Difference-in-Differences when Parallel Trends Holds Conditional on Covariates*

Carolina Caetano† Brantly Callaway‡

March 21, 2026

Abstract

We consider difference-in-differences identification and estimation strategies when the parallel trends assumption holds conditional on covariates, which can be time-varying, time-invariant, or both. We uncover several weaknesses of two-way fixed effects (TWFE) regressions in this context. The most important, which we call *hidden linearity bias*, arises because transformations that eliminate unit fixed effects also transform the covariates, either implicitly changing the identification strategy or relying on correct model specification. We provide diagnostics for assessing a TWFE regression's susceptibility to hidden linearity bias and propose alternative estimation strategies that circumvent these issues.

JEL Codes: C14, C21, C23

Keywords: Difference-in-Differences, Time-Varying Covariates, Time-Invariant Covariates, Hidden Linearity Bias, Two-way Fixed Effects Regression, Doubly Robust Estimation, Conditional Parallel Trends, Treatment Effect Heterogeneity

*Some of the results in this paper were originally in "Difference-in-differences with time-varying covariates" (Caetano et al. (2022)). This paper and our companion paper, "Difference-in-differences with bad controls" (Caetano, Callaway, Payne, and Rodrigues (2025)), replace that paper. The code for the new estimation approaches proposed in the paper is provided in the `R ptetools` package, which is available on CRAN. Code for the TWFE and AIPW diagnostics discussed in the paper is available in the `R twfeweights` package, which is available at https://github.com/bcallaway11/twfeweights. We thank Kyle Butts, Andrew Goodman-Bacon, Pedro Sant'Anna, Tymon Sloczynski, as well as many seminar and conference participants for helpful comments.

†John Munro Godfrey, Sr. Department of Economics, University of Georgia. carol.caetano@uga.edu

‡John Munro Godfrey, Sr. Department of Economics, University of Georgia. brantly.callaway@uga.edu

1 Introduction

In this paper, we study difference-in-differences (DiD) when the parallel trends assumption holds after conditioning on covariates. Researchers often include covariates in the parallel trends assumption to make it more plausible (e.g., Heckman et al. (1998) and Abadie (2005)), with the aim of comparing the change in outcomes over time for treated units to the change in outcomes over time for untreated units with similar observed characteristics. We pay careful attention to the different types of covariates that can show up in the parallel trends assumption: time-varying covariates and/or time-invariant covariates. The econometrics literature and empirical applications often use covariates in substantially different ways. In the econometrics literature, it is common to assume that the covariates are all time-invariant or, if there are time-varying covariates, to use a pre-treatment value of the time-varying covariates as a time-invariant covariate; see, for example, Abadie (2005), Bonhomme and Sauder (2011), Sant'Anna and Zhao (2020), and Callaway and Sant'Anna (2021). On the other hand, in empirical work, the most common way to include covariates is in the following TWFE regression[1]

$$Y_{it} = \theta_t + \eta_i + \alpha D_{it} + X_{it}'\beta + e_{it}, \tag{1}$$

where θ_t is a time fixed effect, η_i is an individual fixed effect, D_{it} is a binary treatment indicator, and X_{it} are time-varying covariates. α is the coefficient of interest in this regression, sometimes interpreted as "the causal effect of the treatment" or, in the presence of treatment effect heterogeneity, often loosely interpreted as some kind of average treatment effect parameter. Being able to include covariates is one of the original main attractions of using a TWFE regression to implement a DiD

[1]Our discussion of limitations of TWFE regressions is specific to this particular TWFE regression. As emphasized in Wooldridge (2025), many of the drawbacks of common versions of TWFE regressions can be addressed by using more flexible TWFE regressions (e.g., including many additional interaction terms). Using a more flexible regression could address some of the issues that we highlight below. We focus on this particular TWFE regression because it is very common in empirical work. In Supplementary Appendix SA5, we systematically reviewed 25 empirical

identification strategy. For example, Angrist and Pischke (2008) write: "A second advantage of regression-DD is that it facilitates empirical work with regressors." Relative to the econometrics literature discussed above, one immediately noticeable difference with the TWFE regression is that it does not include time-invariant covariates. This is because, if time-invariant covariates enter the TWFE model in an analogous way to the time-varying covariates (i.e., with a time-invariant coefficient), then they will be absorbed into the unit fixed effect. This is a reason commonly given for not including time-invariant covariates in difference-in-differences applications.

We uncover several limitations of this TWFE regression. Although TWFE regressions with only two time periods are known to be robust to treatment effect heterogeneity under unconditional parallel trends, we show that TWFE regressions that rely on conditional parallel trends assumptions are susceptible to a number of problems *even in the case with only two time periods*. We show that TWFE regressions can include non-neglible misspecification bias terms for any of three reasons: (1) violations of certain linearity conditions on the model for untreated potential outcomes over time, (2) paths of untreated potential outcomes that depend on the level of time-varying covariates in addition to (or instead of) the change in the covariates over time, and (3) paths of untreated potential outcomes that depend on time-invariant covariates. The first issue is expected, as similar conditions show up in the literature on interpreting cross-sectional regressions under unconfoundedness (Goldsmith-Pinkham et al. (2024), Blandhol et al. (2025), and Hahn (2023)). Also, assuming a linear model for untreated potential outcomes is often a key step for motivating

difference-in-differences papers. Of these, 15 included covariates in the same manner as in Equation (1), 7 did not include time-varying covariates (6 of these did not include any covariates at all, 1 included only time-invariant covariates), 2 included baseline versions of the covariates (i.e., the values of the time-varying covariates in the first period), and 1 included both time-varying covariates as in Equation (1) and baseline versions of the covariates. In addition, the TWFE regression in Equation (1) is also emphasized as the way to introduce covariates in DiD applications in textbook treatments of difference-in-differences (e.g., Wooldridge (2010, Sections 6.5.2 and 10.6.4), Angrist and Pischke (2008, Chapter 5), Cunningham (2021, Chapter 9)).

linear models for the outcome itself (see Angrist and Pischke (2008) for a number of examples).

Issues (2) and (3) are more subtle. We refer to the bias arising from them as *hidden linearity bias*. Unlike Issue (1), there is no analog of hidden linearity bias in cross-sectional settings. This bias arises because, to estimate Equation (1), the unit fixed effect is removed by transforming the model. For example, with two time periods, α is estimated by taking first differences to eliminate the unit fixed effect. A consequence is that the covariates are also differenced, so the estimated model ultimately only controls for *changes* in the time-varying covariates. Below, we consider an application with state-level panel data, a state-level treatment, and time-varying covariates such as a state's population. In this context, controlling for the change in a state's population may not account for the level of a state's population (i.e., states with similar changes in population could have quite different levels of population), or for other time-invariant covariates such as region.[2]

An important question is how much the bias discussed above matters in practice. Since this bias is hard to measure directly, we propose simple diagnostic tools to assess TWFE regressions' sensitivity to hidden linearity bias. Our idea is to recast α from the TWFE regression as a re-weighting estimator (see Aronow and Samii (2016) and Chattopadhyay and Zubizarreta (2023) for related ideas). We recover the "implicit regression weights" (which are straightforward to calculate) and apply them to the levels of time-varying covariates and time-invariant covariates. If the implicit regression weights balance these covariates, hidden linearity bias is likely small. Otherwise, α may be quite sensitive to violations of linearity conditions. Moreover, even if none of these issues arise, TWFE regressions deliver weighted averages of conditional ATTs with non-transparent weights that can be negative or exhibit weight-reversal properties similar to the ones pointed out in Sloczynski (2022) in the cross-sectional case.

We propose several new estimation strategies that do not suffer from *any* of the limitations of

[2]It is also common to motivate the TWFE regression in Equation (1) by a selection-on-observables argument conditional on covariates and unit and time fixed effects. Versions of the same issues we highlight here also apply in this case, as these approaches require the model to be correctly specified.

the TWFE regression discussed above. Essentially, in line with the parallel trends assumption, we first difference the outcome, but we do not apply the same transformation to the covariates, allowing the levels of time-varying covariates and time-invariant covariates to remain in the estimating equation. We operationalize this idea by building on popular estimators in the DiD literature, particularly the augmented inverse propensity score weighting (AIPW) estimators in Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021), though the same idea could be used with other estimators such as matching, inverse probability tilting (Graham et al. (2012)), entropy balancing (Hainmueller (2012)), and covariate balancing propensity score (Imai and Ratkovic (2014)). We also show that AIPW estimators of the ATT under conditional parallel trends can be reformulated as re-weighting estimators. Combined with our TWFE diagnostics, this allows researchers to compare covariate balancing properties across leading ATT estimators during the "design phase," before using the outcome at all (Ho et al. (2007), Rubin (2008), and Imbens and Rubin (2015)).

We conclude the paper by revisiting an application from Cheng and Hoekstra (2013) on the effects of stand-your-ground laws on homicides. We find that including time-varying covariates, such as a state's population and/or median income, in a TWFE regression balances the average of the within-transformed covariates but often does little to improve (and in some cases worsens) covariate balance in terms of the levels of the same covariates or in terms of time-invariant covariates. Using our approach, covariate balance is substantially improved. Our estimates are mostly qualitatively similar to those reported in Cheng and Hoekstra (2013), though we find somewhat weaker evidence that stand-your-ground laws increase homicides.

The paper proceeds as follows. Section 2 presents the main assumptions and some preliminary identification results. Section 3 discusses the limitations of TWFE regressions under conditional parallel trends, and Section 4 proposes simple diagnostics for assessing the extent of hidden linearity bias in TWFE regressions. These sections focus on the baseline two-period, two-group DiD setup, where many of our main insights can already be seen. Section 5 extends the arguments to settings with multiple periods and variation in treatment timing across units. Section 6 proposes alternative estimation strategies that circumvent TWFE regressions' limitations. Section 7 revis-

its an application from Cheng and Hoekstra (2013) on the effects of stand-your-ground laws on homicides. Finally, Section 8 concludes.

2 Setup

For much of the paper, we consider the canonical two-period setting, with periods denoted by t^* and t^*-1, and no treated units in the first period. Let D_i be a binary treatment indicator, whose time subscript is omitted because treatment differences occur only in the second period. Let X_{it} denote a $k \times 1$ vector of time-varying covariates for unit i in time period t, and Z_i an $l \times 1$ vector of time-invariant covariates. Let Y_{it} denote the observed outcome, with $Y_{it}(1)$ and $Y_{it}(0)$ the treated and untreated potential outcomes. Observed and potential outcomes are related by $Y_{it^*} = D_i Y_{it^*}(1) + (1-D_i) Y_{it^*}(0)$ and $Y_{it^*-1} = Y_{it^*-1}(0)$; i.e., in the second period we observe treated (untreated) potential outcomes for treated (untreated) units, and in the first period we observe untreated potential outcomes for all units.[3] We also suppose throughout the paper that all expectations exist and take all statements conditional on covariates to hold almost surely.

2.1 Identification

Following the vast majority of the difference-in-differences literature, we target identifying the average treatment effect on the treated (ATT), which is given by

$$\text{ATT} := \mathbb{E}[Y_{t^*}(1) - Y_{t^*}(0) | D = 1].$$

We also define a conditional-on-covariates version of the ATT as

$$\text{ATT}(X_{t^*}, X_{t^*-1}, Z) := \mathbb{E}[Y_{t^*}(1) - Y_{t^*}(0) | X_{t^*}, X_{t^*-1}, Z, D = 1].$$

We make the following assumptions:

Assumption 1 (Random Sampling)**.** *The observed data* $\{Y_{it^*}, Y_{it^*-1}, X_{it^*}, X_{it^*-1}, Z_i, D_i\}_{i=1}^n$ *are i.i.d.*

[3] The discussion above implicitly imposes a SUTVA assumption and a no-anticipation assumption (that pre-treatment outcomes are not affected by eventually participating in the treatment). These are standard in the DiD literature. We discuss no-anticipation in more detail in Section 5.

Assumption 2 (Overlap)**.** $\mathrm{P}(D=1) > \varepsilon$ *and* $\mathrm{P}(D=1|X_{t^*}, X_{t^*-1}, Z) < 1-\varepsilon$ *for some* $\varepsilon > 0$.

Assumption 3 (Conditional Parallel Trends)**.**

$$\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D=1] = \mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D=0].$$

Assumption 1 assumes we have access to an iid two-period panel. Assumption 2 is a standard version of an overlap condition often invoked in the DiD literature (e.g., Abadie (2005)). In practice, it says that, for all treated units, there exist untreated units with the same characteristics. Assumption 3 says that the path of untreated potential outcomes is the same on average for treated and untreated groups after conditioning on time-varying and time-invariant covariates. Assumption 3 also implicitly restricts the covariates to be unaffected by the treatment, ruling out so-called "bad controls." See Supplementary Appendix SA2.6 and Caetano et al. (2022) for more details.

