

July 2025

“Robustness of Two-Way Fixed Effects Estimators to
Heterogeneous Treatment Effects”

Anaïs Fabre

Robustness of Two-Way Fixed Effects Estimators to Heterogeneous Treatment Effects*

Anaïs Fabre[†]

July 16, 2025

Abstract

This paper studies the Two-Way Fixed Effects (TWFE) estimator in a general setting where multiple groups can enter and exit a binary treatment over time. It establishes necessary and sufficient conditions for this estimator to correspond to a convex combination of Average Treatment Effects (ATEs). I show that the TWFE estimator is a weighted sum of five different types of two-by-two comparisons, with positive weights. Parallel trends assumptions on either the untreated or treated potential outcomes must hold for each comparison to identify the ATE of the group switching treatment status. When treatment effects are contemporaneous but can be heterogeneous across groups and over time, both parallel trends assumptions are thus necessary and sufficient for the TWFE estimator to be a weighted sum of ATEs, with positive weights. Under parallel trends on the untreated potential outcomes and on the first exposure to treatment, the presence of dynamic treatment effects is necessary — but not sufficient — for this result to break.

*This paper was previously circulated under the title ‘A Note on Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects’. I am very grateful for the guidance and helpful comments from my advisors Olivier De Groote and Thierry Magnac, as well as from Matteo Bobba, Hippolyte Boucher, Sylvain Chabé-Ferret, Clément de Chaisemartin, Xavier d’Haultfoeuille, Tim Ederer, Koen Jochmans, Nour Meddahi, and workshop participants at the Toulouse School of Economics. I acknowledge funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d’Avenir) program, grant ANR-17-EURE-0010 and grant MATCHINEQ - ANR-22-CE26-0005-01.

[†]Institute for Fiscal Studies: anaïs.fabre@ifs.org.uk.

1 Introduction

The difference-in-differences strategy is one of the most popular quasi-experimental methods to estimate the effect of a policy. Its core idea is to compare the evolution of the outcome of interest before and after a group receives a treatment to the one of a group which does not receive it. With two groups and two periods, such a comparison is an unbiased estimator of the Average Treatment Effect on the Treated (ATT). However, most empirical applications depart from this simple framework, and instead leverage the variation in treatment exposure across multiple groups and time periods. With several groups and periods, researchers will typically interpret as the ATT the parameter of the treatment dummy in a Two-Way Fixed Effects (TWFE) regression, also including group and time fixed effects.

Despite its popularity, the properties of this estimator remained under-studied until recently. The literature has now concluded that, under the standard common trends assumption, the TWFE estimator corresponds to a weighted sum of the Average Treatment Effects (ATEs) of each group and period, with weights that may be negative in the presence of heterogeneous treatment effects (de Chaisemartin and D’Haultfoeuille, 2023b). The TWFE estimator may then be negative even when all ATEs are positive.

While heterogeneous treatment effects are understood as a source of these potential negative weights, the necessary and sufficient conditions for the TWFE estimator to be a convex combination of ATEs have not been formally derived. This paper fills this gap. Contrasting with the recent literature (de Chaisemartin and D’Haultfoeuille, 2020; Goodman-Bacon, 2021), it establishes this result in a general setting where treatment effects are not restricted to be homogeneous across groups nor over time. When treatment effects are contemporaneous, I show that both the well-known parallel trends assumption on the *untreated* potential outcomes, but also a parallel trends assumption on the *treated* potential outcomes are necessary and sufficient conditions for the TWFE estimator to always estimate a weighted sum of ATEs, all entering with positive weights. Homogeneous treatment effects across groups or over time are sufficient for this result to hold under the standard parallel trends assumption on the *untreated* potential outcomes, but not necessary. Extending the framework to allow for non-contemporaneous treatment effects, I further show that when common trends on the first-exposure and never-treated potential outcomes hold, the absence of dynamic treatment effects is sufficient, but not necessary, for the TWFE estimator to correspond to a convex combination of ATEs.

To derive these results, I consider a general framework where groups are allowed to enter and exit a binary treatment over time. It is thus not restricted to the staggered case, which has received a lot of attention in spite of relatively few applications (de Chaisemartin and D’Haultfoeuille, 2020). I first show that the TWFE estimator is a weighted sum of five different types of two-by-two difference-in-differences comparisons. It may include (i) standard, (ii) reverse, (iii) leaver, (iv) reverse leaver, and (v) double-

switcher difference-in-differences. The four first comparisons contrast a group which switches treatment status (enters or leaves treatment) with a group keeping the same treatment status (either treated or untreated). Comparisons (v) contrast the evolution of outcomes of a group joining treatment to a group leaving it.

In a framework where the effect of the treatment can only be contemporaneous, I first establish that each of these comparisons is an unbiased estimator of the ATE of the group switching treatment status if and only if a parallel trends assumption either on the *untreated* or *treated* potential outcomes holds. These intermediary results broaden the panel of comparisons available to researchers to identify causal effects. In particular, I show how the double-switcher difference-in-differences - comparing, across two time periods, a group switching from being treated to untreated, to a group switching from being untreated to treated - can allow to identify treatment effects of interest under the standard parallel trends assumption.

I then establish the implication of these results for the TWFE estimator: it is a weighted sum of ATEs of different groups and periods, all weighted positively. Each ATE enters proportionally to the number of comparison groups available to identify it. Necessary and sufficient conditions for this result are surprising: the trends in potential outcomes both when *untreated* and *treated* should evolve similarly across groups. Together, these common trends assumptions imply that ATEs should follow the same evolution over time across groups, but do not impose homogeneity. Importantly, the decomposition provided includes reverse difference-in-differences, which have been understood as ‘forbidden comparisons’ at the origin of potential negative weights (Borusyak et al., 2024). This type of comparisons is typically a biased estimator of the ATE for the group switching treatment status under the standard parallel trends assumption on *untreated* potential outcomes. However, they become valid if *treated* potential outcomes evolve similarly across groups.

Finally, I explore whether this result is robust to the presence of dynamic treatment effects. I provide a decomposition of the TWFE estimator in a framework allowing for treatment effects to be non-contemporaneous and provide necessary and sufficient conditions for the TWFE estimator to be a convex combination of ATEs in this setting. I find that when outcomes when both untreated and first-exposed to treatment evolve similarly across groups, dynamic treatment effects are necessary, although not sufficient, for the decomposition of the TWFE estimator as a weighted sum of ATEs to break. Overall, it is important to note that the conditions under which the TWFE estimator always identifies a convex combination of ATEs remain strong. This paper does not advocate for its use in applied work, but rather aims at improving our understanding of this widely adopted estimator.

Related Literature The difference-in-differences literature has recently put the TWFE estimator under increased scrutiny. de Chaisemartin and D’Haultfoeuille (2020) study a general framework with several

groups and time periods where groups can enter and leave treatment. They conclude that the TWFE estimator is a weighted sum of ATEs in each group and period, with weights that may be negative when ATEs are heterogeneous over time or across groups. [Borusyak et al. \(2024\)](#) reach the same conclusion in the staggered case, where groups cannot exit treatment.

To shed light on the origin of such negative weights, both [Goodman-Bacon \(2021\)](#) and [Strezhnev \(2018\)](#) provide decompositions of the TWFE estimator in terms of two-by-two difference-in-differences comparisons. Focusing on the staggered setting, [Goodman-Bacon \(2021\)](#) establishes that negative weights come from the comparisons of late treated units with already-treated units, which are biased estimators of the corresponding ATE in the presence of heterogeneous treatment effects over time. These comparisons are thus understood as ‘forbidden comparisons’. The paper concludes that constant treatment effects are then necessary for the TWFE estimator to identify a convex combination of ATEs. Similarly, [Borusyak et al. \(2024\)](#) conclude that ‘these comparisons are only valid when the homogeneity assumption is true’. While [Strezhnev \(2018\)](#) expands the TWFE decomposition to the general case where groups can leave treatment, he then focuses on the staggered case and reaches the same conclusion.

