have identified at least three issues with this pre-testing approach. First, the absence of a significant pre-trend does not necessarily imply that parallel trends holds; in fact, these tests often have low power. Second, conditioning the analysis on the result of a pre-test can introduce additional statistical distortions from a selection effect known as pre-test bias. Third, if a significant difference in trends is detected, we may still wish to learn something about the treatment effect of interest.

Several recent papers have therefore suggested alternative methods for settings where there is concern that parallel trends may be violated. One class of solutions involves modifications to the common practice of pre-trends testing to ensure that the power of pre-tests is high against relevant violations of parallel trends. This can be done either by running power calculations for economically-relevant violations of parallel trends, or using a “non-inferiority” approach that reverses the role of the null and alternative hypotheses so that the pre-test only “passes” if there is evidence that the pre-trend is small. A second class of solutions has proposed methods that remain valid under certain types of violations of parallel trends, such as when the post-treatment violation of parallel trends is assumed to be no larger than the maximal pre-treatment violation of parallel trends, or when there are non-treated groups that are known to be more/less affected by the confounds as the treated group. These approaches allow for a variety of robustness and sensitivity analyses which are

useful in a wide range of empirical settings, and we discuss them in detail.

Alternative sampling assumptions: A third strand of the DiD literature discusses alternatives to the classical “sampling-based” approach to inference with a large number of clusters. One topic of interest is inference procedures in settings with a small number of treated clusters. Standard cluster-robust methods assume that there is a large number of both treated and untreated clusters, and thus can perform poorly in this case. A variety of alternatives with better properties have been proposed for this case, including permutation and bootstrap procedures. These methods typically either model the dependence of errors across clusters, or alternatively place restrictions on the treatment assignment mechanism. We briefly highlight these approaches and discuss the different assumptions needed for them to perform well.

Another direction that has been explored relates to conducting “design-based” inference for DiD. Canonical approaches to inference view the randomness in the data as coming from the fact that the observed data is drawn from a super-population of units. However, this super-population view may be unnatural in DiD settings where all units in the population are observed, such as when we have data on all 50 US states or administrative data on a full population. Design-based approaches to uncertainty instead view the population as fixed and the assignment of treatment as stochastic. Although design-based approaches have typically been employed in the case of randomized experiments, recent work has extended this to the case of “quasi-experimental” strategies like DiD. Luckily, the message of this literature is positive, in the sense that methods that are valid from the canonical sampling-based view are typically also valid from the design-based view as well. The design-based approach to inference helps to overcome conceptual issues related to defining the super-population, and can also be useful for determining the appropriate level of clustering of standard errors. A useful rule of thumb implied by this approach is that it is typically appropriate to cluster at the level at which treatment is independently assigned.

Other topics: We conclude by briefly touching on some other areas of focus within the DiD literature, as well as highlighting some areas for future research. Examples include using DiD to estimate distributional treatment effects; settings with quasi-random treatment timing; spillover effects; estimating heterogeneous treatment effects; and connections between DiD and other panel data methods.

Overall, the growing DiD econometrics literature emphasizes the importance of clarity and precision in a researcher’s discussion of his or her assumptions, comparison group and time frame selection, causal estimands, estimation methods, and robustness checks. When used in combination with context-specific information, these new methods can both improve

the validity and interpretability of DiD results and more clearly delineate their limitations.

Given the vast literature on DiD, our goal is not to be comprehensive, but rather to give a clean presentation of some of the most important directions the literature has gone. Wherever possible, we try to give clear practical guidance for applied researchers, concluding each section with practical recommendations for applied researchers. For reference, we include Table 1, which contains a checklist for a practitioner implementing a DiD analysis, and Table 2, which lists R and Stata packages for implementing many of the methods described in this paper.

2 The Basic Model

This section describes a simple two-period setting in which the econometrics of DiD are well-understood. Although this “canonical” setting is arguably too simple for most applications, clearly articulating the assumptions in this setup serves as a useful baseline for understanding recent innovations in the DiD literature.

2.1 Treatment Assignment and Timing

Consider a model in which there are two time periods, $t = 1, 2$. Units indexed by i are drawn from one of two populations. Units from the treated population ($D_i = 1$) receive a treatment of interest between period $t = 1$ and $t = 2$, whereas units from the untreated (a.k.a. comparison or control) population ($D_i = 0$) remain untreated in both time periods. The econometrician observes an outcome Y_{it} and treatment status D_i for a panel of units, $i = 1, \dots, N$ and $t = 1, 2$. Although DiD methods also accommodate the case where only repeated cross-sectional data is available, or where the panel is unbalanced, we focus on the simpler setup with balanced panel data for ease of exposition.

2.2 Potential Outcomes and Target Parameter

We adopt a potential outcomes framework for the observed outcome, as in, e.g., [Rubin \(1974\)](#) and [Robins \(1986\)](#). Let $Y_{it}(0, 0)$ denote unit i ’s potential outcome in period t if i remains untreated in both periods. Likewise, let $Y_{it}(0, 1)$ denotes unit i ’s potential outcome in period t if i is untreated in the first period but exposed to treatment by the second period. To simplify notation we will write $Y_{it}(0) = Y_{it}(0, 0)$ and $Y_{it}(1) = Y_{it}(0, 1)$, but it will be useful for our later discussion to make clear that these potential outcomes in fact correspond with a path of treatments. As is usually the case, due to the fundamental problem of causal inference ([Holland, 1986](#)), we only observe one of the two potential outcomes for each unit

i . That is, the observed outcome is given by $Y_{it} = D_i Y_{it}(1) + (1 - D_i) Y_{it}(0)$. This potential outcomes framework implicitly encodes the stable unit treatment value assumption (SUTVA) that unit i 's outcomes do not depend on the treatment status of unit $j \neq i$, which rules out spillover effects.

The causal estimand of primary interest in the canonical DiD setup is the average treatment effect on the treated (ATT) in period $t = 2$,

$$\tau_2 = \mathbb{E}[Y_{i2}(1) - Y_{i2}(0) \mid D_i = 1].$$

It simply measures the average causal effect on treated units in the period that they are treated ($t = 2$).

2.3 The Parallel Trends Assumption and Identification

The challenge in identifying τ_2 is that the untreated potential outcomes, $Y_{i2}(0)$, are never observed for the treated group ($D_i = 1$). Difference-in-differences methods overcome this identification challenge via assumptions that allow us to impute the mean counterfactual untreated outcomes for the treated group by using (1) the change in outcomes for the untreated group and (2) the baseline outcomes for the treated group. The key assumption for identifying τ_2 is the parallel trends assumption, which intuitively states that the average outcome for the treated and untreated populations would have evolved in parallel if treatment had not occurred.

Assumption 1 (Parallel Trends).

$$\mathbb{E}[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 1] = \mathbb{E}[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 0]. \quad (1)$$

The parallel trends assumption can be rationalized by imposing a particular generative model for the untreated potential outcomes. If $Y_{it}(0) = \alpha_i + \phi_t + \epsilon_{it}$, where ϵ_{it} is mean-independent of D_i , then Assumption 1 holds. Note that this model allows treatment to be assigned non-randomly based on characteristics that affect the level of the outcome (α_i), but requires the treatment assignment to be mean-independent of variables that affect the *trend* in the outcome (ϵ_{it}). In other words, parallel trends allows for the presence of selection bias, but the bias from selecting into treatment must be the same in period $t = 1$ as it is in period $t = 2$.

Another important and often hidden assumption required for identification of τ_2 is the no-anticipation assumption, which states that the treatment has no causal effect prior to its implementation. This is important for identification of τ_2 , since otherwise the changes in the

outcome for the treated group between period 1 and 2 could reflect not just the causal effect in period $t = 2$ but also the anticipatory effect in period $t = 1$ (Abbring and van den Berg, 2003; Malani and Reif, 2015).

Assumption 2 (No anticipatory effects). $Y_{i1}(0) = Y_{i1}(1)$ for all i with $D_i = 1$.

Under the parallel trends and no-anticipation assumptions, the ATT in period 2 (τ_2) is identified. To see why this is the case, observe that by re-arranging terms in the parallel trends assumption (see equation (1)), we obtain

$$\mathbb{E}[Y_{i2}(0) | D_i = 1] = \mathbb{E}[Y_{i1}(0) | D_i = 1] + \mathbb{E}[Y_{i2}(0) - Y_{i1}(0) | D_i = 0].$$

Further, by the no anticipation assumption, $\mathbb{E}[Y_{i1}(0) | D_i = 1] = \mathbb{E}[Y_{i1}(1) | D_i = 1]$. It follows that

$$\begin{aligned} \mathbb{E}[Y_{i2}(0) | D_i = 1] &= \mathbb{E}[Y_{i1}(1) | D_i = 1] + \mathbb{E}[Y_{i2}(0) - Y_{i1}(0) | D_i = 0] \\ &= \mathbb{E}[Y_{i1} | D_i = 1] + \mathbb{E}[Y_{i2} - Y_{i1} | D_i = 0], \end{aligned}$$

where the second equality uses the fact that we observe $Y(1)$ for treated units and $Y(0)$ for untreated units. The previous display shows that we can infer the *counterfactual* average outcome for the treated group by taking its observed pre-treatment mean and adding the change in mean for the untreated group. Since we observe $Y(1)$ for the treated group directly, it follows that $\tau_2 = \mathbb{E}[Y_{i2}(1) - Y_{i2}(0) | D_i = 1]$ is identified as

$$\tau_2 = \underbrace{\mathbb{E}[Y_{i2} - Y_{i1} | D_i = 1]}_{\text{Change for } D_i = 1} - \underbrace{\mathbb{E}[Y_{i2} - Y_{i1} | D_i = 0]}_{\text{Change for } D_i = 0}, \quad (2)$$

i.e. the “difference-in-differences” of population means!

2.4 Estimation and Inference

Equation (2) gives an expression for τ_2 in terms of a “difference-in-differences” of population expectations. Therefore, a natural way to estimate τ_2 is to replace expectations with their sample analogs,

$$\hat{\tau}_2 = (\bar{Y}_{t=2, D=1} - \bar{Y}_{t=1, D=1}) - (\bar{Y}_{t=2, D=0} - \bar{Y}_{t=1, D=0}),$$

where $\bar{Y}_{t=t', D=d}$ is the sample mean of Y for treatment group d in period t' .

Although these sample means could be computed “by hand”, an analogous way of computing $\hat{\tau}_2$, which facilitates the computation of standard errors, is to use the two-way fixed effects (TWFE) regression specification

$$Y_{it} = \alpha_i + \phi_t + (1[t = 2] \cdot D_i)\beta + \epsilon_{it}, \quad (3)$$

which regresses the outcome Y_{it} on an individual fixed effect, a time fixed effect, and an interaction of a post-treatment indicator with treatment status.¹ In this canonical DiD setup, it is straightforward to show that the ordinary least squares (OLS) coefficient $\hat{\beta}$ is equivalent to $\hat{\tau}_2$.

OLS estimates of $\hat{\beta}$ from (3) provide consistent estimates and asymptotically valid confidence intervals of τ_2 when Assumptions 1 and 2 are combined with the assumption of independent sampling.

Assumption 3. *Let $W_i = (Y_{i2}, Y_{i1}, D_i)'$ denote the vector of outcomes and treatment status for unit i . We observe a sample of N i.i.d. draws $W_i \sim F$ for some distribution F satisfying parallel trends.*

Under Assumptions 1-3 and mild regularity conditions,

$$\sqrt{n}(\hat{\beta} - \tau_2) \rightarrow_d \mathcal{N}(0, \sigma^2),$$

in the asymptotic as $N \rightarrow \infty$ and T is fixed. The variance σ^2 is consistently estimable using standard clustering methods that allow for arbitrary serial correlation at the unit level (Liang and Zeger, 1986; Arellano, 1987; Wooldridge, 2003; Bertrand, Duflo and Mullainathan, 2004). The same logic easily extends to cases where units are clustered within independent clusters (e.g. states), and the standard errors are clustered at the appropriate level, provided that the number of treated and untreated clusters both grow large. Constructing consistent point estimates and asymptotically valid confidence intervals is thus straightforward via OLS.

Having introduced all of the components of the “canonical” DiD model, we now discuss the ways that different strands of the recent DiD have relaxed each of these components.

¹ With a balanced panel, the OLS coefficient on β is also numerically identical to the coefficient from a regression that replaces α_i and ϕ_t with treatment group and period fixed effects,

$$Y_{it} = \alpha + D_i\theta + 1[t = 2]\zeta + (1[t = 2] \cdot D_i)\beta + \epsilon_{it}.$$

The latter regression generalizes to repeated cross-sectional data.

3 Relaxing assumptions on treatment assignment and timing

Several recent papers have focused primarily on relaxing the baseline assumptions about treatment assignment and timing discussed in Section 2. A topic of considerable attention has been settings where there are more than two periods, and units adopt a treatment of interest at different points in time. For instance, different states may enact a particular type of legislation at different points in time. We provide an overview of some of the key developments in the literature, and refer the reader to the review by [de Chaisemartin and D’Haultfoeuille \(2021a\)](#) for additional details.

3.1 Generalized model with staggered treatment adoption

Several recent papers have focused on relaxing the timing assumptions discussed in Section 2, while preserving the remaining structure of the stylized model (i.e., parallel trends, no anticipation, and independent sampling). Since most of the recent literature considers a setup in which treatment is an absorbing state, we start with that framework; in Section 3.4, we discuss extensions to the case where treatment can turn on and off or is non-binary. We introduce the following assumptions and notation, which captures the primary setting studied in this literature.

Treatment timing. There are T periods indexed by $t = 1, \dots, T$, and units can receive a binary treatment of interest in any of the periods. Treatment is an absorbing state, so that once a unit is treated they remain treated for the remainder of the panel. We denote by D_{it} an indicator for whether unit i receives treatment in period t , and $G_i = \min\{t : D_{it} = 1\}$ the earliest period at which unit i has received treatment. If i is never treated during the sample, then $G_i = \infty$. Treatment is an absorbing state, so that $D_{it} = 1$ for all $t \geq G_i$. We define $R_{it} = t - G_i + 1$ to be the time relative to treatment, so that $R_{it} = 1$ in the first treated period for unit i .

Potential outcomes. We extend the potential outcomes framework introduced above to the multi-period setting. Let $\mathbf{0}_s$ and $\mathbf{1}_s$ denote s -dimensional vectors of zeros and ones, respectively. We denote unit i ’s potential outcome in period t if they were first treated at time g by $Y_{it}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$, and denote by $Y_{it}(\mathbf{0}_T)$ their “never-treated” potential outcome. This notation again makes explicit that potential outcomes can depend on the entire *path* of treatment assignments. Since we have assumed that treatment “stays on” once it is turned

on, the entire path of potential outcomes is summarized by the first treatment date (g), and so to simplify notation we can index potential outcomes by treatment starting time: $Y_{it}(g) = Y_{it}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ and $Y_{it}(\infty) = Y_{it}(\mathbf{0}_T)$.²

Parallel trends. There are several ways to extend the canonical parallel trends assumption to the staggered setting. The simplest extension of the parallel trends assumption to the staggered case requires that the two-group, two-period version of parallel trends holds for all combinations of periods and all combinations of “groups” treated at different times.