Under Assumptions 1 to 3, the ATT is identified, and, in particular, it is given by

$$\text{ATT} = \mathbb{E}[\Delta Y_{t^*}|D=1] - \mathbb{E}\Big[\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]\Big|D=1\Big]. \tag{2}$$

We state this result formally in Proposition SA1 in the Supplementary Appendix. It follows from the same arguments as in existing work on difference-in-differences such as Heckman et al. (1998) and Abadie (2005), up to separately keeping track of the time-varying and time-invariant covariates. This expression says that the ATT can be recovered by comparing the mean path of outcomes for the treated group relative to the path they would have experienced if they had remained untreated. Under conditional parallel trends, the counterfactual path can be recovered by taking the path of outcomes conditional on time-varying and time-invariant covariates for the untreated group and averaging over the covariate distribution of the treated group.

3 Interpreting TWFE Regressions

This section considers how to interpret α in the TWFE regression in Equation (1). We continue to focus on the setting with two time periods where no units are treated in the first time period and some, but not all, units become treated in the second time period. This is a favorable setting for TWFE regressions as it does not introduce problems related to using already-treated units in the

comparison group (de Chaisemartin and D'Haultfoeuille (2020) and Goodman-Bacon (2021)).

For interpreting the TWFE regression, many of our results involve linear projections. Let $\mathrm{L}(D|\Delta X_{t^*})$ denote the (population) linear projection of D on ΔX_{t^*}.[4] That is,

$$\mathrm{L}(D|\Delta X_{t^*}) := \Delta X_{t^*}'\mathbb{E}[\Delta X_{t^*}\Delta X_{t^*}']^{-1}\mathbb{E}[\Delta X_{t^*}D] = \Delta X_{t^*}'\gamma.$$

Similarly, for $d \in \{0,1\}$, define

$$\mathrm{L}_d(\Delta Y_{t^*}|\Delta X_{t^*}) := \Delta X_{t^*}'\mathbb{E}[\Delta X_{t^*}\Delta X_{t^*}'|D=d]^{-1}\mathbb{E}[\Delta X_{t^*}\Delta Y_{t^*}|D=d] = \Delta X_{t^*}'\beta_d,$$

which is the linear projection of ΔY_{t^*} on ΔX_{t^*} for the treated group (when $d=1$) and for the untreated group (when $d=0$), respectively.

Notice that, with two periods, it is helpful to equivalently re-write Equation (1) as

$$\Delta Y_{it^*} = \alpha D_i + \Delta X_{it^*}'\beta + \Delta e_{it^*}. \tag{3}$$

We view Equation (3) as a linear projection model rather than as a linear conditional expectation/structural model, thus allowing for heterogeneous treatment effects. Our interest in this section is in determining what kind of conditions are required to interpret α as the ATT or at least as a weighted average of some underlying treatment effect parameters.

Next, we provide our first main result on interpreting α in terms of underlying causal effect parameters along with some additional bias terms.

[4] All of the linear projections in this section include an intercept. This involves a slight abuse of notation where, for example, we augment ΔX_{t^*} so that it includes an intercept in addition to the change in time-varying covariates over time. Similarly, we also slightly abuse notation in Equation (3) by taking β to include an extra parameter in its first element corresponding to the intercept. For all results below involving linear projections, we assume that they are well-defined. This typically involves a rank condition, such as that $\mathbb{E}[\Delta X_{t^*}\Delta X_{t^*}']$ is positive definite. The vast majority of our results are provided in terms of population (rather than sample) quantities. For expressions that only involve means and linear projections (which applies to many of our results below), analogous results hold for the corresponding sample quantities.

Theorem 1. *Under Assumptions 1 to 3, α from Equation (3) can be expressed as*

$$\begin{aligned}\alpha = \mathbb{E}\Big[w(\Delta X_{t^*})\mathrm{ATT}(X_{t^*}, X_{t^*-1}, Z)\Big|D=1\Big] \\ + \mathbb{E}\Big[w(\Delta X_{t^*})\Big\{\Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0] - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, D=0]\Big) \quad \text{(A)} \\ + \Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, D=0] - \mathbb{E}[\Delta Y_{t^*}|\Delta X_{t^*}, D=0]\Big) \quad \text{(B)} \\ + \Big(\mathbb{E}[\Delta Y_{t^*}|\Delta X_{t^*}, D=0] - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\Big\}\Big|D=1\Big], \quad \text{(C)}\end{aligned}$$

where $w(\Delta X_{t^*}) := \dfrac{1-\mathrm{L}(D|\Delta X_{t^*})}{\mathbb{E}[(1-\mathrm{L}(D|\Delta X_{t^*}))|D=1]}$, *and* $\mathbb{E}[w(\Delta X_{t^*})|D=1] = 1$.

Theorem 1 states that α is a weighted average of underlying conditional ATTs (we discuss the weights in more detail below) plus several undesirable bias terms.[5] Note that the weights $w(\Delta X_{t^*})$ can be negative if there exist values of ΔX_{t^*} among the treated group such that $\mathrm{L}(D|\Delta X_{t^*}) > 1$. Since $w(\Delta X_{t^*})$ has mean one, the bias terms should be a first-order concern for empirical researchers. This differs from several recent papers on interpreting regressions in different contexts, where the regression coefficient ends up including bias terms but the weights have mean zero (Sun and Abraham (2021), de Chaisemartin and D'Haultfoeuille (2023), and Goldsmith-Pinkham et al. (2024)).

The bias in Term (A) arises because the regression in Equation (3) does not include time-invariant covariates. This term suggests that failing to include time-invariant covariates in the TWFE regression when the path of untreated potential outcomes actually depends on time-invariant covariates undesirably contributes to how α is calculated. In our application to stand-your-ground laws, this bias term could arise because of state-level time-invariant covariates that affect the path of untreated potential outcomes (e.g., region indicators, if trends in homicides (absent the policy) are different across different regions of the country).[6]

[5]The proof of Theorem 1 (especially the parts concerning the weights) is mechanically related to work on interpreting cross-sectional regressions under unconfoundedness or other related settings (Angrist (1998), Aronow and Samii (2016), Sloczynski (2022), Goldsmith-Pinkham et al. (2024), Blandhol et al. (2025), and Hahn (2023)). The hidden linearity bias terms in the α expression (discussed in detail below) are specific to the DiD setting.

[6]As a point of clarification, it is common in empirical difference-in-differences applications that

Term (B) is nonzero when the path of untreated potential outcomes depends on the levels of time-varying covariates instead of only on the change in covariates over time. Together, we refer to Terms (A) and (B) as *hidden linearity bias*. Hidden linearity bias arises because an additional implication of linearity is that the estimating equation, as a by-product of differencing out the unit fixed effect, only ends up including the change in covariates over time. In the case where the model is correctly specified, then this transformation of the covariates is appropriate. However, if the model is viewed as an approximation, then an undesirable implication of linearity is that it implicitly changes the identification strategy from one that includes levels of time-varying covariates and time-invariant covariates to one that only includes the change in the covariates over time. For example, when including state population in a TWFE regression, the researcher likely intends to compare treated and untreated states with similar population levels. However, the TWFE regression effectively compares states with similar population changes, which, of course, could have quite different population levels.

Hidden linearity bias does not show up in cross-sectional settings. Modern empirical work often views linear regressions as approximations. In cross-sectional settings, the approximation view can be attractive as linearity itself may be a strong assumption, but using the linear model in estimation is convenient and has other good properties, such as being the best linear approximation to a possibly nonlinear conditional expectation function. In contrast, the discussion above

use state-level data to include region-time fixed effects (i.e., to include a region indicator with a time-varying coefficient). This partially, though not entirely, addresses the issues discussed in this section; see, in particular, Wooldridge (2025, Section 5.2). Perhaps a better example comes from DiD applications that use individual-level data, where it is less common to include time-invariant covariates with time-varying coefficients in the TWFE regression. For example, it is uncommon in labor economics to include a person's race as a covariate in a TWFE regression because it does not vary over time, despite the fact that it seems likely that the path of many labor market outcomes depends on race. Similarly, it is possible, but not common, to include a time-invariant continuous covariate with a time-varying coefficient in a TWFE regression.

highlights that linearity has substantially more bite for DiD. This suggests that linearity should be more carefully examined in panel data settings than in cross-sectional settings.

Term (C) is non-zero when the conditional expectation of the change in untreated potential outcomes conditional on covariates is nonlinear in the change in covariates over time. This type of linearity condition is the one that researchers would likely suspect to be implicit in the TWFE regression. A similar term shows up in cross-sectional settings with different papers discussing various conditions under which it is equal to zero (Angrist (1998), Blandhol et al. (2025), and Hahn (2023)). In general, this term is non-zero, though it may be reasonable to hope that the conditional expectation is close to being linear in many cases.

Next, we provide an assumption to eliminate the bias terms discussed above.

Assumption 4 (Additional Assumptions to Rule Out Bias Terms)**.**

(A) The path of untreated potential outcomes does not depend on time-invariant covariates. That is, $\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D=0] = \mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, D=0]$.

(B) The path of untreated potential outcomes only depends on the change in time-varying covariates. That is, $\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, D=0] = \mathbb{E}[\Delta Y_{t^*}(0)|\Delta X_{t^*}, D=0]$.

(C) The path of untreated potential outcomes is linear in the change in time-varying covariates. That is, $\mathbb{E}[\Delta Y_{t^*}(0)|\Delta X_{t^*}, D=0] = \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})$.

Theorem 2. *Under Assumptions 1 to 3, and if, in addition, Assumption 4 also holds, then*

$$\alpha = \mathbb{E}\Big[w(\Delta X_{t^*})\mathrm{ATT}(X_{t^*}, X_{t^*-1}, Z)\Big|D=1\Big],$$

where the weights $w(\Delta X_{t^*})$ *are the same ones defined in Theorem 1. If, in addition,* $\mathrm{ATT}(X_{t^*}, X_{t^*-1}, Z)$ *is constant across all values of* (X_{t^*}, X_{t^*-1}, Z)*, then*

$$\alpha = \mathrm{ATT}.$$

Theorem 2 provides sufficient conditions for α from Equation (3) to be equal to a weighted average of conditional ATTs under the conditional parallel trends assumption in Assumption 3.

The intuition for the result is that the conditions in Assumption 4 together imply that

$$\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0] = \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*}).$$

This is sufficient for the bias terms in Theorem 1 to be equal to 0, and, thus, α is equal to a weighted average of $\mathrm{ATT}(X_{t^*}, X_{t^*-1}, Z)$.

The result in Theorem 2 suggests several potential issues with the TWFE regression in Equation (3). First, the additional conditions in Assumption 4 are likely to be strong in many applications, and, perhaps more importantly, it is very uncommon for empirical work to grapple with whether or not these types of assumptions are plausible in a given application.

Second, even if one is willing to maintain the additional assumptions in Assumption 4, α from the TWFE regression is still hard to interpret for several reasons. The first issue with interpreting α is that, although the weights have mean one, it is possible to have negative weights for some values of $\mathrm{ATT}(X_{t^*}, X_{t^*-1}, Z)$. This can happen for values of the covariates among the treated group where $\mathrm{L}(D|\Delta X_{t^*}) > 1$, which is possible because $\mathrm{L}(D|\Delta X_{t^*})$ is a linear projection of a binary treatment on ΔX_{t^*} that is not restricted to be between 0 and 1. Negative weights have often been emphasized as being particularly problematic (see, for example, de Chaisemartin and D'Haultfoeuille (2020) and Blandhol et al. (2025)). For example, negative weights imply that it is possible to come up with examples where $\mathrm{ATT}(X_{t^*}, X_{t^*-1}, Z)$ is positive for all values of the covariates, but α could be negative due to the weighting scheme. In empirical work, estimating $\mathrm{L}(D|\Delta X_{t^*})$ and checking if there are negative weights is straightforward. Another issue is that the weights have a *weight-reversal* property (we adapt this terminology from Sloczynski (2022)). Notice that the ideal weighting scheme would be for $w(\Delta X)$ to be uniformly equal to one, in which case, $\alpha = \mathrm{ATT}$. Relative to this natural baseline, the weights in Theorem 2 indicate that α tends to put too much weight on conditional ATTs for values of the covariates that are relatively uncommon among the treated group relative to the untreated group and puts too little weight on conditional ATTs for values of the covariates that are relatively common among the treated group relative to the untreated group.

Finally, if, in addition to all the previous conditions, conditional ATTs are constant across dif-

ferent values of the covariates, then $\alpha = \text{ATT}$. This is a treatment effect homogeneity condition with respect to the covariates.[7] It *is* somewhat weaker than individual-level treatment effect homogeneity, and it allows for treatment effects to be systematically different for treated units relative to untreated units. Instead, it says that, for the treated group, treatment effects cannot be systematically different across different values of the covariates. However, this assumption is likely to be very strong in most economic applications, and it is not commonly considered in empirical work.

These results differ greatly from our earlier result on identifying the ATT in Equation (2). That result did not require any of the additional conditions in Assumption 4.

