Group-specific linear trends and the triple-differences in time design

Anton Strezhnev*

September 2, 2024

Abstract

Differences-in-differences designs for estimating causal effects rely on an assumption of "parallel trends" – that in the absence of the intervention, treated units would have followed the same outcome trajectory as observed in control units. When parallel trends fails, researchers often turn to alternative strategies that relax this identifying assumption. One popular approach is the inclusion of a group-specific linear time trend in the commonly used two-way fixed effects (TWFE) estimator. In a setting with a single post-treatment and two pre-treatment periods it is well known that this is equivalent to a non-parametric "triple-differences" estimator which is valid under a "parallel trends-in-trends" assumption (Egami and Yamauchi 2023). This paper analyzes the TWFE estimator with group-specific linear time trends in the more general setting with many pre- and post-treatment periods. It shows that this estimator can be interpreted as an average over triple-differences terms involving both pre-treatment and post-treatment observations. As a consequence, this estimator does not identify a convex average of post-treatment ATTs without additional effect homogeneity assumptions even when there is no staggering in treatment adoption. A straightforward solution is to make the TWFE specification fully dynamic with a separate parameter for each relative treatment time. However, identification requires that researchers omit at least *two* pre-treatment relative treatment time indicators to estimate a group-specific linear trend. The paper shows how to properly extend this estimator to the staggered adoption setting using the approach of Sun and Abraham (2021), correcting a perfect collinearity error in recent implementations of this method in Hassell and Holbein (2024). It concludes with a note of caution for researchers, showing through a replication of Kogan (2021) how inferences from group-specific time trend specifications can be extremely sensitive to arbitrary specification choices when parallel trends violations are present but do not follow an easily observed functional form.

*Author contact information: Assistant Professor, University of Chicago, Department of Political Science. Email: astrezhnev@uchicago.edu

1 Introduction

The difference-in-differences (DiD) design for estimating causal effects is ubiquitous in the social sciences. Goldsmith-Pinkham (2024) finds that over the last two decades, it has been the identification strategy most responsible for the "credibility revolution" in economics. Political science is also no stranger to the difference-in-difference design and, in particular, the commonly used two-way-fixed-effects (TWFE) regression estimator (Chiu et al. 2023). The popularity of difference-in-differences stems in part from its ability to address certain forms of unobserved confounding in an observational setting. That is, when treatment is not randomly assigned, any observed difference in an outcome between treated and control units could be attributed to fundamental differences between the types of units assigned to these conditions – selection-into-treatment bias – rather than to the actual effect of treatment. Difference-in-differences designs attempt to account for this by subtracting off an estimate of the bias from a time period prior to the initiation of treatment where units in either condition are unaffected by the intervention. Under the assumption that this bias is invariant over time – in other words, that in the absence of treatment, the mean outcomes among treated and control units would follow *parallel trends* – the difference-in-differences comparison identifies the Average Treatment effect on the Treated (ATT).

In settings with many pre-treatment and post-treatment time periods, standard practice for estimating the ATT is to fit a two-way fixed effects regression estimated by ordinary least-squares (OLS). This approach regresses the outcome observed for each unit and time period on an indicator variable for whether the unit is exposed to treatment at that time along with two sets of additive "fixed effects" – separate intercepts for each unit (or equivalently treatment group) and for each time period. It is well known that when there exists a single, common treatment initiation time, this "static" TWFE regression will estimate an average of the post-treatment ATTs without requiring any additional assumptions beyond those standard for differences-in-differences. In addition to the static TWFE specification, researchers will also often estimate "dynamic" regressions which include separate indicators for each post-treatment period to assess the trajectory of treatment effects over time. These regressions also often include indicators for pre-treatment periods in order to conduct placebo tests for the main identifying assumptions. Although such "pre-trends" tests cannot prove that the parallel trends assumption holds, failed pre-trends call

into question the validity of the design.

One common strategy for relaxing the parallel trends assumption upon observing failed pre-trends is to assume that the violation has a particular parametric form which can be estimated. This is commonly done through the inclusion of a "group-specific" time trend term in the standard two-way fixed effects regression. That is, researchers will include an interaction between the unit (or group) fixed effects and a polynomial of time. Typically most specifications assume a linear time trend but higher-order polynomials are not uncommon.

Political science in particular has made substantial use of this extension of the conventional two-way fixed effects setup. Looking at applied work published in the last five years (since 2019) in the *American Political Science Review*, *American Journal of Political Science*, and the *Journal of Politics*, I find 39[1] published articles that mention estimating regressions which include some sort of unit-specific (e.g. county, city, region, etc...) linear time trend. These articles span the major subfields of the discipline and include work on the determinants of political participation (Hall and Yoder 2022; Grumbach and Hill 2022; Benesch et al. 2023; Bøggild and Jensen 2024), the effects of events and policies on electoral outcomes (Arias and Stasavage 2019; Kogan 2021; Ternullo 2022; Hamel 2024; Hassell, Holbein, and Baldwin 2020; Hassell and Holbein 2024; Ward 2020), the effects of electoral outcomes and representation on policies (Dynes and Holbein 2020; Harding 2020; Gulzar, Lal, and Pasquale 2024; Hankinson and Magazinnik 2023), sub-national political institutions (Potter 2022; Payson 2020; Su and Buerger 2024), money and politics (Fouirnaies 2021; Kilborn and Vishwanath 2022), media and politics (Foos and Bischof 2022; Sides, Vavreck, and Warshaw 2022; Esberg and Siegel 2023), education politics (Paglayan 2019, 2021; Marshall 2019), immigration politics (Masterson and Yasenov 2021; Hainmueller and Hangartner 2019; Ferwerda 2021), right-wing populism and intergroup conflict (Ansell et al. 2022; Dipoppa, Grossman, and Zonszein 2023; Abdelgadir and Fouka 2020), sub-national conflict and violence (Blair 2022; Magaloni, Franco-Vivanco, and Melo 2020), historical political economy (Zhang and Lee 2020; Fouka 2019; Rogowski et al. 2022; Pulejo 2023), and international organizations (Egel

1. The specific Google Scholar search query was for any one of the phrases "*-specific linear", "linear *-specific" or "linear time trend for each." To avoid over-counting, I do not include here papers that generally mention unit-specific "time trends" without specifying a parametric form as this phrase also often refers to specifications that allow the time fixed effects to vary arbitrarily by group. These regressions correspond to standard difference-in-differences designs where parallel trends is assumed to hold *within group* rather than to the triple-differences setup discussed here. As such, this number is likely an *undercount* of the total number of papers including group-specific trends.

and Obermeier 2023).

Despite the ubiquity of this approach, there is surprisingly very little written about the underlying identification strategy to which the group-specific time trends regression corresponds.[2] Descriptions of this method are frustratingly vague, particularly in guides for applied researchers. Angrist and Pischke (2009) leaves only a few paragraphs for this specification, describing it as a "check on the DD identification strategy" which "allows treatment and control...to follow different trends in a limited but potentially revealing way." When discussing how to interpret the results from the conventional and the group-specific time trends specification, the text only states that "it is heartening to find that the estimated effects of interest are unchanged by the inclusion of these trends, and discouraging otherwise." Wing, Simon, and Bello-Gomez (2018) describes the group-specific time trend regression as a robustness check for the standard difference-in-differences, noting that "In practice, most researchers interpret the group-specific linear trends model more casually by comparing the treatment effect estimates in the restricted and unrestricted models. If the treatment effect is not sensitive to the alternative specification, most researchers consider the core results more credible." Most recently in political science, two papers in the American Political Science Review: Hassell, Holbein, and Baldwin (2020) and Hassell and Holbein (2024) have argued that researchers should consider including group-specific linear trends to address concerns over failed pre-trends, particularly in the case of recent difference-in-differences studies of the effect of exposure to school shootings on county-level turnout and Democratic party vote share. However, this work still discusses the identification assumptions implied by group-specific trends regressions in very loose and general terms, stating only that "identification comes from sharp deviations from otherwise smooth county-specific trends" (Hassell and Holbein 2024).

What researchers *do* know about the two way fixed-effects regression with group-specific linear time trends comes primarily from the setting with a single post-treatment time period and two pre-treatment time periods. In this case, the coefficient on treatment is equivalent to a non-parametric triple-differences estimator which takes the difference between a "primary" difference-in-difference

2. One notable exception is Mora and Reggio (2019) which formally justifies the standard practice of using regression estimators with group-specific polynomial time trends to facilitate identification under extensions of the parallel trends assumption when there are many pre- and post-treatment periods. However, it does so in a setting that assumes the *entire* regression model is correct (including effect homogeneity in the static TWFE specification). It does not, as this paper does, examine the issues that can arise when fitting a static specification under effect heterogeneity.

between the post-treatment and last pre-treatment period and a "placebo" difference-in-differences between the last pre-treatment period and the next-to-last pre-treatment period (Wooldridge 2021). This approach, which this paper will refer to as the "triple-differences in time" design to distinguish from other uses of the term triple differences, identifies the average treatment effect on the treated under what Egami and Yamauchi (2023) calls a "parallel trends-in-trends" assumption. This assumption allows parallel trends to be violated in that the first-difference in control potential outcomes over time is *not* the same between treated and control units. However, it assumes that this discrepancy can be rectified by a further round of differencing - that even if the trends differ, the "trends in the trend" are the same. Mora and Reggio (2019) generalizes this assumption to an arbitrary number of differences, describing a family of "parallel-q" assumptions that allow for effects to be identified by comparing the q-differenced outcome between the treated and control group.

But much less is known about whether this connection between the group-specific time trend TWFE regression and the non-parametric triple-difference in time holds in the setting where researchers have many pre-treatment and many post-treatment periods. When there is no variation in when treated units initiate treatment – that is, no "staggered adoption" – the conventional static TWFE regression is equivalent to a difference-in-differences estimator that averages over pre- and post-treatment outcomes for treated and control units. However, when treatment adoption is staggered, recent work (Goodman-Bacon 2021; De Chaisemartin and d'Haultfoeuille 2020; Borusyak and Jaravel 2018) has shown that identification of the ATT using the static TWFE regression requires additional constant effects assumptions. Intuitively, as Goodman-Bacon (2021) shows, this is because the static TWFE estimator can be understood as an average over all 2x2 differences-in-differences comparisons in the sample. Some of these comparisons are differences in the difference between a treated group and a control group at one time period and the difference between those same groups at a future time period where both units are under treatment. Unless one of two additional effect homogeneity assumptions is satisfied (Li and Strezhnev 2024), these 2x2s are contaminated by a non-zero difference between two in-sample ATTs. As a consequence, the estimand being identified by the regression is a non-convex average of ATTs with some effects of these effects potentially receiving "negative weights" (De Chaisemartin and d'Haultfoeuille 2020). However, a similar decomposition does not yet exist for the static TWFE regression with

group-specific time trends.

This paper develops such a decomposition, focusing on the setting where there is *no* staggered adoption. Even when this significant constraint on the distribution of treatment is imposed – where conventional TWFE requires no additional assumptions beyond the standard difference-in-differences design – static TWFE regressions that include group-specific time trends do not identify the ATT without assuming effect homogeneity over time. The decomposition shows that the regression coefficient for this estimator can be written as an average over all triple-differences in time comparisons within the sample. While all of the primary "differences-in-differences" involve comparisons across one post-treatment and one pre-treatment period, only some of the corresponding "placebo" differences-in-differences are valid to adjust for the differential pre-trends – those involving two *pre-treatment* periods. However, the regression also incorporates"placebo" DiDs that difference across two *post-treatment* periods. This term identifies the differential pre-trends and the difference in treatment effects across those two time periods. Therefore, it is only valid to adjust for the differential pre-trends when the difference in effects is zero. That is, when treatment effects are constant. Notably, this form of "forbidden comparison" is distinct from those identified in previous work on staggered adoption and appear in the decomposition despite the fact that all treated units initiate treatment at the same time.

The solution in the non-staggered setting is simple – fitting a "dynamic" TWFE regression rather than a static one. This specification involves parameterizing the treatment through a set of "relative treatment time" indicators that take on a value of 1 if a unit is some q periods from initiating treatment. The coefficient on each indicator has a straightforward interpretation as an average over the triple differences involving that period and all time periods associated with omitted indicators. Ensuring that all post-treatment relative treatment time indicators are included in the regression avoids the aforementioned "forbidden comparisons" by ensuring that no pairs of treated periods are used to construct the placebo DiDs. Each of these coefficients identifies the ATT for that particular post-treatment time period without imposing constant effects. Standard placebo tests for the parallel trends-in-trends assumption can be carried out simply by including pre-treatment (negative) relative treatment time indicators, allowing researchers to present all of the dynamic treatment effect and "pre-trends" estimates in the standard "event study plot" (Freyaldenhoven et al. 2021).

It is straightforward to extend this solution to the setting with staggered adoption given that many of the proposed "heterogeneity-robust" estimators can be interpreted as averages over analyses of non-staggered subsets of the data.[3] While each of the new approaches can, in principle, incorporate the triple-differences in time design, this paper focuses on the "interaction-weighted" estimator proposed by Sun and Abraham (2021), hereinafter referred to as SA-TWFE. This estimator can be understood as an average over each non-staggered TWFE between each unique treatment "cohort" (set of units with a common treatment initiation time) and the control group (which never adopts treatment).[4] By correctly specifying each of the component non-staggered regressions (including all post-treatment relative treatment time indicators), the SA-TWFE approach allows researchers to estimate the trajectory of ATTs for each cohort and and aggregate them into summaries of the treatment's impact.

This approach for incorporating group-specific time trends into a the SA-TWFE heterogeneity-robust estimator has been recently proposed in the political science literature by Hassell and Holbein (2024). However, one crucial complication with including group-specific polynomial time-trends in the dynamic TWFE specification is that it requires researchers omit additional relative treatment time periods in order to avoid perfect collinearity with the two-way fixed effects and time trends. While the standard dynamic TWFE specification requires only one omitted (pre-treatment) relative treatment time indicator, group-specific linear trends require two, group-specific quadratic trends require three, and so on (Borusyak and Jaravel 2018). For the SA-TWFE estimator with group-specific linear trends, this means omitting enough relative treatment time indicators such that at least two are omitted for each cohort – typically the two periods just prior to treatment onset. Because the specification presented in Hassell and Holbein (2024) does not do this – the only omitted period that is common to all cohorts is the period prior to treatment onset – the resulting SA-TWFE estimates for many of the cohorts are driven entirely by how the implementing statistical software package chooses to resolve the perfect collinearity problem. Presently, the most common implementations of SA-TWFE in both `Stata` and `R`[5] are not guaranteed to always omit a pre-treatment relative treatment time indicator and, in fact, will typically

3. See Li and Strezhnev (2024) for a discussion of the similarities across these proposed estimators.

4. This is implemented in a single regression by interacting the relative treatment time indicators with indicators for each treatment cohort.

5. In `Stata`, this is the `eventstudyinteract` module (Sun 2022). In `R`, this is the the `sunab` method used by the `feols` function in the `fixest` package (Bergé 2018).

omit a *post-treatment* relative treatment time indicator. This paper corrects this implementation error and warns researchers to be deliberate in how they select the baselines when implementing the dynamic TWFE regressions to estimate their treatment effects under the triple-differences in time design.

Lastly, this paper cautions researchers against treating group-specific time trends regressions as a "robustness check" for the static TWFE regression. Differences between estimates may be attributable to effect heterogeneity rather than to a parallel trends violation while similarities may not guarantee that the DiD identifying assumptions hold. The triple-differences in time design only addresses a particular form of parallel trends violation, but not all such violations take the form of a clean, polynomial trend. It illustrates this through a replication of Kogan (2021) which examines the effect of the staggered roll-out of the U.S. food stamp program during the 1960s and 1970s on county-level vote share for the Democratic presidential candidate. Although the original paper uses a static TWFE estimator with county-specific linear time trends, the replication shows that omitting these county-specific linear trends results in only minimal changes to the estimated positive effect. However, this masks a noticeable pre-trends violation which is revealed by inspecting the event study plots for each of the three treatment cohorts. The failed pre-trend placebo tests appear to be related to a single peculiar time period: the 1964 presidential election. And because these violations lack a clean linear form, dynamic TWFE specifications are extremely sensitive to the choice of omitted pre-treatment baselines. The signs of the effect estimates change depending on which periods are selected for the placebo difference-in-differences. While the static TWFE estimates change minimally when group-specific time trends are included, the *dynamic* TWFE estimates are noticeably different – and neither appears to fully address the divergence in trends between treated and control counties. Overall, the replication suggests that there is no clear evidence for any effect of the food stamp program on Democratic party presidential vote share.

After reviewing the related literature on difference-in-differences designs and two-way fixed effects estimators, the remainder of the paper is organized as follows: Section 2 outlines the identifying assumptions behind the triple-differences in time design. It generalizes the identification strategy to the setting with many pre-treatment and many post-treatment time periods and distinguishes it from other types of triple differences designs that rely on observing unaffected groups

with a similar unit/time structure. It defines the non-parametric TDiT estimator which takes the form of a difference between a "primary" difference-in-difference involving a post-treatment time period and a pre-treatment time period and a "placebo" difference-in-differences that uses two pre-treatment periods. The latter adjusts for the differential pre-trends in the former. Section 3 presents the primary result of this paper – a decomposition of the static TWFE regression with unit-specific time trends into an average over TDiT terms. When there is more than one post-treatment period, the static TWFE will not identify the ATT when there is effect heterogeneity over time even when treatment is not staggered. It then shows how the "dynamic" TWFE regression addresses this problem and explains how to adapt the approach to allow for staggered treatment adoption using the Sun and Abraham (2021) estimator. Section 4 analyzes two papers that implement TWFE regressions with group-specific time trends: Hassell, Holbein, and Baldwin (2020) and Kogan (2021). For the former, I correct an error in the implementation of group-specific time trends with the Sun and Abraham (2021) estimator found in Hassell and Holbein (2024). For the latter, I show how the absence of a clear linear pre-trend can result in estimates that are highly sensitive to baseline selection. Section 5 concludes with a set of general recommendations for practitioners approaching group-specific time trends regressions.