Assumption 4 (Parallel trends for staggered setting). *For all $t \neq t'$ and $g \neq g'$,*

$$\mathbb{E}[Y_{it}(\infty) - Y_{it'}(\infty) | G_i = g] = \mathbb{E}[Y_{it}(\infty) - Y_{it'}(\infty) | G_i = g']. \quad (4)$$

This assumption imposes that in the counterfactual where treatment had not occurred, the outcomes for all adoption groups would have evolved in parallel.

Several variants of Assumption 4 have been considered in the literature. For example, Callaway and Sant’Anna (2021) consider a relaxation of Assumption 4 that imposes (4) only for pairs of groups and time periods such that at least one of the groups has already been treated. Likewise, several papers, including Callaway and Sant’Anna (2021) and Sun and Abraham (2021), consider versions that impose (4) only for groups that are eventually treated, and not for the never-treated group ($g = \infty$). There are tradeoffs between the different forms of Assumption 4: imposing parallel trends for all groups and all periods is a stronger assumption and thus may be less plausible; on the other hand, it may allow one to obtain more precise estimates.³ We return to these tradeoffs in our discussion of different estimators for the staggered case below.

No anticipation. The no-anticipation assumption from the canonical model also extends in a straightforward way to the staggered setting. Intuitively, it imposes that if a unit is untreated in period t , their outcome does not depend on what time period they will be treated in the future.

Assumption 5 (Staggered no anticipation assumption). $Y_{it}(g) = Y_{it}(\infty)$ for all $t < g$.

²If one were to map this staggered potential outcome notation to the one used in canonical DiD setups, we would write $Y_{it}(2)$ and $Y_{it}(\infty)$ instead of $Y_{it}(1)$ and $Y_{it}(0)$ as defined in Section 2.2, respectively. We use the $_{it}Y(0), Y_{it}(1)$ notation in Section 2.2 because it is likely more familiar to the reader, and widely used in the literature on the canonical model.

³If all units are eventually treated, then imposing parallel trends only among treated units also limits the number of periods for which the ATT is identified.

3.2 Interpreting the estimand of two-way fixed effects models

Recall that in the simple two-period model, the estimand (population coefficient) of the two-way fixed effects specification (3) corresponds with the ATT under the parallel trends assumption. A substantial focus of the recent literature has been whether the estimand of commonly-used generalizations of this TWFE model to the multi-period, staggered timing case have a similar, intuitive causal interpretation. In short, the literature has shown that the estimand of TWFE specifications in the staggered setting often does not correspond with an intuitive causal parameter even under the natural extensions of the parallel trends and no-anticipation assumptions described above.

Static TWFE. We begin with a discussion of the “static” TWFE specification, which regresses the outcome on individual and period fixed effects and an indicator for whether the unit i is treated in period t ,

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta_{post} + \epsilon_{it}. \quad (5)$$

The static specification yields a sensible estimand when there is no heterogeneity in treatment effects across either time or units. Formally, let $\tau_{it}(g) = Y_{it}(g) - Y_{it}(\infty)$. Suppose that for all units i , $\tau_{it}(g) = \tau$ whenever $t \geq g$. This imposes that (1) all units have the same treatment effect, and (2) the treatment has the same effect regardless of how long it has been since treatment started. Then, under a suitable generalization of the parallel trends assumption (e.g. Assumption 4) and no anticipation assumption (Assumption 5), the population regression coefficient β_{post} from (5) is equal to τ .

Issues arise with the static specification, however, when there is heterogeneity of treatment effects across either time since treatment or across units, as shown in [Borusyak and Jaravel \(2018\)](#), [de Chaisemartin and D’Haultfoeulle \(2020b\)](#), and [Goodman-Bacon \(2021\)](#), among others. Suppose first that there is heterogeneity in time since treatment only. That is, $\tau_{it}(g) = \sum_{s \geq 0} \tau_s 1[t - g = s]$, so all units have treatment effect τ_s in the s -th period after they receive treatment. In this case, β_{post} corresponds with a potentially non-convex weighted average of the parameters τ_s , i.e. $\beta_{post} = \sum_s \omega_s \tau_s$, where the weights ω_s sum to 1 but may be negative. The possibility of negative weights is concerning because, for instance, all of the τ_s could be positive and yet the coefficient β_{post} may be negative! In particular, longer-run treatment effects will often receive negative weights. Likewise, if treatment effects differ across units but are constant across time, i.e. $\tau_{it}(g) = \tau_i$ for all $t \geq g$, then β_{post} may be a non-convex weighted average of the individual level effects τ_i . Similar issues arise if

there is heterogeneity both across units and time since treatment.⁴

Goodman-Bacon (2021) provides some helpful intuition to understand this phenomenon. He shows that $\hat{\beta}_{post}$ can be written as a convex weighted average of differences-in-differences comparisons between pairs of units and time periods in which one unit changed its treatment status and the other did not. Counterintuitively, however, this decomposition includes difference-in-differences that use as a “control” group units who were treated in earlier periods. Hence, an early-treated unit can get negative weights if it appears as a “control” for many later-treated units. This decomposition further highlights that β_{post} may not be a sensible estimand when treatment effects differ across either units or time, because of its inclusion of “forbidden comparisons”.

We now give some more mathematical intuition for why weighting issues arise in the static specification with heterogeneity. From the Frisch-Waugh-Lovell theorem, the coefficient β_{post} from (5) is equivalent to the coefficient from a univariate regression of Y_{it} on $D_{it} - \hat{D}_{it}$, where \hat{D}_{it} is the predicted value from a regression of D_{it} on the other right-hand side variables in (5), $D_{it} = \tilde{\alpha}_i + \tilde{\phi}_t + u_{it}$. However, a well-known issue with OLS with binary outcomes is that its predictions may fall outside the unit interval. If the predicted value \hat{D}_{it} is greater than 1, then $D_{it} - \hat{D}_{it}$ will be negative even when a unit is treated, and thus that unit’s outcome will get negative weight in $\hat{\beta}_{post}$. To see this more formally, we can apply the formula for univariate OLS coefficients to obtain that

$$\hat{\beta}_{post} = \frac{\sum_{i,t} (D_{it} - \hat{D}_{it}) Y_{it}}{\sum_{i,t} (D_{it} - \hat{D}_{it})^2}. \quad (6)$$

The denominator is positive, and so the weight that $\hat{\beta}_{post}$ places on Y_{it} is proportional to $D_{it} - \hat{D}_{it}$. Thus, if $D_{it} > 0$ and $D_{it} - \hat{D}_{it} < 0$, then $\hat{\beta}_{post}$ will be decreasing in Y_{it} . But because $Y_{it} = Y_{it}(\infty) + \tau_{it}(g)$, it follows that $\tau_{it}(g)$ gets negative weight in $\hat{\beta}_{post}$.

These negative weights will tend to arise for early-treated units in periods late in the sample. To see why this is the case, we note that some algebra shows that $\hat{D}_{it} = \bar{D}_i + \bar{D}_t - \bar{D}$, where $\bar{D}_i = T^{-1} \sum_t D_{it}$ is the time average of D for unit i , $\bar{D}_t = N^{-1} \sum_i D_{it}$ is the cross-sectional average of D for period t , and $\bar{D} = (NT)^{-1} \sum_{i,t} D_{it}$ is the average of D across both periods and units. It follows that if we have a unit that has been treated for almost all periods ($\bar{D}_i \approx 1$) and a period in which almost all units have been treated ($\bar{D}_t \approx 1$), then $\hat{D}_{it} \approx 2 - \bar{D}$, which will be strictly greater than 1 if there is a non-substantial fraction of non-treated units in some period ($\bar{D} < 1$). We thus see that $\hat{\beta}_{post}$ will tend to put negative

⁴We focus in this section on decompositions of the static TWFE model in a standard, sampling-based framework. Athey and Imbens (2022) study the static specification in a finite-sample randomization-based framework.

weight on τ_{it} for early-adopters in late periods within the sample. This decomposition makes clear that the static OLS coefficient $\hat{\beta}_{post}$ is not aggregating natural comparisons of units, and thus will not produce a sensible estimand when there is arbitrary heterogeneity. When treatment effects are homogeneous – i.e. $\tau_{it}(g) \equiv \tau$ – the negative weights on τ for some units cancel out the positive weights on other units, and thus β_{post} recovers the causal effect under a suitable generalization of parallel trends.

Dynamic TWFE. Next, we turn our attention to the “dynamic specification” that regresses the outcome on individual and period fixed effects, as well as dummies for time relative to treatment

$$Y_{it} = \alpha_i + \phi_t + \sum_{\substack{r \neq 0 \\ -\bar{T} \leq r \leq \bar{T}}} 1[R_{it} = r] \beta_r + \epsilon_{it}. \quad (7)$$

Unlike the static specification, the dynamic specification yields a sensible causal estimand when there is heterogeneity only in time since treatment. In particular, the results in [Borusyak and Jaravel \(2018\)](#) and [Sun and Abraham \(2021\)](#) imply that if $\tau_{it}(g) = \sum_{s \geq 0} \tau_s 1[t - g = s]$, so all units have treatment effect τ_s in the s -th period after treatment, then $\beta_s = \tau_s$ under a suitable generalization of the parallel trends assumption such as Assumption 4.⁵ [Sun and Abraham \(2021\)](#) show, however, that when there are heterogeneous treatment effects across adoption cohorts, the coefficients from specification (7) become difficult to interpret. There are two issues. First, as with the “static” regression specification above, the coefficient β_r may put *negative* weight on the treatment effect r periods after treatment for some units. Second, the coefficient β_r can put positive weight on treatment effects at lags $r' \neq r$, so there is cross-lag “contamination.” Like the static specification, the dynamic specification thus fails to yield sensible estimates of dynamic causal effects under heterogeneity across cohorts. The derivation of this result is mathematically more complex, and so we do not pursue it here. The intuition is that, as in the static case, the dynamic OLS specification does not aggregate natural comparisons of units and includes “forbidden comparisons.” An important implication of the results derived by [Sun and Abraham \(2021\)](#) is that if treatment effects are heterogeneous, the “treatment lead” coefficients from (7) are not guaranteed to be zero even if parallel trends is satisfied in all periods, and thus evaluation of pre-trends based on these coefficients can be very misleading.

⁵We note that the homogeneity assumption can be relaxed so that all adoption cohorts have the same expected treatment effect, i.e. $\mathbb{E}[\tau_{i,g+s}(g) | G = g] \equiv \tau_s$ for all s and g . Additionally, these results assume that all possible relative time indicators are included. As discussed in [Sun and Abraham \(2021\)](#) and [Baker, Larcker and Wang \(2021\)](#), among others, problems may arise if one “bins” endpoints (e.g. includes a dummy for 20+ years since treatment).

3.2.1 Diagnostic approaches

Several recent papers introduce diagnostic approaches for understanding the extent of the aggregation issues under staggered treatment timing, with a focus on the static specification (5). [de Chaisemartin and D’Haultfoeuille \(2020b\)](#) propose reporting the degree of heterogeneity in treatment effects that would be necessary for the estimated treatment effect to have the “wrong sign.” [Goodman-Bacon \(2021\)](#) proposes reporting the weights that $\hat{\beta}_{post}$ places on the different 2-group, 2-period difference-in-differences, which allows one to evaluate how much weight is being placed on “forbidden” comparisons of already-treated units and how removing the comparisons would change the estimate. [Jakiela \(2021\)](#) proposes evaluating both whether TWFE places negative weights on some treated units and whether the data rejects the constant treatment effects assumption.

3.3 New Estimators For Staggered Timing

Several recent papers have proposed alternative estimators that more sensibly aggregate heterogeneous treatment effects in settings with staggered treatment timing. The derivation of each of these estimators follows a similar logic to the derivation of the DiD estimator in the motivating example in Section 2. We begin by specifying a causal parameter of interest (analogous to the ATT τ_2). With the help of the (generalized) parallel trends and no-anticipation assumptions, we can infer the counterfactual outcomes for treated units using trends in outcomes for an appropriately chosen “clean” control group of untreated units. This allows us to express the target parameter in terms of identified expectations, analogous to equation (2). Finally, we replace population expectations with sample averages to form an estimator of the target parameter.

We describe in detail the approach taken by [Callaway and Sant’Anna \(2021\)](#), and then discuss the connections to other approaches. They consider as a building block the group-time average treatment effect on the treated, $ATT(g, t) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) | G_{it} = g]$, which gives the average treatment effect at time t for the cohort first treated in time g . They then consider identification and estimation under generalizations of the parallel trends assumption such as Assumption 4 and variants thereof.⁶ Intuitively, under Assumptions 4 (parallel trends) and 5 (no anticipation), we can identify $ATT(g, t)$ by comparing the expected change in outcome for cohort g between periods $g - 1$ and t to that for a control group not-yet treated

⁶[Callaway and Sant’Anna \(2021\)](#) also consider generalizations where the parallel trends assumption holds only conditional on covariates. We discuss this extension in Section 4.2 below, but focus for now on the case without covariates.

at period t . Formally,

$$ATT(g, t) = \mathbb{E}[Y_{it} - Y_{i,g-1} | G_i = g] - \mathbb{E}[Y_{it} - Y_{i,g-1} | G_i = g'], \text{ for any } g' > t$$

which can be viewed as the multi-period analog of the identification result in equation (2).⁷ Since this holds for any comparison group $g' > t$, it also holds if we average over some set of comparisons \mathcal{G} such that $g' > t$ for all $g' \in \mathcal{G}$,

$$ATT(g, t) = \mathbb{E}[Y_{it} - Y_{i,g-1} | G_i = g] - \mathbb{E}[Y_{it} - Y_{i,g-1} | G_i \in \mathcal{G}].$$

We can then estimate $ATT(g, t)$ by replacing expectations with their sample analogs,

$$\widehat{ATT}(g, t) = \frac{1}{N_g} \sum_{i: G_i = g} [Y_{it} - Y_{i,g-1}] - \frac{1}{N_{\mathcal{G}}} \sum_{i: G_i \in \mathcal{G}} [Y_{it} - Y_{i,g-1}]. \quad (8)$$

Specifically, Callaway and Sant’Anna (2021) consider two options for \mathcal{G} . The first uses only never-treated units ($\mathcal{G} = \{\infty\}$) and the second uses all not-yet-treated units ($\mathcal{G} = \{g' : g' > t\}$). When there are a relatively small number of periods and treatment cohorts, reporting $\widehat{ATT}(g, t)$ for all relevant (g, t) may be reasonable.

When there are many treated periods and/or cohorts, however, reporting all the $\widehat{ATT}(g, t)$ may be cumbersome, and each one may be imprecisely estimated. Thankfully, the method described above extends easily to estimating any weighted average of the $ATT(g, t)$. For instance, we may be interested in an “event-study” parameter that gives the (weighted) average of the treatment effect l periods after adoption across different adoption cohorts,

$$ATT_l^w = \sum_g w_g ATT(g, g + l). \quad (9)$$

The weights w_g could be chosen to weight different cohorts equally, or in terms of their relative frequencies in the treated population. It is straightforward to form estimates for ATT_l^w by averaging the estimates $\widehat{ATT}(g, t)$ discussed above. We refer the reader to Callaway and Sant’Anna (2021) for a discussion of a variety of other weighted averages that may be economically relevant. Inference is straightforward using either the delta method or a bootstrap, as described in Callaway and Sant’Anna (2021).