Remark 1 (Alternative conditions on the propensity score for interpreting α)**.** One can also show that α is equal to a weighted average of conditional ATTs under restrictions on the propensity score (rather than restrictions on $\mathbb{E}[\Delta Y_t(0)|X_{t^*}, X_{t^*-1}, Z, D = 0]$ as above); namely, $\text{P}(D = 1|X_{t^*}, X_{t^*-1}, Z) = \text{L}(D|\Delta X_{t^*})$. See Angrist (1998), Aronow and Samii (2016), and Sloczynski (2022) for results along these lines with cross-sectional data under unconfoundedness. In Section SA1.3, we argue that, in the panel data context that we consider, linearity conditions are less plausible on the propensity score than on the outcome models discussed above. Moreover, some leading cases where the propensity score would be linear by construction in cross-sectional settings do not apply in our setting.

Remark 2 (Comparison to conditions for other estimation strategies)**.** Interestingly, very similar restrictions as the ones discussed in Assumption 4 arise in some recently proposed "heterogeneity robust" versions of difference-in-differences. For example, the imputation approaches proposed in Gardner et al. (2023) and Borusyak et al. (2024) involve estimating the model $Y_{it}(0) = \theta_t + \eta_i + X_{it}'\beta + e_{it}$ (see Gardner et al. (2023, Eq. (7)) and Borusyak et al. (2024, Eq. (5))) which, in the two-

[7] Meyer (1995), Abadie (2005), and Sant'Anna and Zhao (2020) all mention that unmodeled treatment effect heterogeneity with respect to the covariates leads α to not be equal to the ATT. The first term in the expression for α in Theorem 1 provides an explicit expression for α when there is treatment effect heterogeneity, and, in agreement with those papers, our result indicates that $\alpha = \text{ATT}$ when there is no treatment effect heterogeneity with respect to the covariates.

period context considered here, implicitly uses the assumption that $\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D = 0] = \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})$—the same condition as implied by Assumption 4. Alternatively, the regression adjustment version of Callaway and Sant'Anna (2021) implicitly uses the assumption that $\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D = 0] = \mathrm{L}_0(\Delta Y_{t^*}|X_{t^*-1}, Z)$—besides linearity, this condition effectively says that the path of untreated potential outcomes does not depend on X_{t^*} once one controls for X_{t^*-1} and Z. The estimators we propose below do not include either of these types of auxiliary assumptions. See Appendix SA3.1 for a more detailed comparison.

4 Covariate Balance Diagnostics

Theorem 1 highlights several potential sources of bias from using the TWFE regression in Equation (3). In this section, we quantitatively assess *how much* these bias terms matter in practice. This is not an easy task as the conditional expectations in Terms (A)-(C) of Theorem 1 are difficult to estimate without imposing additional functional form assumptions. The misspecification bias terms in Terms (A)-(C) amount to violations of linearity that come from differences between $\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D = 0]$ and $\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})$. Below, we propose a simple approach to assess the sensitivity of α from the TWFE regression to possible violations of this linearity condition based on assessing implicit covariate balance. The second part of this section considers related diagnostics of augmented inverse propensity score weighting (AIPW) estimators of the ATT along the lines of the alternative estimators we propose later in the paper.

To motivate this section's results, note that if we could find "balancing weights" $\vartheta_0(X_{t^*}, X_{t^*-1}, Z)$ that re-weight the untreated group such that it has the same distribution of (X_{t^*}, X_{t^*-1}, Z) as the treated group, then it would be the case that

$$\begin{aligned}
\mathbb{E}[\Delta Y_{t^*}(0)|D = 1] &= \mathbb{E}\Big[\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D = 0]\Big|D = 1\Big] \\
&= \mathbb{E}\Big[\vartheta_0(X_{t^*}, X_{t^*-1}, Z)\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D = 0]\Big|D = 0\Big] \\
&= \mathbb{E}[\vartheta_0(X_{t^*}, X_{t^*-1}, Z)\Delta Y_{t^*}|D = 0],
\end{aligned}$$

and, therefore, that we could recover $\text{ATT} = \mathbb{E}[\Delta Y_{t^*}|D = 1] - \mathbb{E}[\vartheta_0(X_{t^*}, X_{t^*-1}, Z)\Delta Y_{t^*}|D = 0]$. In other words, if we could balance the distribution of covariates for the untreated group relative to the

treated group, then we could recover the path of untreated potential outcomes for the treated group by looking at the mean path of outcomes for the untreated group after it has been re-weighted to have the same distribution of covariates as the treated group. These sorts of balancing weights are related to a large number of weighting estimators. For example, in population, the weights from propensity score re-weighting satisfy this property (Rosenbaum and Rubin (1983)).

One important property of balancing weights is that they balance functions of the covariates across groups; i.e., for some function of the covariates g,

$$\mathbb{E}[g(X_{t^*}, X_{t^*-1}, Z)|D=1] = \mathbb{E}[\vartheta_0(X_{t^*}, X_{t^*-1}, Z)g(X_{t^*}, X_{t^*-1}, Z)|D=0]. \tag{4}$$

We show that TWFE and AIPW estimators can be re-expressed as weighting estimators with particular weights. Following the discussion above, we apply these weights to functions of the covariates to check how well these weights balance the covariates across groups. If the weights do not balance the covariates well, the corresponding estimator is more sensitive to violations of modeling assumptions for the outcome than if the weights balance the covariates well. Heuristically, unbalanced covariates that have larger effects on the outcome are more problematic than covariates that have smaller effects on the outcome. Finally, although we emphasize TWFE and AIPW, the same sorts of covariate balance diagnostics could be applied to any DiD estimator that can be expressed as a weighting estimator (e.g., DiD versions of matching, inverse probability tilting, entropy balancing, etc.).

4.1 TWFE Diagnostics

Returning to α from the TWFE regression, a useful insight is that it can be written as a re-weighting estimator. To see this, notice that it follows from Frisch-Waugh-Lovell arguments that

$$\alpha = \mathbb{E}\left[\frac{(D - \mathrm{L}(D|\Delta X_{t^*}))\Delta Y_{t^*}}{\mathbb{E}[(D - \mathrm{L}(D|\Delta X_{t^*}))^2]}\right]. \tag{5}$$

Let $\pi := \mathrm{P}(D=1)$, then it immediately follows from the law of iterated expectations that

$$\alpha = \mathbb{E}[w_1(\Delta X_{t^*})\Delta Y_{t^*}|D=1] - \mathbb{E}[w_0(\Delta X_{t^*})\Delta Y_{t^*}|D=0],$$

where $w_1(\Delta X_{t^*}) = \frac{\pi(1-\mathrm{L}(D|\Delta X_{t^*}))}{\mathbb{E}[(D-\mathrm{L}(D|\Delta X_{t^*}))^2]}$ and $w_0(\Delta X_{t^*}) = \frac{(1-\pi)\mathrm{L}(D|\Delta X_{t^*})}{\mathbb{E}[(D-\mathrm{L}(D|\Delta X_{t^*}))^2]}$. We refer to the weights $w_d(\Delta X_{t^*})$ as *implicit regression weights* below. Notice that these weights are simple to calculate, as the most complicated terms are linear projections. Building on the intuition for weighting estimators discussed earlier in this section, the diagnostics we propose in this section come from applying these weights to functions of the covariates to check how well the weights balance the covariates across groups. In the context of cross-sectional data under the assumption of unconfoundedness, Aronow and Samii (2016) and Chattopadhyay and Zubizarreta (2023) derive related weights and discuss a number of properties of these types of weights. For our purposes, the most notable property is that these weights will balance (in mean) the covariates that show up in the regression; thus, in our case, they will balance ΔX_{t^*} across groups. See Proposition SA3 for a more detailed explanation of why this is the case. Although the weights balance the mean of ΔX_{t^*}, they do not necessarily balance the distribution/means of the levels of time-varying covariates (that is, X_{t^*} or X_{t^*-1}) or of time-invariant covariates Z. Thus, our strategy below is to assess the sensitivity of the TWFE regression to violations of linearity by comparing terms such as

$$\begin{aligned}
&\mathbb{E}[w_1(\Delta X_{t^*})X_{t^*}|D=1] \quad \text{to} \quad \mathbb{E}[w_0(\Delta X_{t^*})X_{t^*}|D=0],\\
&\mathbb{E}[w_1(\Delta X_{t^*})X_{t^*-1}|D=1] \quad \text{to} \quad \mathbb{E}[w_0(\Delta X_{t^*})X_{t^*-1}|D=0], \text{ or}\\
&\mathbb{E}[w_1(\Delta X_{t^*})Z|D=1] \quad \text{to} \quad \mathbb{E}[w_0(\Delta X_{t^*})Z|D=0].
\end{aligned}$$

If these terms are all close to each other, it suggests that the implicit regression weights effectively balance time-invariant covariates and the levels of time-varying covariates between the treated group and the untreated group, and, hence, that α from the TWFE regression is not much affected by hidden linearity bias. On the other hand, if these terms are not close to each other, it suggests that α from the TWFE regression could be sensitive to violations of linearity.

4.2 AIPW Diagnostics

The main class of estimators that we suggest as alternatives to the TWFE regression are augmented inverse propensity score weighting (AIPW) estimators. These estimators involve estimating both an outcome regression model and a model for the propensity score. In this section, we

introduce the particular AIPW estimands that we consider. Following a similar motivation as in the previous section for TWFE regressions, we recast our AIPW approach as a weighting estimator. Then, we can apply these implicit AIPW weights to the covariates or functions of the covariates, allowing us to assess how well this estimation strategy balances covariate distributions for the treated and untreated groups.[8] As a step towards developing an AIPW estimator, it is a straightforward extension of the identification results in Equation (2) to show (see, e.g., Robins et al. (1994), Sloczynski and Wooldridge (2018), and Sant'Anna and Zhao (2020)) that

$$ATT = \mathbb{E}\left[\Delta Y_{t^*} - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]\middle|D=1\right] - \mathbb{E}\left[w_0^{aipw}\left(\Delta Y_{t^*} - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]\right)\middle|D=0\right], \quad (6)$$

where[9] $w_0^{aipw} := \dfrac{\varpi_0^{aipw}}{\mathbb{E}[\varpi_0^{aipw}|D=0]}$, with $\varpi_0^{aipw} := \dfrac{(1-\pi)p(X_{t^*}, X_{t^*-1}, Z)}{\pi\left(1-p(X_{t^*}, X_{t^*-1}, Z)\right)}$.

Estimating the ATT based on this expression requires first the estimation of $\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]$ and $p(X_{t^*}, X_{t^*-1}, Z)$. In this section, we specify a linear working model, $\mathrm{L}_0(\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z)$, for the expectation. Similarly, let $\tilde{p}(X_{t^*}, X_{t^*-1}, Z)$ denote a working model for $p(X_{t^*}, X_{t^*-1}, Z)$ (leading choices include a logit or probit model, but there are other possibilities).[10] We allow for the possibility that either or both of these models are misspecified. Given these working models for the outcome regression and the propensity score, we define

$$\widetilde{\mathrm{ATT}} = \mathbb{E}\left[\Delta Y_{t^*} - \mathrm{L}_0(\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z)\middle|D=1\right] - \mathbb{E}\left[\tilde{w}_0^{aipw}\left(\Delta Y_{t^*} - \mathrm{L}_0(\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z)\right)\middle|D=0\right] \quad (7)$$

[8] The results in this section build on several recent papers that have shown that ostensible outcome models can often be reinterpreted as weighting estimators; these include Robins et al. (2007), Kline (2011), and Chattopadhyay and Zubizarreta (2023), particularly Chattopadhyay and Zubizarreta (2023) though this paper is in the context of cross-sectional data under unconfoundedness. See Supplementary Appendix SA1 for additional discussion.

[9] All of the weights in this section are functions of (X_{t^*}, X_{t^*-1}, Z), but we omit this dependence to simplify notation.

[10] To be clear, the proof of Proposition 1 does not require any substantive restrictions on the model for the propensity score, but it does use linearity of the outcome regression model. That said, the outcome regression model could include interactions, higher order terms, etc.

where $\tilde{w}_0^{aipw} := \dfrac{\tilde{\varpi}_0^{aipw}}{\mathbb{E}[\tilde{\varpi}_0^{aipw}|D=0]}$, with $\tilde{\varpi}_0^{aipw} := \dfrac{(1-\pi)\tilde{p}(X_{t^*}, X_{t^*-1}, Z)}{\pi\big(1-\tilde{p}(X_{t^*}, X_{t^*-1}, Z)\big)}$.

$\widetilde{\text{ATT}}$ is a parametric AIPW estimand corresponding to the ATT expression in Equation (6) but with working models replacing the outcome regression and propensity score. The sample analog of $\widetilde{\text{ATT}}$ is doubly robust, in the sense that $\widetilde{\text{ATT}} = \text{ATT}$ if either $\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0] = \text{L}_0(\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z)$ or $p(X_{t^*}, X_{t^*-1}, Z) = \tilde{p}(X_{t^*}, X_{t^*-1}, Z)$, (i.e., if either the outcome regression model or the propensity score model is correctly specified). The following proposition shows that $\widetilde{\text{ATT}}$ can be written as a re-weighting estimator.