The novel identification result presented in this paper builds on these decompositions and qualifies these claims. By carefully deriving the conditions which are both necessary and sufficient for the TWFE estimator to correspond to a convex combination of ATEs, I show that the so-called ‘forbidden comparisons’ are actually valid if the potential outcomes under treatment evolve similarly across groups. Importantly, and in contrast to the general conclusion of the above papers, this highlights that homogeneous treatment effects across groups or over time are sufficient for the TWFE to be valid under the standard common trends assumption on the untreated potential outcomes, but not necessary.

To derive this result, this paper also contributes to expanding the tool box of the applied economist, by establishing the assumptions under which each of the five types of two-by-two difference-in-differences comparisons which enter the TWFE estimator is an unbiased estimator of its corresponding ATE. In particular, it shows that treatment effects of interest can be identified from the double-switcher difference-in-differences.

It should be noted that I do not consider cases with non-binary, continuous or several treatments ([de Chaisemartin and D’Haultfoeuille, 2018,0](#); [Callaway et al., 2021](#)) nor regressions with control variables. [de Chaisemartin and D’Haultfoeuille \(2023b\)](#) and [Roth et al. \(2023\)](#) provide rich reviews of this recent literature.

2 Framework

I consider a framework with G groups and T periods.¹ Let us denote $D_{g,t}$ the binary treatment status of group g at period t , and $(Y_{g,t}(0), Y_{g,t}(1))$ the potential outcomes when untreated and when treated, respectively. The observed outcome of group g in period t is denoted $Y_{g,t}(D_{g,t})$. Ω collects the potential outcomes and treatment status for all g and t , i.e. $\Omega = \{Y_{g,t}(0), Y_{g,t}(1), D_{g,t}\}_{\forall g,t}$.

I also impose the same assumptions as [de Chaisemartin and D’Haultfoeuille \(2020\)](#). Discussions of these assumptions can be found in their paper.

Assumption 1. (*Balanced Panel of Groups*): $\forall (g, t) \in \{1, \dots, G\} \times \{1, \dots, T\}, N_{g,t} > 0$.

Assumption 2. (*Independent Groups*): *The vectors $(Y_{g,t}(0), Y_{g,t}(1), D_{g,t})_{1 \leq t \leq T}$ are mutually independent.*

Assumption 3. (*Strong Exogeneity*): $\forall (g, t, d) \in \{1, \dots, G\} \times \{2, \dots, T\} \times \{0, 1\}$,
 $E(Y_{g,t}(d) - Y_{g,t-1}(d) | D_{g,1}, \dots, D_{g,T}) = E(Y_{g,t}(d) - Y_{g,t-1}(d))$.

These assumptions impose very few restrictions on Ω , and in particular do not restrict treatment effects to be homogeneous across groups, nor over time. It however rules out dynamic treatment effects, i.e. treatment effects cannot depend on the treatment history of the group. I relax this assumption in [Section 4.3](#).

Finally, the ATE of group g in period t , $\Delta_{g,t}$, writes:

$$\Delta_{g,t} = Y_{g,t}(1) - Y_{g,t}(0)$$

We consider the following TWFE regression:

$$Y_{g,t} = \alpha_g + \gamma_t + \beta_{fe} D_{g,t} + \epsilon_{g,t} \tag{1}$$

We let $\hat{\beta}_{fe}$ denote the OLS estimator of the coefficient of $D_{g,t}$, with $\beta_{fe} = E[\hat{\beta}_{fe}]$. Several recent studies have demonstrated that, under [Assumptions 1-3](#) and a common trends assumption on *untreated* potential outcomes, β_{fe} is equal to the expectation of a weighted sum of $\Delta_{g,t}$, with potentially negative weights when ATEs are heterogeneous across groups or over time (see [de Chaisemartin and D’Haultfoeuille \(2023b\)](#) for a survey). This result implies that β_{fe} may be negative even if all ATEs are positive. The general intuition is that negative weights arise because the TWFE estimator includes so-called ‘forbidden comparisons’, comparing the outcome evolution of some groups to invalid control groups, such as always-treated units. In the next sections, I revisit these results and provide necessary and sufficient conditions for the TWFE

¹Groups can, for example, correspond to geographical locations or schools, but they may also consist of a single individual.

estimator to weigh all ATEs positively, while allowing ATEs to be heterogeneous across groups and over time.

3 A General Decomposition of the TWFE Estimator

What are the necessary and sufficient conditions for the TWFE estimator to weigh all ATEs positively? To answer this question, a crucial first step is to decompose the TWFE estimator as a weighted sum of standard two-by-two difference-in-differences comparisons. Such a decomposition will make it straightforward to understand what each difference estimates, and under which assumptions. In particular, I show that $\hat{\beta}_{fe}$ is a weighted sum of five different types of two-by-two comparisons. Let $\hat{\beta}_{g,k,t,t'}^{DD}$ denote the two-by-two comparison of the outcomes of group g and k between periods t and t' , $\forall g \in \{1, \dots, G\}, \forall k \in \{1, \dots, G\} \setminus g, \forall t \in \{1, \dots, T\}, \forall t' \in \{1, \dots, T\} \setminus t$:

$$\hat{\beta}_{g,k,t,t'}^{DD} = (Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'})$$

We can define five types of two-by-two comparisons, depending on the direction of treatment status change within the switching group and the treatment status of the comparison group:

Standard Difference-in-Differences, $\hat{\beta}_{g,k,t,t'}^S$:

$$\hat{\beta}_{g,k,t,t'}^S = \hat{\beta}_{g,k,t,t'}^{DD} \times \underbrace{\mathbb{1}\{D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 0, t > t'\}}_{\equiv \omega_{g,k,t,t'}^S} \quad (2)$$

Equation (2) corresponds to the standard difference-in-differences comparison, where the evolution of the outcome of a group, g , which becomes treated between period t' and t is compared to the one of a group, k , which is untreated in both periods.

Reverse Difference-in-Differences, $\hat{\beta}_{g,k,t,t'}^R$:

$$\hat{\beta}_{g,k,t,t'}^R = \hat{\beta}_{g,k,t,t'}^{DD} \times \underbrace{\mathbb{1}\{D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 1, D_{k,t'} = 1, t > t'\}}_{\equiv \omega_{g,k,t,t'}^R} \quad (3)$$

Equation (3) describes the reverse difference-in-differences comparison.² It contrasts the evolution of the outcome of a group, g , which becomes treated between period t' and t with the one of a group, k , which is treated in both periods.

²Several studies have used such an empirical strategy, such as Rossi and Villar (2020) and Chabé-Ferret and Voia (2021).

Leaver Difference-in-Differences, $\hat{\beta}_{g,k,t,t'}^L$:

$$\hat{\beta}_{g,k,t,t'}^L = \hat{\beta}_{g,k,t,t'}^{DD} \times \underbrace{\mathbb{1}\{D_{g,t} = 1, D_{g,t'} = 1, D_{k,t} = 0, D_{k,t'} = 1, t > t'\}}_{\equiv \omega_{g,k,t,t'}^L} \quad (4)$$

Equation (4) describes the leaver difference-in-differences comparison. It compares the evolution of the outcome of a group, g , between period t' and t , which is treated in both periods with the one of a group, k , which is treated in period t' but has left treatment in period t . Units which remain treated are thus used as a control group for units leaving treatment. This comparison does not exist in a staggered design, where groups cannot leave treatment.

Reverse Leaver Difference-in-Differences, $\hat{\beta}_{g,k,t,t'}^{RL}$:

$$\hat{\beta}_{g,k,t,t'}^{RL} = \hat{\beta}_{g,k,t,t'}^{DD} \times \underbrace{\mathbb{1}\{D_{g,t} = 0, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 1, t > t'\}}_{\equiv \omega_{g,k,t,t'}^{RL}} \quad (5)$$

Equation (5) describes the reverse leaver difference-in-differences comparison. It compares the evolution of the outcome of a group, g , between period t' and t , which is treated in neither of the two periods, with the one of a group which is treated in period t' but has left treatment in period t . Similarly, this comparison does not appear in staggered designs.