1.1 Related literature

This work builds on the rapidly growing literature studying difference-in-differences in settings that depart from the "canonical" two-period design with a single treated group and single control group.[6] The literature aims to, in some sense, "catch up" our theoretical understanding of the DiD design and its associated estimation strategies with conventional applied practice, which rarely works with datasets that match the canonical set-up. A large branch of this work examines the standard TWFE estimator when units are permitted to initiate treatment at varying times (Borusyak and Jaravel 2018; De Chaisemartin and d'Haultfoeuille 2020; Goodman-Bacon 2021; Sun and Abraham 2021; Imai and Kim 2021). The central insight is that when treatment effects are heterogeneous and treatment adoption is staggered, the TWFE regression is not guaranteed to identify a convex average of ATTs. Goodman-Bacon (2021) decomposes the static TWFE regression coefficient in the staggered adoption setting to show that it can be interpreted as a

6. See Roth et al. (2023) for a comprehensive overview.

weighted average over different types of 2x2 DiD comparisons in the sample. Some of these comparisons are conventional DiDs since they involve taking the difference between post-treatment and pre-treatment outcomes for one unit and subtracting the difference for those same time periods for a unit that remained under control. However, others are "forbidden comparisons" in which the second difference involves outcomes observed for units that remain under *treatment*. This difference captures both the selection-into-treatment bias and the difference in treatment effects between those two periods and are therefore invalid DiDs unless effects are homogeneous.

This paper takes the same general approach as Goodman-Bacon (2021) and applies it to the OLS estimator of the static TWFE regression model with group-specific time trends using the well-known Frisch-Waugh-Lovell (Frisch and Waugh 1933; Lovell 1963) theorem to obtain an expression in terms of differences across outcome means. It diverges from the existing literature by focusing on the more simplified setting with a single treatment group and a common treatment timing. This is done to illustrate how the "negative weighting" problem can arise for certain extensions of the static TWFE regression even in the absence of staggered adoption and to emphasize that the inclusion of group-specific time trends is not guaranteed to make results more "robust." When parallel trends holds, static TWFE in this setting identifies the average of post-treatment ATTs without any additional assumptions, while static TWFE with group-specific time trends only does so if these ATTs are identical.

Previous work in this area has not paid much attention to estimators that augment the TWFE with group-specific polynomial trends, with a few notable exceptions. In an extension, Borusyak and Jaravel (2018) applies its main result on negative weighting in the static TWFE under staggered adoption to the case of unit-specific time trends. It shows that negative weighting of dynamic treatment effects can occur for this estimator even in settings where the conventional static TWFE regression will avoid the problem (no staggering). Kahn-Lang and Lang (2020) illustrates the problem with a numerical example where parallel trends holds but the treatment has a delayed rather than instantaneous effect after adoption. De-trending flips the sign of the TWFE estimate because part of the treatment effect is effectively absorbed in the group-specific linear trends. Wolfers (2006) makes a similar point in a re-analysis of prior work on the effect of no-fault divorce on divorce rates when critiquing a static specification that includes state-specific linear trends, noting that in the post-treatment period, state-specific linear time trends are observation-

ally equivalent to linearly increasing treatment effects. The results in this paper complement these prior findings by providing additional intuition for why the negative weights occur – even when there are no "forbidden comparisons" of the style described in Borusyak and Jaravel (2018), the inclusion of group-specific time trends can introduce new "forbidden comparisons" in the form of invalid triple-differences terms.

The conclusion of Section 3 and the replications show how to correctly incorporate group-specific time-trends in the staggered adoption setting such that both sources of the negative weighting problem are avoided. This paper is not primarily concerned with selecting the "best" heterogeneity-robust DiD estimator and therefore chooses only one for illustration: the Sun and Abraham (2021) "interaction-weighted" estimator, referred to here as SA-TWFE. This is largely due to the fact that SA-TWFE can be easily understood as a weighted average of non-staggered TWFE estimates and is therefore closest in principle to the conventional TWFE regression. SA-TWFE has also been recently proposed by Hassell and Holbein (2024) as a heterogeneity-robust estimator that can easily accommodate group-specific time trends. However, as this paper shows, researchers cannot simply use the same specification that they would for the conventional SA-TWFE. Identification under group-specific linear trends requires at least *two* relative treatment time indicators to be omitted to avoid perfect collinarity between the fixed effects and the treatment indicators (the standard SA-TWFE without trends requires just one). As a consequence, the group-specific time trends specifications presented in Hassell and Holbein (2024) suffer from perfect collinearity problems which statistical software implementations typically resolve in an arbitrary manner. The connection to the triple-differences in time design developed in this paper also makes it clear how heterogeneity-robust estimators aside from SA-TWFE could be adapted to handle group-specific time trends. Any methods that rely on cross-sectional regressions of first-differences in the outcome on a treatment variable such as Callaway and Sant'Anna (2021) or Dube et al. (2023) could be augmented by adding an additional (potentially re-scaled) differencing step relative to another pre-treatment period. For imputation methods (Borusyak, Jaravel, and Spiess 2021; Liu, Wang, and Xu 2022; Gardner 2022), it is already straightforward to include a unit-specific time trend in the TWFE model that fit to the controls and the validity of this approach is discussed in Borusyak, Jaravel, and Spiess (2021).

This paper also focuses on the setting where treatment is *absorbing* and there are no reversal.

While a number of estimation approaches have been suggested for this more general setting where treatment units can subsequently revert to control initiating treatment (e.g. De Chaisemartin and d'Haultfoeuille 2020; Imai and Kim 2021; Imai, Kim, and Wang 2021; Dube et al. 2023), they all partly rely on the imposition of some form of effect homogeneity or "no-spillover" assumption in order to construct any DiD terms which contain post-reversal observations. Notably, the extension of the "sequential DiD" estimator in Egami and Yamauchi (2023) falls into this category as it can be seen as a version of the De Chaisemartin and d'Haultfoeuille (2020) "switchers" estimator that uses two periods prior to treatment onset rather than one in order to construct the triple-difference.

2 Triple-differences in Time (TDiT)

This section formalizes the identifying assumptions underpinning the triple-differences in time (TDiT) identification strategy. As an extension of the conventional differences-in-differences design, this approach allows researchers to relax the core "parallel trends" assumption in a limited manner. Here, I use a similar notation as in Li and Strezhnev (2024) which builds on the characterization of the DiD design in Roth et al. (2023). Assume a balanced panel of N units observed over T time periods. Y_{it} denotes the observed outcome for unit i at time period t. Let $\mathcal{D} = \{D_1, D_2, D_3, \ldots, D_N\}$ denote the treatment assignment across units in the sample. In the non-staggered adoption setting, there are only two possible conditions to which units can be assigned. $D_i \in \{0, 1\}$ denotes whether a unit i receives treatment ($D_i = 1$) or remains under control for the duration of the observed time periods ($D_i = 0$). Treated units initiate treatment at a common time $g, 1 < g \leq T$. All units are untreated at time 1 and there is at least one time period where treated units are under treatment. Let $N_1 = \sum_{i=1}^{N} D_i$ denote the number of treated units and $N_0 = N - N_1$ denote the number of control units. All proofs for the propositions here can be found in Appendix A.

Causal effects are defined as contrasts in potential outcomes (Rubin 1974). Let $Y_{it}(1)$ denote the outcome that one would observe at time t if unit i received treatment and $Y_{it}(0)$ the potential outcome observed at time t under control. A conventional consistency or Stable Unit Treatment Value (SUTVA) assumption links the observed outcomes to the potential outcomes.

Assumption 1 *Consistency*

$$Y_{it}(d) = Y_{it} \text{ if } D_{it} = d$$

The core estimand of interest is the average treatment effect on the treated at time t which is defined as the difference in the potential outcomes under treatment and control at time t among those units that are in the group initiating treatment at time g

Definition 1 *Average Treatment Effect on the Treated at time t*

$$ATT(t) = E[Y_{it}(1)|D_i = 1] - E[Y_{it}(0)|D_i = 1]$$

As a single summary of the treatment effect, researchers may also target the *average* of the $ATT(t)$s in the post-treatment period which I simply label the ATT.

Definition 2 *Average Treatment Effect on the Treated*

$$ATT = \frac{1}{T - g + 1} \sum_{t=g}^{T} ATT(t)$$

The conventional differences-in-differences identification strategy makes two main assumptions about the potential outcomes in order to facilitate identification of the ATTs. The first – "parallel trends" – assumes that the differences in the potential outcomes under control between any two time periods t and t' are the same between the treated units and the control units. In other words, although the overall *levels* of the potential outcomes in the treated group may be systematically higher or lower than those in the control group, this difference is constant over time and can be eliminated by a simple first-differencing step.

Assumption 2 *Parallel trends*

For all $t \neq t'$

$$E[Y_{it}(0) - Y_{it'}(0)|D_i = 1] - E[Y_{it}(0) - Y_{it'}(0)|D_i = 0] = 0$$

The second assumption – "no anticipation" – rules out any effect of treatment in the pre-treatment periods ($t < g$). Essentially, treatments in the "future" cannot affect outcomes observed

in the "past."

Assumption 3 *No anticipation*

$$Y_{it}(1) = Y_{it}(0) \quad \textit{if } t < g$$

With these two assumptions, unbiased estimation of the Average Treatment effect on the Treated at time t is straightforward via a difference-in-differences estimator between t and any $t' < g$. Define the outcome mean for treatment group d at time t as $\bar{Y}_{d,t} = \frac{1}{N_d} \sum_{i:D_i=d} Y_{it}$.

Definition 3 *Non-parametric differences-in-differences estimator*

$$\hat{\tau}_{DiD}(t) = \bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'}$$

for any $t' < g$

However, researchers may be concerned that parallel trends does not hold. Pre-trends placebo tests – differences-in-differences estimates using two periods prior to adoption of treatment – may be non-zero, suggesting that treated and control outcomes are trending in divergent directions for reasons other than treatment. An approach to weakening parallel trends is to permit *some* divergence in the outcome trends that has a known structure which can be eliminated by further adjustments. "Triple-differences" designs allow for parallel trends to be violated by assuming that the *violation* of parallel trends is a constant and can therefore be removed by an additional round of differencing. That is, a "primary" difference-in-differences comparison that is biased for the target treatment effect can be corrected by subtracting a second "placebo" difference-in-differences which identifies only the bias – the violation of parallel trends. For triple-differences in time designs, researchers assume that the parallel trends violation between any two time periods t and t' is a constant, scaled by the time gap between those two periods.

Assumption 4 *Parallel trends-in-trends*

Let Δ *be an unknown constant.*

$$E[Y_{it}(0) - Y_{it'}(0)|D_i = 1] - E[Y_{it}(0) - Y_{it'}(0)|D_i = 0] = \Delta(t - t')$$

for all $t \neq t'$.

The above assumption generalizes the "parallel trends-in-trends" assumption in Egami and Yamauchi (2023) to the setting with arbitrarily many pre-treatment and post-treatment periods and is equivalent to stating that for any two adjacent time periods, t and $t-1$, the treatment and control group trends in the counterfactual $Y_{it}(0) - Y_{i(t-1)}(0)$ differ by a constant Δ.

Notably, this assumption is somewhat more restrictive than what is required for triple-differences designs that instead leverage unaffected observations that share the same unit and time structure as the primary sample to difference-out the parallel trends violation (Olden and Møen 2022; Strezhnev 2023). For this form of triple-differences, the violation of parallel trends can vary arbitrarily depending on which time periods t and t' are used to construct the primary difference-in-difference because the same time periods are also used to construct the placebo. However, in the triple-differences in time design, this is *not* the case. Identification hinges on the ability to *extrapolate* an observed violation estimated using pre-treatment observations to a post-treatment period.

Under the parallel trends-in-trends and no anticipation assumptions, it is possible to construct a simple non-parametric triple-differences in time estimator which takes the form of a difference between a "primary" difference-in-difference between time periods t and t' and a rescaled "placebo" difference-in-difference involving two time periods denoted t_2 and t_1. Time period t is the only post-treatment period ($t \geq g$) while t', t_1 and t_2 are all pre-treatment (less than g).

Definition 4 *Triple-Differences in Time (TDiT) estimator*

$$\hat{\tau}_{TDiT} = \left[\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'}\right] - \left[\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1}\right] \times \frac{(t-t')}{(t_2 - t_1)}$$

for any $t \geq g$, $t', t_1, t_2 < g$, $t_1 \neq t_2$

This extends the three-period difference-in-difference-in-differences estimator to be the setting with many pre-treatment and post-treatment periods. When $t' = t_2 = t-1$ and $t_1 = t-2$, it is equivalent to the "sequential DiD" described in Egami and Yamauchi (2023). However, in principle, any pair of pre-treatment periods could be used to construct the placebo DiD and any post-/pre- pair could be used for the primary DiD. The primary and placebo DiDs can vary in the

distance between their component time periods as the placebo is appropriately re-scaled to match the difference $t - t'$.

It's straightforward to see that under Assumptions 3 4, the TDiT estimator is unbiased for the ATT at time t.

Proposition 1 *Unbiasedness of the TDiT estimator*

Under Assumptions 1, 3, and 4.

$$E[\hat{\tau}_{TDiT}|\mathcal{D}] = ATT(t)$$

Intuitively, the "primary" difference-in-differences identifies a combination of the ATT at time t and a term capturing the divergence in parallel trends between t and t'. Under parallel trends-in-trends, this divergence is assumed to be equal to the product of some unobserved constant Δ and the time gap between t and t'. The "placebo" difference-in-difference does not identify any effect since treatment is assumed to have no impact in periods prior to adoption under the no anticipation assumption. It *only* identifies the bias term Δ multiplied by the difference $t_2 - t_1$. When $t - t' = t_2 - t_1$, the scaling factor is equal to 1 and simply subtracting the placebo DiD from the primary DiD eliminates the bias from the latter. Otherwise, the scaling factor adjusts for the fact that the divergence in parallel trends varies by time in a manner assumed to be known.

3 Decomposing static TWFE with group-specific time trends

In practice, researchers typically estimate treatment effects under a differences-in-differences identification strategy using a regression estimator with two sets of fixed effects parameters. The first, is either a fixed effect for each unit or, equivalently, a fixed effect for each distinct treatment timing group. The second is a fixed effect for each time period. The treatment is parameterized using an indicator that takes on a value of 1 for units exposed to treatment during the periods in which the exposure is active. This is sometimes referred to as the "static" specification (Sun and Abraham 2021). The standard two-way fixed effects specification in the non-staggered adoption

case can therefore be written as

$$Y_{it} = \tau\Big(D_i \times \mathbf{1}(t \geq g)\Big) + \alpha D_i + \delta_t + \varepsilon_{it},$$

where δ_t is the time fixed effect, α is the group fixed effect, τ is the treatment effect parameter of interest and ε_{it} is a mean-zero random error term. Note that in this writing of the TWFE specification, the global intercept and the group fixed effect for the control group ($D_i = 0$) are omitted to avoid perfect collinearity. It is well known that in the case of two time periods, one pre-treatment and one post-treatment, the OLS estimator of τ, $\hat{\tau}$ is equivalent to the non-parametric difference-in-differences estimator (Definition 3). Additionally, when there are multiple pre- and post-treatment periods but still only a single treated group with a common treatment initiation time, the static TWFE is equivalent to the difference-in-difference between the post-treatment averages and the pre-treatment averages. Under parallel trends and no staggered adoption, this identifies the ATT (Definition 2) – the average of post-treatment $ATT(t)$s.

To relax parallel trends to parallel trends-in-trends, researchers augment the conventional static TWFE with a "group-specific" linear time trend (Mora and Reggio 2019). This is typically done by simply interacting the unit fixed effect with the time variable, but again, it is equivalent to simply interact time with an indicator for each unique treatment timing group, which in the non-staggered setting is simply the treated group.

$$Y_{it} = \tau\Big(D_i \times \mathbf{1}(t \geq g)\Big) + \alpha D_i + \delta_t + \beta(D_i \times t) + \varepsilon_{it}, \tag{1}$$

Here β is the coefficient that captures the differential linear time trend between the treated and control groups.