Proposition 1. *To conserve on notation, let $X = (X_{t^*}, X_{t^*-1}, Z)$. Define γ_0 as the linear projection coefficient from projecting $p(X)/\big(1-p(X)\big)$ on X; similarly define $\tilde{\gamma}_0$ as the linear projection coefficient from projecting $\tilde{p}(X)/\big(1-\tilde{p}(X)\big)$ on X. Then, under Assumptions 1 to 3,*

$$\widetilde{\text{ATT}} = \mathbb{E}\left[\vartheta_1^{aipw}\Delta Y_{t^*}\middle|D=1\right] - \mathbb{E}\left[\vartheta_0^{aipw}\Delta Y_{t^*}\middle|D=0\right],$$

where ϑ_1^{aipw} and ϑ_0^{aipw} are weights defined as

$$\vartheta_1^{aipw} := 1 \qquad \textit{and} \qquad \vartheta_0^{aipw} := \tilde{w}_0^{aipw} + \frac{\gamma_0' X}{\mathbb{E}[\gamma_0' X|D=0]} - \frac{\tilde{\gamma}_0' X}{\mathbb{E}[\tilde{\gamma}_0' X|D=0]},$$

such that $\mathbb{E}[\vartheta_1^{aipw}|D=1] = \mathbb{E}[\vartheta_0^{aipw}|D=0] = 1$ and $\mathbb{E}[\vartheta_0^{aipw} X|D=0] = \mathbb{E}[X|D=1]$.

The proof of Proposition 1 is provided in Supplementary Appendix SA1. Proposition 1 shows that the parametric AIPW estimand $\widetilde{\text{ATT}}$ can be re-formulated as a weighting estimator. It is possible for the weights to be negative; in applications, it is straightforward to calculate the sample analog of the weights—see Supplementary Appendix SA1 for more details. The main takeaway from Proposition 1 is that, unlike the implicit TWFE weights discussed above, the implicit AIPW weights balance the levels of time-varying covariates and time-invariant covariates across groups.

Remark 3 (Regression adjustment and IPW as special cases of AIPW)**.** Two special cases of the parametric AIPW estimand are worth mentioning. First, if we set $\tilde{p}(X_{t^*}, X_{t^*-1}, Z) = \pi$ (i.e., no covariates enter the propensity score working model), the second term in Equation (7) equals zero and $\widetilde{\text{ATT}}$ reduces to a regression adjustment estimand. Second, if the outcome regression working model includes only an intercept (i.e., no covariates enter the outcome regression), $\widetilde{\text{ATT}}$ reduces to

an inverse propensity score weighting (IPW) estimand. Our results in this section therefore cover both of these cases as well.

Remark 4 (Additional Diagnostics)**.** Although we emphasize covariate balance with respect to the first moment of the covariates, once we have re-formulated TWFE and AIPW as weighting estimators, all of the tools in the covariate balance checking toolkit become available, e.g., comparing higher order moments of covariates after re-weighting, plotting the distribution of covariates after re-weighting, and calculating the implied target population and effective sample size of the estimator. See Austin and Stuart (2015) and Imbens and Rubin (2015) for substantially more details.

5 Multiple Periods and Variation in Treatment Timing

In this section, we extend the two-period analysis to a setting with multiple periods and variation in treatment timing across units. This setting is common in empirical work in economics and has been studied in several recent papers (de Chaisemartin and D'Haultfoeuille (2020), Goodman-Bacon (2021), Callaway and Sant'Anna (2021), and Sun and Abraham (2021), among others). The proofs of all of the results in this section are provided in Supplementary Appendix SA2.

To start with, we introduce some additional notation and discuss how to extend the assumptions from Section 2 to the setting considered here (we provide formal versions of these assumptions as Assumptions MP-1 to MP-4 in Supplementary Appendix SA2). Let T denote the number of time periods. In this section, we allow for T to be larger than two, but we focus on "short" panels where T is considered to be fixed. We consider a setting with staggered treatment adoption, where (i) no units are treated in the first period (or units that are treated in the first period are dropped) and (ii) treatment timing can vary across units, but once a unit becomes treated, it remains treated in subsequent periods. Under staggered treatment adoption, a unit's entire sequence of treatments is fully characterized by its "group" where group refers to the period when the unit became treated. Let G_i denote a unit's group and denote the full set of groups by $\mathscr{G} \subseteq \{2,\ldots,\mathrm{T}+1\}$. We use the convention of setting $G_i = \mathrm{T}+1$ among units that do not participate in the treatment in any period from $2,\ldots,\mathrm{T}$,[11] and we define $\bar{\mathscr{G}} := \mathscr{G} \setminus \{\mathrm{T}+1\}$ as the set of groups that participate in

[11] In the literature, it is somewhat more common to set $G_i = \infty$ for never-treated units, but setting

the treatment in any period. It is also convenient to define a binary indicator for the never-treated group: let $U_i = 1$ for units that never participate in the treatment and $U_i = 0$ otherwise.

Let Y_{it} denote the observed outcome for unit i in time period t. Under staggered treatment adoption, we can define potential outcomes based on a unit's group; that is, let $Y_{it}(g)$ denote the potential outcome for unit i in time period t if it were in group g. In terms of potential outcomes, the observed outcome is $Y_{it} = Y_{it}(G_i)$. In other words, the observed outcome is the potential outcome according to unit i's actual group. To make the notation more transparent, we also define $Y_{it}(0)$ to be unit i's potential outcome in time period t if it never participated in the treatment. We also make a no-anticipation assumption that says that, in periods before a unit is treated, its observed outcomes are untreated potential outcomes. This rules out that the treatment affects outcomes in periods before the treatment actually occurs. Next, define X_{it} to be a $k \times 1$ vector of time-varying covariates, and let $\mathbf{X}_i := (X_{i1}', X_{i2}', \ldots, X_{i\mathrm{T}}')'$ denote the $Tk \times 1$ vector that stacks the time-varying covariates across periods. Finally, we continue to use Z_i to denote an $l \times 1$ vector of time-invariant covariates. Besides the staggered treatment adoption and no-anticipation assumptions, we also maintain an iid sampling assumption, a multi-period version of overlap, and a multi-period version of conditional parallel trends, the latter of which we provide here:

Assumption MP-PT (Multi-Period Parallel Trends)**.** *For $t = 2, \ldots, T$ and for all $g \in \mathscr{G}$,*

$$\mathbb{E}[\Delta Y_t(0) | \mathbf{X}, Z, G = g] = \mathbb{E}[\Delta Y_t(0) | \mathbf{X}, Z].$$

Following Callaway and Sant'Anna (2021) and Wooldridge (2025), we target the identification of group-time average treatment effects, defined as

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

$\mathrm{ATT}(g,t)$ is the average treatment effect for group g in period t. We also define the conditional-on-covariates version of group-time average treatment effects

$$\mathrm{ATT}_{g,t}(\mathbf{x}, z) := \mathbb{E}[Y_t(g) - Y_t(0) | \mathbf{X} = \mathbf{x}, Z = z, G = g].$$

$G_i = \mathrm{T} + 1$ unifies some of the notation for the TWFE decomposition results presented below.

In Proposition SA4, we show that both of these are identified and can be expressed as

$$\text{ATT}_{g,t}(\mathbf{X},Z) = \mathbb{E}[Y_t - Y_{g-1}|\mathbf{X},Z,G=g] - \mathbb{E}[Y_t - Y_{g-1}|\mathbf{X},Z,U=1]$$

and $$\text{ATT}(g,t) = \mathbb{E}[Y_t - Y_{g-1}|G=g] - \mathbb{E}\Big[\mathbb{E}[Y_t - Y_{g-1}|\mathbf{X},Z,U=1]\Big|G=g\Big].$$

This result generalizes the identification result in Equation (2) from a setting with two time periods to one with staggered treatment adoption. The argument closely follows the identification result for $\text{ATT}(g,t)$ in Callaway and Sant'Anna (2021) except that some covariates can be time-varying.

Group-time average treatment effects are important building blocks for our results below on interpreting TWFE regressions. However, unlike α from the TWFE regression in Equation (1), they are functional parameters in the sense that they can vary arbitrarily across g and t. Therefore, it is more natural to compare α from the TWFE regression to an aggregated causal effect parameter; in particular, we consider the following overall average treatment effect on the treated parameter

$$\text{ATT}^o := \mathbb{E}\Big[\bar{Y}^{post} - \bar{Y}(0)^{post}\Big|U=0\Big],$$

where, for units that ever participate in the treatment, we define

$$\bar{Y}_i^{post} := \frac{1}{\text{T}-G_i+1}\sum_{t=G_i}^{\text{T}} Y_{it} \qquad \text{and} \qquad \bar{Y}_i(0)^{post} := \frac{1}{\text{T}-G_i+1}\sum_{t=G_i}^{\text{T}} Y_{it}(0).$$

These are the average observed outcome and average untreated potential outcome, respectively, across unit i's post-treatment time periods. Thus, ATT^o is the average treatment effect across the population that participates in the treatment in any time period. Callaway and Sant'Anna (2021) show that

$$\text{ATT}^o = \sum_{g\in\bar{\mathscr{G}}}\sum_{t=g}^{\text{T}} w^o(g,t)\text{ATT}(g,t),$$

where $w^o(g,t) := \bar{p}_g/(\text{T}-g+1)$ and $\bar{p}_g := \text{P}(G=g|G\in\bar{\mathscr{G}})$, which is the probability of being in group g conditional on being among the set of groups that ever participates in the treatment.

TWFE Decomposition

Next, we provide a decomposition of α from Equation (1) under staggered treatment adoption. The discussion below uses double-demeaned random variables; for example, $\ddot{Y}_{it} := Y_{it} - \bar{Y}_i - \mathbb{E}[Y_t] +$

$\frac{1}{\mathrm{T}}\sum_{s=1}^{\mathrm{T}}\mathbb{E}[Y_s]$. We focus on estimating α from Equation (1) by fixed effects estimation. Thus, after applying the double-demeaning transformation, we use the following estimating equation:

$$\ddot{Y}_{it} = \alpha \ddot{D}_{it} + \ddot{X}_{it}'\beta + \ddot{e}_{it}.$$

Before providing our main results, we need to introduce more notation. First, notice that a unit's group fully determines $\ddot{D}_{it}$; i.e., $\ddot{D}_{it} = h(G_i, t)$ where

$$h(g,t) := \mathbf{1}\{t \geq g\} - \frac{\mathrm{T}-g+1}{\mathrm{T}} - \mathbb{E}[D_t] + \frac{1}{\mathrm{T}}\sum_{s=1}^{\mathrm{T}}\mathbb{E}[D_s].$$

Next, define the population linear projection of $\ddot{D}_{it}$ on $\ddot{X}_{it}$ as

$$\mathrm{L}(\ddot{D}_t|\ddot{X}_t) = \ddot{X}_t'\mathbb{E}\left[\frac{1}{\mathrm{T}}\sum_{s=1}^{\mathrm{T}}\ddot{X}_s\ddot{X}_s'\right]^{-1}\mathbb{E}\left[\frac{1}{\mathrm{T}}\sum_{s=1}^{\mathrm{T}}\ddot{X}_s\ddot{D}_s\right] =: \ddot{X}_t'\Gamma,$$

and the population linear projection of $(Y_{it}\text{-}Y_{ig-1})$ on $(X_{it}\text{-}X_{ig-1})$ using the never-treated group as

$$\mathrm{L}_0\Big(Y_t\text{-}Y_{g-1}\Big|X_t\text{-}X_{g-1}\Big) =: \lambda_{0,t,g-1} + \big(X_t\text{-}X_{g-1}\big)'\Lambda_{0,t,g-1},$$

where $\lambda_{0,t,g-1}$ is the intercept and $\Lambda_{0,t,g-1}$ is the slope coefficient, both of which can vary by the period t and the base period $(g\text{-}1)$. Furthermore, define Λ_0 as the vector of coefficients from a TWFE regression of Y_{it} on X_{it} using only the never-treated group (see Equation (SA10) for the complete expression), and define $\lambda_t := \mathbb{E}[Y_t - X_t'\Lambda_0|U=1]$. Finally, define

$$\xi_{t,g-1}(\mathbf{X},Z) := \mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X},Z,U\text{=}1] - \Big((\lambda_t - \lambda_{g-1}) + (X_t\text{-}X_{g-1})'\Lambda_0\Big),$$

which will correspond to misspecification bias similar to terms (A), (B), and (C) in Theorem 1 (more details below). Next, we provide our main result relating α to underlying causal effect parameters and misspecification bias terms under staggered treatment adoption.