Double-Switcher Difference-in-Differences, $\hat{\beta}_{g,k,t,t'}^{DS}$:

$$\hat{\beta}_{g,k,t,t'}^{DS} = \hat{\beta}_{g,k,t,t'}^{DD} \times \underbrace{\mathbb{1}\{D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 1, t > t'\}}_{\equiv \omega_{g,k,t,t'}^{DS}} \quad (6)$$

Equation (6) introduces a two-by-two comparison which has received seldom attention in the literature, the double-switcher difference-in-differences. It compares the evolution of outcomes of two groups: group g is not treated in period t' and joins treatment in period t , while group k is treated in period t' but leaves treatment in period t .

We can now rewrite the TWFE estimator as a sum, with positive weights, of these five different types of two-by-two comparisons.

Theorem 1. $\hat{\beta}_{fe}$ is a weighted sum of five different types of two-by-two difference-in-differences:

$$\hat{\beta}_{fe} = \frac{\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [\hat{\beta}_{g,k,t,t'}^S + \hat{\beta}_{k,g,t',t}^R + \hat{\beta}_{g,k,t,t'}^L + \hat{\beta}_{k,g,t',t}^{RL} + 2\hat{\beta}_{g,k,t,t'}^{DS}]}{\sum_t \sum_{g: D_{g,t}=1} \sum_{k: D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]}$$

where, for $g \neq k$, $t \neq t'$ and $c \in \{S, L, R, RL, DS\}$:

$$\hat{\beta}_{g,k,t,t'}^c = \omega_{g,k,t,t'}^c \hat{\beta}_{g,k,t,t'}^{DD}$$

Details of the proof, building on a result from [Strezhnev \(2018\)](#), are provided in [Appendix A.1](#). The above decomposition shows that, in the general case, the TWFE estimator includes five types of two-by-two difference-in-differences comparisons. Each weight simply corresponds to a dummy equal to one each time a relevant comparison group exists. Next section clarifies what each of these comparisons identifies, and under which assumptions.

4 Identification

[Section 4.1](#) studies conditions under which each of the five two-by-two comparisons defined above identifies the ATE of the group switching treatment status. [Section 4.2](#) highlights the implication of these results for the TWFE estimator. Parallel trends assumptions on both *untreated* and *treated* potential outcomes are necessary and sufficient for the latter to be a weighted sum of ATEs, with positive weights. Finally, [Section 4.3](#) explores whether the latter result can hold in a framework allowing for dynamic treatment effects.

4.1 The Five Difference-in-Differences Comparisons: Identification

Let us first study separately each of the two-by-two difference-in-differences comparisons that may be included in the TWFE estimator. In what follows, I consider the set of Ω such that [Assumptions 1-3](#) hold.

Standard Difference-in-Differences, $\hat{\beta}_{g,k,t,t'}^S$: Consider two groups g and k , two periods t, t' , such that $t > t'$, $D_{g,t} = 1$, $D_{g,t'} = 0$, $D_{k,t} = 0$ and $D_{k,t'} = 0$. We are thus in the standard case where we observe a group which remains untreated between period t' and t , and a group which becomes treated. It is well-established that the following common trends assumption is required for the standard difference-in-differences estimator to always identify the ATT, i.e. the ATE of group g at time t .

Assumption 4. (*Common Trends of the Potential Outcome Without Treatment*): For $t \geq 2$, $E(Y_{g,t}(0) - Y_{g,t-1}(0))$ does not vary across g .

We can write:

$$\text{Assumption 4 holds if and only if } \forall g, k \neq g, t, t' \neq t, \forall \Omega, E[\hat{\beta}_{g,k,t,t'}^S | \mathbf{D}] = E[\Delta_{g,t} | \mathbf{D}] \times \omega_{g,k,t,t'}^S.$$

where \mathbf{D} is the vector of treatment status history for every group. Details can be found in Appendix A.2.

Reverse Difference-in-Differences, $\hat{\beta}_{g,k,t,t'}^R$: Consider two groups g and k , two periods t, t' , such that $t > t'$, $D_{g,t} = 1$, $D_{g,t'} = 0$, $D_{k,t} = 1$ and $D_{k,t'} = 1$. We are in a case where always-treated units are used as a control group for late-treated units. These comparisons are shown to be at the origin of negative weights in the TWFE estimator (Goodman-Bacon, 2021), and are presented as ‘forbidden comparisons’. In particular, under Assumption 4, we have:

$$E[\hat{\beta}_{g,k,t,t'}^R | \mathbf{D}] = E[\Delta_{g,t} - (\Delta_{k,t} - \Delta_{k,t'}) | \mathbf{D}] \times \omega_{g,k,t,t'}^R$$

Thus, if $E[\Delta_{k,t} | \mathbf{D}] \neq E[\Delta_{k,t'} | \mathbf{D}]$ and $E[\Delta_{k,t} | \mathbf{D}] \neq E[\Delta_{g,t} | \mathbf{D}]$, i.e. in the presence of heterogeneous treatment effects over time and across groups, these comparisons do not identify an ATE.³ This is why they are typically understood as forbidden.

Yet, Kim and Lee (2019) show that, under a common trends assumption on the potential outcomes under treatment, these two-by-two comparisons are unbiased estimators of the ATE of group g in the pre-treatment period, t' . We thus consider the following assumption:

Assumption 5. (*Common Trends of the Potential Outcome With Treatment*): For $t \geq 2$, $E(Y_{g,t}(1) - Y_{g,t-1}(1))$ does not vary across g .

In particular, adding and subtracting $E[Y_{g,t'}(1) | D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 1, D_{k,t'} = 1]$, we have:

$$\begin{aligned} & E[\hat{\beta}_{g,k,t,t'}^{DD} | D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 1, D_{k,t'} = 1] \\ &= E[\Delta_{g,t'} + (Y_{g,t}(1) - Y_{g,t'}(1) - (Y_{k,t}(1) - Y_{k,t'}(1))) | D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 1, D_{k,t'} = 1] \end{aligned}$$

In a framework allowing for heterogeneous treatment effects, Assumption 5 is thus both necessary and sufficient for the reverse difference-in-differences to always identify the ATE of the switching group in period t' . We can write:

$$\text{Assumption 5 holds if and only if } \forall g, k \neq g, t, t' \neq t, \forall \mathbf{D}, E[\hat{\beta}_{g,k,t,t'}^R | \mathbf{D}] = E[\Delta_{g,t'} | \mathbf{D}] \times \omega_{g,k,t,t'}^R.$$

³Note that this qualifies the statement of Goodman-Bacon (2021) who notes that, when effects vary over time but not across units, time-varying effects bias estimates away from their corresponding ATE. However, in a static framework, when treatment effects are homogeneous across groups in a given time period but not over time, we would also have an unbiased estimator of a relevant ATE:

$$E[\hat{\beta}_{g,k,t,t'}^R | \mathbf{D}] = E[\Delta_{k,t'} | \mathbf{D}] \times \omega_{g,k,t,t'}^R = E[\Delta_{g,t'} | \mathbf{D}] \times \omega_{g,k,t,t'}^R$$

Thus, under a common trends assumption on the untreated potential outcomes only, reverse difference-in-differences are forbidden comparisons if treatment effects vary over time *and* across groups, and not merely over time.

Overall, these comparisons should not be systematically understood as forbidden when ATEs are allowed to be heterogeneous: they require a different common trends assumption to be valid.