It is well-known that when $T = g = 3$ the OLS estimator of $\hat{\tau}$ is equivalent to the non-parametric triple-differences in time estimator (Definition 4) in which $t = 3$, $t' = 2$, $t_2 = 2$, $t_1 = 1$.[7] How does this generalize to the setting where T can be arbitrarily large and g can be less than T? It is useful to first consider the types of triple-differences terms that could be constructed in this setting. Figure 1 presents a simple example with six time periods split evenly between three pre-treatment and three post-treatment periods. The figures present the treated

7. See Wooldridge (2021) for a recent discussion of this equivalence.

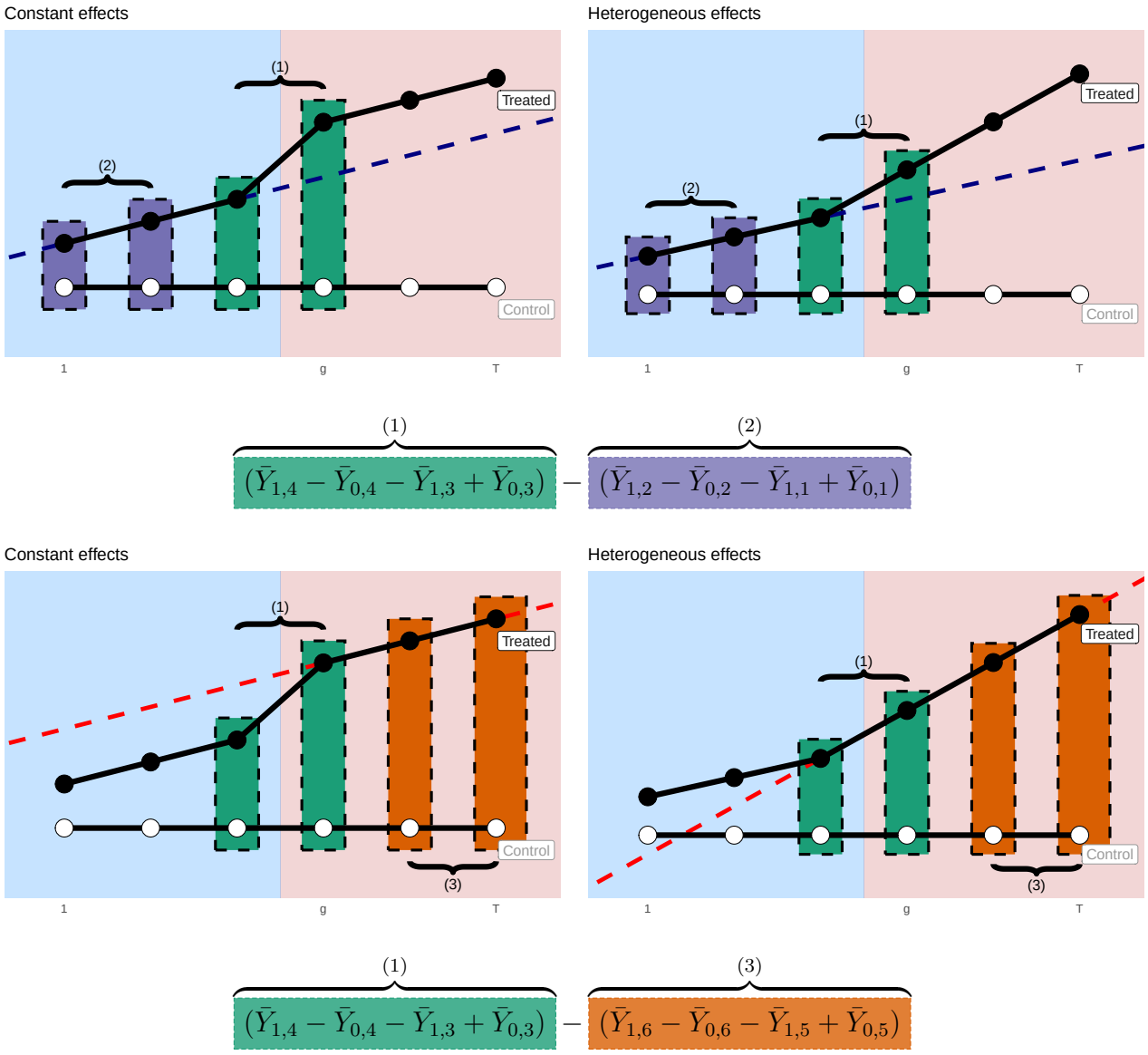

Figure 1: Visualization of primary and placebo difference-in-difference terms for the triple-difference in time – $T = 6, g = 4$

and control outcome means in each time period. All figures show a setting where the treated and control groups exhibit divergent linear trends under control. Figures in the left panel show a case where the effect of treatment is constant over time while those in the right panel show a case with heterogeneous effects.

Suppose the effect of interest is $ATT(4)$ or the immediate effect of treatment in the first period units can be under treatment. Start by taking the difference-in-difference between period 4 and one of the pre-treatment periods – in this example, period 3. This difference-in-difference identifies

the ATT in time 4 plus the divergence from parallel trends. To eliminate the latter, it is necessary to subtract off another difference-in-differences term that identifies this divergence. The upper panel of Figure 1 shows one possible choice for that placebo DiD – periods 2 and 1. Both are pre-treatment so the no anticipation assumption ensures that this term only identifies the deviation from parallel trends irrespective of whether there is effect heterogeneity or not. However, it is also possible to use two *post-treatment* periods to construct the placebo DiD, as shown in the lower panel of Figure 1. When treatment effects are constant, this also identifies the same bias term since the constant treatment effect is eliminated by differencing across time.

However, the right-hand panels of Figure 1 illustrate the potential pitfalls from using post-treatment differences-in-differences to adjust for the parallel trends violation when effects are heterogeneous. Here, the figure presents a treatment effect that also increases linearly over time rather than being constant for all post-treatment periods. While the pre-treatment placebo DiD will still correctly identify the trend in the trends, the post-treatment placebo DiD is contaminated by the presence of a non-zero difference in the ATTs of periods 6 and 5. As a result, while the triple-difference in the first panel will correctly identify the positive treatment effect irrespective of whether effects are constant or heterogeneous, the triple-difference in the second panel equals zero when treatment effects are *also* linear, despite the impact of treatment being positive. Among post-treatment comparisons, the treatment effect heterogeneity is indistinguishable from a divergence from parallel trends. Only by restricting the former can one use post-treatment observations to learn about the latter.

Proposition 2, shows that it is possible to interpret the OLS estimator of the coefficient on treatment in the static TWFE group-specific linear trends regression, $\hat{\tau}$, as an average over both of these types of triple-differences in time: those that use two pre-treatment periods for the placebo DiD and those that use two *post-treatment* periods.

Proposition 2 *The OLS estimator $\hat{\tau}$ for the static TWFE regression with group-specific time*

trends in Equation 1 can be written as an average over non-parametric triple-differences

$$\hat{\tau} = \frac{1}{(g-1)(T-g+1)} \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \Bigg\{ \overbrace{(\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'})}^{(1)} - \Bigg[\frac{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} \overbrace{(\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})}^{(2)} + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} \overbrace{(\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})}^{(3)}}{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \Bigg] (t - t') \Bigg\}$$

Intuitively, the regression averages over every primary DiD involving one post-treatment period and one pre-treatment period ($t : g \leq t \leq T$) and one pre-treatment period ($t' : 1 \leq t' < g$). Under no anticipation and the parallel trends-in-trends assumption, this identifies the ATT at time t plus the linear divergence in trends between those two time periods $\Delta(t - t')$. To eliminate the divergence, the regression constructs an average over all possible placebo DiDs. This includes summing over all pairs of time periods $t_2 > t_1$ where both t_2 and t_1 are less than g, the first treated period, as well as all $t_2 > t_1$ where both periods are greater than or equal to g. This sum is then rescaled to match the gap in time between t and t'.

Assumption 5 *Constant treatment effects over time*

For all t, t', $g \geq t \geq T$, $g \geq t' \geq T$

$$ATT(t) = ATT(t')$$

When will this regression identify the ATT without any additional assumptions beyond what is necessary for the non-parametric TDiT? When there is only a single treated period, it is straightforward to see that there are no placebo DiD terms in group (3) and that all placebo DiD terms in group (2) involve only pre-treatment periods and therefore identify the linear divergence in parallel trends. However, when there are more than two post-treatment periods, unbiased estimation of the ATT using the regression requires an additional assumption on the treatment effects – that they are constant over time (Assumption 5).

Proposition 3 *Under Assumptions 1, 3, and 2, $\hat{\tau}$ identifies a weighted average of $ATT(t)$*

$$E[\hat{\tau}|\mathcal{D}] = \sum_{t=g}^{T} ATT(t) \left[\frac{(T-1)(T+1) + 3(T-g+1)(T-2t+1)}{(T-g+1)(T-1)(T+1) - 3(g-1)(T-g+1)^2} \right]$$

where the weights on each $ATT(t)$ sum to 1 but are not guaranteed to be non-negative.

Adding Asumption 5

$$E[\hat{\tau}|\mathcal{D}] = \frac{1}{(T-g+1)} \sum_{t=g}^{T} ATT(t) = ATT$$

While it is clear that without constant effects, the ATT as a *uniform* average over the post-treatment $ATT(t)$s is not identified by the regression, it is worth asking what quantity *is* identified when effects are heterogeneous. Does the regression still identify a useful average of the post-treatment $ATT(t)$s? Proposition 3 shows that $\hat{\tau}$ can be interpreted as a weighted average of the $ATT(t)$s, but the weights on any given ATT are not guaranteed to be non-negative. Consider the case of $ATT(T)$, the ATT for the last period. The weight on this effect is

$$\begin{aligned}
\frac{(T-1)(T+1) + 3(T-g+1)(T-2T+1)}{(T-g+1)(T-1)(T+1) - 3(g-1)(T-g+1)^2} &= \frac{(T-1)(T+1) - 3(T-g+1)(T-1)}{(T-g+1)(T-1)(T+1) - 3(g-1)(T-g+1)^2} \\
&= \frac{(T-1)[(T+1) - 3(T-g+1)]}{(T-g+1)(T-1)(T+1) - 3(g-1)(T-g+1)^2} \\
&= \frac{(T-1)(3g - 2(T+1))}{(T-g+1)(T-1)(T+1) - 3(g-1)(T-g+1)^2}
\end{aligned}$$

which is negative if $g < \frac{2}{3}(T+1)$. Notably, for even $T, T > 2$, this weight will always be negative under an equal pre-treatment/post-treatment split. In this case, $g = \frac{T}{2} + 1$ and $\frac{T}{2} + 1 < \frac{2}{3}T + \frac{2}{3}$ for any integer $T > 2$.

From the decomposition in Proposition 2, it is therefore possible to recover the "negative weights" result for the group-specific linear time trends regression first shown in Borusyak and Jaravel (2018). What the decomposition helps clarify is the nature of the "forbidden comparisons" that generate this negative weighting problem. Here, they are the result of invalid placebo difference-in-differences terms that are constructed using post-treatment observations.

Even though all of the primary DiD terms are valid under effect heterogeneity, the *placebo* DiD terms are not. As a consequence, the negative weighting problem occurs even when treatment is not staggered, an important difference from the negative weighting results for the conventional static TWFE.

The solution to this problem, as noted in Borusyak and Jaravel (2018) is to estimate a dynamic TWFE specification in which the treatment is parameterized via a set of indicators for whether a particular unit is some number of periods relative to the treatment. Let $\mathcal{Q} = \{-g+1, -g+2, \ldots, T-g\}$ denote the set of all possible "relative treatment times." Negative values indicate pre-treatment periods for a particular unit while non-negative values denote post-treatment periods. Here $q = 0$ denotes the *first* period in which a unit is treated. Let $D_{it}^{(q)}$ be an indicator that takes on a value of 1 if unit i is q periods from initiating treatment.[8] The dynamic TWFE regression with group-specific time trends is specified as:

$$Y_{it} = \sum_{q \in \mathcal{Q}^*} \tau_q D_{it}^{(q)} + \alpha D_i + \delta_t + \beta(D_i \times t) + \varepsilon_{it}, \tag{2}$$

where $\mathcal{Q}^*$ denotes the set of relative treatment times included in the regression. Note that it is impossible to fit a regression where $\mathcal{Q}^* = \mathcal{Q}$ as the relative treatment time indicators will be perfectly collinear with the unit fixed effect for the treated group. Typical convention for the standard TWFE is to omit $q = -1$, the last pre-treatment period. However, when adding the linear group-specific time trends, one omitted period does not suffice to address perfect collinearity. This is because the remaining relative time indicators and the unit fixed effect will still be perfectly collinear with $D_i \times t$. Therefore an additional relative treatment time indicator needs to be removed for the linear trends specification, two more need to be removed for the quadratic specification, and so on for higher-order polynomials. Any additional omitted relative treatment time indicators should also still correspond to *pre-treatment* periods only. Intuitively, these are the time periods that are used as the baselines for constructing the primary and placebo difference-in-differences. As shown in Sun and Abraham (2021), omitting an indicator for a particular relative treatment time is equivalent to imposing an assumption that the treatment effect for that time is zero. "No anticipation" (Assumption 3) guarantees this for pre-treatment periods but obviously not for

8. This will always be 0 for the control units for any q.

post-treatment periods.

It is easy to see how the specification that is fully-saturated in post-treatment relative treatment time indicators eliminates the problematic placebo DiDs. The OLS estimator $\hat{\tau}_q$ is equivalent to the regression coefficient from a static TWFE specification that only includes one post-treatment period $g + q$ and the pre-treatment periods associated with the omitted indicators. Because there is only a single "treatment" period in this regression, the estimator only uses placebo DiDs from group (2) (Proposition 2). As a result, $\hat{\tau}_q$ identifies the $ATT(g + q)$ without any additional assumptions on effect homogeneity.

It is also straightforward to generalize this approach to estimate dynamic treatment effects in settings where treated units initiate treatment at different times using the SA-TWFE estimator (Sun and Abraham 2021). In short, this approach estimates a separate set of dynamic treatment effects for each distinct treatment timing group or "cohort" using a non-staggered TWFE regression involving only that cohort and the set of units that never adopt treatment. These estimates are then aggregated for each relative treatment time period by weighting them in proportion to the size of each cohort among cohorts that have that particular relative treatment time. Sun and Abraham (2021) refer to this approach as an "interaction-weighted" estimator since each of the component non-staggered TWFE regressions can be estimated from a single fully-saturated regression that includes interactions between the relative treatment time indicators and indicators for each cohort. For the standard SA-TWFE estimator, researchers need to omit enough relative treatment time periods such that at least one is omitted for each unique cohort to avoid perfect collinearity in any one of the component non-staggered TWFE regressions. Conventionally, this indicator is $q = -1$ since it appears for all treated cohorts in the sample except for the "always treated" units (which are dropped).

This last part slightly complicates the extension of the SA-TWFE estimator to the setting with group-specific time trends. As discussed above, in order to include a linear trend for each cohort, researchers need to omit additional relative treatment time periods to avoid perfect collinearity. Each cohort must have at least two omitted pre-treatment relative treatment time indicators in order to estimate the SA-TWFE regression with cohort-specific linear trends. A natural period to omit would be $q = -2$, which would only require dropping one additional cohort from the analysis: those units that have only a single pre-treatment period (as it is impossible to estimate a time

trend for these units without imposing effect homogeneity assumptions). When this is not done intentionally and manually, most modern software implementations of OLS will selectively omit one regressor among the collinear regressors in order to ensure that estimates can be computed. This is done with varying degrees of warning to the researcher. For the purposes of the SA-TWFE estimator, current popular implementations in both `R` and `Stata` will drop the *last* relative treatment time indicator for each cohort – problematically, a *post-treatment* period. This is the case with the SA-TWFE estimates presented in Hassell and Holbein (2024), which are examined further in the reanalysis in the next section.

4 Replications

4.1 Hassell, Holbein, and Baldwin (2020)

Hassell, Holbein, and Baldwin (2020) (here referred to as HHB) examine the effect of gun violence in the United States on political outcomes. Specifically, the article's difference-in-differences analyses look at the impact of school shootings on turnout and Democratic party vote share in federal, state and local elections. This replication focuses on the effect of shootings on Democratic party congressional vote share from 2000 to 2018. This particular analysis was one of three re-analyses conducted in Hassell and Holbein (2024), which aimed to resolve conflicts between the null results presented in HHB and positive findings of the effect of shootings on Democratic party vote share in García-Montoya, Arjona, and Lacombe (2022) and Yousaf (2021). Hassell and Holbein (2024) argue that any purported positive effects are a consequence of violations of the parallel trends assumption. Including group-specific linear trends - as is done in HHB but not in the other two studies - eliminates any positive treatment effect. Since the patterns of pre-trends are comparable in all three of these studies, this replication focuses only on the HHB data.

Figure 3 plots the distribution of treatment among observations in the dataset (Mou, Liu, and Xu 2023). Congressional elections are observed every two years and counties that experience a school shooting are considered "treated" throughout all elections after the shooting. Most counties are "never-treated" in that they are unexposed to a school shooting (by the HHB definition) during the 2000 to 2018 period. However, 4 counties are exposed in 2000 and so are considered "always treated" during the period under analysis. In the original replication in HHB, these units are

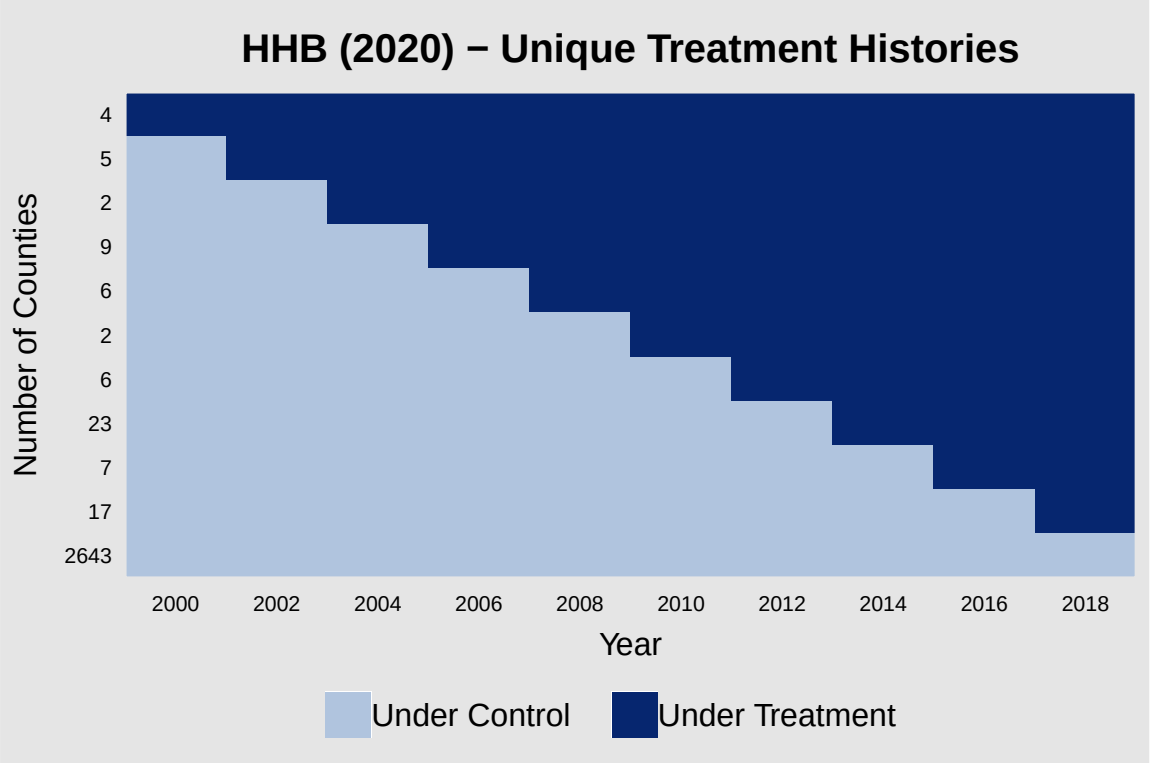

Figure 2: Distribution of treatment in Hassell, Holbein, and Baldwin (2020)

included in all analyses. However, they should be dropped from a "heterogeneity-robust" analysis since they lack any observed pre-treatment period (in the included data).