Theorem 3. *Under Assumptions MP-PT and MP-1 to MP-4,*

$$\alpha = \sum_{g\in\bar{\mathscr{G}}}\sum_{t=g}^{T}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\Big(\mathrm{ATT}_{g,t}(\mathbf{X},Z) + \xi_{t,g-1}(\mathbf{X},Z)\Big)\Big|G=g\Big] + \sum_{g\in\bar{\mathscr{G}}}\sum_{t=1}^{g-1}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\xi_{t,g-1}(\mathbf{X},Z)\Big|G=g\Big], \quad (8)$$

$$\textit{where } w_{g,t}^{twfe}(\ddot{X}_t) := \frac{\Big(h(g,t) - \ddot{X}_t'\Gamma\Big)\pi_g}{\sum_{l\in\bar{\mathscr{G}}}\sum_{s=l}^{T}\mathbb{E}\Big[\big(h(l,s) - \ddot{X}_{is}'\Gamma\big)\Big|G\text{=}l\Big]\pi_l}, \ \sum_{g\in\bar{\mathscr{G}}}\sum_{t=g}^{T}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\Big|G\text{=}g\Big]\text{=}1, \textit{ and } \sum_{g\in\bar{\mathscr{G}}}\sum_{t=1}^{g-1}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\Big|G\text{=}g\Big]\text{=-}1.$$

Theorem 3 shows that α from the TWFE regression in Equation (1) is equal to a weighted average of conditional-on-covariates group-time average treatment effects plus misspecification bias terms. The first term in Equation (8) covers post-treatment periods while the second term covers pre-treatment periods. The misspecification bias terms arise in both pre- and post-treatment periods; if there are violations of conditional parallel trends, these would also show up in the second term (see Proposition SA7). This result is analogous to (and extends) the result in Theorem 1 in the case with exactly two periods. Like the earlier case, the weights on conditional group-time average treatment effects are (i) driven by the estimation method, (ii) can be negative, and (iii) sum to one across post-treatment periods.

The following result decomposes the misspecification bias terms in Theorem 3.

Proposition 2. *Under Assumptions MP-1 to MP-4 and MP-PT, the misspecification bias terms in Theorem 3 can be decomposed as*

$$\begin{aligned}
\xi_{t,g-1}(\mathbf{X},Z) = {} & \mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X},Z,U=1] - \mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X},U=1]\Big) && \text{(MB-1)}\\
& + \Big(\mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X},U=1] - \mathbb{E}[Y_t\text{-}Y_{g-1}|X_t,X_{g-1},U=1]\Big) && \text{(MB-2)}\\
& + \Big(\mathbb{E}[Y_t\text{-}Y_{g-1}|X_t,X_{g-1},U=1] - \mathbb{E}[Y_t\text{-}Y_{g-1}|(X_t\text{-}X_{g-1}),U=1]\Big) && \text{(MB-3)}\\
& + \Big(\mathbb{E}[Y_t\text{-}Y_{g-1}|(X_t\text{-}X_{g-1}),U=1] - \big(\lambda_{0,t,g-1} + (X_t\text{-}X_{g-1})'\Lambda_{0,t,g-1}\big)\Big) && \text{(MB-4)}\\
& + \Big(\big(\lambda_{0,t,g-1} - (\lambda_t - \lambda_{g-1})\big) + (X_t\text{-}X_{g-1})'(\Lambda_{0,t,g-1} - \Lambda_0)\Big). && \text{(MB-5)}
\end{aligned}$$

Next, we discuss the components of the misspecification bias terms in Proposition 2 along with a set of sufficient conditions to eliminate them from the expression for α in Theorem 3. These conditions rationalize interpreting α from Equation (1) as a weighted average of $\text{ATT}_{g,t}(\mathbf{X},Z)$. The conditions are stated formally in Assumption MP-5.

Conditions to Eliminate Misspecification Bias

(1) The path of untreated potential outcomes does not depend on time-invariant covariates.

(2) The path of untreated potential outcomes does not depend on time-varying covariates in other periods besides $(g-1)$ and t.

(3) The path of untreated potential outcomes only depends on the change in time-varying covariates between periods $(g-1)$ and t.

(4) The path of untreated potential outcomes is linear in the change in time-varying covariates.

(5) The effect of the change in time-varying covariates over time on the path of untreated potential outcomes is constant across time periods.

Each condition serves to set the corresponding term in Proposition 2 to 0. Conditions (1)-(3) are all required to deal with the multiple-period version of hidden linearity bias: that transforming the model to eliminate the unit fixed effect also changes the functional form of the time-varying covariates and eliminates the time-invariant covariates, and, hence, effectively results in changing the parallel trends assumption. Condition (4) is a linearity requirement, similar to Condition (C) in Assumption 4 in the two-period case. Condition (5) has no immediate analog in the two-period case. It says that the path of untreated potential outcomes can depend on the magnitude of changes in time-varying covariates, but the impact of a given change should not vary across time periods.

Although these conditions eliminate misspecification bias, we show in the Supplementary Appendix (Theorem SA1) that, similarly to the two period case, even if these conditions hold, α is still equal to a weighted average of $\text{ATT}_{g,t}(\mathbf{X}, Z)$ with weights that are (i) non-transparently driven by the estimation method, (ii) difficult to rationalize, and (iii) can be negative (as pointed out in de Chaisemartin and D'Haultfoeuille (2020)). We also show that, under additional restrictions on treatment effect heterogeneity with respect to covariates, groups, and time periods, $\alpha = \text{ATT}$. However, like the case with two periods, these extra conditions are likely to be very strong in most applications. The results in Theorems 3 and SA1 are related to results in several other papers, including Goodman-Bacon (2021), Lin and Zhang (2022), and Ishimaru (2022). See Supplementary Appendix SA2 for a detailed discussion.

AIPW Estimands with Multiple Periods

Next, we consider parametric AIPW estimands for group-time average treatment effects and the overall average treatment effect with multiple periods and variation in treatment timing. This is the population version of our main alternative estimator to the TWFE regression; see the next section for further details. Define the parametric AIPW estimand for $\text{ATT}(g,t)$ as

$$\begin{aligned}\widetilde{ATT}^{aipw}(g,t) = \mathbb{E}\Big[(Y_t - Y_{g-1}) - \tilde{\text{L}}^0_{g,t}(Y_t - Y_{g-1}|\mathbf{X},Z)\Big|G=g\Big] \\ - \mathbb{E}\Big[\tilde{w}^{0,aipw}_{g,t}(\mathbf{X},Z)\big((Y_t - Y_{g-1}) - \tilde{\text{L}}^0_{g,t}(Y_t - Y_{g-1}|\mathbf{X},Z)\big)\Big|U=1\Big],\end{aligned} \tag{9}$$

where $\tilde{\text{L}}^0_{g,t}(Y_t - Y_{g-1}|\mathbf{X},Z)$ is the linear projection of $Y_t - Y_{g-1}$ onto $\mathbf{W}_{g,t}$ among comparison units, with $\mathbf{W}_{g,t}$ a possibly (g,t)-varying vector of regressors that may use covariate values from selected time periods (e.g., X_t, X_{g-1}, and Z) and may include interactions and higher-order terms, and $\tilde{w}^{0,aipw}_{g,t} := \dfrac{\tilde{\varpi}^{0,aipw}_{g,t}(\mathbf{X},Z)}{\mathbb{E}[\tilde{\varpi}^{0,aipw}_{g,t}(\mathbf{X},Z)|U=1]}$, with $\tilde{\varpi}^{0,aipw}_{g,t}(\mathbf{X},Z) = \dfrac{\pi_0 \tilde{p}_{g,t}(\mathbf{X},Z)}{\pi_g\big(1-\tilde{p}_{g,t}(\mathbf{X},Z)\big)}$, $\pi_g := \text{P}(G=g)$, $\pi_0 := \text{P}(U=1)$, and $\tilde{p}_{g,t}(\mathbf{X},Z)$ is a parametric working model for the generalized propensity score

$$p_g(\mathbf{X},Z) := \text{P}(G=g|\mathbf{X},Z,\mathbf{1}\{G=g\}+U=1),$$

which is the conditional probability of being in group g conditional on being in group g or the never-treated group. We index $\tilde{p}_{g,t}(\mathbf{X},Z)$ by g and t to allow the working model for the generalized propensity score to change across time periods, particularly with respect to which time-varying covariates are included in the model. Note that $\widetilde{\text{ATT}}^{aipw}(g,t)$ is a model-dependent quantity that depends on the choice of working models $\tilde{\text{L}}^0_{g,t}$ and $\tilde{p}_{g,t}$, and therefore differs in general from the causal target parameter $\text{ATT}(g,t)$. It is straightforward to show, analogously to the two-period case, that the sample analog of $\widetilde{\text{ATT}}^{aipw}(g,t)$ is doubly robust for $\text{ATT}(g,t)$ (see the next section for more details). In addition, define the following parametric AIPW estimand for ATT^o:

$$\widetilde{\text{ATT}}^{aipw,o} = \sum_{g\in\bar{\mathcal{G}}}\sum_{t=g}^{\text{T}} \widetilde{\text{ATT}}^{aipw}(g,t) w^o(g,t). \tag{10}$$

Covariate Balance Diagnostics

In Supplementary Appendix SA2, we extend the covariate balance diagnostics for TWFE and AIPW with two periods (Section 4) to the staggered treatment adoption setting considered in this section. In particular, in Proposition SA9, we show that α can be rewritten in terms of implicit regression weights:

$$\alpha = \sum_{g\in\mathcal{G}}\sum_{t=g}^{\mathrm{T}} \bar{w}^{twfe}(g,t)\left\{\mathbb{E}\left[w_{g,t}^{1,twfe}(\mathbf{X},Z)(Y_t - Y_{g-1})\middle| G=g\right] - \mathbb{E}\left[w_{g,t}^{0,twfe}(\mathbf{X},Z)(Y_t-Y_{g-1})\middle| U=1\right]\right\} \quad (11)$$
$$+ \sum_{g\in\mathcal{G}}\sum_{t=1}^{g-1} \bar{w}^{twfe}(g,t)\left\{\mathbb{E}\left[w_{g,t}^{1,twfe}(\mathbf{X},Z)(Y_t - Y_{g-1})\middle| G=g\right] - \mathbb{E}\left[w_{g,t}^{0,twfe}(\mathbf{X},Z)(Y_t-Y_{g-1})\middle| U=1\right]\right\} + r,$$

where the sum weights $\bar{w}^{twfe}(g,t) := \mathbb{E}[w_{g,t}^{twfe}(\ddot{X}_t)|G=g]$, the expectation weights $w_{g,t}^{1,twfe}(\mathbf{X},Z) := \frac{(\ddot{D}_t - \ddot{X}_t'\Gamma)}{\mathbb{E}[(\ddot{D}_t - \ddot{X}_t'\Gamma)|G=g]}$ and $w_{g,t}^{0,twfe}(\mathbf{X},Z) := \frac{(\ddot{D}_t - \ddot{X}_t'\Gamma)}{\mathbb{E}[(\ddot{D}_t - \ddot{X}_t'\Gamma)|U=1]}$, and r is a remainder term.[12]

Similarly, we show in Proposition SA10 that $\widetilde{\mathrm{ATT}}^{aipw,o}$ can be rewritten in terms of weighted averages of paths of outcomes for each group relative to the never-treated group. In particular,

$$\widetilde{ATT}^{aipw,o} = \sum_{g\in\mathcal{G}}\sum_{t=g}^{\mathrm{T}} w^o(g,t)\left\{\mathbb{E}\left[\vartheta_{g,t}^{1,aipw}(\mathbf{X},Z)(Y_t - Y_{g-1})\middle| G=g\right] - \mathbb{E}\left[\vartheta_{g,t}^{0,aipw}(\mathbf{X},Z)(Y_t-Y_{g-1})\middle| U=1\right]\right\}, \quad (12)$$

where $w^o(g,t)$ are the same weights as in ATT^o above, $\vartheta_{g,t}^{1,aipw}(\mathbf{X},Z) := 1$, and

[12] The remainder term is a byproduct of using $(g-1)$ as the base period in the decomposition of α presented here. In the discussion after Proposition SA9, we argue that this term is likely to be small in most applications, and, indeed, this term is negligible in all diagnostics we report in our application. Very similar arguments can rationalize a different base period choice, such as $(g-2)$ or $(g-3)$, which could be attractive in applications with anticipation effects (see Callaway and Sant'Anna (2021) and Sun and Abraham (2021)). In Proposition SA8, we also provide a decomposition that uses period one as the base period, and it involves exactly the same weights but does not include a remainder term. We prefer the decomposition using $(g-1)$ as the base period because (i) it allows a direct comparison with the parametric AIPW estimand discussed below, where differences are due entirely to differences in the implicit weighting schemes, and (ii) it allows us to quantify how much pre-treatment violations of parallel trends contribute to α.

$$\vartheta_{g,t}^{0,aipw}(\mathbf{X},Z) := \tilde{w}_{g,t}^{0,aipw}(\mathbf{X},Z) + \frac{\gamma_{g,t}'\mathbf{W}_{g,t}}{\mathbb{E}[\gamma_{g,t}'\mathbf{W}_{g,t}|U=1]} - \frac{\tilde{\gamma}_{g,t}'\mathbf{W}_{g,t}}{\mathbb{E}[\tilde{\gamma}_{g,t}'\mathbf{W}_{g,t}|U=1]}.$$

where $\mathbf{W}_{g,t}$ denotes the covariates used in $\tilde{\mathrm{L}}_{g,t}^{0}(Y_t\text{–}Y_{g-1}|\mathbf{X},Z)$, $\gamma_{g,t}$ is the linear projection coefficient from projecting $p_g(\mathbf{X},Z)/(1-p_g(\mathbf{X},Z))$ on $\mathbf{W}_{g,t}$, and $\tilde{\gamma}_{g,t}$ is the linear projection coefficient from projecting $\tilde{p}_{g,t}(\mathbf{X},Z)/(1-\tilde{p}_{g,t}(\mathbf{X},Z))$ on $\mathbf{W}_{g,t}$.