This alternative approach to estimation has two primary advantages over standard static or dynamic TWFE regressions. The first is that it provides sensible estimands even un-

⁷This derivation only relies on equation (4) to hold for pairs of periods and groups such that at least one unit is treated, and thus Callaway and Sant’Anna (2021) can impose a weaker variant of Assumption 4 that imposes (4) only for such groups and periods.

der arbitrary heterogeneity of treatment effects. The second is that it makes transparent exactly which units are being used as a control group to infer the unobserved potential outcomes. This contrasts with standard TWFE models, which we have seen make unintuitive comparisons under staggered timing.

Several other recent papers propose similar estimation strategies. [de Chaisemartin and D’Haultfoeuille \(2020b\)](#) propose an estimator that can be applied when treatment turns on and off (see Section 3.4 below), but in the context of the staggered setting here corresponds with the Callaway and Sant’Anna estimator for ATT_0^w and a particular choice of weights. [Sun and Abraham \(2021\)](#) propose an estimator that uses the last-to-be-treated units as the comparison ($\mathcal{G} = \{\max_i G_i\}$), rather than the not-yet-treated. [Marcus and Sant’Anna \(2021\)](#) propose a recursive estimator that more efficiently exploits the identifying assumptions in [Callaway and Sant’Anna \(2021\)](#). See, also, [Imai and Kim \(2021\)](#) and [Strezhnev \(2018\)](#) for closely related ideas.

Another related approach is to run a stacked regression where each treated unit is matched to ‘clean’ (i.e. not-yet-treated) controls and there are separate fixed effects for each set of treated units and its control, as in [Cengiz, Dube, Lindner and Zipperer \(2019\)](#) among others. [Gardner \(2021\)](#) shows that this approach estimates a convex weighted average of the $ATT(g, t)$ under parallel trends and no anticipation, although the weights are determined by the number of treated units and variance of treatment within each stacked event, rather than by economic considerations.

Another related estimator is what [Borusyak, Jaravel and Spiess \(2021\)](#) refer to as an imputation estimator (see, also, [Gardner \(2021\)](#), [Liu, Wang and Xu \(2021\)](#) and [Wooldridge \(2021\)](#) for similar proposals). Specifically, they fit a TWFE regression, $Y_{it}(\infty) = \alpha_i + \lambda_t + \epsilon_{it}$, using observations only for units and time periods that are not-yet-treated. They then infer the never-treated potential outcome for each unit using the predicted value from this regression, $\hat{Y}_{it}(\infty)$. This provides an estimate of the treatment effect for each unit, $Y_{it} - \hat{Y}_{it}(\infty)$, and these individual-level estimates can be aggregated to form summary parameters like the $ATT(g, t)$ described above. Although the [Borusyak et al. \(2021\)](#) estimator is difficult to characterize in closed form in general, in the special case where units are either first-treated at some g^* or never-treated, the imputation estimator for $ATT(g^*, t)$ is given by

$$\widehat{ATT}(g^*, t)_{Imp} = \frac{1}{N_{g^*}} \sum_{i: G_i = g^*} [Y_{it} - \bar{Y}_{i, 1: g^* - 1}] - \frac{1}{N_\infty} \sum_{i: G_i = \infty} [Y_{it} - \bar{Y}_{i, 1: g^* - 1}],$$

where $\bar{Y}_{i, 1: g^* - 1} = \frac{1}{g^* - 1} \sum_{t=1}^{g^* - 1} Y_{it}$ is i ’s average outcome over periods 1 to $g^* - 1$. The imputation estimator thus differs from $\widehat{ATT}(g, t)$ in that it uses the average pre-treatment outcome over all periods before period g^* , rather than the outcome in period $g^* - 1$ only. This

can potentially improve efficiency since it averages over more periods, and indeed [Borusyak et al. \(2021\)](#) show that this is necessarily the case in a model for $Y_{it}(\infty)$ with homoskedastic errors. On the other hand, the validity of the estimator depends on a stronger parallel trends assumption, which in this example requires parallel trends for all periods rather than beginning only in period $g^* - 1$. Although it might seem unlikely that parallel trends would hold exactly for all post-treatment periods and not for all pre-treatment periods in some applications, relying on this stronger assumption may lead to larger biases if the parallel trends assumption holds only approximately. As discussed in [Roth \(2018\)](#) and [de Chaisemartin and D’Haultfœuille \(2021a\)](#), for example, the imputation estimator may be more susceptible to bias if there is a monotonic violation of parallel trends, since it puts more weight on earlier periods where the treatment and comparison groups are less similar. See [Marcus and Sant’Anna \(2021\)](#) for additional discussion of the tradeoffs between the strength of the identifying assumptions and the precision of one’s estimates in staggered treatment settings.

3.4 Further extensions to treatment timing/assignment

Our discussion so far has focused on the case where there is a binary treatment that is adopted at a particular date and remains on afterwards. Several recent papers have studied settings with more complicated forms of treatment assignment. We briefly highlight a few of the recent contributions, and refer the reader to the review in [de Chaisemartin and D’Haultfœuille \(2021a\)](#) for more details.

[de Chaisemartin and D’Haultfœuille \(2020b\)](#) and [Imai and Kim \(2021\)](#) adopt similar assumptions about treatment timing, but do not necessarily require that treatment is an absorbing state. Their estimators intuitively compare changes in outcomes for units whose treatment status changed to other units whose treatment status remained constant over the same periods. This approach yields an interpretable causal effect under generalizations of the parallel trends assumption and a “no carryover” assumption that imposes that potential outcomes only depend on current treatment status and not on the full treatment histories. [de Chaisemartin and D’Haultfœuille \(2020a\)](#) relax the no-carryover assumption, allowing potential outcomes to depend on the full path of treatments, at the expense of imposing a parallel trends assumption that requires parallel trends in untreated potential outcomes regardless of a unit’s path of treatment. [de Chaisemartin and D’Haultfœuille \(2018\)](#) study “fuzzy” DiD settings in which all groups are treated in both time periods, but the proportion of units exposed to treatment increases in one group but not in the other. Finally, [de Chaisemartin and D’Haultfœuille \(2021b\)](#) and [Callaway, Goodman-Bacon and Sant’Anna \(2021\)](#)

study settings with multi-valued or continuous treatments.

3.5 Recommendations

The results discussed above show that while conventional TWFE specifications make sensible comparisons of treated and untreated units in the canonical two-period DiD model, in the staggered case they typically make “forbidden comparisons” between already-treated units. As a result, treatment effects for some units and time periods receive negative weights in the TWFE estimand. In extreme cases, this can lead the TWFE estimand to have the “wrong sign” — e.g., the estimand may be negative even if all the treatment effects are positive. Even if the weights are not so extreme as to create sign reversals, it may nevertheless be difficult to interpret which comparisons the TWFE estimator is making, as the “control group” is not transparent, and the weights it chooses are unlikely to be those most relevant for economic policy.

In our view, the most direct remedy for this problem is to use the methods discussed in Section 3.3 that explicitly specify the comparisons to be made between treatment and control groups, as well as the desired weights in the target parameter. These methods allow one to estimate a well-defined causal parameter (under parallel trends), with transparent weights and transparent comparison groups (e.g. not-yet-treated or never-treated units). This approach, in our view, provides a more complete solution to the problem than the diagnostic approaches discussed in Section 3.2.1. Although it is certainly valuable to have a sense of the extent to which conventional TWFE specifications are making bad comparisons, eliminating these undesirable comparisons seems to us a better approach than diagnosing the extent of the issue. Using a TWFE specification may be justified for efficiency reasons if one is confident that treatment effects are homogeneous, but we suspect that this will be rare in practice.

The question of which of the many heterogeneity-robust DiD methods discussed in Section 3.3 to use is trickier. As described above, there are some tradeoffs between the efficiency of the estimator and the strength of the parallel trends assumption needed for identification. The best estimator to use may therefore depend on the context. Nevertheless, it is our practical experience that the various heterogeneity-robust DiD estimators typically (although not always) produce similar answers. The first-order consideration is therefore to use an approach that makes clear what the target parameter is and which groups are being compared for identification. Thankfully, there are now statistical packages that make implementing (and comparing) the results from these estimators straightforward in practice (see Table 2).

We acknowledge that these new methods may initially appear complicated to researchers

accustomed to analyzing seemingly simple regression specifications such as (5) or (7). However, while traditional TWFE regressions are easy to *specify*, as discussed above they are actually quite difficult to *interpret*, since they make complicated and unintuitive comparisons across groups. By contrast, the methods that we recommend have a simple interpretation using a coherent control group. And while more complex to express in regression format, they can be viewed as simple aggregations of comparisons of group means. We suspect that once researchers gain experience using the newer heterogeneity-robust DiD methods, they will not seem so scary after all!

4 Relaxing or allowing the parallel trends assumption to be violated

A second strand of the literature has focused on the possibility that the canonical parallel trends assumption may not hold exactly. Approaches to this problem include relaxing the parallel trends assumption to hold only conditional on covariates, testing for pre-treatment violations of the parallel trends assumption, and various tools for robust inference and sensitivity analysis that explore the possibility that parallel trends may be violated in certain ways.

4.1 Why might parallel trends be violated?

The canonical parallel trends assumption requires that the mean outcome for the treated group would have evolved in parallel with the mean outcome for the untreated group if the treatment had not occurred. As discussed in Section 2, this allows for confounding factors that affect treatment status, but these must have a constant additive effect on the mean outcome.

In practice, however, we will often be unsure of the validity of the parallel trends assumption for several reasons. First, there will often be concern about *time-varying* confounding factors. For example, Democratic-leaning states may be more likely to adopt a particular policy (e.g., the minimum wage) and be exposed to different macro-economic shocks. If the state of the macro-economy is time-varying, then we would expect this confound to have a time-varying effect on the outcome. A second concern relates to the potential sensitivity of the parallel trends assumption to the chosen function form of the outcome. If the mean of $Y_{it}(0)$ evolves in parallel for both groups, then it will generally not be the case that the mean of $\log(Y_{it}(0))$ evolves in parallel. Indeed, Roth and Sant’Anna (2021b) show that parallel trends can hold for all monotonic transformations of the outcome $g(Y_{it}(0))$ essentially only if

the population can be partitioned into two groups, the first of which is effectively randomly assigned between treatment and control and the second for which the outcome distribution is stable over time. Although there are some cases where these conditions may be (approximately) met — the most prominent of which is random assignment of treatment — they are likely not to hold in most settings where DiD is used, and thus parallel trends will be sensitive to functional form. It will often not be obvious that parallel trends should hold for the particular functional form chosen for our analysis, and thus we may be skeptical of its validity.

4.2 Parallel trends conditional on covariates

One way to increase the credibility of the parallel trends assumption is to require that it holds only conditional on covariates. Indeed, if we condition on a rich enough set of covariates X_i , we may be willing to believe that treatment is nearly randomly assigned conditional on X_i . Imposing only parallel trends conditional on X_i gives us an extra degree of robustness, since conditional random assignment can fail so long as the remaining unobservables have a time-invariant additive effect on the outcome.

In the canonical model discussed in Section 2, the parallel trends assumption can be naturally extended to incorporate covariates as follows.

Assumption 6 (Conditional Parallel Trends).

$$\mathbb{E}[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 1, X_i] = \mathbb{E}[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 0, X_i] \text{ (almost surely) }, \quad (10)$$

for X_i a pre-treatment vector of observable covariates.

For simplicity, we will focus first on the conditional parallel trends assumption in the canonical two-period model, although several papers have also extended this idea to the case of staggered treatment timing, as we will discuss towards the end of this subsection. We furthermore limit our discussion to covariates measured prior to treatment and which are time-invariant (although they may have a time-varying impact on the outcome); relevant extensions to this are also discussed below.

In addition to the conditional parallel trends assumption, we will also impose an overlap condition (a.k.a. positivity condition), which guarantees that for each treated unit with covariates X_i , there are at least some untreated units in the population with the same value of X_i . This overlap assumption is particularly important for using standard inference procedures (Khan and Tamer, 2010).

Assumption 7 (Strong overlap). *The conditional probability of belonging to the treatment group, given observed characteristics, is uniformly bounded away from one, and the proportion of treated units is bounded away from zero. That is, for some $\epsilon > 0$, $P(D_i = 1|X_i) < 1 - \epsilon$, almost surely, and $\mathbb{E}[D] > 0$.*

Given the conditional parallel trends assumption, no anticipation assumption, and overlap condition, the ATT conditional on $X_i = x$,

$$\tau_2(x) = \mathbb{E}[Y_{i2}(1) - Y_{i2}(0)|D_i = 1, X_i = x],$$

is identified for all x with $P(D_i = 1|X_i = x) > 0$. In particular,

$$\tau_2(x) = \underbrace{\mathbb{E}[Y_{i2} - Y_{i1} | D_i = 1, X_i = x]}_{\text{Change for } D_i = 1, X_i = x} - \underbrace{\mathbb{E}[Y_{i2} - Y_{i1} | D_i = 0, X_i = x]}_{\text{Change for } D_i = 0, X_i = x}. \quad (11)$$

Note that equation (11) is analogous to (2) in the canonical model, except it conditions on $X_i = x$. Intuitively, among the sub-population with $X_i = x$, we have parallel trends, and so we can take the same steps as in Section 2 to infer the conditional ATT for that sub-population. The unconditional ATT can then be identified by averaging $\tau_2(x)$ over the distribution of X_i in the treated population. Using the law of iterated expectations, we have

$$\tau_2 = \mathbb{E}[Y_{i2}(1) - Y_{i2}(0)|D_i = 1] = \mathbb{E}\left[\underbrace{\mathbb{E}[Y_{i2}(1) - Y_{i2}(0)|D_i = 1, X_i]}_{\tau_2(X_i)} | D_i = 1\right].$$

When X_i is discrete and takes a small number of values — for example, if X_i is an indicator for whether someone has a college degree — then estimation is straightforward. We can just run an unconditional DiD for each value of X_i , and then aggregate the estimates to form an estimate for the overall ATT, using the delta method or bootstrap for the standard errors.

When X_i is continuously distributed (or discrete with a very large number of support points), estimation becomes more complicated, because we will typically not have a large enough sample to do an unconditional DiD within each possible value of X_i . Thankfully, there are several available econometric approaches to semi-/non-parametrically estimate the ATT even with continuous covariates. We first discuss the limitations of using TWFE regressions in this setting, and then discuss several alternative approaches.

Standard linear regression. Given that the TWFE specification (3) yielded consistent estimates of the ATT under Assumptions 1-3 in the canonical DiD model, it may be tempting

to augment this specification with controls for a time-by-covariate interaction,

$$Y_{it} = \alpha_i + \phi_t + (1[t = 2] \cdot D_i)\beta + (X_i \cdot 1[t = 2])\gamma + \varepsilon_{it}, \quad (12)$$

for estimation under conditional parallel trends. Unfortunately, this augmented specification need not yield consistent estimates of the ATT under conditional parallel trends without additional homogeneity assumptions. The intuition is that equation (12) implicitly models the conditional expectation function (CEF) of $Y_{i2} - Y_{i1}$ as depending on X_i with a constant slope of γ , regardless of i 's treatment status. If there are heterogeneous treatment effects that depend on X_i — e.g., the ATT varies by age of participants — the derivative of the CEF with respect to X_i may depend on treatment status D_i as well. In these practically relevant setups, estimates of β can be biased for the ATT; see Meyer (1995) and Abadie (2005) for additional discussion. Fortunately, there are several semi-/non-parametric methods available that allow for consistent estimation of the ATT under conditional parallel trends under weaker homogeneity assumptions.