Dropping the always-treated units and estimating the static TWFE regressions both with and without the linear time trends yields results consistent with the argument of Hassell and Holbein (2024). Table 1 shows that what appears as a strong positive treatment effect on Democratic party vote share becomes statistically indistinguishable from zero when county-specific linear time trends are included.

Figure 3 plots the estimated dynamic treatment effects using the SA-TWFE estimator with a single, common baseline period of -2, the election prior to the start of treatment. The event study plot shows clear evidence of a roughly linear pre-trend which is comparable in magnitude and functional form to the post-treatment estimates. The plot strongly suggests that any observed treatment effects can be plausibly explained by divergence from the parallel trends assumption.

Before turning to the SA-TWFE estimator with group-specific time trends, it is worth examining the implementation of this estimator in Hassell and Holbein (2024) both without and with group-specific linear trends. Both of these regressions are mis-specified in the original pa-

Table 1: Static TWFE estimates of the effect of school shootings on county-level Democratic party congressional vote share (Hassell, Holbein, and Baldwin 2020)

Outcome:	Democratic party Congressional vote share	
Model:	(1)	(2)
Variables		
Any prior school shooting	0.0754***	0.0031
	(0.0126)	(0.0123)
County FE	Yes	Yes
Year FE	Yes	Yes
County-specific linear trends		Yes
Observations	27,200	27,200
Counties	2,720	2,720

Clustered (County) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

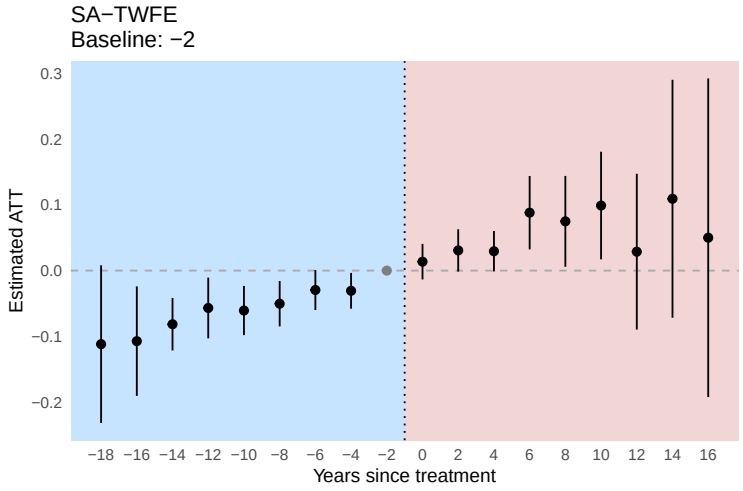

Figure 3: SA-TWFE estimates of the effect of school shootings on Democratic party vote share.

per. First, they include "always-treated" units. Second, some of the omitted baseline periods are *post-treatment*. Specifically, the regressions omit periods $-2, -16, -18, 16$ and 18. As discussed in Sun and Abraham (2021), omitting a relative treatment time period from the regression imposes an assumption that the treatment effect for this period is zero. While this is guaranteed by "no anticipation" for pre-treatment (negative) periods, it is *not* guaranteed for a post-treatment period without additional substantive knowledge. Additionally, with multiple omitted pre-treatment periods, it becomes more difficult to interpret the pre-trends estimates, which average over the DiDs

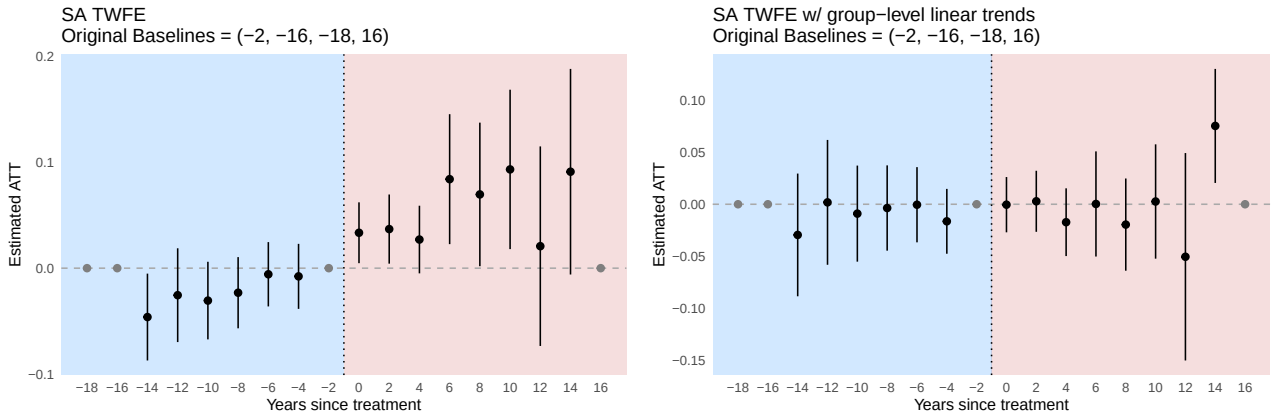

Figure 4: SA-TWFE estimates without and with group-specific time trends. Original baselines in Hassell and Holbein (2024)

between each of the omitted baselines. The first panel of Figure 4 plots the dynamic SA-TWFE estimates under these original baselines.[9] The estimated pre-trends are noticeably attenuated, as are the estimated treatment effects compared to the SA-TWFE estimates with the single baseline of -2. Hassell and Holbein (2024) compares these smaller SA-TWFE estimates to a conventional dynamic TWFE event study regression, and remarks that this is evidence that "effect heterogeneity plays some role" in the apparent inflated effect of school shootings. However, this interpretation is incorrect and appears to be mainly due to the fact that the SA-TWFE specification in the paper uses a different set of baseline comparison periods than the dynamic TWFE. SA-TWFE with a single common baseline period (Figure 3) yields comparable results to the dynamic TWFE figure in the original paper. This is also generally more consistent with the core argument that school shootings have no effect since effect heterogeneity by definition cannot be a relevant factor when all treatment effects are actually zero.

Irrespective of whether one looks at the standard dynamic TWFE regression or the "heterogeneity-robust" SA-TWFE results, it is clear that parallel trends is likely violated. Additionally, the violation seems to take the form of a clear linear trend over the entire period being studied. As such, it may seem sensible to include a "group-specific linear trend" as part of our dynamic treatment effect estimator. As discussed in the previous section, estimating a dynamic TWFE regression with all post-treatment indicators included avoids the "forbidden comparisons" problem that occurs in the static TWFE with group-specific trends. However, it is necessary to omit two relative

9. The results presented here omit "always-treated" units which are the only units with relative treatment time period 18.

time indicators rather than just one to avoid perfect collinearity. To do this correctly with the SA-TWFE estimator, at least two periods need to be dropped for *each* cohort in the sample. However, Hassell and Holbein (2024) use the same set of omitted indicators for all SA-TWFE specifications: including those with linear and quadratic group-specific time trends. This is insufficient for the latter two regressions. While all cohorts (except for the always-treateds) have a period −2, most cohorts lack periods −16 and −18. For some cohorts, identification depends on the linear trend estimated between a post-treatment (16) and pre-treatment period (−2), which is invalid unless, again, it is assumed that there are no treatment effects for post-treatment period 16. In all other cases, the software implementation addresses the perfect collinearity problem by arbitrarily dropping a relative time indicator for each cohort.

To fix perfect collinearity, the `fixest::feols` implementation in R implicitly drops the *last* relative treatment time, which will always be post-treatment. The `eventstudyinteract` implementation in Stata also appears to do the same but, interestingly, still includes that omitted period (and the assumed zero effect) when taking the average over the cohort-specific effects – inherently attenuating some of the estimates towards zero. Either way, it is clear that the software defaults are not guaranteed to "fail safely." By dropping a post-treatment period, the resulting placebo "difference-in-difference" in the underlying triple-difference involve comparisons between post-treatment and pre-treatment (-2) periods and are therefore contaminated by the treatment effects from that omitted post-treatment period. The second panel of Figure 4 plots the estimated effects from this specification, which uses the default settings of `fixest::feols` in R to resolve perfect collinearity.

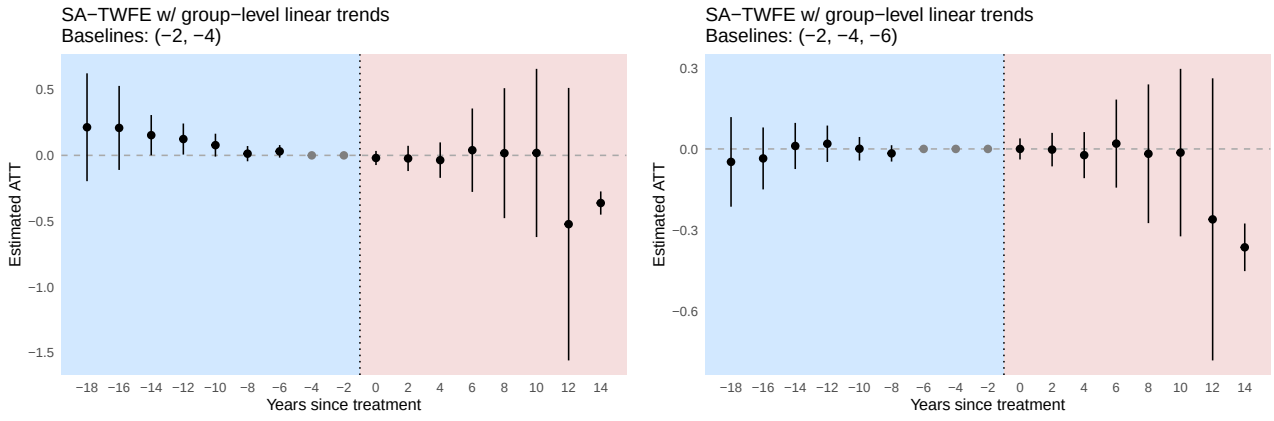

Figure 5: SA-TWFE estimates with group-specific linear time trends. Multiple baseline choices.

To correctly implement the SA-TWFE interaction-weighted estimator with linear trends and not have to impose implausible "no effects" assumptions, the omitted baselines need to be pre-treatment and at least two need to be common to every cohort. Figure 5 plots the results from two sets of baseline choices. The first (left panel) omits periods -2 and -4 to construct the linear trend. Note that for this estimator it is necessary to omit not only the always treated units, but also the "nearly-always" treated (those under treatment for all but one time period). While it's possible to construct a 2x2 DiD for these units, it is not possible to construct a valid triple-difference.

These estimates appear to have somewhat over-corrected for the linear trend! This is the challenge with polynomial time trends - if researchers are unwilling to make constant effects assumptions, estimating the time trend requires extrapolation quite far out of the support of the actual data. For example, the effect at 14 years post-treatment (8 elections) can only be estimated for the 2004 cohort. But for that cohort, there are only have two pre-treatment periods to use to estimate the trend. Therefore the linear trend from two periods one election apart is essentially scaled up to adjust for the parallel trends violation 8 elections out. Adding an additional held-out baseline, -6, changes the estimates somewhat (right panel). There are no longer statistically significant estimates for the pre-treatment placebo periods. However, note that this *does not* change the estimate for the "14 years since treatment" effect since that cohort does not have a -6 relative treatment time period to include into the triple-differences average. One should still probably not take that estimated negative effect particularly seriously.

Ultimately, Hassell, Holbein, and Baldwin (2020) and Hassell and Holbein (2024) are correct on the substance. Estimates suggesting school shootings increase Democratic party vote share are entirely driven by pre-trends. And conveniently, this pre-trend appears to be roughly linear, making the adjustment reasonable. However, it would only be possible to know that the linear trend is sensible through inspection of the dynamic TWFE event study regressions and the pre-trends. The static TWFE alone does not tell researchers whether the pre-trends violation is likely to have this convenient functional form. Even then, the pre-trends are always primarily diagnostic - researchers cannot guarantee that what looks like a constant linear violation of parallel trends continues into the post-treatment period. In some sense, this is a particularly "easy" data setting for polynomial time trends – the functional form appears clear and not only is the treatment effect constant over time, it is very likely *zero*. Therefore, despite the corrections to the mis-specified

estimators, very little changes. This may not be the case for all datasets and research designs.

4.2 Kogan (2021)

To illustrate the pitfalls of mis-specified group-specific time trends regressions when a linear pre-trend is not clearly evident, I turn to a replication of Kogan (2021) which adapts a treatment originally introduced in Hoynes and Schanzenbach (2009) to estimate the effects of the staggered roll-out of the U.S. food stamp program on Democratic party presidential vote share. Kogan (2021) estimates a specification with state-specific time fixed effects and group-specific linear time trends. It finds a strong positive effect of adoption of the program on Democratic party vote share. Unfortunately, this appears to be driven by a very peculiar parallel trends violation related to the onset of the program and the election of 1964 that does not fit the linear form, but nevertheless shows up in failed pre-trends diagnostics. Additionally, specifying the dynamic SA-TWFE regression with group-specific time trends yields inferences that are extremely sensitive to the choice of omitted pre-treatment baselines.

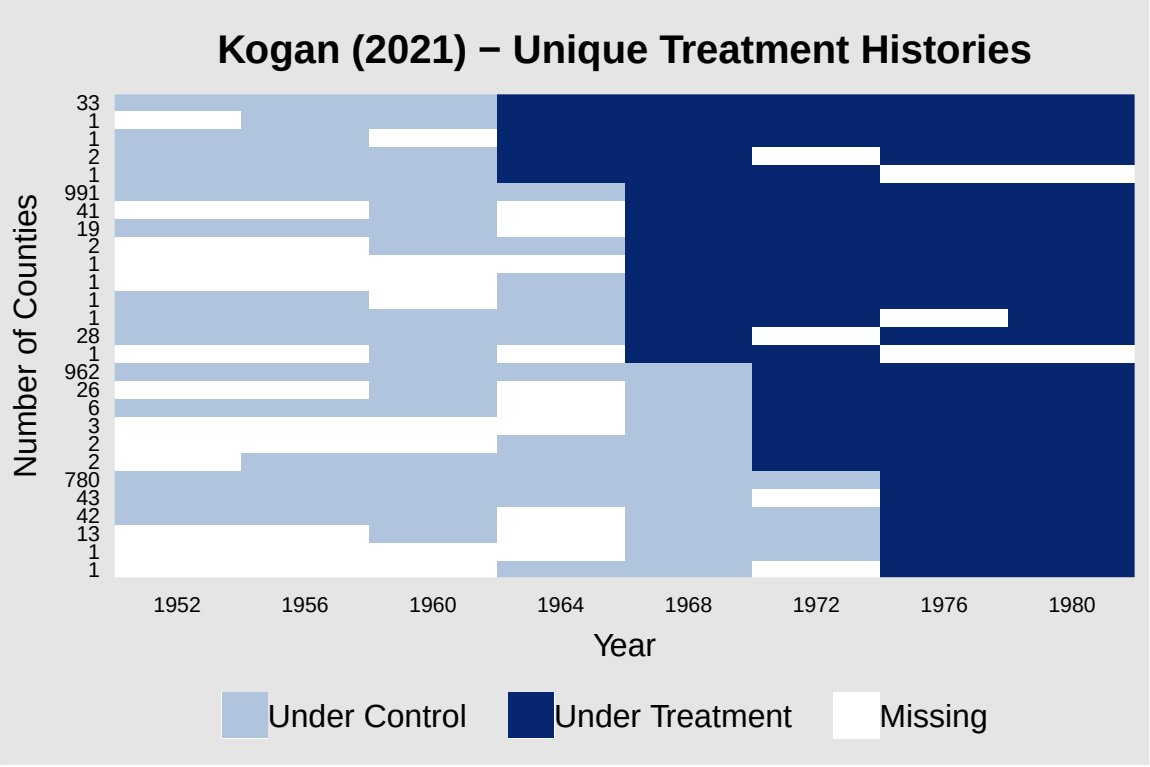

Figure 6: Distribution of treatment in Kogan (2021)

Plotting the distribution of treatment (Figure 6) shows three distinct treatment cohorts defined by the first election after exposure: 1964, 1968 and 1972. A handful of counties have outcome missingness, but for the most part this is negligible. Most of the treated counties fall into the latter two cohorts since the Food Stamp Act itself was only introduced in 1964. A handful of pilot programs had been implemented in the years prior. Notably, the original analysis included time periods (1976 and 1980) where all units were under treatment as the program had been fully rolled out by the 1976 election. It is not possible to estimate effects for these periods without strong additional effect homogeneity assumptions as there is no way to construct a "clean" difference-in-differences comparison for these periods. Therefore, the replication focuses on the period from 1952 to 1972 with the counties that did not implement the food stamp program until after the 1972 election acting as the "never-treated" units.