Applying the TWFE or AIPW weights to (functionals of) $\mathbf{X}$ and Z allows the researcher to assess how well each approach implicitly balances covariates. The key differences between TWFE and AIPW covariate balance properties are: (i) TWFE weights depend only on transformed time-varying covariates, not levels or time-invariant covariates, while AIPW balances whatever covariates are included in the model; (ii) AIPW weights balance towards group g (the correct target for ATT(g,t)), while TWFE re-weights both groups; (iii) TWFE can be affected by pre-treatment violations of parallel trends, while AIPW is not; and (iv) both can have negative weights.

6 Alternative Estimation Strategies

This section discusses alternative estimation strategies that avoid the limitations of TWFE regressions discussed above. First, we discuss AIPW estimators, which are attractive and straightforward to adapt to our context. Second, from an empirical perspective, the main complication is that, with panel data and time-varying covariates, the dimension of the covariates can be very large. We discuss several different dimension reduction ideas in the second part of this section.

To start, define $m_{g,t}(\mathbf{X},Z) := \mathbb{E}[Y_t(0) - Y_{g-1}(0)|\mathbf{X},Z,U=1]$. We refer to $m_{g,t}(\mathbf{X},Z)$ as an outcome regression model. Let $\hat{m}_{g,t}(\mathbf{X},Z)$ and $\hat{p}_{g,t}(\mathbf{X},Z)$ denote estimators of $m_{g,t}(\mathbf{X},Z)$ and the generalized propensity score $p_g(\mathbf{X},Z)$, respectively (in line with the discussion above, we allow the model for the generalized propensity score to change across time periods). Then, we consider AIPW estimators of ATT(g,t) of the form

$$\widehat{\mathrm{ATT}}^{aipw}(g,t) := \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{w}_{g,t}^{1,aipw}(\mathbf{X}_i,Z_i) - \hat{w}_{g,t}^{0,aipw}(\mathbf{X}_i,Z_i)\Big)\Big((Y_t - Y_{g-1}) - \hat{m}_{g,t}(\mathbf{X}_i,Z_i)\Big),$$

which, after slightly re-arranging terms, is the sample analog of Equation (9) (where here we also

allow for the possibility of a nonlinear model for the outcome regression), and where

$$\hat{w}_{g,t}^{1,aipw}(\mathbf{X}_i, Z_i) := \frac{\mathbf{1}\{G_i = g\}}{\hat{\pi}_g} \quad \text{and} \quad \hat{w}_{g,t}^{0,aipw}(\mathbf{X}_i, Z_i) := \frac{\mathbf{1}\{U_i = 1\}\frac{\hat{p}_{g,t}(\mathbf{X}_i, Z_i)}{1-\hat{p}_{g,t}(\mathbf{X}_i, Z_i)}}{\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}\{U_j = 1\}\frac{\hat{p}_{g,t}(\mathbf{X}_j, Z_j)}{1-\hat{p}_{g,t}(\mathbf{X}_j, Z_j)}}.$$

AIPW estimators have been well-studied and have several known properties, as we discuss next.

Remarks on AIPW Estimation

1. If we specify parametric models for $m_{g,t}(\mathbf{X}, Z)$ and $p_g(\mathbf{X}, Z)$, then $\widehat{\text{ATT}}^{aipw}(g,t)$ is doubly robust for $\text{ATT}(g,t)$. Thus, if either the outcome regression or the propensity score model is correctly specified, then the estimator is consistent for $\text{ATT}(g,t)$. See Robins et al. (1994), Scharfstein et al. (1999), and Sloczynski and Wooldridge (2018) for general results on the double robustness property of AIPW estimators and Sant'Anna and Zhao (2020) for the specific case of DiD.
2. Given parametric models for $m_{g,t}(\mathbf{X}, Z)$ and $p_g(\mathbf{X}, Z)$, asymptotic normality of $\widehat{\text{ATT}}^{aipw}(g,t)$ holds under Assumptions MP-PT and MP-1 to MP-4 and weak regularity conditions following the same arguments as in Callaway and Sant'Anna (2021).
3. The estimator $\widehat{\text{ATT}}^{aipw}(g,t)$ can be used to construct an estimator of the overall average treatment effect, ATT^o, by averaging over all groups and time periods. In particular,

$$\widehat{\text{ATT}}^{o} = \sum_{g \in \bar{\mathcal{G}}} \sum_{t=g}^{\text{T}} \hat{w}^o(g,t)\widehat{\text{ATT}}^{aipw}(g,t),$$

   where $\hat{w}^o(g,t) = \frac{\hat{\text{P}}(G=g|U=0)}{T-g+1}$. This estimator is consistent for ATT^o and is asymptotically normal under the same conditions discussed above. Similar results hold for event studies or other aggregated parameters that can be expressed as weighted averages of $\text{ATT}(g,t)$. These results follow directly from those provided in Callaway and Sant'Anna (2021).
4. Regression adjustment is a special case of AIPW when no covariates are included in the generalized propensity score model. In this case $\widehat{\text{ATT}}^{aipw}(g,t)$ simplifies to

$$\widehat{\text{ATT}}^{ra}(g,t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{1}\{G_i = g\}}{\hat{\pi}_g}\Big(Y_{it} - Y_{ig-1} - \hat{m}_{g,t}(\mathbf{X}_i, Z_i)\Big),$$

   where $\widehat{\text{ATT}}^{ra}(g,t)$ denotes the regression adjustment estimator of $\text{ATT}(g,t)$. In this case, consistent and asymptotically normal estimation of $\text{ATT}(g,t)$ hinges on correct specification of

the outcome regression $m_{g,t}(\mathbf{X}, Z)$. Regression adjustment is an important special case because small groups (in terms of the number of observations) are fairly common in empirical work. In such cases, estimating the generalized propensity score can be highly unstable, making regression adjustment a possibly more appropriate alternative.

5. Several extensions to the AIPW estimator above apply in our case, such as accounting for anticipation effects and using alternative comparison groups (e.g., not-yet-treated versus never-treated); see Callaway and Sant'Anna (2021) for details. Since these issues are standard, we do not elaborate, though they are often important in empirical work and are supported in our code.
6. In cases where the researcher does not wish to specify parametric models for $m_{g,t}(\mathbf{X}, Z)$ and $p_g(\mathbf{X}, Z)$, essentially the same estimator can be used but with machine learners or nonparametric estimators replacing the parametric models (following similar arguments as those in Chernozhukov et al. (2018)).

Dimension Reduction

An important practical challenge is that, by construction, the dimension of the covariates in $m_{g,t}(\mathbf{X}, Z)$ and $p_g(\mathbf{X}, Z)$ is likely to be high as $\mathbf{X}$ is of dimension Tk, where k is the number of time-varying covariates. In most applications, reducing the dimension of the covariates will be desirable.[13] One leading dimension-reducing assumption is that

$$m_{g,t}(\mathbf{X}, Z) = m_{g,t}(X_t - X_{g-1}, X_{g-1}, Z) \quad \text{and} \quad p_g(\mathbf{X}, Z) = p_{g,t}(X_t - X_{g-1}, X_{g-1}, Z),$$

which says that, in terms of time-varying covariates, the outcome regressions and generalized propensity scores only depend on (i) the change in the time-varying covariates from the base period to the current period and (ii) the level of the time-varying covariates in the base period, rather than the covariates across all time periods. This type of specification includes both types of covariates that show up in Callaway and Sant'Anna (2021) and in imputation approaches such as Gardner et

[13] Related dimension reduction assumptions also appear in the literature on marginal structural models (e.g., Robins et al. (2000)), where outcome models and propensity scores are typically specified as parsimonious functions of covariate history rather than the full covariate trajectory.

al. (2023) and Borusyak et al. (2024). Other options are to (i) assume that $m_{g,t}(\mathbf{X}, Z) = m_{g,t}(\bar{X}, Z)$ and $p_g(\mathbf{X}, Z) = p_g(\bar{X}, Z)$ where $\bar{X}$ is the average of each time-varying covariate, or (ii) choose the covariates in the outcome regression and generalized propensity score in a data-driven way (see Supplementary Appendix SA4 for an example). Rather than necessarily advocating a particular approach to dimension reduction, we instead emphasize that any approach to dimension reduction should be a carefully and transparently considered step of the analysis rather than being inherited from the estimation strategy, as is the case with TWFE regressions.

7 Application

In this section, we apply our methods to study the effect of stand-your-ground laws on homicides, building on Cheng and Hoekstra (2013). Stand-your-ground laws remove the duty to retreat in violent altercations. Cheng and Hoekstra (2013) study 2000-2010, when 20 states implemented such laws in a staggered fashion. There are contrasting theoretical implications of the effect of stand-your-ground policies on homicides: possible deterrence effects (fewer altercations, fewer homicides) versus increased deadliness (more homicides per altercation). Cheng and Hoekstra (2013) find that stand-your-ground laws increased homicides.

[Table 1 about here.]

Below, we provide two sets of results using two different subsets of the data. First, we use a subset of the data that only includes the years 2000 and 2010. This first dataset is in line with our arguments above for the case of exactly two periods. While we report estimates of the effects of stand-your-ground policies on homicides, much of our main interest is in how different estimation strategies (based on *the same identification strategies*) balance the distribution of covariates. We are able to assess this using the covariate balance diagnostics that we proposed for TWFE and AIPW earlier in the paper. Table 1 provides summary statistics using the two-period subset of the data and for the full set of covariates used in Cheng and Hoekstra (2013). There are notable differences in covariates between treated and untreated states. Treated states were more likely to be Southern/Midwestern, had lower median income, more prisoners, and higher population,

poverty, and unemployment rates. Some differences also appear in covariate changes: poverty and prisoners increased more in treated states, while median income decreased more. For our second set of results, we mimic Cheng and Hoekstra (2013)'s setting much more closely: we use the full data of all 50 states across all available years, the same set of covariates as in one of their main specifications, and the same sampling weights.

Many of our results in this section are reported in figures that summarize covariate balance after applying the implicit weighting schemes for different estimation strategies. The figures report the standardized difference between the treated and comparison group for each covariate considered. The standardized difference is the difference between the average value of the covariate for the treated group relative to the untreated group, scaled by the pooled standard deviation of the covariate. We report both raw covariate balance and covariate balance after applying the implicit TWFE or AIPW weights.[14] To give a sense of magnitude: typically, standardized differences of around 0.1 or smaller are considered small; around 0.3 are considered medium; and around 0.5 or larger are considered large. See, for example, Imbens and Rubin (2015) for a textbook discussion.

Results with Two Periods and only Population and Region as Covariates

For the first set of results, we consider a highly simplified setting. The outcome is the log of the number of homicides in a state. We consider one time-varying covariate: the log of a state's population, and one time-invariant covariate: the Census region (Midwest, Northeast, South, or West). The intuition for this identification strategy is that a researcher would like to estimate the impact of the policy by comparing the change in log homicides among treated and untreated states that have similar populations and are located in the same region of the country (although region is not included in the regression, the argument would be that it is implicitly controlled for with the unit fixed effect). We also assess balance of $\mathbf{1}\{\log(\text{population}) \leq 15\}$ to see how well different approaches balance the fraction of small states, as none of the approaches considered

[14]The figures assess covariate balance, but not whether covariates have the same distribution as for the treated group; the latter holds by construction for AIPW and regression adjustment estimators but not for TWFE. See Remark SA1 for more details.

below mechanically force this variable to be balanced.

[Figure 1 about here.]

Figure 1 summarizes treatment effects and covariate balance. Both panels show qualitatively similar point estimates (TWFE: 0.115, AIPW: 0.157). The AIPW estimate is roughly 40% larger in magnitude, though neither estimate is statistically different from zero, nor are the estimates statistically different from each other. Because we are only using two time periods, these results are unsurprising.

The covariate balance results are more interesting. In the raw data (the red circles in the figure), there is a small imbalance in population changes, a moderate imbalance in population levels, and a large imbalance in regional distribution. TWFE (Panel a) balances changes in log population (which aligns with the theoretical properties of TWFE), but, in terms of balancing the levels of log population, the regression essentially has no effect: the standardized difference is only 3% smaller using the implicit TWFE regression weights than in the raw data. In other words, *controlling for the change in the log of population in the TWFE regression does not result in any more similar treated and untreated groups in terms of log population levels*, which was one of the main goals of including the state's population in the model to begin with. Similarly, the TWFE regression essentially does not affect the balance of the region indicators or the fraction of small states. In contrast, AIPW (Panel b) perfectly balances all the covariates included in the model, which is in line with its theoretical properties, and improves balance with respect to the fraction of small states.