Regression adjustment. An alternative approach to allow for covariate-specific trends in DiD settings is the regression adjustment procedure proposed by Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998). Their main idea exploits the fact that under conditional parallel trends, strong overlap, and no anticipation we can write the ATT as

$$\begin{aligned} \tau_2 &= \mathbb{E}[\mathbb{E}[Y_{i2} - Y_{i1} \mid D_i = 1, X_i] - \mathbb{E}[Y_{i2} - Y_{i1} \mid D_i = 0, X_i] \mid D_i = 1], \\ &= \mathbb{E}[Y_{i2} - Y_{i1} \mid D_i = 1] - \mathbb{E}[\mathbb{E}[Y_{i2} - Y_{i1} \mid D_i = 0, X_i] \mid D_i = 1], \end{aligned}$$

where the second equality follows from the law of iterated expectations. Thus, to estimate the ATT under conditional parallel trends, one simply needs to estimate the conditional expectation among untreated units given covariate values, and then average these “predictions” using the empirical distribution of X_i among treated units. That is, we estimate τ_2 with

$$\hat{\tau}_2 = \frac{1}{N_1} \sum_{i:D_i=1} \left((Y_{i2} - Y_{i1}) - \hat{\mathbb{E}}[Y_{i2} - Y_{i1} \mid D_i = 0, X_i] \right), \quad (13)$$

where $\hat{\mathbb{E}}[Y_{i2} - Y_{i1} \mid D_i = 0, X_i]$ is the estimated conditional expectation function fitted on the control units (but evaluated at X_i for a treated unit). We note that if one uses a linear model for $\hat{\mathbb{E}}[Y_{i2} - Y_{i1} \mid D_i = 0, X_i]$, then this would be similar to a modification of (12) that interacts X_i with both treatment group and time dummies, although the two are not quite

identical because the outcome regression approach re-weights using the distribution of X_i among units with $D_i = 1$. The researcher need not restrict themselves to linear models for the CEF, however, and can use more flexible semi-/non-parametric methods instead. One popular approach in empirical practice is to match each treated unit to a “nearest neighbor” untreated unit with similar (or identical) covariate values, and then estimate $\hat{\mathbb{E}}[Y_{i2} - Y_{i1}|D_i = 0, X_i]$ using $Y_{l(i)2} - Y_{l(i)1}$, where $l(i)$ is the unit matched to i , in which case $\hat{\tau}_2$ reduces to the simple DiD estimator between treated units and the matched comparison group.

The outcome regression approach will generally be consistent for the ATT when the outcome model used to estimate $\hat{\mathbb{E}}[Y_{i2} - Y_{i1}|D_i = 0, X_i]$ is correctly specified. Inference can be done using the delta-method for parametric models, and there are also several methods available for semi-/non-parametric models (under some additional regularity conditions), including the bootstrap, as described in Heckman et al. (1998). Inference is more complicated, however, when one models the outcome evolution of untreated units using a nearest-neighbor approach with a fixed number of matches: the resulting estimator is no longer asymptotically linear and thus standard bootstrap procedures are not asymptotically valid (e.g., Abadie and Imbens, 2006, 2008, 2011, 2012). Ignoring the matching step can also cause problems, and one therefore needs to use inference procedures that accommodate matching as described in the aforementioned papers.⁸

Inverse probability weighting An alternative to modeling the conditional expectation function is to instead model the propensity score, i.e. the conditional probability of belonging to the treated group given covariates, $p(X_i) = P(D_i = 1|X_i)$. Indeed, as shown by Abadie (2005), under Assumptions 2, 6 and 7, the ATT is identified using the following inverse probability weighting (IPW) formula:

$$\tau_2 = \frac{\mathbb{E} \left[\left(D_i - \frac{(1 - D_i)p(X_i)}{1 - p(X_i)} \right) (Y_{i2} - Y_{i1}) \right]}{\mathbb{E}[D_i]}. \quad (14)$$

As in the regression adjustment approach, researchers can use the “plug-in principle” to estimate the ATT by plugging in an estimate of the propensity score to the equation above,

⁸Although these nearest-neighbor procedures have been formally justified for cross-sectional data, they are easily adjustable to the canonical 2x2 DiD setup with balanced panel data. We are not aware of formal extensions that allows for unbalanced panel data, repeated cross-sectional data, or more general DiD designs.

$$\hat{\tau}_2 = \frac{\frac{1}{N} \sum_{i=1}^N \left(D_i - \frac{(1 - D_i)\hat{p}(X_i)}{1 - \hat{p}(X_i)} \right) (Y_{i2} - Y_{i1})}{\frac{1}{N} \sum_{j=1}^N D_j}. \quad (15)$$

Similar to the regression adjustment approach, the propensity score model can be estimated using parametric models or semi-/non-parametric models (under suitable regularity conditions). The IPW approach will generally be consistent if the model for the propensity scores is correctly specified. Inference can be conducted using standard tools; see, e.g., [Abadie \(2005\)](#).⁹

Doubly-robust estimators The outcome regression and IPW approaches described above can also be combined to form “doubly-robust” (DR) methods that are valid if either the outcome model or the propensity score model is correctly specified. Specifically, [Sant’Anna and Zhao \(2020\)](#) show that under Assumptions 2, 6 and 7, the ATT is identified as:

$$\tau_2 = \mathbb{E} \left[\left(\frac{D_i}{\mathbb{E}[D_i]} - \frac{\frac{(1 - D_i)p(X_i)}{1 - p(X_i)}}{\mathbb{E} \left[\frac{(1 - D_i)p(X_i)}{1 - p(X_i)} \right]} \right) (Y_{i2} - Y_{i1} - \mathbb{E}[Y_{i2} - Y_{i1} | D_i = 0, X_i]) \right]. \quad (16)$$

As before, one can then estimate the ATT by plugging in estimates of both the propensity score and the CEF,

$$\hat{\tau}_2 = \frac{1}{N} \sum_{i=1}^N \left(\frac{D_i}{\frac{1}{N} \sum_{j=1}^N D_j} - \frac{\frac{(1 - D_i)\hat{p}(X_i)}{1 - \hat{p}(X_i)}}{\frac{1}{N} \sum_{j=1}^N \frac{(1 - D_j)\hat{p}(X_j)}{1 - \hat{p}(X_j)}} \right) (Y_{i2} - Y_{i1} - \hat{\mathbb{E}}[Y_{i2} - Y_{i1} | D_i = 0, X_i]).$$

The outcome equation and the propensity score can be modeled with either parametric or semi-/non-parametric methods, and DR methods will generally be consistent if either of these models is correctly specified. In addition, [Chang \(2020\)](#) shows that data-adaptive/machine-

⁹One drawback of the IPW approach is that it can lead to very imprecise inference if the estimated propensity scores are close to zero, which may occur if either the estimates of the propensity score are imprecise or Assumption 7 is close to being violated. [Sant’Anna and Zhao \(2020\)](#) show that this instability can be partially addressed by using “normalized” weights.

learning methods can also be used with DR methods. Standard inference tools can be used as well; see, e.g., [Sant’Anna and Zhao \(2020\)](#). Finally, under some regularity conditions, the DR estimator achieves the semi-parametric efficiency bound when both outcome evolution and propensity score models are correctly specified ([Sant’Anna and Zhao, 2020](#)).

Extensions to staggered treatment timing: Although the discussion above focused on DiD setups with two groups and two periods, these different procedures have been extended to staggered DiD setups when treatments are binary and non-reversible. More precisely, [Callaway and Sant’Anna \(2021\)](#) extend the regression adjustment, IPW and DR procedures above to estimate the family of $ATT(g, t)$ ’s discussed in Section 3.3. They then aggregate these estimators to form different treatment effect summary measures. [Wooldridge \(2021\)](#) proposes an alternative regression adjustment procedure that is suitable for staggered setups. His proposed estimator differs from the [Callaway and Sant’Anna \(2021\)](#) regression adjustment estimator as he exploits additional information from pre-treatment periods, which, in turn, can lead to improvements in precision. On the other hand, if these additional assumptions are violated, [Wooldridge \(2021\)](#)’s estimator may be more biased than [Callaway and Sant’Anna \(2021\)](#)’s. [de Chaisemartin and D’Haultfœuille \(2020b,a\)](#) consider an estimator which includes covariates in a linear manner.

Caveats. Throughout, we assume that the covariates X_i were measured prior to the introduction of the intervention and are, therefore, unaffected by it. If X_i can in fact be affected by treatment, then conditioning on it induces a “bad control” problem that can induce bias; see [Zeldow and Hatfield \(2021\)](#) for additional discussion.

Another important question relates to whether researchers should condition on pre-treatment outcomes. Proponents of including pre-treatment outcomes argue that controlling for lagged outcomes can reduce bias from unobserved confounders ([Ryan, 2018](#)).¹⁰ However, conditioning on lagged outcomes need not necessarily reduce bias. For example, [Daw and Hatfield \(2018\)](#) show that when the treated and comparison groups have different outcome distributions, matching the treated and control groups on lagged outcomes selects control units with a particularly large “shock” in the pre-treatment period. This can then induce bias owing to a mean-reversion effect, when in fact not conditioning on lagged outcomes would have produced parallel trends. See, also [Chabé-Ferret \(2015\)](#) and [Ding and Li \(2019\)](#) for related discussion.

¹⁰We note that when lagged outcomes are included, the conditional parallel trends assumption actually reduces to a conditional mean independence assumption for the untreated potential outcome, since the $Y_{i1}(0)$ terms on both sides of (10) cancel out.

4.3 Testing for pre-existing trends

Although conditioning on pre-existing covariates can help increase the plausibility of the parallel trends assumption, researchers typically still worry that there remain unobserved time-varying confounders. An appealing feature of the DiD design is that it allows for a natural plausibility check on the identifying assumptions: did outcomes for the treated and comparison groups (possibly conditional on covariates) move in parallel prior to the time of treatment? It has therefore become common practice to check, both visually and using statistical tests, whether there exist pre-existing differences in trends (“pre-trends”) as a test of the plausibility of the parallel trends assumption.

To fix ideas, consider a simple extension of the canonical non-staggered DiD model in Section 2 in which we observe outcomes for an additional period $t = 0$ during which no units were treated. (These ideas will extend to the case of staggered treatment or conditional parallel trends). By the no-anticipation assumption, $Y_{it} = Y_{it}(0)$ for all units in periods $t = 0$ and $t = 1$. We can thus check whether the analog to the parallel trends assumption held between periods 0 and 1 — that is, is

$$\underbrace{\mathbb{E}[Y_{i1} - Y_{i0} | D_i = 1]}_{\text{Pre-treatment change for } D_i = 1} - \underbrace{\mathbb{E}[Y_{i1} - Y_{i0} | D_i = 0]}_{\text{Pre-treatment change for } D_i = 0} = 0?$$

In the non-staggered setting, this hypothesis can be conveniently tested using a TWFE specification that includes leads and lags of treatment,

$$Y_{it} = \alpha_i + \phi_t + \sum_{r \neq 0} 1[R_{it} = r] \beta_r + \epsilon_{it}, \quad (17)$$

where the coefficient on the lead of treatment $\hat{\beta}_{-1}$ is given by

$$\hat{\beta}_{-1} = \frac{1}{N_1} \sum_{i:D_i=1} Y_{i0} - Y_{i1} - \frac{1}{N_0} \sum_{i:D_i=0} Y_{i0} - Y_{i1}.$$

Testing for pre-treatment trends thus is equivalent to testing the null hypothesis that $\beta_{-1} = 0$. This approach is convenient to implement and extends easily to the case with additional pre-treatment periods and non-staggered treatment adoption. When there are multiple pre-treatment periods, it is common to plot the $\hat{\beta}_r$ in what is called as an “event-study” plot. If all of the pre-treatment coefficients (i.e., $\hat{\beta}_r$ for $r < 0$) are insignificant, this is usually interpreted as a sign in favor of the validity of the design, since we cannot reject the null that parallel trends was satisfied in the pre-treatment period.

This pre-testing approach extends easily to settings with staggered adoption and/or con-

ditional parallel trends assumptions. For example, the Callaway and Sant’Anna (2021) estimator can be used to construct “placebo” estimates of ATT_l^w for $l < 0$, i.e. the ATT l periods *before* treatment. The estimates \widehat{ATT}_l^w can be plotted for different values of l (corresponding to different lengths of time before/after treatment) to form an event-study plot analogous to that for the non-staggered case. This illustrates that the idea of testing for pre-trends extends easily to the settings with staggered treatment adoption or conditional parallel trends, since the Callaway and Sant’Anna (2021) can be applied for both of these settings. These results are by no means specific to the Callaway and Sant’Anna (2021) estimator, though, and event-study plots can be created in a similar fashion using other estimators for either staggered or conditional DiD settings. We caution, however, against using dynamic TWFE specifications like (17) in settings with staggered adoption, since as noted by Sun and Abraham (2021), the coefficients β_r may be contaminated by treatment effects at relative time $r' > 0$, so with heterogeneous treatment effects the pre-trends test may reject even if parallel trends holds in the pre-treatment period.

4.4 Issues with testing for pre-trends

Although tests of pre-existing trends are a natural and intuitive plausibility check of the parallel trends assumption, recent research has highlighted that they also have several limitations. First, even if pre-trends are exactly parallel, this need not guarantee that the post-treatment parallel trends assumption is satisfied. Kahn-Lang and Lang (2020) give an intuitive example: the average height of boys and girls evolves in parallel until about age 13 and then diverges, but we should not conclude from this that there is a causal effect of bar mitzvahs (which occur for boys at age 13) on children’s height!

A second issue is that even if there are pre-existing differences in trends, the tests described above may fail to reject owing to low power (Bilinski and Hatfield, 2018; Freyaldenhoven, Hansen and Shapiro, 2019; Kahn-Lang and Lang, 2020; Roth, Forthcoming). That is, even if there is a pre-existing trend, it may not be significant in the data if our pre-treatment estimates are imprecise.

To develop some intuition for why power may be low, suppose that there is no true treatment effect but there is a pre-existing linear difference in trends between the treatment and comparison groups. Then in the simple example from above, the pre-treatment and post-treatment event-study coefficients will have the same magnitude, $|\beta_{-1}| = |\beta_1|$. If the two estimated coefficients $\hat{\beta}_{-1}$ and $\hat{\beta}_1$ also have the same sampling variance, then by symmetry the chance that we find a significant pre-trend will be the same as the probability we find a significant post-treatment effect. But this means that a pre-trend that we detect only half the

time will also lead us to find a significant treatment effect half the time — that is, 10 times more often than we expect to find a spurious effect using a nominal 95% confidence interval! Another intuition for this phenomenon, given by [Bilinski and Hatfield \(2018\)](#), is that pre-trends tests reverse the traditional roles of type I and type II error: they set the assumption of parallel trends (or no placebo pre-intervention effect) as the null hypothesis and only “reject” the assumption if there is strong evidence against it. This controls the probability of finding a violation when parallel trends holds tightly at 5% (or another chosen α -level), but the probability of failing to identify a violation can be much higher, corresponding to type II error of the test.

In addition to being concerning from a theoretical point of view, the possibility of low power appears to be relevant in practice: in simulations calibrated to papers published in three leading economics journals, [Roth \(Forthcoming\)](#) found that linear violations of parallel trends that conventional tests would detect only 50% of the time often produce biases as large as (or larger than) the estimated treatment effect.