Table 2: Static TWFE estimates of the effect of the food stamp program on county-level Democratic party presidential vote share (Kogan 2021)

Dependent Variable:	Democratic Presidential vote share (0-100)	
Model:	(1)	(2)
Variables		
Food stamp program adopted	0.7832***	0.7246**
	(0.3002)	(0.2993)
County FE	Yes	Yes
State-Year FE	Yes	Yes
County-year linear trend		Yes
Fit statistics		
Observations	17,605	17,605
Counties	3,005	3,005

Clustered (County) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Although Kogan (2021) primarily presents results from regressions with county-specific time trends, it is instructive to compare these estimates to those from the conventional static TWFE regression. Table 2 presents the results from these two regressions with no additional covariates. I retain the original study's choice to use state-specific year fixed effects – which effectively limits all DiD comparisons to counties within the same state. Overall, the results from both specifications

appear to be extremely similar to one another.

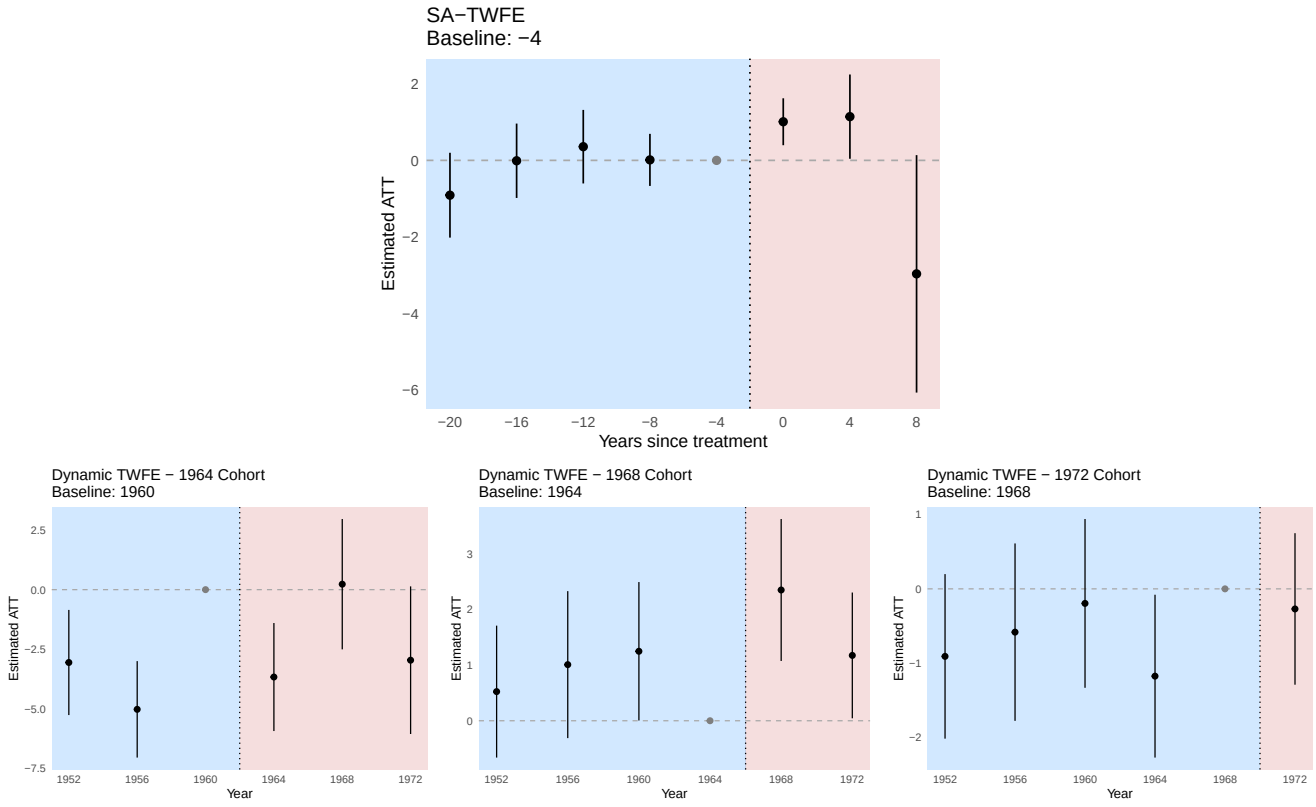

Figure 7: SA-TWFE estimates of the effect of food stamp program adoption on Democratic party vote share (Kogan 2021)

Setting aside the group-specific linear trends and looking at just the SA-TWFE dynamic estimates (Figure 7), there appears to be *some* positive effect in the post-treatment period with no *clear* pre-trends violation when averaging all of the cohort-specific estimates for each pre-treatment relative treatment time. However, these flat pre-trends appear to masking strong but countervailing positive and negative pre-trends violations across the different treatment timing groups. Plotting the cohort-specific TWFE estimates for each of the three timing groups, I find no effect for the 1972 cohort but some evidence of a pre-trends violation when taking the difference-in-differences between 1968 and 1964. For the 1968 cohort, I do see some evidence of a post-treatment effect, but again, also some evidence of a pre-trends violation. Were the baseline period 1960 or 1956, it does not appear that there would be a significant effect in 1968 or 1972. Indeed, projecting the linear pre-treatment trend forward (and excluding 1964) strongly suggests the absence of an effect for this cohort. Finally, estimates for the 1964 cohort suggest a *negative*

effect of the program, but with extremely clear pre-trends violations. Here, the estimates are essentially the converse of what is seen for the 1968 cohort.

In all three of these cohorts, there does not appear to be a clear and consistent linear pre-trend in the pre-treatment period, in contrast with HHB. Rather, at least for the largest two cohorts, it appears to be more consistent with a "blip" violation in parallel trends related specifically to the 1964 presidential election. Notably, 1964 was a significant landslide for Democrat Lyndon Johnson over Republican Barry Goldwater. As a consequence, one potential explanation for the parallel trends violation is a "compression" in the vote share differences between treated/control counties in 1964 due to ceiling effects.

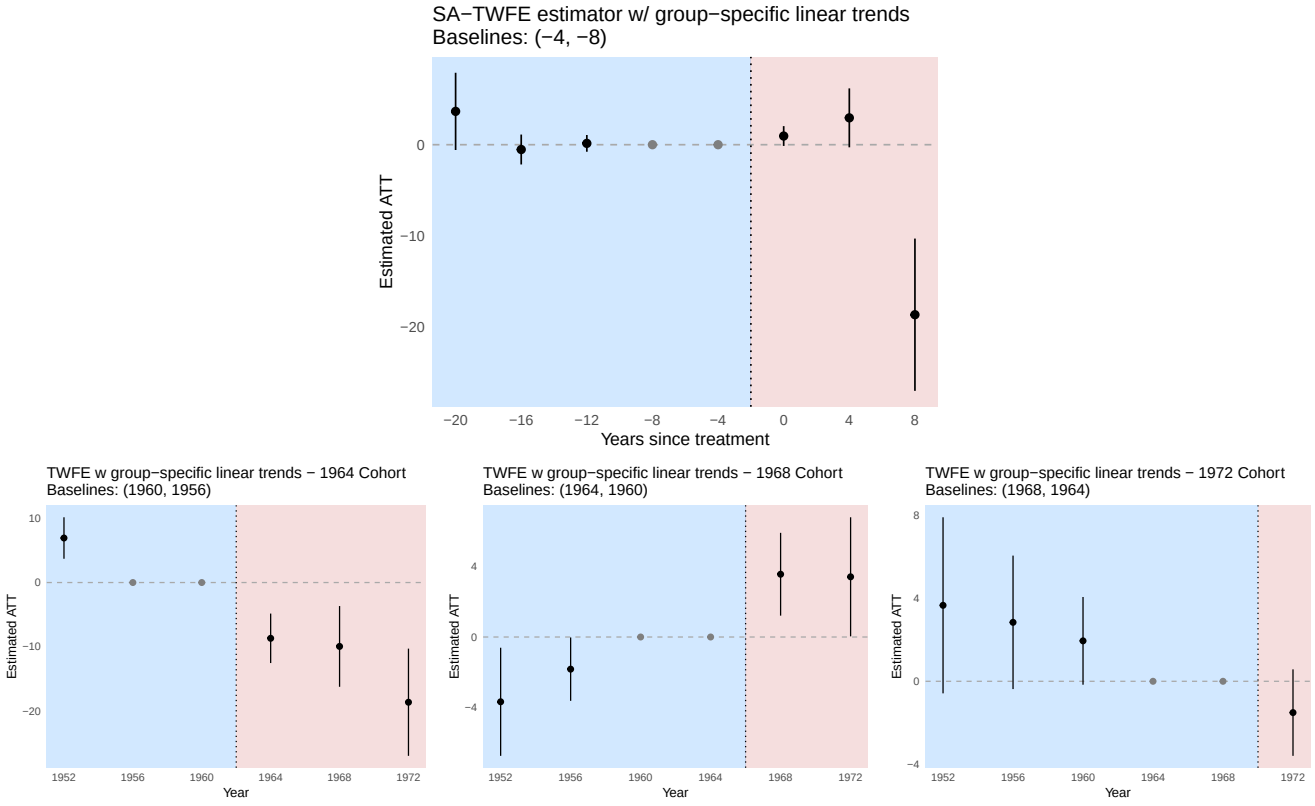

Figure 8: SA-TWFE w/ group-specific time trends estimates of the effect of food stamp program adoption on Democratic party vote share (Kogan 2021) – Baselines: (-4, -8)

Without an obvious pre-treatment linear trend, it is also the case that any linear trend adjustment is likely to be sensitive to the selection of the baseline periods. First, let's consider the intuitive baseline choice of -4 and -8 – the last and next-to-last elections before treatment. Figure 8 shows that this specification estimates a significant 10 percentage point negative effect

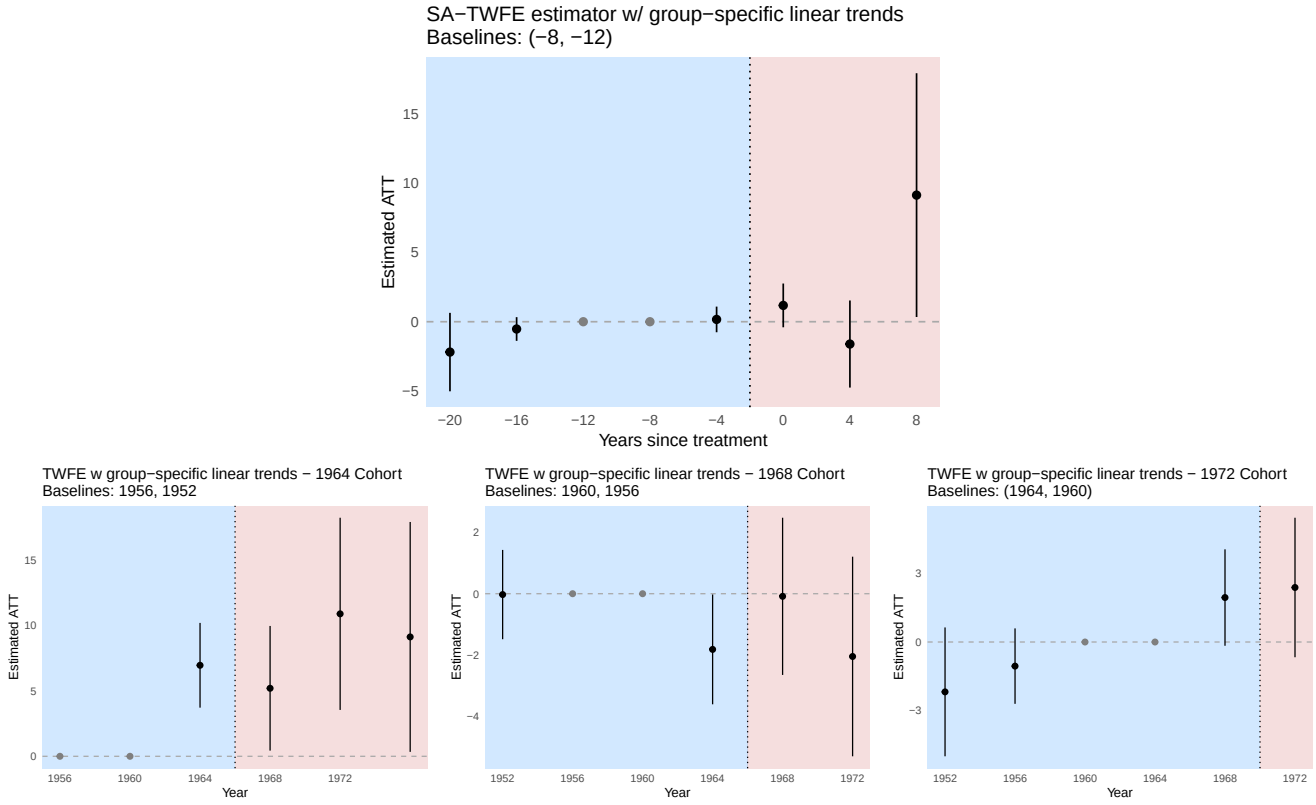

Figure 9: SA-TWFE w/ group-specific time trends estimates of the effect of food stamp program adoption on Democratic party vote share (Kogan 2021) – Baselines: (-8, -12)

of the program after three elections. Again, a closer look at the cohort-specific estimates suggests that this may be driven by idiosyncrasies (and continued violations of the identifying assumptions) with the 1964 cohort in particular. For all three of the cohorts, the pre-treatment placebos suggest that the triple-differences strategy is failing to address the potential parallel trends violation. Choosing an alternative baseline pair: -8 and -12 yields estimates (Figure 9 that instead show a statistically significant 10 percentage point *positive* effect of the program after three elections. While under the parallel trends-in-trends identifying assumption either set of baselines should be valid, the results are qualitatively distinct. Again, the cohort-specific estimates show evidence of failed "pre-trends," particularly for the 1964 cohort.

Overall, the results suggest that the parallel trends violation in Kogan (2021) is not of the form that would be amenable to a group-specific linear trends adjustment. It also highlights how the conventional practice of comparing changes in the regression coefficients under alternate model specifications can mask actual violations of the underlying identifying assumptions. Including

group-specific time trends into the standard static TWFE regression involves moving multiple assumptions. While it relaxes parallel trends (in a very particular way) it introduces a set of comparisons that require greater reliance on a constant effects assumption in the effects over time.

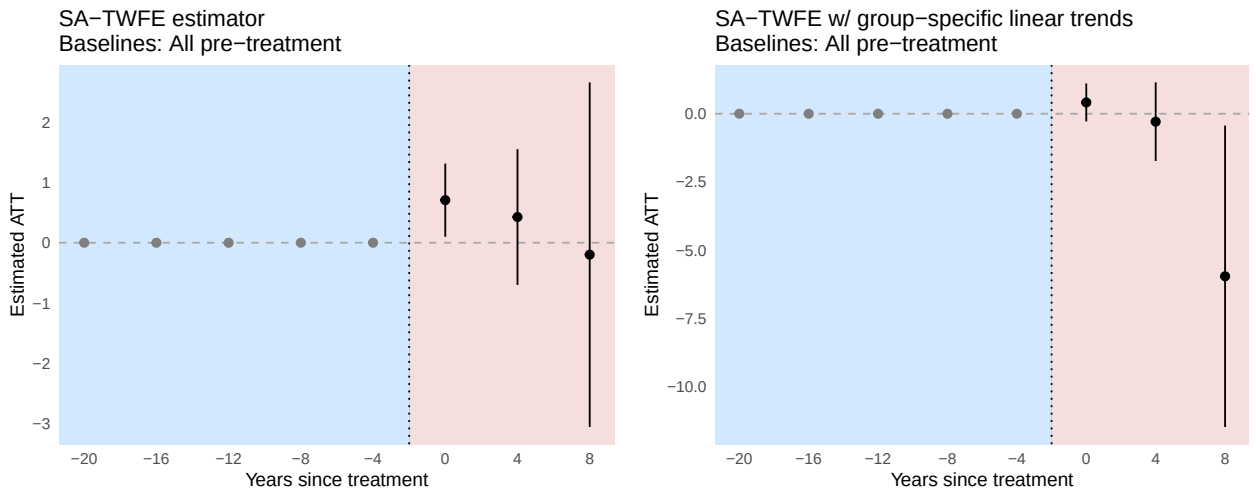

Figure 10: SA-TWFE estimates with and without group-specific time trends – All pre-treatment periods as baselines (Kogan 2021)

Consider what happens when the SA-TWFE regression is estimated with all possible pre-treatment baselines and compare the results when group-specific linear trends are included and not included (Figure 10). The normal SA-TWFE results suggest a small effect in the first election under treatment while the SA-TWFE with group-specific linear trends estimates show no immediate effect and a *negative* (albeit noisy) effect in the third election under treatment. Contrast this sign-flip with the minimally changed results from the static TWFE in Table 2. Unless researchers have a compelling reason to believe that the parallel trends violation has an obvious polynomial functional form, regressions with group-specific linear trends can convey a false sense of robustness.