Given these results, an interesting follow-up question is: what is the main driver of the improved covariate balance across estimators? We investigate this question in Supplementary Appendix SA4, where we provide covariate balance statistics for eight different specifications varying both the estimator (between regression adjustment and AIPW) and the covariates included (across (i) changes in time varying covariates only, (ii) pre-treatment levels of time-varying covariates only, (iii) both time-varying and pre-treatment levels of covariates, and (iv) pre-treatment levels of covariates and time-invariant covariates). The two main takeaways are, first, that the choice of estimator (regression adjustment vs. AIPW) matters much less than the covariate specification. In

fact, both of these give similar estimates to TWFE and have similar covariate balancing properties as TWFE when only the change in time-varying covariates is included. Second, specifications that directly include pre-treatment level of log population and region indicators do well at balancing the level of log population in each period, suggesting large gains from including the level of a time-varying covariate in at least one period.

Results with More Periods and More Covariates

Next, we use a much closer specification to the one used in Cheng and Hoekstra (2013). We take the log of homicides per 100,000 people in the state as the outcome, and use sampling weights based on the state's average population.[15] We also include additional covariates: log of the number of police per 100,000 population, log of the number of incarcerated persons per 100,000, log of government spending on assistance and subsidies per capita, log of government spending on public welfare per capita, median household income, poverty rate, unemployment rate, and demographic shares of black and white males ages 15–24 and 25–44. Because our unit of observation is the state, the size of many of our groups is very small. This results in AIPW estimation being infeasible (as it is impossible to estimate a generalized propensity score). Therefore, we only report TWFE estimates and regression adjustment estimates. Finally, in line with Cheng and Hoekstra (2013) (but unlike the results above), all the estimates in this section include region-by-year fixed effects.

[Figure 2 about here.]

Figure 2 provides the results. Relative to the two-period results, there are larger differences across specifications. The TWFE estimate is 0.067 (statistically significant). Among the regression adjustment specifications, estimates vary considerably: the change-in-covariates specification (Panel (b)) gives a somewhat larger estimate than TWFE that is marginally significant; the level-of-covariates specification (Panel (c)) estimate is close to zero; and the specification with both levels and changes (Panel (d)) produces an estimate similar to TWFE but less precisely estimated. Part of the difference between TWFE and regression adjustment arises because TWFE is affected by

[15] See Remark SA14 for a discussion on extending our results to include sampling weights.

pre-treatment violations of parallel trends (the second term in Equation (11)). Manually zeroing out this pre-treatment contribution increases the TWFE estimate by 31% to 0.088.

The remaining differences are explained by different implicit weighting schemes. In the figure, covariate balance is assessed in terms of how well the implicit weights across all post-treatment periods balance the average of each covariate. Relative to the raw data, the TWFE regression does improve covariate balance, though there are still some covariates that are severely unbalanced, particularly median income, log of incarceration rate, and poverty rate. Regression adjustment that includes the change in covariates over time (Panel (b)) does not perform much better. The last two specifications perform better in terms of covariate balance. We also calculated the effective untreated group sample size across post-treatment periods for each regression adjustment specification (see Supplementary Appendix SA18 for the specific calculation), finding values of 63.1, 26.7, and 9.9 for the specifications in Panels (b), (c), and (d), respectively. The sharp drop for Panel (d) is reflected in its larger standard errors. Taken together, the covariate balance diagnostics and effective sample sizes (both of which can be computed during the "design phase" of the analysis) provide an argument for preferring the specification in Panel (c) in this application.

Discussion

A main takeaway from our application is that the functional form under which covariates enter the model is a first-order concern when controlling for particular covariates in the parallel trends assumption. Approaches that inherit transformed covariates as a byproduct of the estimation strategy, whether it be TWFE, imputation/regression adjustment, or AIPW, performed poorly in terms of balancing the levels of time-varying covariates or time-invariant covariates. On the other hand, regression adjustment and AIPW approaches that include *any* level of a time-varying covariate and time-invariant covariates (such as the default implementation of Callaway and Sant'Anna (2021)) performed substantially better. In some cases, including time-invariant covariates and levels *and* changes in time-varying covariates performed better, although this was not uniformly true. We conjecture that a good heuristic for empirical work is to always include some version of the level of time-varying covariates (e.g., a pre-treatment value of each covariate) and time-invariant covari-

ates. In applications with enough data, one should also consider including changes in time-varying covariates as well.

8 Conclusion

We have considered difference-in-differences identification strategies when (i) the identification strategy hinges on comparing treated and untreated units with the same observed covariates and (ii) these covariates include time-varying and/or time-invariant variables. In this empirically common setting, researchers have most often implemented this identification strategy using a TWFE regression as in Equation (1). In this paper, we demonstrated a number of potential weaknesses of TWFE regressions in this context. Some of these weaknesses, such as lack of robustness to multiple periods and variation in treatment timing, or reliance on certain linearity conditions, are likely not surprising given existing work in the difference-in-differences literature. However, we also documented several other weaknesses that we termed *hidden linearity bias*, which arise because the transformations used to eliminate the unit fixed effect in the TWFE regression also change the functional form of the covariates. This transformation thus either effectively changes the identification strategy (to one where only the change in time-varying covariates is included in the parallel trends assumption) or relies heavily on a correctly specified linear model. Empirical work rarely examines whether these conditions are reasonable, and, in most applications, they are likely to be considered too strong. We proposed several diagnostic tools for assessing the sensitivity of TWFE regressions that include covariates to hidden linearity bias. We also proposed an alternative estimation strategy, building on recent work in the DiD literature, that does not suffer from hidden linearity bias, does not require any auxiliary assumptions along the lines mentioned above, and is effectively no more complicated to implement in practice than the TWFE regression.

References

Abadie, Alberto (2005). "Semiparametric difference-in-differences estimators". *The Review of Economic Studies* 72.1, pp. 1–19.

Angrist, Joshua (1998). "Estimating the labor market impact of voluntary military service using Social Security data on military applicants". *Econometrica* 66.2, pp. 249–288.

Angrist, Joshua D and Jorn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Aronow, Peter and Cyrus Samii (2016). "Does regression produce representative estimates of causal effects?" *American Journal of Political Science* 60.1, pp. 250–267.

Austin, Peter C and Elizabeth A Stuart (2015). "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies". *Statistics in Medicine* 34.28, pp. 3661–3679.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky (2025). "When is TSLS actually LATE?". Working Paper.

Bonhomme, Stephane and Ulrich Sauder (2011). "Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling". *Review of Economics and Statistics* 93.2, pp. 479–494.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2024). "Revisiting event-study designs: Robust and efficient estimation". *Review of Economic Studies*, rdae007.

Caetano, Carolina, Brantly Callaway, Robert Payne, and Hugo Sant'Anna (2022). "Difference in differences with time-varying covariates". Working Paper.

Callaway, Brantly and Pedro HC Sant'Anna (2021). "Difference-in-differences with multiple time periods". *Journal of Econometrics* 225.2, pp. 200–230.

Chattopadhyay, Ambarish and José R Zubizarreta (2023). "On the implied weights of linear regression for causal inference". *Biometrika* 110.3, pp. 615–629.

Cheng, Cheng and Mark Hoekstra (2013). "Does strengthening self-defense law deter crime or escalate violence? Evidence from expansions to Castle Doctrine". *Journal of Human Resources* 48.3, pp. 821–854.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018). "Double/debiased machine learning for treatment and

structural parameters". *The Econometrics Journal* 21.1, pp. C1–C68.

Cunningham, Scott (2021). *Causal Inference: The Mixtape*. Yale University Press.

de Chaisemartin, Clement and Xavier D'Haultfoeuille (2020). "Two-way fixed effects estimators with heterogeneous treatment effects". *American Economic Review* 110.9, pp. 2964–2996.

— (2023). "Two-way fixed effects and differences-in-differences estimators with several treatments". *Journal of Econometrics* 236.2, p. 105480.

Gardner, John, Neil Thakral, Linh T Tô, and Luther Yap (2023). "Two-stage differences in differences". Working Paper.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár (2024). "Contamination bias in linear regressions". *American Economic Review* 114.12, pp. 4015–4051.

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". *Journal of Econometrics* 225.2, pp. 254–277.

Graham, Bryan S, Cristine Campos de Xavier Pinto, and Daniel Egel (2012). "Inverse probability tilting for moment condition models with missing data". *The Review of Economic Studies* 79.3, pp. 1053–1079.

Hahn, Jinyong (2023). "Properties of least squares estimator in estimation of average treatment effects". *SERIEs* 14.3, pp. 301–313.

Hainmueller, Jens (2012). "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies". *Political Analysis*, pp. 25–46.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd (1998). "Characterizing selection bias using experimental data". *Econometrica* 66.5, pp. 1017–1098.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart (2007). "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference". *Political Analysis* 15.3, pp. 199–236.

Imai, Kosuke and Marc Ratkovic (2014). "Covariate balancing propensity score". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1, pp. 243–263.

Imbens, Guido W and Donald B Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Ishimaru, Shoya (2022). "What do we get from a two-way fixed effects estimator? Implications from a general numerical equivalence". Working Paper.

Kline, Patrick (2011). "Oaxaca-Blinder as a reweighting estimator". *American Economic Review* 101.3, pp. 532–37.

Lin, Lihua and Zhengyu Zhang (2022). "Interpreting the coefficients in dynamic two-way fixed effects regressions with time-varying covariates". *Economics Letters* 216, p. 110604.

Meyer, Bruce D. (1995). "Natural and quasi-experiments in economics". *Journal of Business & Economic Statistics* 13.2, pp. 151–161.

Robins, James, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky (2007). "Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable". *Statistical Science* 22.4, pp. 544–559.

Robins, James M, Miguel Angel Hernán, and Babette Brumback (2000). "Marginal structural models and causal inference in Epidemiology". *Epidemiology* 11.5, p. 551.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of regression coefficients when some regressors are not always observed". *Journal of the American Statistical Association* 89.427, pp. 846–866.

Rosenbaum, Paul and Donald Rubin (1983). "The central role of the propensity score in observational studies for causal effects". *Biometrika* 70.1, pp. 41–55.

Rubin, Donald B (2008). "For objective causal inference, design trumps analysis". *The Annals of Applied Statistics* 2.3, pp. 808–840.

Sant'Anna, Pedro H. C. and Jun Zhao (2020). "Doubly robust difference-in-differences estimators". *Journal of Econometrics* 219.1, pp. 101–122.

Scharfstein, Daniel O., Andrea Rotnitzky, and James M. Robins (1999). "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models". *Journal of the American Statistical Association* 94.448, pp. 1096–1120.

Sloczynski, Tymon (2022). "Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights". *The Review of Economics and Statistics* 104.3, pp. 501–509.

Sloczynski, Tymon and Jeffrey M Wooldridge (2018). "A general double robustness result for estimating average treatment effects". *Econometric Theory* 34.1, pp. 112–133.

Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". *Journal of Econometrics* 225.2, pp. 175–199.

Wooldridge, Jeffrey M (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT press.

— (2025). "Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators". *Empirical Economics* 69.5, pp. 2545–2587.

A Proofs of Results with Two Periods

Lemma 1. *Under Assumptions 1 and 2, for* $d \in \{0,1\}$,

$$\mathbb{E}\Big[\mathrm{L}(D|\Delta X_{t^*})\mathrm{L}_d(\Delta Y_{t^*}|\Delta X_{t^*})\Big|D=d\Big] = \mathbb{E}\Big[\mathrm{L}(D|\Delta X_{t^*})\Delta Y_{t^*}\Big|D=d\Big].$$

The proof of Lemma 1 is provided in the Supplementary Appendix.

Lemma 2. *Under Assumptions 1 and 2,*

$$\mathbb{E}\Big[\big(D-\mathrm{L}(D|\Delta X_{t^*})\big)^2\Big] = \mathbb{E}\Big[1-\mathrm{L}(D|\Delta X_{t^*})\Big|D=1\Big]\pi.$$

The proof of Lemma 2 is provided in the Supplementary Appendix.

Lemmas 1 and 2 are used in some of our arguments below involving linear projections.