A third issue with pre-trends testing is that conditioning the analysis on “passing” a pre-trends test induces a selection bias known as pre-test bias ([Roth, Forthcoming](#)). Intuitively, if there is a pre-existing difference in trends in population, the draws from the DGP in which we fail to detect a significant pre-trend are a selected sample from the true DGP. [Roth \(Forthcoming\)](#) shows that in many cases, this additional selection bias can exacerbate the bias from a violation of parallel trends.

A final issue with the current practice of pre-trends testing relates to what happens if we do detect a significant pre-trend. In this case, the pre-trends test suggests that parallel trends is likely not to hold exactly, but researchers may still wish to learn something about the treatment effect of interest. Indeed, it seems likely that with enough precision, we will nearly always reject that the parallel trends assumption holds exactly in the pre-treatment period. Nevertheless, we may still wish to learn something about the treatment effect, especially if the violation of parallel trends is “small” in magnitude. However, the conventional approach does not make clear how to proceed in this case.

4.4.1 Improved diagnostic tools

A few papers have proposed alternative tools for detecting pre-treatment violations of parallel trends that take into account some of the limitations discussed above. [Roth \(Forthcoming\)](#) developed tools to conduct power analyses and calculate the likely distortions from pre-testing under researcher-hypothesized violations of parallel trends. These tools allow the researcher to assess whether the limitations described above are likely to be severe for potential violations of parallel trends deemed economically relevant.

Bilinski and Hatfield (2018) and Dette and Schumann (2020) propose “non-inferiority” approaches to pre-testing that help address the issue of low power by reversing the roles of the null and alternative hypotheses. That is, rather than test the null that pre-treatment trends are zero, they test the null that the pre-treatment trend is large, and reject only if the data provides strong evidence that the pre-treatment trend is small. For example, Dette and Schumann (2020) consider null hypotheses of the form $H_0 : \max_{r < 0} |\beta_r| \geq c$, where β_r are the (population) pre-treatment coefficients from regression (17). This ensures that the test “detects” a pre-trend with probability at least $1 - \alpha$ when in fact the pre-trend is large (i.e. has magnitude at least c).

These non-inferiority approaches are an improvement over standard pre-testing methods, since they guarantee by design that the pre-test is powered against large pre-treatment violations of parallel trends. However, using these approaches does not provide any formal guarantees that ensure the validity of confidence intervals for the treatment effect, the main object of interest. They also do not avoid statistical issues related to pre-testing (Roth, Forthcoming), and do not provide clear guidance on what to do when the test fails to reject the null of a large pre-trend. This has motivated more formal robust inference and sensitivity analysis approaches that consider inference on the ATT when parallel trends may be violated.

4.5 Robust inference and sensitivity analysis

Bounds using pre-trends. Rambachan and Roth (2021) propose an approach for robust inference and sensitivity analysis when parallel trends may be violated, building on earlier work by Manski and Pepper (2018). Their approach attempts to formalize the intuition motivating pre-trends testing: that the counterfactual post-treatment trends cannot be “too different” from the pre-trends. To fix ideas, consider the non-staggered treatment adoption setting described in Section 4.4. Denote by δ_1 the violation of parallel trends in the first post-treatment period:

$$\delta_1 = \mathbb{E}[Y_{i2}(0) - Y_{i1}(0)|D_i = 1] - \mathbb{E}[Y_{i2}(0) - Y_{i1}(0)|D_i = 0].$$

The bias δ_1 is unfortunately not directly identified, since we do not observe the untreated potential outcomes, $Y_{i2}(0)$, for the treated group. However, by the no anticipation assumption, we can identify the *pre-treatment* analog to δ_1 ,

$$\delta_{-1} = \mathbb{E}[Y_{i0}(0) - Y_{i1}(0)|D_i = 1] - \mathbb{E}[Y_{i0}(0) - Y_{i1}(0)|D_i = 0],$$

which looks at pre-treatment differences in trends between the groups, with $\delta_{-1} = \mathbb{E}[\hat{\beta}_{-1}]$

from the event study regression (17).

Rambachan and Roth (2021) then consider robust inference under assumptions that restrict the possible values of δ_1 given the value of δ_{-1} — or more generally, given $\delta_{-1}, \dots, \delta_{-K}$ if there are K pre-treatment coefficients. For example, one type of restriction they consider states that the magnitude of the post-treatment violation of parallel trends can be no larger than a constant \overline{M} times the maximal pre-treatment violation, i.e. $|\delta_1| \leq \overline{M} \max_{r < 0} |\delta_r|$. They also consider restrictions that bound the extent that δ_1 can deviate from a linear extrapolation of the pre-treatment differences in trends. Rambachan and Roth (2021) use tools from the partial identification and sensitivity analysis literature (Armstrong and Kolesár, 2018; Andrews, Pakes and Roth, 2019) to construct confidence sets for the ATT that are uniformly valid under the imposed restrictions. These confidence sets take into account the fact that we do not observe the true pre-treatment difference in trends δ_{-1} , only an estimate $\hat{\beta}_{-1}$. In contrast to conventional pre-trends tests, the Rambachan and Roth (2021) confidence sets thus tend to be larger when there is more uncertainty about the pre-treatment difference in trends (i.e. when the standard error on $\hat{\delta}_{-1}$ is large).

This approach enables a natural form of sensitivity analysis. For example, a researcher might report that the conclusion of a positive treatment effect is robust up to the value $\overline{M} = 2$. This indicates that to invalidate the conclusion of a positive effect, we would need to allow for a post-treatment violation of parallel trends two times larger than the maximal pre-treatment violation.

It is worth highlighting that although we’ve described these tools in the context of non-staggered treatment timing and an unconditional parallel trends assumption, they extend easily to the case of staggered treatment timing and conditional parallel trends as well. Indeed, under mild regularity conditions, these tools can be used anytime the researcher has a treatment effect estimate $\hat{\beta}_{post}$ and a placebo estimate $\hat{\beta}_{pre}$, so long as she is willing to bound the possible bias of $\hat{\beta}_{post}$ given the expected value of $\hat{\beta}_{pre}$. For example, in the staggered setting, $\hat{\beta}_{post}$ could be an estimate of ATT_l^w for $l > 0$ using one of the estimators described in Section 3.3, and $\hat{\beta}_{pre}$ could be placebo estimates of $ATT_{-1}^w, \dots, ATT_{-k}^w$. See https://github.com/pedrohcgcs/CS_RR for examples on how these sensitivity analyses can be combined with the Callaway and Sant’Anna (2021) estimator in R.

Bounds using bracketing. Ye, Keele, Hasegawa and Small (2021) consider an alternative partial identification approach where there are two control groups whose trends are assumed to “bracket” that of the treatment group. Consider the canonical model from Section 2, and suppose the untreated units can be divided into two control groups, denoted $C_i = a$ and $C_i = b$. For ease of notation, let $C_i = trt$ denote the treated group, i.e. units with $D_i = 1$.

Let $\Delta(c) = \mathbb{E}[Y_{i2}(0) - Y_{i1}(0)|C_i = c]$. Instead of the parallel trends assumption, [Ye et al. \(2021\)](#) impose that

$$\min\{\Delta(a), \Delta(b)\} \leq \Delta(trt) \leq \max\{\Delta(a), \Delta(b)\}, \quad (18)$$

so that the trend in $Y(0)$ for the treated group is bounded above and below (“bracketed”) by the minimum and maximum in groups a and b . An intuitive example where we may have such bracketing is if each of the groups corresponds with a set of industries, and one of the control groups (say group a) is more cyclical than the treated group while the other (say group b) is less cyclical. If the economy was improving between periods $t = 1$ and $t = 2$, then we would expect group a to have the largest change in the outcome and group b to have the smallest change; whereas if the economy was getting worse, we would expect the opposite. Under equation (18) and the no anticipation assumption, the ATT is bounded,

$$\mathbb{E}[Y_{i1}|D_i = 1] + \min\{\Delta(a), \Delta(b)\} \leq \tau_2 \leq \mathbb{E}[Y_{i1}|D_i = 1] + \max\{\Delta(a), \Delta(b)\}.$$

This reflects that if we knew the true counterfactual trend for the treated group we could learn the ATT exactly, and therefore that bounding this trend means we can obtain bounds on the ATT. [Ye et al. \(2021\)](#) further show how one can construct confidence intervals for the ATT, and extend this logic to settings with multiple periods (but non-staggered treatment timing). See, also, [Hasegawa, Webster and Small \(2019\)](#) for a related, earlier approach.

4.5.1 Other approaches.

[Keele, Small, Hsu and Fogarty \(2019\)](#) propose a sensitivity analysis in the canonical two-period DiD model that summarizes the strength of confounding factors that would be needed to induce a particular bias. [Freyaldenhoven, Hansen, Pérez Pérez and Shapiro \(2021\)](#) propose a visual sensitivity analysis in which one plots the “smoothest” trend through an event-study plot that could rationalize the data under the null of no effect. Finally, [Freyaldenhoven et al. \(2019\)](#) propose a GMM-based estimation strategy that allows for parallel trends to be violated when there exists a covariate assumed to be affected by the same confounds as the outcome but not by the treatment itself.

4.6 Recommendations

We suspect that in most practical applications of DiD, researchers will not be confident ex ante that the parallel trends assumption holds exactly, owing to concerns about time-varying confounds and sensitivity to functional form. The methods discussed in this setting

for relaxing the parallel trends assumption and/or assessing sensitivity to violations of the parallel trends assumption will therefore be highly relevant in most contexts where DiD is applied.

A natural starting point for these robustness checks is to consider whether the results change meaningfully when imposing parallel trends only conditional on covariates. Among the different estimation procedures we discussed, we view doubly-robust procedures as a natural default, since they are valid if either the outcome model or propensity score is well-specified and have desirable efficiency properties. A potential exception to this recommendation arises in settings with limited overlap, i.e., when the estimated propensity score is close to 0 or 1, in which case regression adjustment estimators may be preferred.

Whether one includes covariates into the DiD analysis or not, we encourage researchers to continue to plot “event-study plots” that allow for a visual evaluation of pre-existing trends. These plots convey useful information for the reader to assess whether there appears to have been a break in the outcome for the treatment group around the time of treatment. In contexts with a common treatment date, such plots can be created using TWFE specifications like (17); in contexts with staggered timing, we recommend plotting estimates of ATT_l^w for different values of l using one of the estimators for the staggered setting described in Section 3.3 to avoid negative weighting issues with TWFE. See Section 4.3 for additional discussion. We also refer the reader to Freyaldenhoven et al. (2021) regarding best-practices for creating such plots, such as displaying simultaneous (rather than pointwise) confidence bands for the path of the event-study coefficients (Olea and Plagborg-Møller, 2019; Callaway and Sant’Anna, 2021).

While event-study plots play an important role in evaluating the plausibility of the parallel trends assumption, we think it is important to appreciate that tests of pre-trends may be underpowered to detect relevant violations of parallel trends, as discussed in Section 4.4. At minimum, we recommend that researchers assess the power of pre-trends tests against economically relevant violations of parallel trends, or conduct non-inferiority tests against pre-trends that are determined to be “economically large”, as described in Section 4.4.1.

We also think it should become standard practice for researchers to report sensitivity analyses that formalize the extent to which their conclusions are sensitive to violations of parallel trends, using the tools described in Section 4.5. We note that employing such sensitivity analyses requires the researcher to evaluate what types of violations of parallel trends are plausible (e.g, how much larger could the counterfactual difference in trends be than the pre-trend?). We encourage researchers to use context-specific knowledge about the potential confounding factors to inform such judgments. For example, in contexts where other factors were quite stable around the time of the treatment, it may be implausible that

the counterfactual difference in trends could be more than twice as large as the magnitude of the pre-trend. On the other hand, if the treatment occurred at the beginning of a recession larger than anything seen in the pre-treatment period, then a pre-trend of that magnitude may indeed be plausible. We therefore think that the most scientific approach to dealing with possible violations of parallel trends involves a combination of state-of-the-art econometric tools and context-specific knowledge about the types of plausible confounding factors.

5 Relaxing sampling assumptions

We now discuss a third strand of the DiD literature, which considers inference under deviations from the canonical assumption that we have sampled a large number of independent clusters from a super-population.

5.1 Inference procedures with few clusters

As described in Section 2, standard DiD inference procedures rely on researchers having access to data on a large number of treated and untreated clusters. Confidence intervals are then based on the central limit theorem, which states that with independently-sampled clusters, the DiD estimator has an asymptotically normal distribution as the number of treated and untreated clusters grows large. In many practical DiD settings, however, the number of independent clusters (and, in particular, treated clusters) may be small, so that the central limit theorem based on a growing number of clusters may provide a poor approximation. For example, many DiD applications using state-level policy changes may only have a handful of treated states. We note that the central limit theorem may provide a poor approximation with few clusters, even if the number of units within each cluster is large. This is because the standard sampling-based view of clustering allows for arbitrary correlations of the outcome within each cluster, and thus there may be common components at the cluster level (a.k.a. cluster-level “shocks”) that do not wash out when averaging over many units within the same cluster. Since we only observe a few observations of the cluster-specific shocks, the average of these shocks will generally not be approximately normally distributed.

Model-based approaches. Several papers have made progress on the difficult problem of conducting inference with a small number of clusters by modeling the dependence within clusters. These papers typically place some restrictions on the common cluster-level shocks, although the exact restrictions differ across papers. The starting point for these papers is

typically a structural equation of the form

$$Y_{ijt} = \alpha_j + \phi_t + D_{jt}\beta + (\nu_{jt} + \epsilon_{ijt}), \quad (19)$$

where Y_{ijt} is the (realized) outcome of unit i in cluster j at time t , α_j and ϕ_t are cluster and time fixed effects, D_{jt} is an indicator for whether cluster j is treated in period t , ν_{jt} is a common cluster-by-time error term, and ϵ_{ijt} is an idiosyncratic unit-level error term. Here, the “cluster-level” error term, ν_{jt} , induces correlation among units within the same cluster. It is usually assumed that ϵ_{ijt} are *iid* mean-zero across i and j (and sometimes t); see, e.g., Donald and Lang (2007), Conley and Taber (2011), and Ferman and Pinto (2019). Letting $Y_{ij} = n_j^{-1} \sum_{i:j(i)=j} Y_{ijt}$ be the average outcome among units in cluster j , where n_j is the number of units in cluster j , we can take averages to obtain

$$Y_{jt} = \alpha_j + \phi_t + D_{jt}\beta + \eta_{jt}, \quad (20)$$

where $\eta_{jt} = \nu_{jt} + n_j^{-1} \sum_{i=1}^{n_j} \epsilon_{ijt}$. Assuming the canonical set-up with two periods where no clusters are treated in period $t = 1$ and some clusters are treated in period $t = 2$, the canonical DiD estimator at the cluster level is equivalent to the OLS estimated coefficient $\hat{\beta}$ from (20), and is given by

$$\begin{aligned} \hat{\beta} &= \beta + \frac{1}{N_1} \sum_{j:D_j=1} \Delta\eta_j - \frac{1}{N_0} \sum_{j:D_j=0} \Delta\eta_j \\ &= \beta + \frac{1}{N_1} \sum_{j:D_j=1} \left(\Delta\nu_j + n_j^{-1} \sum_{i=1}^{n_j} \Delta\epsilon_{ij} \right) - \frac{1}{N_0} \sum_{j:D_j=0} \left(\Delta\nu_j + n_j^{-1} \sum_{i=1}^{n_j} \Delta\epsilon_{ij} \right), \end{aligned} \quad (21)$$

where now N_d corresponds with the number of *clusters* with treatment d , and $\Delta\eta_j = \eta_{j2} - \eta_{j1}$ (and likewise for the other variables). The equation in the previous display highlights the challenge in this setup: with few clusters, the averages of the $\Delta\nu_j$ among treated and untreated clusters will tend not to be approximately normally distributed, and their variance may be difficult to estimate.