5 Concluding Remarks

Political scientists are right to be concerned about parallel trends violations in difference-in-differences studies as recent large-scale replications of published panel data studies have shown (Chiu et al. 2023). Since many of these studies implement a conventional two-way fixed effects regression estimator it is simple to suggest and simple to include a group-specific time trend

as a "robustness check." Unfortunately, our understanding of how this changes the underlying assumptions required for TWFE to identify a useful treatment effect parameter has been limited. This paper shows that TWFE with group-specific linear time trends implies a triple-differences in time identification strategy. Researchers are implicitly assuming that any parallel trends violation in the sample exhibits a known functional form with respect to time and that the divergence from parallel trends can be eliminated simply through another round of differencing using an additional DiD from two pre-treatment periods.

Even if this identification strategy is plausible, this paper shows that its implementation in the form of a TWFE regression estimator requires additional assumptions in all but the simplest settings. Because the estimator uses post-treatment observations to adjust for the divergence in trends, any heterogeneity in treatment effects will be absorbed as part of the process for estimating the parallel trends violation. If parallel trends holds and treatment adoption is not staggered but treatment effects vary over time, then the standard two-way fixed effects estimator identifies a convex average of ATTs, while the regression with additional group-specific time trends *may not*.

Although it is easy to correct this by estimating separate parameters for each post-treatment period, doing so makes clear how sensitive regressions with group-specific time trends are to choices in specification. By giving up effect homogeneity, researchers must rely heavily on *extrapolation*, using the trend estimated *only* in the pre-treatment period to estimate the deviation from parallel trends in the post-treatment period. While it is tempting to suggest that this can be solved simply by fitting more flexible models, this is fundamentally a *data* problem and an *overlap* problem. As is well known in political science, even small deviations from the assumed parametric form can have significant consequences when extrapolating far outside the support of the data (King and Zeng 2006), an issue that is particularly acute for higher-order polynomials.

To conclude, researchers should not treat group-specific time trends in the same way that they would the inclusion of additional covariates in a regression model. The two imply different relaxations of the parallel trends assumption and use wildly different approaches to adjustment. In the case of covariates, adjustment takes the form of *stratification* – carrying out the same DiD analysis in subsets of the data. Absent any complications from lack of overlap or the inclusion of post-treatment variables, covariate adjustment introduces no additional problems and allows researchers to assume that parallel trends need only hold *within* covariate strata rather than in

the sample as a whole. Group-specific time trends, by contrast, adjust not by stratification but by introducing an additional *differencing* step that extrapolates the observed pre-treatment trend according to an assumed parametric form. Unfortunately, credibly extrapolating an observed pre-treatment divergence to the post-treatment period typically requires substantive knowledge that is lacking in most settings. Rather than trying to estimate the parallel trends violation exactly (and potentially inducing greater biases) a more reliable path to assessing a DiD's robustness to violations of the identifying assumption would be a partial identification/sensitivity analysis approach, such as the method developed in Rambachan and Roth (2023), which permits researchers to construct confidence sets for the treatment effect of interest under a theoretically motivated range of possible parallel trends violations.

References

Abdelgadir, Aala, and Vasiliki Fouka. 2020. "Political secularism and Muslim integration in the West: Assessing the effects of the French headscarf ban." *American Political Science Review* 114 (3): 707–723.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Ansell, Ben, Frederik Hjorth, Jacob Nyrup, and Martin Vinæs Larsen. 2022. "Sheltering populists? House prices and the support for populist parties." *The Journal of Politics* 84 (3): 1420–1436.

Arias, Eric, and David Stasavage. 2019. "How large are the political costs of fiscal austerity?" *The Journal of Politics* 81 (4): 1517–1522.

Benesch, Christine, Rino Heim, Mark Schelker, and Lukas Schmid. 2023. "Do voting advice applications change political behavior?" *The Journal of Politics* 85 (2): 684–700.

Bergé, Laurent. 2018. "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm." *CREA Discussion Papers,* no. 13.

Blair, Christopher W. 2022. "The fortification dilemma: border control and rebel violence." *American Journal of Political Science.*

Bøggild, Troels, and Carsten Jensen. 2024. "When politicians behave badly: Political, democratic, and social consequences of political incivility." *American Journal of Political Science.*

Borusyak, Kirill, and Xavier Jaravel. 2018. *Revisiting event study designs.* SSRN.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2021. "Revisiting event study designs: Robust and efficient estimation." *arXiv preprint arXiv:2108.12419.*

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

Chiu, Albert, Xingchen Lan, Ziyi Liu, and Yiqing Xu. 2023. "What To Do (and Not to Do) with Causal Panel Analysis under Parallel Trends: Lessons from A Large Reanalysis Study." *Available at SSRN 4490035.*

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

Dipoppa, Gemma, Guy Grossman, and Stephanie Zonszein. 2023. "Locked down, lashing out: COVID-19 effects on Asian hate crimes in Italy." *The Journal of Politics* 85 (2): 389–404.

Dube, Arindrajit, Daniele Girardi, Oscar Jorda, and Alan M Taylor. 2023. *A local projections approach to difference-in-differences event studies.* Technical report. National Bureau of Economic Research.

Dynes, Adam M, and John B Holbein. 2020. "Noisy retrospection: The effect of party control on policy outcomes." *American Political Science Review* 114 (1): 237–257.

Egami, Naoki, and Soichiro Yamauchi. 2023. "Using multiple pretreatment periods to improve difference-in-differences and staggered adoption designs." *Political Analysis* 31 (2): 195–212.

Egel, Naomi, and Nina Obermeier. 2023. "A friend like me: The effect of international organization membership on state preferences." *The Journal of Politics* 85 (1): 340–344.

Esberg, Jane, and Alexandra A Siegel. 2023. "How exile shapes online opposition: Evidence from Venezuela." *American Political Science Review* 117 (4): 1361–1378.

Ferwerda, Jeremy. 2021. "Immigration, voting rights, and redistribution: Evidence from local governments in Europe." *The Journal of Politics* 83 (1): 321–339.

Foos, Florian, and Daniel Bischof. 2022. "Tabloid media campaigns and public opinion: Quasi-experimental evidence on Euroscepticism in England." *American Political Science Review* 116 (1): 19–37.

Fouirnaies, Alexander. 2021. "How do campaign spending limits affect elections? Evidence from the United Kingdom 1885–2019." *American Political Science Review* 115 (2): 395–411.

Fouka, Vasiliki. 2019. "How do immigrants respond to discrimination? The case of Germans in the US during World War I." *American Political Science Review* 113 (2): 405–422.

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse M Shapiro. 2021. *Visualization, identification, and estimation in the linear panel event-study design.* Technical report. National Bureau of Economic Research.

Frisch, Ragnar, and Frederick V Waugh. 1933. "Partial time regressions as compared with individual trends." *Econometrica: Journal of the Econometric Society,* 387–401.

García-Montoya, Laura, Ana Arjona, and Matthew Lacombe. 2022. "Violence and voting in the United States: How school shootings affect elections." *American Political Science Review* 116 (3): 807–826.

Gardner, John. 2022. "Two-stage differences in differences." *arXiv preprint arXiv:2207.05943.*

Goldsmith-Pinkham, Paul. 2024. "Tracking the Credibility Revolution across Fields." *arXiv preprint arXiv:2405.20604.*

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

Grumbach, Jacob M, and Charlotte Hill. 2022. "Rock the registration: Same day registration increases turnout of young voters." *The Journal of Politics* 84 (1): 405–417.

Gulzar, Saad, Apoorva Lal, and Benjamin Pasquale. 2024. "Representation and forest conservation: Evidence from India's scheduled areas." *American Political Science Review* 118 (2): 764–783.

Hainmueller, Jens, and Dominik Hangartner. 2019. "Does direct democracy hurt immigrant minorities? Evidence from naturalization decisions in Switzerland." *American Journal of Political Science* 63 (3): 530–547.

Hall, Andrew B, and Jesse Yoder. 2022. "Does homeownership influence political behavior? Evidence from administrative data." *The Journal of Politics* 84 (1): 351–366.

Hamel, Brian T. 2024. "Traceability and Mass Policy Feedback Effects." *American Political Science Review,* 1–16.

Hankinson, Michael, and Asya Magazinnik. 2023. "The supply-equity trade-off: The effect of spatial representation on the local housing supply." *The Journal of Politics* 85 (3): 1033–1047.

Harding, Robin. 2020. "Who is democracy good for? Elections, rural bias, and health and education outcomes in sub-Saharan Africa." *The Journal of Politics* 82 (1): 241–254.

Hassell, Hans JG, and John B Holbein. 2024. "Navigating potential pitfalls in difference-in-differences designs: reconciling conflicting findings on mass shootings' effect on electoral outcomes." *American Political Science Review,* 1–21.

Hassell, Hans JG, John B Holbein, and Matthew Baldwin. 2020. "Mobilize for our lives? School shootings and democratic accountability in US elections." *American Political Science Review* 114 (4): 1375–1385.

Hoynes, Hilary W, and Diane Whitmore Schanzenbach. 2009. "Consumption responses to in-kind transfers: Evidence from the introduction of the food stamp program." *American Economic Journal: Applied Economics* 1 (4): 109–139.

Imai, Kosuke, and In Song Kim. 2021. "On the use of two-way fixed effects regression models for causal inference with panel data." *Political Analysis* 29 (3): 405–415.

Imai, Kosuke, In Song Kim, and Erik H Wang. 2021. "Matching methods for causal inference with time-series cross-sectional data." *American Journal of Political Science.*

Kahn-Lang, Ariella, and Kevin Lang. 2020. "The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications." *Journal of Business & Economic Statistics* 38 (3): 613–620.

Kilborn, Mitchell, and Arjun Vishwanath. 2022. "Public money talks too: How public campaign financing degrades representation." *American Journal of Political Science* 66 (3): 730–744.

King, Gary, and Langche Zeng. 2006. "The dangers of extreme counterfactuals." *Political analysis* 14 (2): 131–159.

Kogan, Vladimir. 2021. "Do welfare benefits pay electoral dividends? Evidence from the national food stamp program rollout." *The Journal of Politics* 83 (1): 58–70.

Li, Zikai, and Anton Strezhnev. 2024. "A Guide to Dynamic Difference-in-Differences Regressions for Political Scientists."

Liu, Licheng, Ye Wang, and Yiqing Xu. 2022. "A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data." *American Journal of Political Science.*

Lovell, Michael C. 1963. "Seasonal adjustment of economic time series and multiple regression analysis." *Journal of the American Statistical Association* 58 (304): 993–1010.

Magaloni, Beatriz, Edgar Franco-Vivanco, and Vanessa Melo. 2020. "Killing in the slums: Social order, criminal governance, and police violence in Rio de Janeiro." *American Political Science Review* 114 (2): 552–572.

Marshall, John. 2019. "The anti-Democrat diploma: How high school education decreases support for the Democratic Party." *American Journal of Political Science* 63 (1): 67–83.

Masterson, Daniel, and Vasil Yasenov. 2021. "Does halting refugee resettlement reduce crime? Evidence from the US refugee ban." *American Political Science Review* 115 (3): 1066–1073.

Mora, Ricardo, and Iliana Reggio. 2019. "Alternative diff-in-diffs estimators with several pretreatment periods." *Econometric Reviews* 38 (5): 465–486.

Mou, Hongyu, Licheng Liu, and Yiqing Xu. 2023. "Panel Data Visualization in R (panelView) and Stata (panelview)." *Journal of Statistical Software* 107 (7): 1–20. https://doi.org/10.18637/jss.v107.i07.

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

Paglayan, Agustina S. 2019. "Public-sector unions and the size of government." *American Journal of Political Science* 63 (1): 21–36.

———. 2021. "The non-democratic roots of mass education: evidence from 200 years." *American Political Science Review* 115 (1): 179–198.

Payson, Julia A. 2020. "Cities in the statehouse: How local governments use lobbyists to secure state funding." *The Journal of Politics* 82 (2): 403–417.

Potter, Rachel Augustine. 2022. "Macro Outsourcing: Evaluating Government Reliance on the Private Sector." *The Journal of Politics* 84 (2): 960–974.

Pulejo, Massimo. 2023. "Religious mobilization and the selection of political elites: Evidence from postwar italy." *American Journal of Political Science.*

Rambachan, Ashesh, and Jonathan Roth. 2023. "A more credible approach to parallel trends." *Review of Economic Studies* 90 (5): 2555–2591.

Rogowski, Jon C, John Gerring, Matthew Maguire, and Lee Cojocaru. 2022. "Public infrastructure and economic development: Evidence from postal systems." *American Journal of Political Science* 66 (4): 885–901.

Roth, Jonathan, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature." *Journal of Econometrics.*

Rubin, Donald. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66 (5): 688.

Sides, John, Lynn Vavreck, and Christopher Warshaw. 2022. "The effect of television advertising in united states elections." *American Political Science Review* 116 (2): 702–718.

Strezhnev, Anton. 2023. "Decomposing triple-differences regression under staggered adoption." *arXiv preprint arXiv:2307.02735.*

Su, Min, and Christian Buerger. 2024. "Playing politics with traffic fines: Sheriff elections and political cycles in traffic fines revenue." *American Journal of Political Science.*

Sun, Liyang. 2022. "EVENTSTUDYINTERACT: Stata module to implement the interaction weighted estimator for an event study."

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

Ternullo, Stephanie. 2022. "The Electoral Effects of Social Policy: Expanding Old-Age Assistance, 1932–1940." *The Journal of Politics* 84 (1): 226–241.

Ward, George. 2020. "Happiness and voting: Evidence from four decades of elections in Europe." *American Journal of Political Science* 64 (3): 504–518.

Wing, Coady, Kosali Simon, and Ricardo A. Bello-Gomez. 2018. "Designing Difference in Difference Studies: Best Practices for Public Health Policy Research." *Annual Review of Public Health* 39:453–469.

Wolfers, Justin. 2006. "Did unilateral divorce laws raise divorce rates? A reconciliation and new results." *American Economic Review* 96 (5): 1802–1820.

Wooldridge, Jeffrey M. 2021. "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators." *Available at SSRN 3906345.*

Yousaf, Hasin. 2021. "Sticking to one's guns: Mass shootings and the political economy of gun control in the United States." *Journal of the European Economic Association* 19 (5): 2765–2802.

Zhang, Nan, and Melissa M Lee. 2020. "Literacy and State–Society Interactions in Nineteenth-Century France." *American Journal of Political Science* 64 (4): 1001–1016.

Appendix

A Proofs

A.1 Proof of Proposition 1

Consider the triple-differences in time estimator with four time periods: t, t', t_2 and t_1. Assume $t \geq g$, $t' < g$, $t_2 < g$, $t_1 < g$, $t_2 \neq t_1$.

Take the expectation of the triple-differences in time estimator conditional on the distribution of treatment $\mathcal{D}$

$$E[\hat{\tau}_{\text{TDiT}}|\mathcal{D}] = E[\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'}|\mathcal{D}] - E[\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1}|\mathcal{D}]\frac{(t-t')}{(t_2-t_1)}$$

Start with the first difference-in-difference. Under consistency (Assumption 1)

$$E[\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'}|\mathcal{D}] = E[Y_{it}(1)|D_i = 1] - E[Y_{it'}(1)|D_i = 1] - E[Y_{it}(0)|D_i = 0] + E[Y_{it'}(0)|D_i = 0]$$

Adding/subtracting the unobserved terms $E[Y_{it}(0)|D_i = 1]$ and $E[Y_{it'}(0)|D_i = 1]$

$$\begin{aligned} &E[\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'}|\mathcal{D}] = \\ &\qquad \Big(E[Y_{it}(1)|D_i = 1] - E[Y_{it}(0)|D_i = 1]\Big) - \Big(E[Y_{it'}(1)|D_i = 1] - E[Y_{it'}(0)|D_i = 1]\Big) \\ &\qquad + E[Y_{it}(0)|D_i = 1] - E[Y_{it'}(0)|D_i = 1] - E[Y_{it}(0)|D_i = 0] + E[Y_{it'}(0)|D_i = 0] \end{aligned}$$

Under the constant linear violation of parallel trends (Assumption 4), the second line equals

$\Delta(t - t')$

$$E[\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'}|\mathcal{D}] =$$
$$\Big(E[Y_{it}(1)|D_i = 1] - E[Y_{it}(0)|D_i = 1]\Big) - \Big(E[Y_{it'}(1)|D_i = 1] - E[Y_{it'}(0)|D_i = 1]\Big)$$
$$+ \Delta(t - t')$$

Under no anticipation (Assumption 3), since $t' < g$, $Y_{it'}(1) = Y_{it'}(0)$ and $ATT(t') = 0$.

$$\begin{aligned} E[\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'}|\mathcal{D}] &= \Big(E[Y_{it}(1)|D_i = 1] - E[Y_{it}(0)|D_i = 1]\Big) + \Delta(t - t') \\ &= ATT(t) + \Delta(t - t') \end{aligned}$$

Now consider the second difference-in-difference term. Following the same approach as above, assumptions 1 and 4 yield

$$E[\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1}|\mathcal{D}] =$$
$$\Big(E[Y_{it_2}(1)|D_i = 1] - E[Y_{it_2}(0)|D_i = 1]\Big) - \Big(E[Y_{it_1}(1)|D_i = 1] - E[Y_{it_1}(0)|D_i = 1]\Big)$$
$$+ \Delta(t_2 - t_1)$$

Since both t_2 and t_1 are less than g, under no anticipation (Assumption 3), both ATT terms are equal to zero, leaving only the bias term.

$$E[\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1}|\mathcal{D}] = \Delta(t_2 - t_1)$$

Substituting back into $E[\hat{\tau}_{\text{TDiT}}|\mathcal{D}]$

$$\begin{aligned} E[\hat{\tau}_{\text{TDiT}}|\mathcal{D}] &= ATT(t) + \Delta(t - t') - \Delta(t_2 - t_1)\frac{(t - t')}{(t_2 - t_1)} \\ &= ATT(t) + \Delta(t - t') - \Delta(t - t') \\ &= ATT(t) \end{aligned}$$

Note that if both t_2 and t_1 are greater than or equal to g, no anticipation *does not* rule out

non-zero $ATT(t_2)$ and $ATT(t_1)$ since both periods could be impacted by treatment. In this case, the second difference-in-difference term identifies $ATT(t_2) - ATT(t_1) + \Delta(t_2 - t_1)$. Identification of $ATT(t)$ would require an additional effect homogeneity assumption that $ATT(t_2)$ and $ATT(t_1)$ are equivalent.