Leaver Difference-in-Differences, $\hat{\beta}_{g,k,t,t'}^L$: Consider two groups g and k , two periods t, t' , such that $t > t'$, $D_{g,t} = 1, D_{g,t'} = 1, D_{k,t} = 0$ and $D_{k,t'} = 1$. We are in a case where always-treated units are used as a control group for a group which is initially treated, and leaves treatment in period t . Taking the expectation and adding and subtracting $E(Y_{k,t}(1)|D_{g,t} = 1, D_{g,t'} = 1, D_{k,t} = 0, D_{k,t'} = 1)$, we find:

$$\begin{aligned} & E[\hat{\beta}_{g,k,t,t'}^{DD} | D_{g,t} = 1, D_{g,t'} = 1, D_{k,t} = 0, D_{k,t'} = 1] \\ &= E[\Delta_{k,t} + (Y_{g,t}(1) - Y_{g,t'}(1) - (Y_{k,t}(1) - Y_{k,t'}(1)))] | D_{g,t} = 1, D_{g,t'} = 1, D_{k,t} = 0, D_{k,t'} = 1 \end{aligned}$$

That is:

$$\text{Assumption 5 holds if and only if } \forall g, k \neq g, t, t' \neq t, \forall \mathbf{\Omega}, E[\hat{\beta}_{g,k,t,t'}^L | \mathbf{D}] = E[\Delta_{k,t} | \mathbf{D}] \times \omega_{g,k,t,t'}^L.^4$$

Reverse Leaver Difference-in-Differences, $\hat{\beta}_{g,k,t,t'}^{RL}$: Consider two groups g and k , two periods t, t' , such that $t > t'$, $D_{g,t} = 0, D_{g,t'} = 0, D_{k,t} = 0$ and $D_{k,t'} = 1$. In this case, never-treated units are used as a control group for a group which is initially treated, and leaves treatment in period t . Taking the expectation and adding and subtracting $E(Y_{k,t'}(0)|D_{g,t} = 0, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 1)$, we can show that this comparison identifies the ATE of group k in period t' under Assumption 4:

$$\begin{aligned} & E[\hat{\beta}_{g,k,t,t'}^{DD} | D_{g,t} = 0, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 1] \\ &= E[\Delta_{k,t'} + (Y_{g,t}(0) - Y_{g,t'}(0) - (Y_{k,t}(0) - Y_{k,t'}(0)))] | D_{g,t} = 0, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 1 \end{aligned}$$

We thus have:

$$\text{Assumption 4 holds if and only if } \forall g, k \neq g, t, t' \neq t, \forall \mathbf{\Omega}, E[\hat{\beta}_{g,k,t,t'}^{RL} | \mathbf{D}] = E[\Delta_{k,t'} | \mathbf{D}] \times \omega_{g,k,t,t'}^{RL}.$$

Double-Switcher Difference-in-Differences, $\hat{\beta}_{g,k,t,t'}^{DS}$: Consider two groups g and k , two periods t, t' , such that $t > t'$, $D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 0$ and $D_{k,t'} = 1$. We are in a case where we compare the outcomes of two groups which treatment status change over time in different directions. Group g is initially not treated, and becomes treated in period t . In contrast, group k is initially treated, and leaves treatment in period t . In this case, the sum of the ATE of group g in period t and of group k in period t' can be

⁴Note that the heterogeneity-robust estimator provided by de Chaisemartin and D'Haultfoeuille (2020) relies on Assumption 5 to identify the ATE of groups leaving treatment.

identified if the standard common trends assumption on potential outcomes when untreated holds:

$$\begin{aligned}
& E[\hat{\beta}_{g,k,t,t'}^{DD} | D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 1] \\
&= E[(Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'}) \\
&+ (Y_{g,t}(0) - Y_{g,t}(0)) + (Y_{k,t'}(0) - Y_{k,t'}(0)) | D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 1] \\
&= E[\Delta_{g,t} + \Delta_{k,t'} | D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 1]
\end{aligned}$$

Overall, we thus have:⁵

$$\textit{Assumption 4} \text{ if and only if } \forall g, k \neq g, t, t' \neq t, \forall \mathbf{\Omega}, E[\hat{\beta}_{g,k,t,t'}^{DS} | \mathbf{D}] = E[\Delta_{g,t} + \Delta_{k,t'} | \mathbf{D}] \times \omega_{g,k,t,t'}^{DS}.$$

Note that this result is of interest in itself. It implies that, in such a setting with two groups and two periods, if one is willing to assume that ATEs are homogeneous across time and across groups, a simple two-by-two difference-in-differences would allow to recover the ATT under the standard common trends assumption even in cases where we observe groups changing treatment status in opposite directions over time.⁶

On top of this, this result implies that one can recover an additional treatment effect which may be of interest in certain settings, even under heterogeneous treatment effects. For example, consider two groups and three periods, such that $D_{1,1} = D_{1,2} = 0$, $D_{1,3} = 1$, $D_{2,1} = 0$, $D_{2,2} = 1$, and $D_{2,3} = 0$. The observations of the two first periods can be used to recover $\Delta_{2,2}$ under Assumption 4. The information contained in the last period would usually be lost. Yet, comparing the changes in outcomes of the two groups in periods 2 and 3 allows to identify the sum of $\Delta_{2,2}$ and $\Delta_{1,3}$. Subtracting the first from the second comparison would thus allow to identify $\Delta_{1,3}$.

4.2 A General Decomposition Result

Section 3 shows that the TWFE estimator is a weighted sum of five different objects. Section 4.1 establishes the assumptions under which each of these objects is an unbiased estimator of its corresponding ATE. Overall, we obtain the following decomposition:

⁵Note that, alternatively, we can show:

$$\textit{Assumption 5} \text{ holds if and only if } \forall g, k \neq g, t, t' \neq t, \forall \mathbf{\Omega}, E[\hat{\beta}_{g,k,t,t'}^{DS} | \mathbf{D}] = E[\Delta_{g,t'} + \Delta_{k,t} | \mathbf{D}] \times \omega_{g,k,t,t'}^{DS}.$$

⁶Assuming that ATEs are homogeneous will not be necessary to show that under Assumptions 4 and 5 none of the ATEs that enter the TWFE estimator are weighted negatively.

Theorem 2.

$$E[\hat{\beta}_{fe}|\mathbf{D}] = \frac{\sum_t \sum_g [E[\Delta_{g,t}|\mathbf{D}] \sum_{t' \neq t} \sum_{k \neq g} [\omega_{g,k,t,t'}^S + \omega_{g,k,t',t}^R + \omega_{k,g,t,t'}^L + \omega_{k,g,t',t}^{RL} + 2\omega_{g,k,t,t'}^{DS} + 2\omega_{k,g,t',t}^{DS}]]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]}$$

$\forall \Omega$ such that Assumptions 1-3 hold, if and only if, Assumptions 4 and 5 hold.

Theorem 2 is obtained by taking the expectation of the expression of $\hat{\beta}_{fe}$ provided in Theorem 1, and plugging in each term derived in Section 4.1. Details of the proof are provided in Appendix A.2. Theorem 2 establishes that, in a general framework where there exist groups switching on and off treatment and where treatment effects are not restricted to be homogeneous across groups nor over time, common trends conditions on both treated and untreated potential outcomes are necessary and sufficient for the TWFE estimator to always identify a convex combination of ATEs.⁷ In particular, the TWFE estimator will never be negative when all ATEs are positive.

Assuming homogeneous treatment effects across groups or over time on top of the standard parallel trends assumption on the untreated potential outcome is a sufficient condition for this result to hold, as it implies that Assumption 5 holds as well.⁸ However, this result revisits previous statements made in the literature by showing that treatment effect homogeneity is not a necessary condition for the TWFE estimator to correspond to a convex combination of ATEs. Importantly, while common trends conditions on both treated and untreated potential outcomes imply that ATEs should evolve similarly across groups — that is, $E(\Delta_{g,t}) = \kappa_g + \gamma_t$ — they do not require that treatment effects are homogeneous across groups or over time. Note also that Assumptions 4 and 5 are both imposed by de Chaisemartin and D’Haultfoeuille (2020) for the estimator they propose as an alternative to the TWFE estimator to be unbiased: the above results show that Assumption 5 has important identifying content for the TWFE estimator as well.