It is worth highlighting that the model described above starts from the structural equation (19) rather than a model where the primitives are potential outcomes as in Section 2. We think that connecting the assumptions on the errors in the structural model (19) to restrictions on the potential outcomes is an interesting open topic for future work. Although a general treatment is beyond the scope of this paper, in Appendix A we derive how the errors in the structural model (19) map to primitives based on potential outcomes in the canonical model from Section 2. Specifically, we show that under the set-up of Section 2,

Assumptions 1 and 2 imply that the canonical DiD estimator takes the form given in (21), where $\beta = \tau_2$ is the ATT at the cluster level, $\nu_{jt} = \nu_{jt,0} + D_j \nu_{jt,1}$ and $\epsilon_{ijt} = \epsilon_{ijt,0} + D_j \epsilon_{ijt,1}$, where¹¹

$$\begin{aligned}\epsilon_{ijt,0} &= Y_{ijt}(0) - \mathbb{E}[Y_{ijt}(0)|j(i) = j] \\ \epsilon_{ijt,1} &= Y_{ijt}(1) - Y_{ijt}(0) - \mathbb{E}[Y_{ijt}(1) - Y_{ijt}(0)|j(i) = j] \\ \nu_{jt,0} &= \mathbb{E}[Y_{ijt}(0)|j(i) = j] - \mathbb{E}[Y_{ijt}(0)|D_j] \\ \nu_{jt,1} &= \mathbb{E}[Y_{ijt}(1) - Y_{ijt}(0)|j(i) = j] - \tau_t.\end{aligned}$$

Thus, in the canonical set-up, restrictions on ν_{jt} and ϵ_{ijt} can be viewed as restrictions on primitives that are functions of the potential outcomes. For the remainder of the sub-section, however, we focus primarily on the restrictions placed on ν_{jt} and ϵ_{ijt} directly — rather than the implications of these assumptions for the potential outcomes — since this simplifies exposition and matches how these assumptions are stated in the literature.

When the number of treated and untreated clusters is fixed, Donald and Lang (2007) assume that the “cluster-specific” shocks ν_{jt} are mean-zero Gaussian, homoskedastic with respect to cluster and treatment status, and independent of other unit-and-time specific shocks. When each cluster is sufficiently large ($n_j \rightarrow \infty$ for all j), these conditions imply that η_{jt} in (20) are (asymptotically) normal, with mean zero and a homoskedastic variance. This allows one to conduct inference using critical values from a t -distribution with $J - 2$ degrees of freedom, where J is the total number of clusters. The key restriction in this approach is the assumption that the cluster-specific shocks ν_{jt} are *iid* normal. In addition to imposing a potentially strong distributional assumption on the shape of the errors, the homoskedasticity assumption also rules out many forms of treatment effect heterogeneity. For example, suppose the cluster-level means of $Y_{it}(0)$ have the same distribution among treated and control clusters. Then if the average treatment effect at the cluster level is heterogeneous, this will tend to lead ν_{jt} to have higher variance among treated clusters, thus violating the homoskedasticity assumption.

Conley and Taber (2011) consider an alternative inference procedure to Donald and Lang (2007) that does not rely on the cluster-specific shocks ν_{jt} being Gaussian. Instead, they consider a setup where the number of treated clusters, J_1 , is fixed and potentially equal to one, but there are a large number of untreated clusters, J_0 , available. The main insight of Conley and Taber (2011) is that, if the cluster-specific error terms η_{jt} from the untreated group are informative about the cluster-specific error terms for the treated group, one can

¹¹In what follows, we write $\mathbb{E}[Y_{ijt}(0)|D_j = d]$ to denote the expectation where one first draws j from the population with $D_j = d$ and then draws $Y_{ijt}(0)$ from that cluster.

conduct inference about β using the estimated distribution of the untreated errors. In order to satisfy the “informativeness” condition, [Conley and Taber \(2011\)](#) impose the following restrictions: the idiosyncratic errors ϵ_{ijt} are *iid* across i and independent of clusters and treatment status, the cluster-specific shocks ν_{jt} are *iid* across j , independent of treatment status, and have mean zero for all t , and that all clusters grow at the same rate as J_0 . Based on these conditions, one can construct confidence intervals for β using the estimated distribution of the clustered error-terms. Although the conditions imposed by [Conley and Taber \(2011\)](#) are arguably weaker than those in [Donald and Lang \(2007\)](#), they still rule out important types of treatment effect heterogeneity. For instance, if average treatment effects differ across clusters, then this will tend to violate the assumption that ν_{jt} is *iid* across j . Another limitation of the [Conley and Taber \(2011\)](#) procedure is that it does not accommodate settings with heterogeneous cluster sizes, a situation that often arises in practice.

[Ferman and Pinto \(2019\)](#) build on [Conley and Taber \(2011\)](#) and show how one can use bootstrap-based inference procedures to allow for some types of heteroskedasticity, paying particular attention to the case where heteroskedasticity arises due to variation in cluster sizes. The key insight of [Ferman and Pinto \(2019\)](#) is that as long as a time-average of the cluster-specific error terms is *iid* across clusters, independent of treatment status, and that one is able to correctly model the conditional variance of these shocks (conditional on cluster size), one can rescale the untreated error terms to account for heteroskedasticity with respect to cluster-size and then use these rescaled residuals in a bootstrap procedure. Here, it is also worth mentioning that cluster size is assumed to affect the outcome only through the variance term, ruling out other types of heterogeneity.

[Hagemann \(2020\)](#) considers a rearrangement/permutation-based method that is applicable to DiD setups with a single large treated cluster and a fixed number of large untreated clusters, provided the average evolution of the untreated outcomes is the same across all untreated clusters. [Hagemann \(2020\)](#)’s proposal allows for heteroskedastic error terms and, in contrast to [Ferman and Pinto \(2019\)](#), it does not require one to directly estimate (or model) the heteroskedasticity. The key insight of [Hagemann \(2020\)](#) is that if we place a bound on the maximal relative heterogeneity across clusters, then we can bound the probability of type I error from a permutation approach. He also shows how one can use this measure of relative heterogeneity to do sensitivity analysis. Like the other proposals above, though, [Hagemann \(2020\)](#)’s approach must also place some strong restrictions on certain types of heterogeneity. In particular, his approach essentially requires that, as cluster size grows large, any single untreated cluster could be used to infer the counterfactual trend for the treated group, and thus his approach rules out cluster-specific heterogeneity in trends in untreated potential outcomes.

Another popular solution with few clusters is the cluster wild bootstrap. In an influential paper, [Cameron, Gelbach and Miller \(2008\)](#) presented simulation evidence that the cluster wild bootstrap procedure can work well in settings with as few as five clusters. More recently, however, [Canay, Santos and Shaikh \(2021\)](#) provided a formal analysis about the conditions under which the cluster wild bootstrap procedure would be asymptotically valid in settings with a few large clusters. Importantly, [Canay et al. \(2021\)](#) show that the reliability of these bootstrap procedures depends on imposing certain homogeneity conditions on treatment effects, as well as the type of bootstrap weights one uses and the estimation method adopted (e.g., restricted vs. unrestricted OLS). In particular, these restrictions are commonly violated when one uses TWFE regressions with cluster-specific and time fixed effects like [\(20\)](#) or when treatment effects are allowed to be heterogeneous across clusters – see Examples 2 and 3 in [Canay et al. \(2021\)](#). Thus, while the wild bootstrap may perform well in certain scenarios with a small number of clusters, it too requires strong homogeneity assumptions.

Finally, in settings with a large number of time periods, it may be feasible to conduct reliable inference with less stringent homogeneity assumptions about treatment effects. For instance, [Canay, Romano and Shaikh \(2017\)](#), [Ibragimov and Müller \(2016\)](#), and [Hagemann \(2021\)](#), respectively propose permutation-based, t -test based, and adjusted permutation based methods procedures that allow one to relax distributional assumptions about common shocks and accommodate richer forms of heterogeneity. The key restriction is that one is comfortable limiting the time-series dependence of the cluster-specific-shocks, and strengthening the parallel trends assumption to hold in many pre- and post-treatment time periods. All of these papers rely on asymptotics where the number of pre-treatment and post-treatment periods grows large, however, and thus may not perform well when in fact that number of periods is small. When the number of time periods is fixed, one can still use some of these methods, but the underlying assumptions are stronger — see, e.g., Remark 4.5 and Section 4.2 of [Canay et al. \(2017\)](#).

Alternative approaches. We now briefly discuss two alternative approaches in settings with a small number of clusters. First, while all of the “model-based” papers above treat ν_{jt} as random, an alternative perspective would be to condition on the values of ν_{jt} and view the remaining uncertainty as coming from the sampling of the individual units within clusters, constructing standard errors by clustering only at the unit level. This will generally produce a violation of parallel trends, but the violation may be relatively small if the cluster-specific shocks are small relative to the idiosyncratic variation. The violation of parallel trends could then be accounted for using the methods described in Section 4. To make this concrete, consider the setting of [Card and Krueger \(1994\)](#) that compares employment in NJ and PA

after NJ raised its minimum wage. The aforementioned papers would consider NJ and PA as drawn from a super-population of treated and untreated states, where the state-level shocks are mean-zero, whereas the alternative approach would treat the two states as fixed and view any state-level shocks between NJ and PA as a violation of the parallel trends assumption. One could then explore the sensitivity of one’s conclusions to the magnitude of this violation, potentially benchmarking it relative to the magnitude of the pre-treatment violations.

Second, a large literature in statistics and a growing literature in econometrics has considered Fisher Randomization Tests (FRTs), otherwise known as permutation tests. The basic idea is to calculate some statistic of the data (e.g. the t -statistic of the DiD estimator), and then recompute this statistic under many permutations of the treatment assignment (at the cluster level). We then reject the null hypothesis of no effect if the test statistic using the original data is larger than 95% of the draws of the test statistics under the permuted treatment assignment. Such tests have a long history in statistics, dating to [Fisher \(1935\)](#). If treatment is randomly assigned, then FRTs have exact finite-sample validity under the sharp null of no treatment effects for all units. The advantage of these tests is that they place no restrictions on the values of $Y(0)$, and thus allow arbitrary heterogeneity in $Y(0)$ across clusters. On the other hand, the assumption of random treatment assignment may often be questionable in DiD settings, as it is substantially stronger than parallel trends. Moreover, the “sharp” null of no effects for all units may not be as economically interesting as the “weak” null of no average effects. [Roth and Sant’Anna \(2021a\)](#) extend the idea of FRTs to settings where there is staggered adoption and (quasi-)random timing of treatment, and show that an FRT with a studentized statistic is both finite-sample valid for the sharp null and asymptotically valid (as the number of clusters grows) for the weak null.

Recommendations. In sum, recent research has made progress on the problem of conducting inference with relatively few clusters, but all of the available approaches require the researcher to impose some potentially strong additional assumptions. Most of the literature has focused on model-based approaches, which require the researcher to impose some homogeneity assumptions across clusters. Different homogeneity assumptions may be more reasonable in different contexts, and so we encourage researchers using these approaches to choose a method relying on a dimension of homogeneity that is most likely to hold (approximately) in their context. Permutation-based methods offer an intriguing alternative which require fewer assumptions about homogeneity in $Y(0)$, but require stronger assumptions on the assignment of treatment and test a potentially less interesting null hypothesis when the number of clusters is small.

5.2 Design-based inference

The canonical approach to inference in DiD views each unit (or cluster) as having been sampled from an infinite larger super-population. This view is natural when, for example, our sample is a labor market survey of workers drawn from a much larger population. The super-population view can be somewhat harder to conceptualize in other applications, for instance when all 50 US states are observed or we have administrative data on an entire population (Manski and Pepper, 2018). In such settings, it is often conceptually easier to view the N units in the sample as the fixed population of interest, and to think of the uncertainty in the data as arising only from the randomness in the treatment assignment vector $D = (D_1, \dots, D_N)'$. This so-called “design-based” view has a long history in statistics, dating to Neyman (1923), but has focused on experiments where treatment is assigned with known probabilities.

Recent work by Rambachan and Roth (2020) has extended this design-based view to settings like DiD. Rambachan and Roth (2020) consider a setting similar to the canonical two-period model in Section 2. However, rather than viewing the units as drawn from a super-population, they treat them as fixed (or conditioned on) – that is, the potential outcomes $Y_{it}(\cdot)$ and number of treated and untreated units (N_1, N_0) are treated as fixed. Only the treatment assignment vector D is considered to be random. In this set-up, they show that the usual DiD estimator is unbiased for a finite-population analog to the ATT under a finite-population analog to the parallel trends assumption. In particular, let π_i denote the probability that $D_i = 1$, and suppose that

$$\sum_{i=1}^N \left(\pi_i - \frac{N_1}{N} \right) (Y_{i2}(0) - Y_{i1}(0)) = 0,$$

so that treatment probabilities are uncorrelated with trends in $Y(0)$ (a finite-population version of parallel trends). Then $E_D[\hat{\tau}_2] = \tau_2^{Fin}$, where $\tau_2^{Fin} = E_D[\frac{1}{N_1} \sum_{i:D_i=1} (Y_{i2}(1) - Y_{i2}(0))]$ is a finite-population analog to the ATT, i.e. the expected average treatment effect on the treated, and the expectation is taken over the distribution of the treatment assignment vector.

Rambachan and Roth (2020) also show that when treatment assignment occurs independently at the unit level,¹² cluster-robust standard errors at the unit level provide valid (but potentially conservative) inference from the design-based perspective in large populations. Similarly, their results imply that if units i are nested in clusters j , and treatment is assigned

¹²Formally, when treatment follows a Poisson assignment mechanism, where units are assigned independently before we condition on the number treated, N_1 .

independently at the cluster level, then standard errors clustered at the j level are valid from the design-based perspective. These results suggest that it may not actually be a problem if it is difficult to conceptualize a super-population. Rather, the “usual” approach remains valid if we can conceptualize the treatment assignment as being stochastic.

These results in the design-based view can also be useful when determining the appropriate level of clustering. When the need to cluster arises from correlation in treatment assignment, then these results imply it is appropriate to cluster at the level at which treatment is independently assigned. Clustering at the level of treatment assignment thus seems like a good rule of thumb in many cases. [Abadie, Athey, Imbens and Wooldridge \(2017\)](#) study a model in which both treatment is random and units are sampled from a larger population, and suggest that one should cluster among units if either their treatment assignments are correlated or the event that they are included in the sample is correlated. Although the [Abadie et al. \(2017\)](#) results are for cross-sectional settings and do not directly apply to DiD, we suspect that a similar heuristic would apply in DiD as well in light of the results in [Rambachan and Roth \(2020\)](#) for the case where treatment is viewed as random. Formalizing this intuition strikes us as an interesting area for future research.

Recommendations. If it is difficult to conceptualize a super-population, fear not! Your DiD analysis can likely still be sensible from a finite-population perspective where we think of the treatment assignment as stochastic. Furthermore, if you are unsure about the appropriate level of clustering, a good rule of thumb (at least from the design-based perspective) is to cluster at the level at which treatment assignment is independently assigned (or at which units are independently sampled).