Proposition A1. *Under Assumptions 1 and 2,* α *from the regression in Equation* (3) *can be decomposed as*

$$\alpha = \mathbb{E}\left[w(\Delta X_{t^*})\Big(\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*}) - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\Big|D=1\right],$$

where $w(\Delta X_t)$ *are the same weights as in Theorem 1.*

Proof. Starting with the numerator from Equation (5), we have that

$$\mathbb{E}\Big[\big(D-\mathrm{L}(D|\Delta X_{t^*})\big)\Delta Y_{t^*}\Big]$$

$$
\begin{aligned}
&= \mathbb{E}\Big[\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\Delta Y_{t^*}\Big|D=1\Big]\pi - \mathbb{E}\Big[\mathrm{L}(D|\Delta X_{t^*})\Delta Y_{t^*}\Big|D=0\Big](1-\pi) \\
&= \mathbb{E}\Big[\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*})\Big|D=1\Big]\pi \\
&\quad - \mathbb{E}\Big[\mathrm{L}(D|\Delta X_{t^*})\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big|D=0\Big](1-\pi) \\
&= \mathbb{E}\Big[D\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*})\Big] - \mathbb{E}\Big[(1-D)\mathrm{L}(D|\Delta X_{t^*})\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big] \qquad (13) \\
&= \mathbb{E}\left[D\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\Big(\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*}) - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\right] \\
&\quad + \mathbb{E}\left[\big(D-\mathrm{L}(D|\Delta X_{t^*})\big)\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\right] \\
&= \mathbb{E}\left[\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\Big(\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*}) - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\Big|D=1\right]\pi, \qquad (14)
\end{aligned}
$$

where the first equality holds by the law of iterated expectations, the second equality holds by Lemma 1, the third equality holds by applying the law of iterated expectations to both terms, the fourth equality holds by adding and subtracting $\mathbb{E}\Big[D\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big]$, and the last equality holds by applying the law of iterated expectations for the first term and because

$$
\mathbb{E}\left[\big(D-\mathrm{L}(D|\Delta X_{t^*})\big)\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\right] = \mathbb{E}\left[\big(D-\mathrm{L}(D|\Delta X_{t^*})\big)\Delta X_{t^*}'\right]\beta_0 = 0,
$$

where the first equality holds by the definition of $\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})$, and the second equality holds because ΔX_{t^*} is uncorrelated with the projection error $\big(D-\mathrm{L}(D|\Delta X_{t^*})\big)$.

Combining the expression in Equation (14) with the expression for the denominator from Lemma 2 completes the proof, given the definition of the weights $w(\Delta X_{t^*})$. □

Proposition A1 says that α is equal to a weighted average of the linear projection of the change in outcomes over time on the change in covariates over time for the treated group relative to the same linear projection for the untreated group. Both the weights and the linear projection terms in the proposition depend only on linear projections, making them straightforward to compute in practice. This proposition serves as an important intermediate step for our main results.

Proposition A2. *Under Assumptions 1 and 2, α in Equation (3) can be decomposed as*

$$
\begin{aligned}
\alpha &= \mathbb{E}\Big[w(\Delta X_{t^*})\Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=1] - \mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=0]\Big)\Big|D=1\Big] \qquad (15) \\
&+ \mathbb{E}\Big[w(\Delta X_{t^*})\Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=0] - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\Big|D=1\Big], \qquad (16)
\end{aligned}
$$

where $w(\Delta X_{t^*})$ are the same weights as in Theorem 1.

Proof. Starting from the numerator of the expression in Proposition A1, we have that

$$\begin{aligned}
&\mathbb{E}\left[\left(1-\mathrm{L}(D|\Delta X_{t^*})\right)\left(\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*})-\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\right)\Big|D=1\right]\pi \\
&\quad=\mathbb{E}\left[\left(1-\mathrm{L}(D|\Delta X_{t^*})\right)\left(\mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=1]-\mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=0]\right)\Big|D=1\right]\pi \\
&\quad-\mathbb{E}\left[\left(1-\mathrm{L}(D|\Delta X_{t^*})\right)\Big\{\left(\mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=1]-\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*})\right)\right. \qquad (17) \\
&\qquad\qquad\left.-\left(\mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=0]-\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\right)\Big\}\Big|D=1\right]\pi,
\end{aligned}$$

which holds by adding and subtracting

$$\mathbb{E}\left[\left(1-\mathrm{L}(D|\Delta X_{t^*})\right)\left(\mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=1]-\mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=0]\right)\Big|D=1\right]\pi.$$

Next, notice that

$$\mathbb{E}\left[\left(1-\mathrm{L}(D|\Delta X_{t^*})\right)\mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=1]\Big|D=1\right]=\mathbb{E}\left[\left(1-\mathrm{L}(D|\Delta X_{t^*})\right)\Delta Y_{t^*}\Big|D=1\right], \qquad (18)$$

which holds by the law of iterated expectations. Further, notice that

$$\mathbb{E}\left[\left(1-\mathrm{L}(D|\Delta X_{t^*})\right)\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*})\Big|D=1\right]=\mathbb{E}\left[\left(1-\mathrm{L}(D|\Delta X_{t^*})\right)\Delta Y_{t^*}\Big|D=1\right], \qquad (19)$$

which holds by Lemma 1. That the terms in Equations (18) and (19) are equal to each other implies that the first line of Equation (17) is equal to 0. Combining the second line of Equation (17) with the expression for the denominator in Equation (5) from Lemma 2 completes the proof. □

Proposition A2 decomposes α into two terms: (i) a weighted average of the average path of outcomes for the treated group relative to the path of outcomes for the untreated group (conditional on covariates), and (ii) a misspecification bias term that is a weighted average of the difference between the average path of outcomes for the untreated group conditional on time-varying and time-invariant covariates and the linear projection of ΔY_{t^*} on ΔX_{t^*} for the untreated group.

Proof of Theorem 1. First, it immediately follows from Assumptions 1 to 3 that $\mathrm{ATT}(X_{t^*},X_{t^*-1},Z)=\mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=1]-\mathbb{E}[\Delta Y_{t^*}|X_{t^*},X_{t^*-1},Z,D=0]$. This implies that Equation (15) in Proposition A2 is equal to $\mathbb{E}[w(\Delta X_{t^*})\mathrm{ATT}(X_{t^*},X_{t^*-1},Z)|D=1]$. The second term comes from adding and subtracting terms to the expression in Equation (16) in Proposition A2.

Note that the decomposition of this misspecification bias term is non-unique, as related terms could be added and subtracted in a different order (this is a similar property to many other decompositions). The properties of the weights hold immediately by their definitions. □

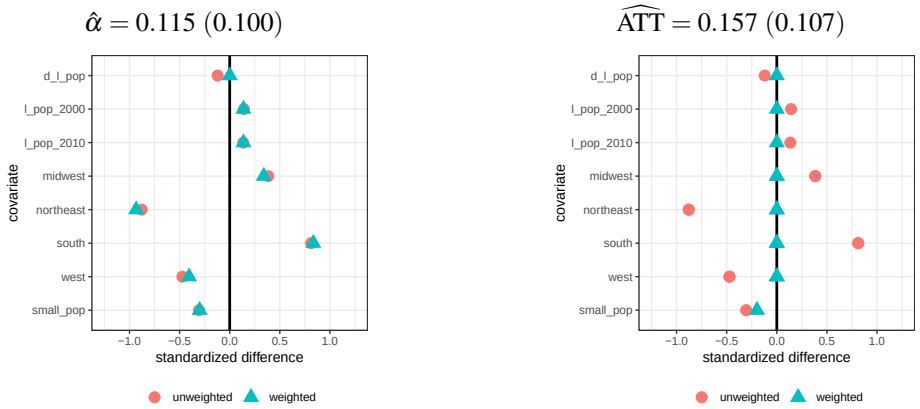
Proof of Theorem 2. The result holds immediately from Theorem 1 by noticing that Assumption 4 directly implies that Equations (A) to (C) are all equal to 0. □

Table 1: Summary Statistics

	Levels (2000)				Changes (2010-2000)			
	Tr.	Untr.	Diff.	Std. Δ	Tr.	Untr.	Diff.	Std. Δ
Outcome								
log homicides	5.19	4.70	0.49	0.34	0.08	-0.03	0.11	0.34
Time-Invariant Covariates								
Midwest	0.33	0.17	0.16	0.38				
Northeast	0.00	0.31	-0.31	-0.88				
South	0.52	0.17	0.35	0.81				
West	0.14	0.35	-0.20	-0.47				
Time-Varying Covariates								
log population	15.11	14.97	0.14	0.14	0.11	0.12	-0.01	-0.12
log police	5.75	5.71	0.03	0.17	-0.01	0.00	-0.01	-0.07
log prisoners	6.14	5.82	0.32	0.78	0.10	0.04	0.06	0.43
log welfare exp. per capita	6.84	6.97	-0.13	-0.44	0.40	0.36	0.04	0.20
log subsidies per capita	4.58	4.69	-0.11	-0.21	0.27	0.17	0.10	0.23
log median income	10.78	10.93	-0.15	-1.08	-0.08	-0.04	-0.04	-0.43
poverty rate	12.59	9.99	2.60	1.06	2.68	2.11	0.57	0.46
unemployment rate	4.05	3.69	0.36	0.42	4.93	4.86	0.08	0.04
% black males 15-24	1.99	2.83	-0.85	-0.20	0.15	0.18	-0.03	-0.07
% black males 25-44	3.51	5.31	-1.80	-0.22	-0.27	-0.66	0.39	0.34
% white males 15-24	11.31	10.43	0.88	0.05	-0.36	0.27	-0.63	-0.34
% white males 25-44	24.52	24.58	-0.06	-0.00	-3.59	-3.90	0.31	0.04

Notes: The table provides summary statistics for the outcomes, time-invariant covariates, and levels and changes in time-varying covariates. States are classified as being treated or untreated based on their treatment status in 2010. The column 'Diff.' reports the difference between the average of each variable for the treated group relative to the untreated group. The column 'Std. Δ.' reports the standardized difference of each variable for the treated group relative to the untreated group, which is the difference divided by the pooled standard deviation.

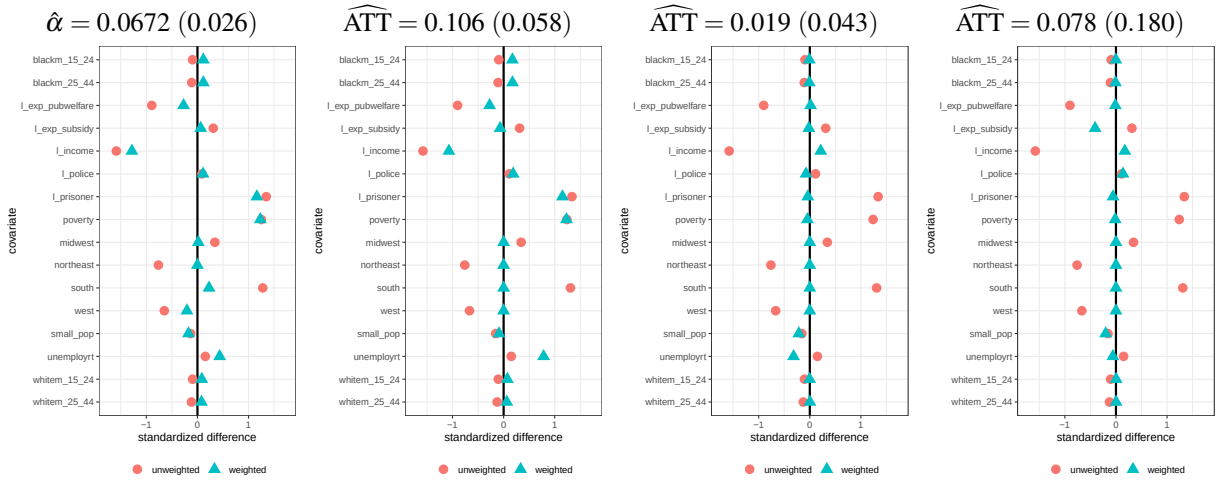
Figure 1: Two Period Covariate Balance using TWFE and AIPW



(a) TWFE (b) AIPW

Notes: The figure reports estimates of the effects of stand-your-ground laws on homicides and covariate balance statistics using the two-period data discussed in the main text. The balance statistics are invariant to the outcome. Different covariates are displayed along the y-axis. `d_l_pop` is the change in the log of state-level population from 2000 to 2010; `l_pop_2000` and `l_pop_2010` are the level of the log of state-level population in 2000 and 2010, respectively; `small_pop` is an indicator for $\log(\text{population}) < 15$; and `midwest`, `northeast`, `south`, `west` are indicators of Census region. The x-axis reports standardized differences for the mean of each covariate between the treated group and the untreated group. The red circles provide the standardized difference for the raw difference, and the blue triangles show the standardized difference after applying the implicit weighting scheme from each estimation method. Panel (a) comes from regressing ΔY_{t^*} on D_{t^*} and ΔX_{t^*}. Panel (b) uses the AIPW estimation strategy discussed in the paper that includes $\Delta X_{t^*}, X_{t^*-1}$, and Z in both the outcome regression model and the propensity score.

Figure 2: Multiple Period Covariate Balance with Additional Covariates



(a) TWFE (b) Reg. Adj.: $\Delta X, Z$ (c) Reg. Adj.: X_{g-1}, Z (d) Reg. Adj.: ΔX, X_{g-1}, Z

Notes: The figure reports estimates of the effects of stand-your-ground laws on homicides and covariate balance statistics using all available data from 2000-2010. See the main text as well as Table 1 for a detailed explanation of each covariate. We also use state-specific average population as sampling weights, as in the main results in Cheng and Hoekstra (2013). The results in Panel (a) come from a TWFE regression that includes all the covariates listed in the figure. Panels (b)-(d) report regression adjustment results with different specifications for the covariates, as described in the main text.