The weights derived in Theorem 2 are very intuitive: they correspond to dummies equal to one each time a relevant comparison group, as defined by the five types of two-by-two comparisons in Section 3, exists. Each of the ATEs thus enters proportionally to the number of comparison groups available to identify it.⁹

⁷Note indeed that $\frac{\sum_t \sum_g \sum_{t' \neq t} \sum_{k \neq g} [\omega_{g,k,t,t'}^S + \omega_{g,k,t',t}^R + \omega_{k,g,t,t'}^L + \omega_{k,g,t',t}^{RL} + 2\omega_{g,k,t,t'}^{DS} + 2\omega_{k,g,t',t}^{DS}]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]} = 1$.

⁸Relatedly Theorem S2 in de Chaisemartin and D’Haultfoeuille (2020) shows in the staggered case that, under the standard parallel trends assumption and when treatment effects are stable across time post-treatment, the TWFE estimator does not weigh any ATEs negatively.

⁹Note that, if we restrict attention to a framework where there exist no groups entering nor leaving treatment while some remain treated, then conditions for the TWFE estimator not to weigh any ATEs negatively are weaker. In particular, only the common trends assumption on the untreated potential outcomes must hold. This, however, rules out the staggered difference-in-differences framework. When, in contrast, there exist no groups entering nor leaving treatment while some remain untreated, the common trends assumption on the untreated potential outcomes can be violated as long as its counterpart for treated potential outcomes holds.

4.3 Introducing Dynamic Treatment Effects

The above decomposition is derived in the case where treatment can only have a contemporaneous effect on the outcome. Let us now introduce the dynamic potential outcome framework, following [Robins \(1986\)](#) and [de Chaisemartin and D'Haultfoeulle \(2024\)](#). In particular, we denote $Y_{g,t}(\mathbf{d})$ the average potential outcome for group g in period t when facing the treatment sequence $\mathbf{d} \in \{0, 1\}^T$. The realized outcome of group g writes $Y_{g,t} = Y_{g,t}(D_{g,1}, D_{g,2}, \dots, D_{g,T}) = Y_{g,t}(\mathbf{D}_{g,T})$, where $\mathbf{D}_{g,t}$ is the vector of treatment status for group g up to period t . Now, $\mathbf{\Omega}$ is the vector $\{(Y_{g,t}(\mathbf{d}))_{\mathbf{d} \in \{0,1\}^T}, \mathbf{D}_{g,t}\}_{\forall g,t}$. We only maintain Assumption 1. Let us follow [de Chaisemartin and D'Haultfoeulle \(2024\)](#) and impose that a group's current outcome does not depend on its future treatment status, as well as a common trends assumption on the never-treated potential outcomes:

Assumption 6. For all g , for all $\mathbf{d} \in \{0, 1\}^T$, $Y_{g,t}(\mathbf{d}) = Y_{g,t}(d_1, \dots, d_t)$.

Assumption 7. $\forall t \geq 2, \forall g, E(Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1}) | \mathbf{D})$, where $\mathbf{0}_t$ corresponds to a vector of zeros of length t , does not vary across g .

Potential outcomes can now depend on past treatments: $Y_{g,t}(\mathbf{D}_{t-1}, d_t)$ may be different from $Y_{g,t}(\mathbf{D}'_{t-1}, d_t)$ where $\mathbf{D}_{t-1} \neq \mathbf{D}'_{t-1}$. I investigate necessary and sufficient conditions for the static TWFE estimator (Equation (1)) to be a convex combination of ATEs in this framework by extending the decomposition provided in Theorem 2.¹⁰ Let us assume that untreated potential outcomes do not depend on the treatment history of a group, that is:

Assumption 8. (*No Dynamics for Untreated Potential Outcomes*):

$$Y_{g,t}(\mathbf{D}_{t-1}, 0) = Y_{g,t}(\mathbf{D}'_{t-1}, 0), \quad \forall \mathbf{D}_{t-1}, \mathbf{D}'_{t-1} \in \{0, 1\}^{t-1}$$

Given Assumptions 7 and 8, the standard, reverse leaver and double-switcher difference-in-differences remain unbiased estimators of their corresponding ATEs even in the presence of dynamics. Let us now consider the reverse difference-in-differences, where we compare group g switching from untreated to treated between period t and t' , to group k remaining treated across the two periods. Adding and subtracting $E[Y_{g,t'}(\mathbf{D}_{g,t'-1}, 1) | \mathbf{D}]$, we have:

$$E[\hat{\beta}_{g,k,t,t'}^{DD} | \mathbf{D}] = E[\Delta_{g,t'}(\mathbf{D}_{g,t'-1}) + Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 1) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 1) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) | \mathbf{D}]$$

where $\Delta_{g,t'}(\mathbf{D}_{g,t'-1}) = Y_{g,t'}(\mathbf{D}_{g,t'-1}, 1) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 0)$, the treatment effect of group g in time t' conditional on having faced treatment history $\mathbf{D}_{g,t'-1}$. A necessary and sufficient condition for the reverse two-by-two

¹⁰The decomposition provided in Theorem 1 does not rely on any assumptions regarding the true data generating process for $Y_{g,t}$, and thus remains valid in the presence of dynamic treatment effects.

comparisons included in the TWFE estimator is thus that, conditional on the treatment history of each group, their treated potential outcomes would follow the same trend. The same assumption is required for the leaver difference-in-differences. The counterpart of Theorem 2 in this framework is thus as follows:

Theorem 3.

$$E[\hat{\beta}_{fe}|\mathbf{D}] = \frac{\sum_t \sum_g [E[\Delta_{g,t}(\mathbf{D}_{g,t-1})|\mathbf{D}] \sum_{t' \neq t} \sum_{k \neq g} [\omega_{g,k,t,t'}^S + \omega_{g,k,t',t}^R + \omega_{k,g,t,t'}^L + \omega_{k,g,t',t}^{RL} + 2\omega_{g,k,t,t'}^{DS} + 2\omega_{k,g,t',t}^{DS}]]}{\sum_t \sum_{g: D_{g,t}=1} \sum_{k: D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]}$$

$\forall \Omega$ such that Assumptions 1, 6-8 hold, if and only if, $\forall t > 2, E[Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t-1}(\mathbf{D}_{g,t-2}, 1)|\mathbf{D}]$ does not vary across g .

The proof follows the same argument as for Theorem 2. It can be shown that the condition that $\forall t > 2, E[Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t-1}(\mathbf{D}_{g,t-2}, 1)|\mathbf{D}]$ does not vary across g does not rule out the presence of non-contemporaneous treatment effects. Yet, it allows for very restricted dynamics, such as the ones presented in Appendix A.2.

Remark Note that a natural counterpart to Assumption 5, the parallel trends assumptions on the treated potential outcomes in the static framework, can be imposed in the framework allowing for non-contemporaneous treatment effects:

Assumption 9. (*Common Trends on the Potential Outcome of First Exposure to Treatment:*) For $t \geq 2, E(Y_{g,t}(\mathbf{0}_{t-1}, 1) - Y_{g,t-1}(\mathbf{0}_{t-2}, 1)|\mathbf{D})$ does not vary across g .

Assumption 9 is a parallel trends assumption on *treated* potential outcomes, conditional on not having been treated before. This simply means that the difference in outcomes when being treated for the first time in period t or in period $t - 1$ would be the same across groups. Together with the parallel trends assumption on untreated potential outcomes, this implies that the treatment effect associated to the first exposure to treatment evolve similarly across groups, but can differ across groups and over time.