6 Other topics and areas for future research

In this section, we briefly touch on some other areas of interest in the DiD literature, and highlight some open areas for future research.

Distributional treatment effects. The DiD literature typically focuses on estimation of the ATT, but researchers may often be interested in the effect of a treatment on the entire distribution of an outcome. [Athey and Imbens \(2006\)](#) propose the Changes-in-Changes model, which allows one to infer the full counterfactual distribution of $Y(0)$ for the treated group in DiD setups. The key assumption is that the mapping between quantiles of $Y(0)$ for the treated and comparison groups remains stable over time – e.g., if the 30th percentile of the outcome for the treated group was the 70th percentile for the comparison group prior

to treatment, this relationship would have been preserved in the second period if treatment had not occurred. [Bonhomme and Sauder \(2011\)](#) propose an alternative distributional DiD model based on a parallel trends assumption for the (log of the) characteristic function, which is motivated by a model of test scores. [Callaway and Li \(2019\)](#) propose a distributional DiD model based on a copula stability assumption. Finally, [Roth and Sant’Anna \(2021b\)](#) show that parallel trends holds for all functional forms under a “parallel trends”-type assumption for the cumulative distribution of $Y(0)$, and this assumption also allows one to infer the full counterfactual distribution for the treated group.

Quasi-random treatment timing. In settings with staggered treatment timing, the generalized parallel trends assumption is often justified by arguing that the timing of treatment is random or quasi-random. [Roth and Sant’Anna \(2021a\)](#) show that if one is willing to assume treatment timing is as good as random, one can obtain more efficient estimates than using the staggered DiD methods discussed in [Section 3.3](#). This builds on earlier work by [McKenzie \(2012\)](#), who highlighted that DiD is typically inefficient in an RCT where lagged outcomes are observed, as well as a large literature in statistics on efficient covariate adjustment in randomized experiments (e.g., [Lin, 2013](#)). [Shaikh and Toulis \(2021\)](#) propose a method for observational settings where treatment timing is random conditional on fixed observable characteristics. We think that developing methods for observational settings where treatment timing is approximately random, possibly conditional on covariates and lagged outcomes, is an interesting area for further study in the years ahead.

Sequential random assignment. As discussed in [Section 3.4](#), an exciting new literature in DiD has begun to focus on settings where treatment can turn on and off and potential outcomes depend on the full path of treatments. A similar setting has been studied extensively in biostatistics, beginning with the pioneering work of [Robins \(1986\)](#). The key difference is that the biostatistics literature has focused on sequential random assignment (a.k.a. sequential ignorability) assumptions that impose that treatment in each period is random conditional on the path of covariates and realized outcomes, rather than parallel trends. We think that integrating these two literatures — e.g., understanding when parallel trends is preferable to sequential ignorability and vice versa, and whether there are any fruitful combinations of the two assumptions — is an interesting area for future research.

Spillover effects. The vast majority of the DiD literature imposes the SUTVA assumption, which rules out spillover effects. However, spillover effects may be important in many economic applications, such as when policy in one area affects neighboring areas, or when

individuals are connected in a network. Butts (2021) provides some initial work in this direction by extending the framework of Callaway and Sant’Anna (2021) to allow for local spatial spillovers. Huber and Steinmayr (2021) also consider extensions to allow for spillover effects. We suspect that in the coming years, we will see more work on DiD with spillovers.

Conditional treatment effects. The DiD literature has placed a lot of emphasis on learning about the ATT’s of different groups. However, in many situations, it may also be desirable to better understand how these ATT’s vary between subpopulations defined by covariate values. For instance, how does the average treatment effect of a training program on earnings vary according to the age of its participants? Abadie (2005) provides re-weighting methods to tackle these types of questions using linear approximations. However, recent research has shown that data-adaptive/machine-learning procedures can be used to more flexibly estimate treatment effect heterogeneity in the context of RCTs or cross-sectional observational studies with unconfoundedness (e.g., Lee, Okui and Whang, 2017; Wager and Athey, 2018; Chernozhukov, Demirer, Duflo and Fernández-Val, 2020). Whether such tools can be adapted to estimate treatment effect heterogeneity in DiD setups is a promising area for future research.

Connections to other panel data methods. DiD is of course one of many possible panel data methods. One of the most prominent alternatives is the synthetic control (SC) method, pioneered by Abadie, Diamond and Hainmueller (2010). Much of the DiD and SC literatures have evolved separately, using different data-generating processes as the baseline (Abadie, 2021). Recent work has begun to try to combine insights from the two literatures (e.g., Arkhangelsky, Athey, Hirshberg, Imbens and Wager, 2021; Ben-Michael, Feller and Rothstein, 2021a,b; Doudchenko and Imbens, 2016). We think that exploring further connections between the literatures — and in particular, providing clear guidance for practitioners on when one we should expect one method to perform better than the other, or whether one should consider a hybrid of the two — is an interesting direction for future research.

7 Conclusion

This paper synthesizes the recent literature on DiD. Some key themes are that researchers should be clear about the comparison group used for identification, match the estimation and inference methods to the identifying assumptions, and explore robustness to possible violations of those assumptions. We emphasize that context-specific knowledge will often

be needed to choose the right identifying assumptions and accompanying methods. We are hopeful that these recent developments will help to make DiD analyses more transparent and credible in the years to come.

Table 1: A Checklist for DiD Practitioners

– **Is everyone treated at the same time?**

If yes, and panel is balanced, estimation with TWFE specifications such as (5) or (7) yield easily interpretable estimates.

If no, consider using a “heterogeneity-robust” estimator for staggered treatment timing as described in Section 3. The appropriate estimator will depend on whether treatment turns on/off and which parallel trends assumption you’re willing to impose. Use TWFE only if you’re confident in treatment effect homogeneity.

– **Are you sure about the validity of the parallel trends assumption?**

If yes, explain why, including a justification for your choice of functional form. If the justification is (quasi-)random treatment timing, consider using a more efficient estimator as discussed in Section 6.

If no, consider the following steps:

1. If parallel trends would be more plausible conditional on covariates, consider a method that conditions on covariates, as described in Section 4.2.
2. Assess the plausibility of the parallel trends assumption by constructing an event-study plot. If there is a common treatment date and you’re using an unconditional parallel trends assumption, plot the coefficients from a specification like (17). If not, then see Section 4.3 for recommendations on event-plot construction.
3. Accompany the event-study plot with diagnostics of the power of the pre-test against relevant alternatives and/or non-inferiority tests, as described in Section 4.4.1.
4. Report formal sensitivity analyses that describe the robustness of the conclusions to potential violations of parallel trends, as described in Section 4.5.

– **Do you have a large number of treated and untreated clusters sampled from a super-population?**

If yes, then use cluster-robust methods at the cluster level. A good rule of thumb is to cluster units whose treatment assignment is correlated (e.g. residents of the same state, when policy is at the state level); see Section 5.2.

If you have a small number of treated clusters, consider using one of the alternative inference methods described in Section 5.1.

If you can’t imagine the super-population, consider a design-based justification for inference instead, as discussed in Section 5.2.

Table 2: Statistical Packages for Recent DiD Methods

Heterogeneity Robust Estimators for Staggered Timing		
<u>Package</u>	<u>Software</u>	<u>Description</u>
did, csdid	R, Stata	Implements Callaway and Sant'Anna (2021)
did2s	R, Stata	Implements Gardner (2021), Borusyak et al (2021), Sun and Abraham (2020), Callaway and Sant'Anna (2021), Roth and Sant'Anna (2021)
didimputation, did_imputation	R, Stata	Implements Borusyak, Jaravel, and Spiess (2021)
DIDmultiplegt, did_multiplegt	R, Stata	Implements de Chaisemartin and D'Haultfoeuille (2020)
eventstudyinteract	Stata	Implements Sun and Abraham (2020)
flexpaneldid	Stata	Implements Dettmann (2020), based on Heckman et al (1998)
fixest	R	Implements Sun and Abraham (2020)
stackedev	Stata	Implements stacking approach in Cengiz et al (2019)
staggered	R	Implements Roth and Sant'Anna (2021), Callaway and Sant'Anna (2020), and Sun and Abraham (2020)
xtevent	Stata	Implements Freyaldenhoven et al (2019)
DiD with Covariates		
<u>Package</u>	<u>Software</u>	<u>Description</u>
DRDID, drdid	R, Stata	Implements Sant'Anna and Zhao (2020)
Diagnostics for TWFE with Staggered Timing		
<u>Package</u>	<u>Software</u>	<u>Description</u>
bacondecomp	R, Stata	Diagnostics from Goodman-Bacon (2021)
TwoWayFEWeights	R	Diagnostics from de Chaisemartin and D'Haultfoeuille (2020)
Diagnostic / Sensitivity for Violations of Parallel Trends		
<u>Package</u>	<u>Software</u>	<u>Description</u>
honestDiD	R	Implements Rambachan and Roth (2021)
pretrends	R	Diagnostics from Roth (2021)

This table lists R and Stata packages for recent DiD methods, and is based on Asjad Naqvi's repository at <https://asjadnaqvi.github.io/DiD/>. Several of the packages listed under "Heterogeneity Robust Estimators" also accommodate covariates.

References

- Abadie, Alberto**, “Semiparametric Difference-in-Differences Estimators,” *The Review of Economic Studies*, 2005, 72 (1), 1–19.
- , “Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects,” *Journal of Economic Literature*, June 2021, 59 (2), 391–425.
- , **Alexis Diamond**, and **Jens Hainmueller**, “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program,” *Journal of the American Statistical Association*, June 2010, 105 (490), 493–505.
- and **Guido W. Imbens**, “Large sample properties of matching estimators for average treatment effects,” *Econometrica*, 2006, 74 (1), 235–267.
- and – , “On the Failure of the Bootstrap for Matching Estimators,” *Econometrica*, 2008, 76 (6), 1537–1557.
- and – , “Bias-Corrected Matching Estimators for Average Treatment Effects,” *Journal of Business & Economic Statistics*, 2011, 29 (1), 1–11.
- and – , “A Martingale Representation for Matching Estimators,” *Journal of the American Statistical Association*, 2012, 107 (498), 833–843.
- , **Susan Athey**, **Guido W. Imbens**, and **Jeffrey Wooldridge**, “When Should You Adjust Standard Errors for Clustering?,” Working Paper 24003, National Bureau of Economic Research 2017.
- Abbring, Jaap H. and Gerard J. van den Berg**, “The nonparametric identification of treatment effects in duration models,” *Econometrica*, 2003, 71 (5), 1491–1517.
- Andrews, Isaiah, Ariel Pakes, and Jonathan Roth**, “Inference for Linear Conditional Moment Inequalities,” Working Paper 2019.
- Arellano, M.**, “Practitioners’ Corner: Computing Robust Standard Errors for Within-groups Estimators,” *Oxford Bulletin of Economics and Statistics*, 1987, 49 (4), 431–434.
- Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager**, “Synthetic Difference-in-Differences,” *American Economic Review*, 2021, 111 (12), 4088–4118.
- Armstrong, Timothy B. and Michal Kolesár**, “Optimal Inference in a Class of Regression Models,” *Econometrica*, 2018, 86 (2), 655–683.
- Athey, Susan and Guido Imbens**, “Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption,” *Journal of Econometrics*, 2022, 226 (1), 62–79.
- and **Guido W. Imbens**, “Identification and Inference in Nonlinear Difference-in-Differences Models,” *Econometrica*, 2006, 74 (2), 431–497.
- Baker, Andrew, David F. Larcker, and Charles C. Y. Wang**, “How Much Should We

- Trust Staggered Difference-In-Differences Estimates?,” *SSRN Electronic Journal*, 2021.
- Ben-Michael, Eli, Avi Feller, and Jesse Rothstein**, “The Augmented Synthetic Control Method,” *Journal of the American Statistical Association*, 2021, *116* (536), 1789–1803.
- , – , and – , “Synthetic Controls with Staggered Adoption,” *Journal of the Royal Statistical Society Series B*, 2021, *Forthcoming*.
- Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan**, “How Much Should We Trust Differences-In-Differences Estimates?,” *The Quarterly Journal of Economics*, 2004, *119* (1), 249–275.
- Bilinski, Alyssa and Laura A. Hatfield**, “Seeking evidence of absence: Reconsidering tests of model assumptions,” *arXiv:1805.03273 [stat]*, May 2018.
- Bonhomme, Stéphane and Ulrich Sauder**, “Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling,” *Review of Economics and Statistics*, 2011, *93* (2), 479–494.
- Borusyak, Kirill and Xavier Jaravel**, “Revisiting Event Study Designs,” SSRN Scholarly Paper ID 2826228, Social Science Research Network, Rochester, NY 2018.
- , – , and **Jann Spiess**, “Revisiting Event Study Designs: Robust and Efficient Estimation,” *arXiv:2108.12419 [econ]*, 2021.
- Butts, Kyle**, “Difference-in-Differences Estimation with Spatial Spillovers,” *arXiv:2105.03737 [econ]*, 2021.
- Callaway, Brantly and Pedro H. C. Sant’Anna**, “Difference-in-Differences with multiple time periods,” *Journal of Econometrics*, 2021, *225* (2), 200–230.
- and **Tong Li**, “Quantile treatment effects in difference in differences models with panel data,” *Quantitative Economics*, 2019, *10* (4), 1579–1618.
- , **Andrew Goodman-Bacon**, and **Pedro H. C. Sant’Anna**, “Difference-in-Differences with a Continuous Treatment,” *arXiv:2107.02637 [econ]*, 2021.
- Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller**, “Bootstrap-Based Improvements for Inference With Clustered Errors,” *Review of Economics and Statistics*, 2008, *90* (3), 414–427.
- Canay, Ivan A., Andres Santos, and Azeem M. Shaikh**, “The wild bootstrap with a “small” number of “large” clusters,” *Review of Economics and Statistics*, 2021, *103* (2), 346–363.
- , **Joseph P. Romano**, and **Azeem M. Shaikh**, “Randomization Tests Under an Approximate Symmetry Assumption,” *Econometrica*, 2017, *85* (3), 1013–1030.
- Card, David and Alan B Krueger**, “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania,” *American Economic Review*, 1994, *84* (4), 772–793.

- Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer**, “The Effect of Minimum Wages on Low-Wage Jobs,” *The Quarterly Journal of Economics*, 2019, 134 (3), 1405–1454.
- Chabé-Ferret, Sylvain**, “Analysis of the bias of Matching and Difference-in-Difference under alternative earnings and selection processes,” *Journal of Econometrics*, 2015, 185 (1), 110–123.
- Chang, Neng-Chieh**, “Double/debiased machine learning for difference-in-differences,” *Econometrics Journal*, 2020, 23, 177–191.
- Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val**, “Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments,” *arXiv: 1712.04802*, 2020, pp. 1–52.
- Conley, Timothy G. and Christopher R. Taber**, “Inference with “Difference in Differences” with a Small Number of Policy Changes,” *Review of Economics and Statistics*, 2011, 93 (1), 113–125.
- Daw, Jamie R. and Laura A. Hatfield**, “Matching and Regression to the Mean in Difference-in-Differences Analysis,” *Health Services Research*, 2018.
- de Chaisemartin, Clément and Xavier D’Haultfœuille**, “Fuzzy Differences-in-Differences,” *The Review of Economic Studies*, 2018, 85 (2), 999–1028.
- **and** –, “Difference-in-Differences Estimators of Intertemporal Treatment Effects,” SSRN Scholarly Paper ID 3731856, Social Science Research Network, Rochester, NY 2020.
- **and** –, “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects,” *American Economic Review*, 2020, 110 (9), 2964–2996.
- **and** –, “Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey,” SSRN Scholarly Paper ID 3980758, Social Science Research Network, Rochester, NY 2021.
- **and** –, “Two-way Fixed Effects Regressions with Several Treatments,” SSRN Scholarly Paper ID 3751060, Social Science Research Network, Rochester, NY 2021.
- Dette, Holger and Martin Schumann**, “Difference-in-Differences Estimation Under Non-Parallel Trends,” *Working Paper*, 2020.
- Ding, Peng and Fan Li**, “A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment,” *Political Analysis*, 2019, 27, 605–615.
- Donald, Stephen G. and Kevin Lang**, “Inference with Difference-in-Differences and Other Panel Data,” *The Review of Economics and Statistics*, 2007, 89 (2), 221–233.
- Doudchenko, Nikolay and Guido W. Imbens**, “Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis,” Working Paper 22791, National Bureau of Economic Research 2016.