A.2 Proof of Proposition 2

Consider the two-way fixed effects regression under no staggering

$$Y_{it} = \tau D_{it} + \alpha D_i + \delta_t + \beta(D_i \times t) + \varepsilon_{it}$$

where $D_{it} = D_i \times \mathbf{1}(t \geq g)$

By Frisch-Waugh-Lovell, the OLS estimator $\hat{\tau}$ is equivalent to the bivariate regression of Y_{it} on $\tilde{D}_{it} = D_{it} - \hat{D}_{it}$ where

$$\hat{D}_{it} = \hat{\alpha} D_i + \hat{\delta}_t + \hat{\beta}(D_i \times t)$$

The coefficients in the auxiliary regression: $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\delta}_t$, minimize the least-squares objective function:

$$\begin{aligned}
\hat{\alpha}, \hat{\beta}, \hat{\delta} &= \underset{\alpha,\beta,\delta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(D_i \times \mathbf{1}(t \geq g) - \alpha D_i - \delta_t - \beta(D_i \times t) \Big)^2 \\
&= \underset{\alpha,\beta,\delta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(D_i \times [\mathbf{1}(t \geq g) - \alpha - \beta \times t] - \delta_t \Big)^2
\end{aligned}$$

Using the normal equations and solving for $\hat{\delta}_t$

$$\hat{\delta}_t = \frac{N_1}{N} \mathbf{1}(t \geq g) - \frac{N_1}{N} \hat{\alpha} - \frac{N_1}{N} (\hat{\beta} \times t)$$

Substituting the time-fixed effect into $\hat{D}_{it}$

$$\begin{aligned}\hat{D}_{it} &= \hat{\alpha}D_i + \frac{N_1}{N}\mathbf{1}(t \geq g) - \frac{N_1}{N}\hat{\alpha} - \frac{N_1}{N}(\hat{\beta} \times t) + \hat{\beta}(D_i \times t) \\ &= \frac{N_1}{N}\mathbf{1}(t \geq g) + \hat{\alpha}\left(D_i - \frac{N_1}{N}\right) + \hat{\beta}t\left(D_i - \frac{N_1}{N}\right)\end{aligned}$$

Next, solving for $\hat{\alpha}$

$$\hat{\alpha} = \frac{T_1}{T} - \frac{\sum_{t=1}^{T}\hat{\delta}_t}{T} - \hat{\beta}\frac{\sum_{t=1}^{T}t}{T}$$

and for $\hat{\beta}$

$$\hat{\beta} = \frac{\sum_{t=g}^{T}t}{\sum_{t=1}^{T}t^2} - \hat{\alpha}\frac{\sum_{t=1}^{T}t}{\sum_{t=1}^{T}t^2} - \frac{\sum_{t=1}^{T}t\hat{\delta}_t}{\sum_{t=1}^{T}t^2}$$

The sums over the time fixed effects can be written as:

$$\sum_{t=1}^{T}\hat{\delta}_t = \frac{N_1T_1}{N} - \frac{N_1T}{N}\hat{\alpha} - \hat{\beta}\frac{N_1}{N}\sum_{t=1}^{T}t$$

$$\sum_{t=1}^{T}t\hat{\delta}_t = \frac{N_1}{N}\sum_{t=g}^{T}t - \frac{N_1}{N}\hat{\alpha}\sum_{t=1}^{T}t - \hat{\beta}\frac{N_1}{N}\sum_{t=1}^{T}t^2$$

Using Faulhaber's formula, sums over t and t^2 can be written in closed form. The following

identities are used throughout the proof.

$$\sum_{t=1}^{T} t = \frac{T(T+1)}{2}$$

$$\sum_{t=1}^{T} t^2 = \frac{T(T+1)(2T+1)}{6}$$

$$\sum_{t=g}^{T} t = \sum_{t=1}^{T} t - \sum_{t=1}^{g-1} t = \frac{(T+g)(T-g+1)}{2}$$

$$\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T} t_2 - t_1 = \frac{(T-1)T(T+1)}{6}$$

$$\sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T} t_2 - t_1 = \sum_{t_1=1}^{T-g+1}\sum_{t_2=t_1}^{T-g+1} t_2 - t_1 = \frac{(T-g+2)(T-g+1)(T-g)}{6}$$

$$\sum_{t_1=1}^{g-1}\sum_{t_2=g}^{T} t_2 - t_1 = \frac{T(g-1)(T-g+1)}{2}$$

$$\sum_{t_1=1}^{g-1}\sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T} t_2 - t_1 = \frac{(T-1)T(T+1) - 3T(g-1)(T-g+1)}{6}$$

Substituting into $\hat{\alpha}$ yields:

$$\begin{aligned}
\hat{\alpha} &= \frac{T_1}{T} - \frac{1}{T}\left(\frac{N_1 T_1}{N} - \frac{N_1 T}{N}\hat{\alpha} - \hat{\beta}\frac{N_1}{N}\sum_{t=1}^{T} t\right) - \hat{\beta}\frac{\sum_{t=1}^{T} t}{T} \\
&= \frac{T_1}{T} - \frac{N_1 T_1}{NT} + \frac{N_1}{N}\hat{\alpha} + \hat{\beta}\frac{N_1}{NT}\sum_{t=1}^{T} t - \hat{\beta}\frac{\sum_{t=1}^{T} t}{T} \\
&= \frac{(N-N_1)T_1}{NT} + \frac{N_1}{N}\hat{\alpha} - \hat{\beta}\frac{N-N_1}{NT}\sum_{t=1}^{T} t \\
\frac{N-N_1}{N}\hat{\alpha} &= \frac{(N-N_1)T_1}{NT} - \hat{\beta}\frac{N-N_1}{NT}\sum_{t=1}^{T} t \\
\hat{\alpha} &= \frac{T_1}{T} - \hat{\beta}\frac{1}{T}\sum_{t=1}^{T} t \\
\hat{\alpha} &= \frac{T_1}{T} - \hat{\beta}\frac{T+1}{2} \\
\hat{\alpha} &= \frac{T-g+1}{T} - \hat{\beta}\frac{T+1}{2}
\end{aligned}$$

And substituting into $\hat{\beta}$ yields:

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{t=g}^{T} t}{\sum_{t=1}^{T} t^2} - \hat{\alpha}\frac{\sum_{t=1}^{T} t}{\sum_{t=1}^{T} t^2} - \frac{1}{\sum_{t=1}^{T} t^2}\left[\frac{N_1}{N}\sum_{t=g}^{T} t - \frac{N_1}{N}\hat{\alpha}\sum_{t=1}^{T} t - \hat{\beta}\frac{N_1}{N}\sum_{t=1}^{T} t^2\right] \\
&= \frac{\sum_{t=g}^{T} t}{\sum_{t=1}^{T} t^2} - \hat{\alpha}\frac{\sum_{t=1}^{T} t}{\sum_{t=1}^{T} t^2} - \frac{N_1}{N}\frac{\sum_{t=g}^{T} t}{\sum_{t=1}^{T} t^2} + \frac{N_1}{N}\frac{\sum_{t=1}^{T} t}{\sum_{t=1}^{T} t^2}\hat{\alpha} + \hat{\beta}\frac{N_1}{N} \\
\frac{N-N_1}{N}\hat{\beta} &= \frac{N-N_1}{N} \times \frac{\sum_{t=g}^{T} t}{\sum_{t=1}^{T} t^2} - \hat{\alpha} \times \frac{N-N_1}{N} \times \frac{\sum_{t=1}^{T} t}{\sum_{t=1}^{T} t^2} \\
\hat{\beta} &= \frac{\sum_{t=g}^{T} t}{\sum_{t=1}^{T} t^2} - \hat{\alpha} \times \frac{\sum_{t=1}^{T} t}{\sum_{t=1}^{T} t^2} \\
\hat{\beta} &= \frac{3}{2T+1}\left[\frac{(T+g)(T-g+1)}{T(T+1)} - \hat{\alpha}\right]
\end{aligned}$$

Solving for $\hat{\beta}$ in closed form

$$\begin{aligned}
\hat{\beta} &= \frac{3}{2T+1}\left[\frac{(T+g)(T-g+1)}{T(T+1)} - \frac{T-g+1}{T}\right] + \hat{\beta}\frac{3(T+1)}{2(2T+1)} \\
\hat{\beta} &= \frac{3}{2T+1}\left[\frac{(g-1)(T-g+1)}{T(T+1)}\right] + \hat{\beta}\frac{3(T+1)}{2(2T+1)} \\
\frac{T-1}{2(2T+1)}\hat{\beta} &= \frac{3}{2T+1}\left[\frac{(g-1)(T-g+1)}{T(T+1)}\right] \\
\frac{T-1}{2}\hat{\beta} &= 3\left[\frac{(g-1)(T-g+1)}{T(T+1)}\right] \\
\hat{\beta} &= 6\left[\frac{(g-1)(T-g+1)}{(T-1)T(T+1)}\right]
\end{aligned}$$

Substituting into $\hat{\alpha}$

$$\begin{aligned}
\hat{\alpha} &= \frac{T-g+1}{T} - 3\left[\frac{(g-1)(T-g+1)}{T(T-1)}\right] \\
&= \frac{(T-3g+2)(T-g+1)}{T(T-1)}
\end{aligned}$$

This allows $\tilde{D}_{it}$ to be written in terms of the closed-form expressions for $\hat{\alpha}$ and $\hat{\beta}$

$$\begin{aligned}\tilde{D}_{it} &= \left(D_i - \frac{N_1}{N}\right)\mathbf{1}(t \geq g) - \hat{\alpha}\left(D_i - \frac{N_1}{N}\right) - \hat{\beta}t\left(D_i - \frac{N_1}{N}\right) \\ &= \left(D_i - \frac{N_1}{N}\right)\mathbf{1}(t \geq g) - \frac{(T-3g+2)(T-g+1)}{T(T-1)}\left(D_i - \frac{N_1}{N}\right) \\ &\quad - 6\left[\frac{(g-1)(T-g+1)}{T(T-1)(T+1)}\right]\left(D_i - \frac{N_1}{N}\right) \times t\end{aligned}$$

Re-arrange terms to write this in the form of a "triple-difference"

$$\begin{aligned}\tilde{D}_{it} &= \left(D_i - \frac{N_1}{N}\right)\left(\mathbf{1}(t \geq g) - \frac{T-g+1}{T}\right) \\ &\quad + \left(D_i - \frac{N_1}{N}\right)\left(\frac{3(g-1)(T-g+1)}{T(T-1)} - \frac{6(g-1)(T-g+1)}{T(T-1)(T+1)}t\right) \\ \tilde{D}_{it} &= \left(D_i - \frac{N_1}{N}\right)\left(\mathbf{1}(t \geq g) - \frac{T-g+1}{T}\right) + \left(\frac{3(g-1)(T-g+1)}{T(T-1)}\right)\left(D_i - \frac{N_1}{N}\right)\left(1 - \frac{2t}{T+1}\right)\end{aligned}$$

Now, write the bivariate FWL regression coefficient of Y_{it} on $\tilde{D}_{it}$

$$\hat{\tau} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}\tilde{D}_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{D}_{it})^2}$$

Start with the numerator first and multiply Y_{it} through each of the two terms in $\tilde{D}_{it}$

$$\begin{aligned}\hat{\tau} &\propto \sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}\left(D_i - \frac{N_1}{N}\right)\left(\mathbf{1}(t \geq g) - \frac{T-g+1}{T}\right) - \\ &\qquad \sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}\left(\frac{3(g-1)(T-g+1)}{T(T-1)}\right)\left(D_i - \frac{N_1}{N}\right)\left(\frac{2t}{T+1} - 1\right) \\ &\propto \sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}\left(D_i - \frac{N_1}{N}\right)\left(\mathbf{1}(t \geq g) - \frac{T-g+1}{T}\right) - \\ &\qquad \sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}\left(\frac{6(g-1)(T-g+1)}{(T+1)T(T-1)}\right)\left(D_i - \frac{N_1}{N}\right)\left(t - \frac{T+1}{2}\right)\end{aligned}$$

The first part simplifies into a sum over differences-in-differences terms. Write the outcome

means at time t under treatment condition d as $\bar{Y}_{d,t} = \frac{1}{N_d} \sum_{D_i=d} Y_{it}$

$$
\begin{aligned}
&\sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it} \left(D_i - \frac{N_1}{N} \right) \left(\mathbf{1}(t \geq g) - \frac{T-g+1}{T} \right) \\
&= \sum_{t=g}^{T} \sum_{i:D_i=1} \frac{(N-N_1)(g-1)}{NT} Y_{it} - \sum_{t=g}^{T} \sum_{i:D_i=0} \frac{N_1(g-1)}{NT} Y_{it} \\
&\qquad - \sum_{t=1}^{g-1} \sum_{i:D_i=1} \frac{(N-N_1)(T-g+1)}{NT} Y_{it} + \sum_{t=1}^{g-1} \sum_{i:D_i=0} \frac{(N_1)(T-g+1)}{NT} Y_{it} \\
&= \frac{N_1(N-N_1)(g-1)}{NT} \sum_{t=g}^{T} \left[\bar{Y}_{1,t} - \bar{Y}_{0,t} \right] - \frac{N_1(N-N_1)(T-g+1)}{NT} \sum_{t=1}^{g-1} \left[\bar{Y}_{1,t} - \bar{Y}_{0,t} \right] \\
&= \frac{N_1(N-N_1)}{NT} \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \left[\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'} \right]
\end{aligned}
$$

The second part also simplifies into an average over "differences-in-differences" but involving all time periods.

$$
\begin{aligned}
&\sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it} \left(\frac{6(g-1)(T-g+1)}{(T+1)T(T-1)} \right) \left(D_i - \frac{N_1}{N} \right) \left(t - \frac{T+1}{2} \right) = \\
&\qquad \left(\frac{6N_1(N-N_1)(g-1)(T-g+1)}{N(T+1)T(T-1)} \right) \left[\sum_{t=1}^{T} \left[\bar{Y}_{1,t} - \bar{Y}_{0,t} \right] \left(t - \frac{T+1}{2} \right) \right] \\
&= \left(\frac{6N_1(N-N_1)(g-1)(T-g+1)}{N(T+1)T(T-1)} \right) \left[\sum_{t=1}^{T} \left[\bar{Y}_{1,t} - \bar{Y}_{0,t} \right] \left(t - \frac{\sum_{t'=1}^{T} t'}{T} \right) \right] \\
&= \left(\frac{6N_1(N-N_1)(g-1)(T-g+1)}{N(T+1)T(T-1)} \right) \left(\frac{1}{T} \right) \sum_{t=1}^{T} \sum_{t'=1}^{T} \left[\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'} \right] t \\
&= \left(\frac{6N_1(N-N_1)(g-1)(T-g+1)}{N(T+1)T(T-1)} \right) \left(\frac{1}{T} \right) \sum_{t'=1}^{T} \sum_{t=t'}^{T} \left(\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'} \right) \left(t - t' \right) \\
&= \left(\frac{N_1(N-N_1)(g-1)(T-g+1)}{NT} \right) \frac{\sum_{t'=1}^{T} \sum_{t=t'}^{T} \left(\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'} \right) \left(t - t' \right)}{\sum_{t'=1}^{T} \sum_{t=t'}^{T} t - t'}
\end{aligned}
$$

Combine these two expressions into a sum over four time indices: t and t' (for the "first"

difference-in-differences) and t_1 and t_2 (for the placebo or "second" difference-in-differences).

$$= \left(\frac{N_1(N-N_1)}{NT}\right) \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \left\{ \bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'} - \frac{\sum_{t_1=1}^{T} \sum_{t_2=t_1}^{T} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})(t_2 - t_1)}{\sum_{t_1=1}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \right\}$$

Splitting the sum over t_1 and t_2 into three separate sums: both post-treatment, both pre-treatment and t_2 post-treatment/t_1 pre-treatment.

$$\begin{aligned} = \left(\frac{N_1(N-N_1)}{NT}\right) \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \Bigg\{ & \bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'} - \\ & \Bigg[\frac{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})(t_1 - t_2)}{\sum_{t_1=1}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} + \\ & \frac{\sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})(t_1 - t_2)}{\sum_{t_1=1}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} + \\ & \frac{\sum_{t_1=1}^{g-1} \sum_{t_2=g}^{T} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})(t_1 - t_2)}{\sum_{t_1=1}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \Bigg] \Bigg\} \end{aligned}$$

Adding and subtracting $t - t'$ and combining terms

$$\begin{aligned} & = \left(\frac{N_1(N-N_1)}{NT}\right) \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \Bigg\{ \Bigg[\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'} \Bigg] - \\ & \Bigg[\frac{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})(t - t')}{\sum_{t_1=1}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} + \frac{\sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})(t - t')}{\sum_{t_1=1}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} + \\ & \frac{\sum_{t_1=1}^{g-1} \sum_{t_2=g}^{T} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})(t - t')}{\sum_{t_1=1}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} + \\ & \frac{\sum_{t_1=1}^{T} \sum_{t_2=t_1}^{T} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})(t_2 - t_1) - \sum_{t_1=1}^{T} \sum_{t_2=t_1}^{T} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} - \bar{Y}_{0,t_1})(t - t')}{\sum_{t_1=1}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \Bigg] \Bigg\} \end{aligned}$$