Note that under this assumption, the absence of dynamic treatment effects implies that the condition that, $\forall t > 2, E[Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t-1}(\mathbf{D}_{g,t-2}, 1)|\mathbf{D}]$ does not vary across g holds. This result has two implications. First, it implies that the absence of dynamic treatment effects is a sufficient condition for the TWFE estimator to correspond to a convex combination of ATEs under Assumptions 1 and 6-9. This may be a plausible assumption in a restricted set of settings, when individuals stay within a group g only one time period. For example, when evaluating educational policies, the treatment unit is often a grade within a school: students within this group will be treated a single period, limiting the potential for dynamic treatment effects. Second, in a framework where untreated potential outcomes as well as potential outcomes when first exposed to treatment evolve similarly across groups, it highlights that the

presence of dynamic treatment effects is a necessary condition for the decomposition presented in Theorem 3 to break. As these conditions allow for treatment effect heterogeneity, it underscores the critical role of non-contemporaneous treatment effects, rather than heterogeneous treatment effects across groups and over time, in explaining when and why the TWFE estimator may fail to be a convex combination of ATEs.

5 Conclusion

Difference-in-differences is one of the most popular quasi-experimental methods to estimate causal effects. Most empirical applications have yet departed from the traditional two-group two-period setting, for which it is established that comparing the evolution of the outcome of interest before and after a group receives a treatment to the one of a never-treated group identifies the ATT. With several groups and periods, researchers will typically interpret the parameter of the treatment dummy in a TWFE regression as the ATT. Yet, recent developments in the difference-in-differences literature have concluded that, under the standard common trends assumption, the TWFE estimator may weigh negatively some ATEs. Although heterogeneous treatment effects are understood as a source of these potential negative weights, necessary and sufficient conditions for the TWFE estimator to be a convex combination of ATEs have not been formally derived.

This paper fills this gap. When only contemporaneous treatment effects are considered, I show that it requires a common trends assumptions on both the *treated* and *untreated* potential outcomes. I further show that under parallel trends on the untreated potential outcomes and on the first exposure to treatment, the absence of dynamic treatment effects is sufficient for this result to hold, but not necessary.

To derive this result, I decompose the TWFE estimator and show that it may include five different types of standard two-by-two comparisons, all entering positively. I then study these comparisons separately and find that each is an unbiased estimator of the ATE of the group switching treatment status under either a common trends assumption on potential outcomes when treated or when untreated. Under these assumptions, I show that the TWFE estimator weighs each ATE proportionally to the number of comparison groups available to identify it.

As noted by [de Chaisemartin and D’Haultfoeuille \(2023b\)](#), ‘understanding the circumstances where TWFE and heterogeneity-robust difference-in-differences estimators are more likely to differ is an important question’. Results derived above are key to understand why the TWFE and heterogeneity-robust difference-in-differences estimators may sometimes be very similar in practice, even when the presence of heterogeneous treatment effects is likely. The valid comparisons highlighted in the paper may also open the way to developing heterogeneity-robust estimators exploiting the variation present in the data in a more comprehensive manner.

A Appendix

A.1 Proof of Theorem 1

We take as a point of departure the decomposition of [Strezhnev \(2018\)](#):

$$\hat{\beta}_{fe} = \frac{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [(Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'})]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]} \quad (7)$$

Let us focus on the numerator:

$$\begin{aligned} & \sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [(Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'})] \\ = & \underbrace{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' < t} [(Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'})]}_A - \underbrace{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' > t} [(Y_{g,t'} - Y_{g,t}) - (Y_{k,t'} - Y_{k,t})]}_B \end{aligned}$$

Let us start by decomposing A:

$$\begin{aligned} A = & \sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' < t} \left[\sum_{g:D_{g,t'}=1} \sum_{k:D_{k,t'}=1} [(Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'})] \right. \\ & + \sum_{g:D_{g,t'}=1} \sum_{k:D_{k,t'}=0} [(Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'})] \\ & + \sum_{g:D_{g,t'}=0} \sum_{k:D_{k,t'}=0} [(Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'})] \\ & \left. + \sum_{g:D_{g,t'}=0} \sum_{k:D_{k,t'}=1} [(Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'})] \right] \end{aligned}$$

Similarly, we can decompose B:

$$\begin{aligned}
B = \sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t'>t} & \left[\sum_{g:D_{g,t'}=1} \sum_{k:D_{k,t'}=1} [(Y_{g,t'} - Y_{g,t}) - (Y_{k,t'} - Y_{k,t})] \right. \\
& + \sum_{g:D_{g,t'}=1} \sum_{k:D_{k,t'}=0} [(Y_{g,t'} - Y_{g,t}) - (Y_{k,t'} - Y_{k,t})] \\
& + \sum_{g:D_{g,t'}=0} \sum_{k:D_{k,t'}=0} [(Y_{g,t'} - Y_{g,t}) - (Y_{k,t'} - Y_{k,t})] \\
& \left. + \sum_{g:D_{g,t'}=0} \sum_{k:D_{k,t'}=1} [(Y_{g,t'} - Y_{g,t}) - (Y_{k,t'} - Y_{k,t})] \right]
\end{aligned}$$

The second term of A and B are the same, they will thus disappear when computing A-B. We thus have, using the notations defined in Section 3:

$$\begin{aligned}
& \sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [(Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'})] \\
& = \sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [\hat{\beta}_{g,k,t,t'}^S + \hat{\beta}_{k,g,t',t}^R + \hat{\beta}_{g,k,t,t'}^L + \hat{\beta}_{k,g,t',t}^{RL} + 2\hat{\beta}_{g,k,t,t'}^{DS}]
\end{aligned}$$

Hence, we can write:

$$\hat{\beta}_{fe} = \frac{\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [\hat{\beta}_{g,k,t,t'}^S + \hat{\beta}_{k,g,t',t}^R + \hat{\beta}_{g,k,t,t'}^L + \hat{\beta}_{k,g,t',t}^{RL} + 2\hat{\beta}_{g,k,t,t'}^{DS}]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]} \quad (8)$$

A.2 Section 4: Proofs

A.2.1 The Five Difference-in-Differences Comparisons: Identification

Let us focus on the standard difference-in-differences. We want to prove that Assumption 4 holds if and only if, $\forall g, k \neq g, t, t' \neq t, \forall \Omega, E[\hat{\beta}_{g,k,t,t'}^S | \mathbf{D}] = E[\Delta_{g,t} | \mathbf{D}] \times \omega_{g,k,t,t'}^S$.

First, it is well known that Assumption 4 is a sufficient condition for the standard difference-in-differences to identify the ATE of the group joining treatment. Let us now show that when Assumption 4 does not hold, it is not true that $E[\hat{\beta}_{g,k,t,t'}^S | \mathbf{D}] = E[\Delta_{g,t} | \mathbf{D}] \times \omega_{g,k,t,t'}^S, \forall g, k \neq g, t, t' \neq t$ and $\forall \Omega$.

Let us assume that Assumption 4 does not hold. Then, $\exists g, k \neq g, t, t' \neq t$ such that $E[Y_{g,t}(0) -$

$Y_{g,t'}(0)|\mathbf{D}] \neq E[Y_{k,t}(0) - Y_{k,t'}(0)|\mathbf{D}]$. Let us consider $\mathbf{\Omega}$ such that $\omega_{g,k,t,t'} = 1$. Then, we have:

$$\begin{aligned} E[\hat{\beta}_{g,k,t,t'}^S|\mathbf{D}] &= E[\Delta_{g,t} + (Y_{g,t}(0) - Y_{g,t'}(0) - (Y_{k,t}(0) - Y_{k,t'}(0)))]|\mathbf{D}] \times \omega_{g,k,t,t'}^S \\ &\neq E[\Delta_{g,t}|\mathbf{D}] \times \omega_{g,k,t,t'}^S \end{aligned}$$

The proofs corresponding to the other two-by-two comparisons follow the same argument.