- Ferman, Bruno and Cristine Pinto**, “Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity,” *The Review of Economics and Statistics*, 2019, 101 (3), 452–467.
- Fisher, R. A.**, *The design of experiments* The design of experiments, Oxford, England: Oliver & Boyd, 1935. Pages: xi, 251.
- Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro**, “Pre-event Trends in the Panel Event-Study Design,” *American Economic Review*, 2019, 109 (9), 3307–3338.
- , – , **Jorge Pérez Pérez, and Jesse M. Shapiro**, “Visualization, Identification, and Estimation in the Linear Panel Event-Study Design,” Working Paper 29170, National Bureau of Economic Research 2021.
- Gardner, John**, “Two-stage differences in differences,” *Working Paper*, 2021.
- Goodman-Bacon, Andrew**, “Difference-in-differences with variation in treatment timing,” *Journal of Econometrics*, 2021, 225 (2), 254–277.
- Hagemann, Andreas**, “Inference with a single treated cluster,” *arXiv:2010.04076 [econ.EM]*, 2020, pp. 1–23.
- , “Permutation inference with a finite number of heterogeneous clusters,” *arXiv:1907.01049 [econ.EM]*, 2021.
- Hasegawa, Raiden B., Daniel W. Webster, and Dylan S. Small**, “Evaluating Missouri’s Handgun Purchaser Law: A Bracketing Method for Addressing Concerns About History Interacting with Group,” *Epidemiology*, 2019, 30 (3), 371–379.
- Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd**, “Characterizing Selection Bias Using Experimental Data,” *Econometrica*, 1998, 66 (5), 1017–1098.
- Heckman, James J., Hidehiko Ichimura, and Petra Todd**, “Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme,” *The Review of Economic Studies*, 1997, 64 (4), 605–654.
- Holland, Paul W.**, “Statistics and Causal Inference,” *Journal of the American Statistical Association*, 1986, 81 (396), 945–960.
- Huber, Martin and Andreas Steinmayr**, “A Framework for Separating Individual-Level Treatment Effects From Spillover Effects,” *Journal of Business & Economic Statistics*, 2021, 39 (2), 422–436.
- Ibragimov, Rustam and Ulrich K. Müller**, “Inference with few heterogeneous clusters,” *Review of Economics and Statistics*, 2016, 98 (1), 83–96.
- Imai, Kosuke and In Song Kim**, “On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data,” *Political Analysis*, 2021, 29 (3), 405–415.
- Jakiela, Pamela**, “Simple Diagnostics for Two-Way Fixed Effects,” *arXiv:2103.13229 [econ, q-fin]*, March 2021.

- Kahn-Lang, Ariella and Kevin Lang**, “The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications,” *Journal of Business & Economic Statistics*, 2020, *38* (3), 613–620.
- Keele, Luke J., Dylan S. Small, Jesse Y. Hsu, and Colin B. Fogarty**, “Patterns of Effects and Sensitivity Analysis for Differences-in-Differences,” *arXiv:1901.01869 [stat]*, February 2019. arXiv: 1901.01869.
- Khan, Shakeeb and Elie Tamer**, “Irregular Identification, Support Conditions, and Inverse Weight Estimation,” *Econometrica*, 2010, *78* (6), 2021–2042.
- Lee, Sokbae, Ryo Okui, and Yoon-Jae Jae Whang**, “Doubly robust uniform confidence band for the conditional average treatment effect function,” *Journal of Applied Econometrics*, 2017, *32* (7), 1207–1225.
- Liang, Kung-Yee and Scott L. Zeger**, “Longitudinal data analysis using generalized linear models,” *Biometrika*, 1986, *73* (1), 13–22.
- Lin, Winston**, “Agnostic notes on regression adjustments to experimental data: Reexamining Freedman’s critique,” *Annals of Applied Statistics*, 2013, *7* (1), 295–318.
- Liu, Licheng, Ye Wang, and Yiqing Xu**, “A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data,” SSRN Scholarly Paper ID 3555463, Social Science Research Network, Rochester, NY 2021.
- Malani, Anup and Julian Reif**, “Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform,” *Journal of Public Economics*, 2015, *124*, 1–17.
- Manski, Charles F. and John V. Pepper**, “How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions,” *The Review of Economics and Statistics*, 2018, *100* (2), 232–244.
- Marcus, Michelle and Pedro H. C. Sant’Anna**, “The role of parallel trends in event study settings : An application to environmental economics,” *Journal of the Association of Environmental and Resource Economists*, 2021, *8* (2), 235–275.
- McKenzie, David**, “Beyond baseline and follow-up: The case for more T in experiments,” *Journal of Development Economics*, 2012, *99* (2), 210–221.
- Meyer, Bruce D.**, “Natural and Quasi-Experiments in Economics,” *Journal of Business & Economic Statistics*, 1995, *13* (2), 151–161.
- Neyman, Jerzy**, “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9,” *Statistical Science*, 1923, *5* (4), 465–472.
- Olea, José Luis Montiel and Mikkel Plagborg-Møller**, “Simultaneous confidence bands: Theory, implementation, and an application to SVARs,” *Journal of Applied Econometrics*, 2019, *34* (1), 1–17.
- Rambachan, Ashesh and Jonathan Roth**, “Design-Based Uncertainty for Quasi-

- Experiments,” *arXiv:2008.00602 [econ, stat]*, 2020.
- **and** –, “An Honest Approach to Parallel Trends,” *Working paper*, 2021.
- Robins, J. M.**, “A New Approach To Causal Inference in Mortality Studies With a Sustained Exposure Period - Application To Control of the Healthy Worker Survivor Effect,” *Mathematical Modelling*, 1986, 7, 1393–1512.
- Roth, Jonathan**, “Should We Adjust for the Test for Pre-trends in Difference-in-Difference Designs?,” *arXiv:1804.01208 [econ, math, stat]*, 2018.
- , “Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends,” *American Economic Review: Insights*, Forthcoming.
- **and Pedro H. C. Sant’Anna**, “Efficient Estimation for Staggered Rollout Designs,” *arXiv:2102.01291 [econ, math, stat]*, 2021.
- **and** –, “When Is Parallel Trends Sensitive to Functional Form?,” *arXiv:2010.04814 [econ, stat]*, 2021.
- Rubin, Donald B.**, “Estimating causal effects of treatments in randomized and nonrandomized studies,” *Journal of Educational Psychology*, 1974, 66 (5), 688–70.
- Ryan, Andrew M.**, “Well-Balanced or too Matchy–Matchy? The Controversy over Matching in Difference-in-Differences,” *Health Services Research*, 2018, 53 (6), 4106–4110.
- Sant’Anna, Pedro H. C. and Jun Zhao**, “Doubly robust difference-in-differences estimators,” *Journal of Econometrics*, 2020, 219 (1), 101–122.
- Shaikh, Azeem and Panos Toulis**, “Randomization Tests in Observational Studies with Staggered Adoption of Treatment,” *arXiv:1912.10610 [stat]*, 2021.
- Strezhnev, Anton**, “Semiparametric weighting estimators for multi-period difference-in-differences designs,” *Working Paper*, 2018.
- Sun, Liyang and Sarah Abraham**, “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,” *Journal of Econometrics*, 2021, 225 (2), 175–199.
- Wager, Stefan and Susan Athey**, “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests,” *Journal of the American Statistical Association*, 2018, 113 (523), 1228–1242.
- Wooldridge, Jeffrey M.**, “Cluster-Sample Methods in Applied Econometrics,” *American Economic Review P&P*, 2003, 93 (2), 133–138.
- , “Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators,” *Working Paper*, 2021, pp. 1–89.
- Ye, Ting, Luke Keele, Raiden Hasegawa, and Dylan S. Small**, “A Negative Correlation Strategy for Bracketing in Difference-in-Differences,” *arXiv:2006.02423 [econ, stat]*, 2021.
- Zeldow, Bret and Laura A. Hatfield**, “Confounding and regression adjustment in

difference-in-differences studies,” *Health Services Research*, 2021, 56 (5), 932–941.

A Connecting model-based assumptions to potential outcomes

This section formalizes connections between the model-based assumptions in Section 5.1 and the potential outcomes framework.

Adopt the notation and set-up in Section 2, except now each unit i belongs to a cluster j and treatment is assigned at the cluster level D_j . We assume clusters are drawn *iid* from a superpopulation of clusters and then units are drawn *iid* within the sampled cluster. We write N_d to denote the number of clusters with treatment d , and n_j the number of units in cluster j . As in the main text, let $Y_{jt} = n_j^{-1} \sum_{i:j(i)=j} Y_{ijt}$ be the sample mean within cluster j . The canonical DiD estimator at the cluster level can then be written as:

$$\begin{aligned}\hat{\tau} &= \frac{1}{N_1} \sum_{j:D_j=1} (Y_{j2} - Y_{j1}) - \frac{1}{N_0} \sum_{i:D_j=0} (Y_{j2} - Y_{j1}) \\ &= \frac{1}{N_1} \sum_{j:D_j=1} \frac{1}{n_j} \sum_{i:j(i)=j} (Y_{ij2} - Y_{ij1}) - \frac{1}{N_0} \sum_{i:D_j=0} \frac{1}{n_j} \sum_{i:j(i)=j} (Y_{ij2} - Y_{ij1}).\end{aligned}$$

Since the observed outcome is $Y(1)$ for treated units and $Y(0)$ for control units, under the no anticipation assumption it follows that

$$\hat{\tau} = \frac{1}{N_1} \sum_{j:D_j=1} \frac{1}{n_j} \sum_{i:j(i)=j} (Y_{ij2}(1) - Y_{ij1}(0)) - \frac{1}{N_0} \sum_{j:D_j=0} \frac{1}{n_j} \sum_{i:j(i)=j} (Y_{ij2}(0) - Y_{ij1}(0)),$$

or equivalently,

$$\begin{aligned}\hat{\tau} &= \frac{1}{N_1} \sum_{j:D_j=1} \frac{1}{n_j} \sum_{i:j(i)=j} (Y_{ij2}(1) - Y_{ij2}(0)) + \\ &\quad \frac{1}{N_1} \sum_{j:D_j=1} \frac{1}{n_j} \sum_{i:j(i)=j} (Y_{ij2}(0) - Y_{ij1}(0)) - \\ &\quad \frac{1}{N_0} \sum_{j:D_j=0} \frac{1}{n_j} \sum_{i:j(i)=j} (Y_{ij2}(0) - Y_{ij1}(0)).\end{aligned}$$

Adding and subtracting terms of the form $\mathbb{E}[Y_{ijt}|j(i) = 1]$, we obtain

$$\begin{aligned}
\hat{\tau} = & \tau_2 + \frac{1}{N_1} \sum_{j:D_j=1} (\mathbb{E}[Y_{ij2}(1) - Y_{ij2}(0)|j(i) = j] - \tau_2) + \\
& \frac{1}{N_1} \sum_{j:D_j=1} \frac{1}{n_j} \sum_{i:j(i)=j} (Y_{ij2}(1) - Y_{ij2}(0) - \mathbb{E}[Y_{ij2}(1) - Y_{ij2}(0)|j(i) = j]) + \\
& \frac{1}{N_1} \sum_{j:D_j=1} \frac{1}{n_j} \sum_{i:j(i)=j} (Y_{ij2}(0) - Y_{ij1}(0) - \mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|j(i) = j]) - \\
& \frac{1}{N_0} \sum_{j:D_j=0} \frac{1}{n_j} \sum_{i:j(i)=j} (Y_{ij2}(0) - Y_{ij1}(0) - \mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|j(i) = j]) + \\
& \frac{1}{N_1} \sum_{j:D_j=1} \mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|j(i) = j] - \frac{1}{N_0} \sum_{j:D_j=0} \mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|j(i) = j],
\end{aligned}$$

where $\tau_2 = \mathbb{E}\left[N_1^{-1} \sum_{j:D_j=1} \mathbb{E}[Y_{ij2}(1) - Y_{ij2}(0)|j(i) = j]\right] = \mathbb{E}[Y_{ij2}(1) - Y_{ij2}(0)|D_j = 1]$ is the ATT among treated clusters (weighting all clusters equally).

Now, we assume parallel trends at the cluster level, so that

$$\mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|D_j = 1] - \mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|D_j = 0] = 0,$$

which implies that

$$\begin{aligned}
\hat{\tau} = & \tau_2 + \frac{1}{N_1} \sum_{j:D_j=1} \underbrace{(\mathbb{E}[Y_{ij2}(1) - Y_{ij2}(0)|j(i) = j] - \tau_2)}_{=\Delta\nu_{j,1}} + \\
& \frac{1}{N_1} \sum_{j:D_j=1} \frac{1}{n_j} \sum_{i:j(i)=j} \underbrace{(Y_{ij2}(1) - Y_{ij2}(0) - \mathbb{E}[Y_{ij2}(1) - Y_{ij2}(0)|j(i) = j])}_{=\Delta\epsilon_{ij,1}} + \\
& \frac{1}{N_1} \sum_{j:D_j=1} \frac{1}{n_j} \sum_{i:j(i)=j} \underbrace{(Y_{ij2}(0) - Y_{ij1}(0) - \mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|j(i) = j])}_{=\Delta\epsilon_{ij,0}} - \\
& \frac{1}{N_0} \sum_{j:D_j=0} \frac{1}{n_j} \sum_{i:j(i)=j} \underbrace{(Y_{ij2}(0) - Y_{ij1}(0) - \mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|j(i) = j])}_{=\Delta\epsilon_{ij,0}} + \\
& \frac{1}{N_1} \sum_{j:D_j=1} \underbrace{(\mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|j(i) = j] - \mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|D_j])}_{\Delta\nu_{j,0}} - \\
& \frac{1}{N_0} \sum_{j:D_j=0} \underbrace{(\mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|j(i) = j] - \mathbb{E}[Y_{ij2}(0) - Y_{ij1}(0)|D_j])}_{\Delta\nu_{j,0}}.
\end{aligned}$$

Letting $\Delta\nu_j = \Delta\nu_{j,0} + D_j\Delta\nu_{j,1}$ and $\Delta\epsilon_{ij} = \Delta\epsilon_{ij,0} + D_j\Delta\epsilon_{ij,1}$, it follows that $\hat{\tau}$ takes the form (21) with $\beta = \tau_2$.