First, combine the third term in the square brackets with the first difference-in-difference. Since the sums are over the same range indices t and t' can be swapped ith t_1 and t_2. Then apply

the results from above using Faulhaber's formula.

$$\begin{aligned}
\frac{\sum_{t=g}^{T}\sum_{t'=1}^{g-1}\sum_{t_1=1}^{g-1}\sum_{t_2=g}^{T}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2}-\bar{Y}_{1,t_1}+\bar{Y}_{0,t_1})(t-t')}{\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}t_2-t_1} &= \frac{\sum_{t_2=g}^{T}\sum_{t_1=1}^{g-1}\sum_{t'=1}^{g-1}\sum_{t=g}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t}-\bar{Y}_{1,t'}+\bar{Y}_{0,t'})(t_2-t_1)}{\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}t_2-t_1} \\
&= \frac{\sum_{t'=1}^{g-1}\sum_{t=g}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t}-\bar{Y}_{1,t'}+\bar{Y}_{0,t'})\sum_{t_2=g}^{T}\sum_{t_1=1}^{g-1}(t_2-t_1)}{\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}t_2-t_1} \\
&= \sum_{t=g}^{T}\sum_{t'=1}^{g-1}(\bar{Y}_{1,t}-\bar{Y}_{0,t}-\bar{Y}_{1,t'}+\bar{Y}_{0,t'})\frac{3T(g-1)(T-g+1)}{T(T-1)(T+1)}
\end{aligned}$$

Then, the fourth term in the brackets is equal to zero.

$$
\begin{aligned}
&\sum_{t=g}^{T}\sum_{t'=1}^{g-1}\Bigg[\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2}-\bar{Y}_{1,t_1}+\bar{Y}_{0,t_1})(t_2-t_1)-\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2}-\bar{Y}_{1,t_1}+\bar{Y}_{0,t_1})(t-t')\Bigg]\\
&=\sum_{t=g}^{T}\sum_{t'=1}^{g-1}\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2})t_2-\sum_{t=g}^{T}\sum_{t'=1}^{g-1}\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2})t_1\\
&\quad-\sum_{t=g}^{T}\sum_{t'=1}^{g-1}\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_1}-\bar{Y}_{0,t_1})t_2+\sum_{t=g}^{T}\sum_{t'=1}^{g-1}\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_1}-\bar{Y}_{0,t_1})t_1\\
&\quad-\sum_{t=g}^{T}\sum_{t'=1}^{g-1}\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2})(t-t')+\sum_{t=g}^{T}\sum_{t'=1}^{g-1}\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_1}-\bar{Y}_{0,t_1})(t-t')\\
&=(g-1)(T-g+1)\sum_{t_2=1}^{T}\sum_{t_1=1}^{t_2}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2})t_2-(g-1)(T-g+1)\sum_{t_2=1}^{T}\sum_{t_1=1}^{t_2}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2})t_1\\
&\quad-(g-1)(T-g+1)\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_1}-\bar{Y}_{0,t_1})t_2+(g-1)(T-g+1)\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_1}-\bar{Y}_{0,t_1})t_1\\
&\quad-\frac{T(g-1)(T-g+1)}{2}\sum_{t_2=1}^{T}\sum_{t_1=1}^{t_2}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2})+\frac{T(g-1)(T-g+1)}{2}\sum_{t_1=1}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_1}-\bar{Y}_{0,t_1})\\
&=(g-1)(T-g+1)\sum_{t_2=1}^{T}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2})t_2^2-(g-1)(T-g+1)\sum_{t_2=1}^{T}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2})\frac{t_2^2+t_2}{2}\\
&\quad-(g-1)(T-g+1)\sum_{t_1=1}^{T}(\bar{Y}_{1,t_1}-\bar{Y}_{0,t_1})\frac{(T+t_1)(T-t_1+1)}{2}\\
&\quad+(g-1)(T-g+1)\sum_{t_1=1}^{T}(\bar{Y}_{1,t_1}-\bar{Y}_{0,t_1})t_1(T-t_1+1)\\
&\quad-\frac{T(g-1)(T-g+1)}{2}\sum_{t_2=1}^{T}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2})t_2+\frac{T(g-1)(T-g+1)}{2}\sum_{t_1=1}^{T}(\bar{Y}_{1,t_1}-\bar{Y}_{0,t_1})(T-t_1+1)
\end{aligned}
$$

Combining terms

$$
\begin{aligned}
&= (g-1)(T-g+1)\sum_{t=1}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t})t^2 - (g-1)(T-g+1)\sum_{t=1}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t})\frac{t^2+t}{2} \\
&\quad - (g-1)(T-g+1)\sum_{t=1}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t})\frac{(T+t)(T-t+1)}{2} \\
&\quad + (g-1)(T-g+1)\sum_{t=1}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t})t(T-t+1) \\
&\quad - \frac{T(g-1)(T-g+1)}{2}\sum_{t=1}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t})t + \frac{T(g-1)(T-g+1)}{2}\sum_{t=1}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t})(T-t+1) \\
&= (g-1)(T-g+1)\sum_{t=1}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t})\left[t^2 - \frac{t^2+t}{2} - \frac{(T+t)(T-t+1)}{2} + t(T-t+1)\right] \\
&\quad + \frac{T(g-1)(T-g+1)}{2}\sum_{t=1}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t})(T-2t+1) \\
&= -\frac{T(g-1)(T-g+1)}{2}\sum_{t=1}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t})(T-2t+1) + \frac{T(g-1)(T-g+1)}{2}\sum_{t=1}^{T}(\bar{Y}_{1,t}-\bar{Y}_{0,t})(T-2t+1) \\
&= 0
\end{aligned}
$$

Substituting back into the numerator and factoring

$$
\begin{aligned}
&= \left(\frac{N_1(N-N_1)}{NT}\right) \times \left(\frac{(T-1)T(T+1) - 3T(g-1)(T-g+1)}{(T-1)T(T+1)}\right) \times \sum_{t=g}^{T}\sum_{t'=1}^{g-1}\Bigg\{(\bar{Y}_{1,t}-\bar{Y}_{0,t}-\bar{Y}_{1,t'}+\bar{Y}_{0,t'}) - \\
&\qquad \left[\frac{\sum_{t_1=1}^{g-1}\sum_{t_2=t_1}^{g-1}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2}-\bar{Y}_{1,t_1}+\bar{Y}_{0,t_1}) + \sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T}(\bar{Y}_{1,t_2}-\bar{Y}_{0,t_2}-\bar{Y}_{1,t_1}+\bar{Y}_{0,t_1})}{\sum_{t_1=1}^{g-1}\sum_{t_2=t_1}^{g-1}t_2-t_1 + \sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T}t_2-t_1}\right](t-t')\Bigg\}
\end{aligned}
$$

Next, simplify the denominator

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{D}_{it})^2 = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(D_i - \frac{N_1}{N}\right)^2\left(\mathbf{1}(t \geq g) - \frac{T-g+1}{T}\right)^2 + \left(\frac{3(g-1)(T-g+1)}{T(T-1)}\right)\sum_{i=1}^{N}\sum_{t=1}^{T}\left(D_i - \frac{N_1}{N}\right)^2\left(1 - \frac{2t}{T+1}\right)^2 + 2\left(\frac{3(g-1)(T-g+1)}{T(T-1)}\right)\sum_{i=1}^{N}\sum_{t=1}^{T}\left(D_i - \frac{N_1}{N}\right)^2\left(\mathbf{1}(t \geq g) - \frac{T-g+1}{T}\right)\left(1 - \frac{2t}{T+1}\right)$$

With some algebra, it can be shown that:

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\left(D_i - \frac{N_1}{N}\right)^2\left(\mathbf{1}(t \geq g) - \frac{T-g+1}{T}\right)^2 = \frac{N_1(N-N_1)(g-1)(T-g+1)}{NT}$$

$$\left(\frac{3(g-1)(T-g+1)}{T(T-1)}\right)\sum_{i=1}^{N}\sum_{t=1}^{T}\left(D_i - \frac{N_1}{N}\right)^2\left(1 - \frac{2t}{T+1}\right)^2 = \left[\frac{N_1(N-N_1)(g-1)(T-g+1)}{NT}\right]\left[\frac{3(g-1)(T-g+1)}{(T+1)(T-1)}\right]$$

$$2\left(\frac{3(g-1)(T-g+1)}{T(T-1)}\right)\sum_{i=1}^{N}\sum_{t=1}^{T}\left(D_i - \frac{N_1}{N}\right)^2\left(\mathbf{1}(t \geq g) - \frac{T-g+1}{T}\right)\left(1 - \frac{2t}{T+1}\right) = -2\left[\frac{N_1(N-N_1)(g-1)(T-g+1)}{NT}\right]\left[\frac{3(g-1)(T-g+1)}{(T+1)(T-1)}\right]$$

Combining, yields an expression for the denominator that partly cancels with the numerator

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{D}_{it})^2 = \left(\frac{N_1(N-N_1)}{NT}\right) \times \left(\frac{(T-1)T(T+1) - 3T(g-1)(T-g+1)}{(T-1)T(T+1)}\right) \times (g-1)(T-g+1)$$

Substituting the numerator and denominator back into the FWL formula for $\hat{\tau}$ yields the final

expression of the regression coefficient as an average over every triple-difference.

$$\hat{\tau} = \frac{1}{(g-1)(T-g+1)} \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \Bigg\{ (\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'}) - \left[\frac{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1}) + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})}{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \right] (t - t') \Bigg\}$$

A.3 Proof of Proposition 3

From 2,

$$\hat{\tau} = \frac{1}{(g-1)(T-g+1)} \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \Bigg\{ (\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'}) - \left[\frac{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1}) + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} (\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1})}{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \right] (t - t') \Bigg\}$$

Taking the expectation conditional on the design $\mathcal{D}$

$$E[\hat{\tau}|\mathcal{D}] = \frac{1}{(g-1)(T-g+1)} \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \Bigg\{ E[\bar{Y}_{1,t} - \bar{Y}_{0,t} - \bar{Y}_{1,t'} + \bar{Y}_{0,t'}|\mathcal{D}] - \left[\frac{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} E[\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1}|\mathcal{D}] + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} E[\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1}|\mathcal{D}]}{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \right] (t - t') \Bigg\}$$

Under the no anticipation assumption (Assumption 3) and the parallel trends-in-trends assumption (Assumption 4), the results in Proposition 1 show that each difference-in-differences between t and t' identifies the $ATT(t)$ plus a bias term.

$$E[\hat{\tau}|\mathcal{D}] = \frac{1}{(g-1)(T-g+1)} \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \Bigg\{ \Big[ATT(t) + \Delta(t - t') \Big] - \left[\frac{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} E[\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1}|\mathcal{D}] + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} E[\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1}|\mathcal{D}]}{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \right] (t - t') \Bigg\}$$

The second difference-in-differences between $t_1 < g$ and $t_2 < g$ – both pre-treatment periods

– identifies the parallel trends violation Δ under Assumptions 3 and 4. However, this is scaled to the difference between t_2 and t_1 rather than t and t'. Again, this follows from the results in Proposition 1.

$$E[\hat{\tau}|\mathcal{D}] = \frac{1}{(g-1)(T-g+1)} \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \Bigg\{ \Bigg[ATT(t) + \Delta(t-t') \Bigg] - \Bigg[\frac{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} \Delta(t_2 - t_1) + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} E[\bar{Y}_{1,t_2} - \bar{Y}_{0,t_2} - \bar{Y}_{1,t_1} + \bar{Y}_{0,t_1} | \mathcal{D}]}{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \Bigg] (t-t') \Bigg\}$$

However, for post-treatment periods $t_1 \geq g$, $t_2 \geq g$, there is no guarantee that the ATT is zero from Assumption 3. Therefore, this "placebo" difference-in-difference identifies both the parallel trends violation Δ and a difference in treatment effects at t_2 and t_1.

$$E[\hat{\tau}|\mathcal{D}] = \frac{1}{(g-1)(T-g+1)} \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \Bigg\{ \Bigg[ATT(t) + \Delta(t-t') \Bigg] - \Bigg[\frac{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} \Delta(t_2 - t_1) + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} ATT(t_2) - ATT(t_1) + \Delta(t_2 - t_1)}{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \Bigg] (t-t') \Bigg\}$$

Re-arranging terms

$$\begin{aligned} E[\hat{\tau}|\mathcal{D}] = \frac{1}{(g-1)(T-g+1)} \sum_{t=g}^{T} \sum_{t'=1}^{g-1} \Bigg\{ & \Bigg[ATT(t) + \Delta(t-t') \Bigg] \\ & - \Bigg[\frac{\Delta(t-t') \Big[\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} (t_2 - t_1) + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} (t_2 - t_1) \Big]}{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \Bigg] \\ & - \Bigg[(t-t') \frac{\sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} ATT(t_2) - ATT(t_1)}{\sum_{t_1=1}^{g-1} \sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T} \sum_{t_2=t_1}^{T} t_2 - t_1} \Bigg] \Bigg\} \end{aligned}$$

Simplifying yields a uniform average over post-treatment $ATT(t)$ and a bias term

$$E[\hat{\tau}|\mathcal{D}] = \frac{1}{(T-g+1)}\sum_{t=g}^{T} ATT(t) - \frac{1}{(g-1)(T-g+1)}\sum_{t=g}^{T}\sum_{t'=1}^{g-1}\left[(t-t')\frac{\sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T} ATT(t_2) - ATT(t_1)}{\sum_{t_1=1}^{g-1}\sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T} t_2 - t_1}\right]$$

Under the assumption of constant effects over time (Assumption 5) $ATT(t_2) = ATT(t_1)$ for $t_2 \geq g$, $t_1 \geq g$. Therefore, the bias term is equal to zero and $\hat{\tau}$ identifies the ATT

$$E[\hat{\tau}|\mathcal{D}] = \frac{1}{(T-g+1)}\sum_{t=g}^{T} ATT(t) = ATT$$

However, in the absence of a constant effects assumption, $\hat{\tau}$ is no longer a convex average over treatment effects. Summing over $t - t'$

$$\begin{aligned} E[\hat{\tau}|\mathcal{D}] = \frac{1}{(T-g+1)}\sum_{t=g}^{T} ATT(t) \\ - \frac{\sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T} ATT(t_2)}{(g-1)(T-g+1)}\left[\frac{\sum_{t=g}^{T}\sum_{t'=1}^{g-1}(t-t')}{\sum_{t_1=1}^{g-1}\sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T} t_2 - t_1}\right] \\ + \frac{\sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T} ATT(t_1)}{(g-1)(T-g+1)}\left[\frac{\sum_{t=g}^{T}\sum_{t'=1}^{g-1}(t-t')}{\sum_{t_1=1}^{g-1}\sum_{t_2=t_1}^{g-1} t_2 - t_1 + \sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T} t_2 - t_1}\right] \end{aligned}$$

Using Faulhaber's formula

$$\begin{aligned} E[\hat{\tau}|\mathcal{D}] = \frac{1}{(T-g+1)}\sum_{t=g}^{T} ATT(t) \\ - \frac{\sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T} ATT(t_2)}{(g-1)(T-g+1)}\left[\frac{3T(g-1)(T-g+1)}{(T-1)T(T+1) - 3T(g-1)(T-g+1)}\right] \\ + \frac{\sum_{t_1=g}^{T}\sum_{t_2=t_1}^{T} ATT(t_1)}{(g-1)(T-g+1)}\left[\frac{3T(g-1)(T-g+1)}{(T-1)T(T+1) - 3T(g-1)(T-g+1)}\right] \end{aligned}$$

Summing over t_1 and t_2 yields three sums over the post-treatment $ATT(t)$ from g to T.

$$\begin{aligned}
E[\hat{\tau}|\mathcal{D}] = \frac{1}{(T-g+1)} \sum_{t=g}^{T} ATT(t) \\
&- \sum_{t=g}^{T} ATT(t)(t-g+1)\left[\frac{3}{(T-1)(T+1)-3(g-1)(T-g+1)}\right] \\
&+ \sum_{t=g}^{T} ATT(t)(T-t+1)\left[\frac{3}{(T-1)(T+1)-3(g-1)(T-g+1)}\right]
\end{aligned}$$

Combining terms yields a weighted average of ATTs.

$$\begin{aligned}
E[\hat{\tau}|\mathcal{D}] &= \frac{1}{(T-g+1)} \sum_{t=g}^{T} ATT(t)\left[1 + \frac{3(T-g+1)(T+g-2t)}{(T-1)(T+1)-3(g-1)(T-g+1)}\right] \\
&= \sum_{t=g}^{T} ATT(t)\left[\frac{(T-1)(T+1)+3(T-g+1)(T-2t+1)}{(T-g+1)(T-1)(T+1)-3(g-1)(T-g+1)^2}\right]
\end{aligned}$$

The weights sum to 1 but are not guaranteed to be non-negative.

$$\begin{aligned}
&\sum_{t=g}^{T} \frac{(T-1)(T+1)+3(T-g+1)(T-2t+1)}{(T-g+1)(T-1)(T+1)-3(g-1)(T-g+1)^2} = \\
&\qquad \frac{(T-1)(T+1)+3(T-g+1)\sum_{t=g}^{T}(T-2t+1)}{(T-g+1)(T-1)(T+1)-3(g-1)(T-g+1)^2} \\
&= \frac{(T-1)(T+1)(T-g+1)+3(T-g+1)^2(1-g)}{(T-g+1)(T-1)(T+1)-3(g-1)(T-g+1)^2} \\
&= \frac{(T-1)(T+1)(T-g+1)-3(T-g+1)^2(g-1)}{(T-g+1)(T-1)(T+1)-3(g-1)(T-g+1)^2} \\
&= 1
\end{aligned}$$