A.2.2 Proof of Theorem 2

Let us impose Assumptions 1-5. Taking the expectation of $\hat{\beta}_{fe}$ conditional on \mathbf{D} , we have:

$$\begin{aligned} E[\hat{\beta}_{fe}|\mathbf{D}] &= \frac{\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} E[\hat{\beta}_{g,k,t,t'}^S + \hat{\beta}_{k,g,t',t}^R + \hat{\beta}_{g,k,t,t'}^L + \hat{\beta}_{k,g,t',t}^{RL} + 2\hat{\beta}_{g,k,t,t'}^{DS}|\mathbf{D}]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]} \\ &= \frac{\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [\beta_{g,k,t,t'}^S + \beta_{k,g,t',t}^R + \beta_{g,k,t,t'}^L + \beta_{k,g,t',t}^{RL} + 2\beta_{g,k,t,t'}^{DS}]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]} \end{aligned}$$

where $\beta_{g,k,t,t'}^C \equiv E[\hat{\beta}_{g,k,t,t'}^C|\mathbf{D}]$,

We can rewrite the numerator:

$$\begin{aligned} &\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [\beta_{g,k,t,t'}^S + \beta_{k,g,t',t}^R + \beta_{g,k,t,t'}^L + \beta_{k,g,t',t}^{RL} + 2\beta_{g,k,t,t'}^{DS}] \\ &= \sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [E[\Delta_{g,t}|\mathbf{D}] \times [\omega_{g,k,t,t'}^S + \omega_{k,g,t',t}^{RL} + 2\omega_{g,k,t,t'}^{DS}] \\ &\quad + E[\Delta_{k,t}|\mathbf{D}] \times [\omega_{k,g,t',t}^R + \omega_{g,k,t,t'}^L] + E[\Delta_{k,t'}|\mathbf{D}] \times [2\omega_{g,k,t,t'}^{DS}]] \end{aligned}$$

The first term of the sum rewrites:

$$\begin{aligned} &\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} \left[E[\Delta_{g,t}|\mathbf{D}] \times [\omega_{g,k,t,t'}^S + \omega_{k,g,t',t}^{RL} + 2\omega_{g,k,t,t'}^{DS}] \right] \\ &= \sum_t \sum_g [E[\Delta_{g,t}|\mathbf{D}] \sum_{t' \neq t} \sum_{k \neq g} [\omega_{g,k,t,t'}^S + \omega_{k,g,t',t}^{RL} + 2\omega_{g,k,t,t'}^{DS}]] \end{aligned}$$

Let us focus on the term $\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [E[\Delta_{k,t} | \mathbf{D}] \times [\omega_{k,g,t',t}^R + \omega_{g,k,t,t'}^L]]$:

$$\begin{aligned} & \sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [E[\Delta_{k,t} | \mathbf{D}] \times [\omega_{k,g,t',t}^R + \omega_{g,k,t,t'}^L]] \\ &= \sum_t \sum_g \left[E[\Delta_{g,t} | \mathbf{D}] \times \sum_{t' \neq t} \sum_{k \neq g} [\omega_{g,k,t',t}^R + \omega_{k,g,t,t'}^L] \right] \end{aligned}$$

And, focusing on the term $\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [E[\Delta_{k,t'} | \mathbf{D}] \times 2\omega_{k,g,t,t'}^{DS}]$:

$$\begin{aligned} & \sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [E[\Delta_{k,t'} | \mathbf{D}] \times 2\omega_{k,g,t,t'}^{DS}] \\ &= \sum_t \sum_g \left[E[\Delta_{g,t} | \mathbf{D}] \times \sum_{t' \neq t} \sum_{k \neq g} 2\omega_{k,g,t,t'}^{DS} \right] \end{aligned}$$

The numerator thus writes:

$$\begin{aligned} & \sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [\beta_{g,k,t,t'}^S + \beta_{k,g,t',t}^R + \beta_{g,k,t,t'}^L + \beta_{k,g,t',t}^{RL} + 2\beta_{g,k,t,t'}^{DS}] \\ &= \sum_t \sum_g \left[E[\Delta_{g,t} | \mathbf{D}] \sum_{t' \neq t} \sum_{k \neq g} [\omega_{g,k,t,t'}^S + \omega_{k,g,t',t}^R + \omega_{g,k,t,t'}^L + \omega_{k,g,t',t}^{RL} + 2\omega_{g,k,t,t'}^{DS} + 2\omega_{k,g,t',t}^{DS}] \right] \end{aligned}$$

We thus have, $\forall \Omega$:

$$E[\hat{\beta}_{fe} | \mathbf{D}] = \frac{\sum_t \sum_g [E[\Delta_{g,t} | \mathbf{D}] \sum_{t' \neq t} \sum_{k \neq g} [\omega_{g,k,t,t'}^S + \omega_{k,g,t',t}^R + \omega_{g,k,t,t'}^L + \omega_{k,g,t',t}^{RL} + 2\omega_{g,k,t,t'}^{DS} + 2\omega_{k,g,t',t}^{DS}]]}{\sum_t \sum_{g: D_{g,t}=1} \sum_{k: D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t}]}$$

Following Section 4.1, if Assumption 4 or 5 does not hold, then we would be able to find an Ω such that there would exist a two-by-two comparison between two groups and two time periods $g, k \neq g, t, t' \neq t$ entering with a positive weight in $E[\hat{\beta}_{fe} | \mathbf{D}]$, while being a biased estimator of its corresponding ATE. Hence, Assumption 4 and 5 are both necessary and sufficient for the above statement to hold.

A.2.3 Introduction of Dynamics: Details

A General Decomposition Result in the Dynamic Framework First, note that Assumptions 7 and 8 are such that the standard, reverse leaver and double-switcher difference-and-differences are unbiased

estimators of their corresponding ATEs. Focusing on the standard difference-in-differences, we have:

$$\begin{aligned}
& E[Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 0) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 0) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 0)) | \mathbf{D}] \\
&= E[Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 0) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 0) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 0)) + Y_{g,t}(\mathbf{D}_{g,t-1}, 0) - Y_{g,t'}(\mathbf{D}_{g,t-1}, 0) | \mathbf{D}] \\
&= E[\Delta_{g,t}(\mathbf{D}_{g,t-1}) + Y_{g,t}(\mathbf{D}_{g,t-1}, 0) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 0) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 0) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 0)) | \mathbf{D}] \\
&= E[\Delta_{g,t}(\mathbf{D}_{g,t-1}) | \mathbf{D}]
\end{aligned}$$

Similarly, for the reverse leaver difference-in-differences, under Assumptions 7 and 8, we have:

$$\begin{aligned}
& E[Y_{g,t}(\mathbf{D}_{g,t-1}, 0) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 0) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 0) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) | \mathbf{D}] \\
&= E[Y_{g,t}(\mathbf{D}_{g,t-1}, 0) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 0) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 0) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) + Y_{k,t'}(\mathbf{D}_{k,t'-1}, 0) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 0) | \mathbf{D}] \\
&= E[\Delta_{k,t'}(\mathbf{D}_{k,t'-1}) + Y_{g,t}(\mathbf{D}_{g,t-1}, 0) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 0) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 0) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 0)) | \mathbf{D}] \\
&= E[\Delta_{k,t'}(\mathbf{D}_{k,t'-1}) | \mathbf{D}]
\end{aligned}$$

While, for the double switcher difference-in-difference, we have, under Assumptions 7 and 8:

$$\begin{aligned}
& E[Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 0) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 0) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) | \mathbf{D}] \\
&= E[\Delta_{g,t}(\mathbf{D}_{g,t-1}) + \Delta_{k,t'}(\mathbf{D}_{k,t'-1}) | \mathbf{D}]
\end{aligned}$$

Finally, while details for the reverse difference-in-differences are given in the body of the paper, the leaver difference-in-differences is such that:

$$\begin{aligned}
& E[Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 1) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 0) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) | \mathbf{D}] \\
&= E[Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 1) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 0) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) + Y_{k,t}(\mathbf{D}_{k,t-1}, 1) - Y_{k,t}(\mathbf{D}_{k,t-1}, 1) | \mathbf{D}] \\
&= E[\Delta_{k,t}(\mathbf{D}_{k,t-1}) + Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 1) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 1) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) | \mathbf{D}]
\end{aligned}$$

It is thus necessary and sufficient for $E[Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 1) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 1) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) | \mathbf{D}]$ to be zero for the reverse difference-in-difference to be an unbiased estimator of the ATE of group k in period t , $\forall \Omega$.

Sufficient Conditions in the Presence of Dynamic Treatment Effects One can show that the decomposition provided in Theorem 3 hold even in the presence of some dynamic treatment effects. Let us impose the following assumption:

Assumption 10. (*Homogeneous Effect from n -period Treatment Exposure:*) $\forall g, t$, the ATE of group g in

period t conditional on history $D_{g,t-1}$ writes: $\Delta_{g,t}(D_{g,t-1}) = \Delta_{g,t}(\mathbf{0}_{t-1}) + \tau_t(\sum_{\ell=1}^{t-1} D_{g,\ell})$, with $\tau_t(0) = 0$.

Assumption 10 defines the ATE of group g in period t as being equal to the sum of the ATE of the group if it was first exposed to treatment in t , and the effect of having been exposed to treatment for $\sum_{\ell=1}^{t-1} D_{g,\ell}$ periods prior to t , $\tau_t(\sum_{\ell=1}^{t-1} D_{g,\ell})$. The latter is a period- t specific function, taking as argument the group's number of periods of exposure to treatment prior to t . This implies that the incremental effect of having been exposed to treatment for n periods is homogeneous across groups for a given t . Note that this assumption does not rule out treatment effect heterogeneity across groups, nor over time. Indeed, the first-exposure ATE is g, t -specific. Further, the incremental effect $\tau_t(n)$ is also allowed to vary across time for a given n .

Overall, Assumptions 7 to 10 imply that $E[Y_{g,t}(\mathbf{D}_{k,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{k,t'-1}, 1) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 1) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) | \mathbf{D}]$ is equal to zero. Under these assumptions, we can derive the following theorem:

Theorem 4. *Suppose Assumptions 1, 6-10 hold. If:*

1. $\nexists g, k \neq g, t, t' \neq t$ such that $\omega_{g,k,t',t}^R = 1$, or if $g, k \neq g, t, t' \neq t$ for which $\omega_{g,k,t',t}^R = 1$ are such that $\sum_{\ell=1}^{t'-1} D_{g,\ell} = \sum_{\ell=1}^{t-1} D_{k,\ell}$ and,
2. $\nexists g, k \neq g, t, t' \neq t$ such that $\omega_{k,g,t,t'}^L = 1$, or if $g, k \neq g, t, t' \neq t$ for which $\omega_{k,g,t,t'}^L = 1$ are such that $\sum_{\ell=1}^{t'-1} D_{g,\ell} = \sum_{\ell=1}^{t-1} D_{k,\ell}$,

then,

$$E[\hat{\beta}_{fe} | \mathbf{D}] = \frac{1}{\sum_t \sum_{g: D_{g,t}=1} \sum_{k: D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]} \times \left[\sum_t \sum_g \left[E[\Delta_{g,t}(D_{g,t-1}) | \mathbf{D}] \sum_{t' \neq t} \sum_{k \neq g} [\omega_{g,k,t,t'}^S + \omega_{k,g,t',t}^{RL} + 2\omega_{g,k,t,t'}^{DS} + 2\omega_{k,g,t',t}^{DS}] \right] + \sum_t \sum_g \left[E[\Delta_{g,t}(D_{k,t-1}) | \mathbf{D}] \sum_{t' \neq t} \sum_{k \neq g} [\omega_{g,k,t',t}^R + \omega_{k,g,t,t'}^L] \right] \right]$$

The proof is detailed below. This result stems from the fact that, first, reverse or leaver difference-in-differences between t' and t when both groups have been exposed to treatment the same number of periods up to $t' - 1$ and up to period $t - 1$ identifies the ATE of the group switching treatment status under its own history. This would for example happen when comparing periods 2 and 4 of group g with $\mathbf{D}_{g,t} = (0, 0, 1, 1)$ and group k with $\mathbf{D}_{k,t} = (0, 1, 0, 1)$. Second, it relies on the fact that when the number of periods under treatment prior to the treated period of the switching group is the same as in the comparison group, reverse and leaver difference-in-differences identify the ATE of the switching group under the history of

the comparison group. For example, comparing a group g such that $\mathbf{D}_{g,t} = (1, 0, 0, 1)$ to a group k with $\mathbf{D}_{k,t} = (0, 0, 1, 1)$ in periods 3 and 4 allows to identify the ATE of group g in period 3 when being exposed to treatment for the first time.

Proof. Let us focus on the reverse difference-in-differences and derive sufficient conditions for those to estimate their corresponding ATE without bias. For that, we would need the following equation to be equal to zero:

$$\begin{aligned} & E[Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{g,t'-1}, 1) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 1) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) | \mathbf{D}] \\ = & E[Y_{g,t}(\mathbf{D}_{k,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{k,t'-1}, 1) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 1) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1))] \\ & + [Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t}(\mathbf{D}_{k,t-1}, 1)] - [Y_{g,t'}(\mathbf{D}_{g,t'-1}, 1) - Y_{g,t'}(\mathbf{D}_{k,t'-1}, 1)] | \mathbf{D} \end{aligned} \quad (9)$$

Let us first show that, under Assumptions 7-10 we have that $E[Y_{g,t}(\mathbf{D}_{k,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{k,t'-1}, 1) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 1) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) | \mathbf{D}] = 0$. Indeed, we have:

$$\begin{aligned} & E[Y_{g,t}(\mathbf{D}_{k,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{k,t'-1}, 1) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 1) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1)) | \mathbf{D}] \\ = & E[Y_{g,t}(\mathbf{D}_{k,t-1}, 1) - Y_{g,t'}(\mathbf{D}_{k,t'-1}, 1) - (Y_{k,t}(\mathbf{D}_{k,t-1}, 1) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 1) \\ & + Y_{g,t}(\mathbf{D}_{k,t-1}, 0) - Y_{g,t}(\mathbf{D}_{k,t-1}, 0) + Y_{g,t'}(\mathbf{D}_{k,t'-1}, 0) - Y_{g,t'}(\mathbf{D}_{k,t'-1}, 0) \\ & + Y_{k,t}(\mathbf{D}_{k,t-1}, 0) - Y_{k,t}(\mathbf{D}_{k,t-1}, 0) + Y_{k,t'}(\mathbf{D}_{k,t'-1}, 0) - Y_{k,t'}(\mathbf{D}_{k,t'-1}, 0)) | \mathbf{D}] \\ = & E[\Delta_{g,t}(D_{k,t-1}) - \Delta_{g,t'}(D_{k,t'-1}) - \Delta_{k,t}(D_{k,t-1}) + \Delta_{k,t'}(D_{k,t'-1}) \\ & + (Y_{g,t}(D_{k,t-1}, 0) - Y_{g,t'}(D_{k,t'-1}, 0) - (Y_{k,t}(D_{k,t-1}, 0) - Y_{k,t'}(D_{k,t'-1}, 0))) | \mathbf{D}] \\ = & E[Y_{g,t}(\mathbf{0}_{t-1}, 1) - Y_{g,t}(\mathbf{0}_{t-1}, 0) + \tau_t(\sum_{\ell=1}^{t-1} D_{k,\ell}) - (Y_{g,t'}(\mathbf{0}_{t'-1}, 1) - Y_{g,t'}(\mathbf{0}_{t'-1}, 0) + \tau_{t'}(\sum_{\ell=1}^{t'-1} D_{k,\ell})) \\ & - (Y_{k,t}(\mathbf{0}_{t-1}, 1) - Y_{k,t}(\mathbf{0}_{t-1}, 0) + \tau_t(\sum_{\ell=1}^{t-1} D_{k,\ell})) + (Y_{k,t'}(\mathbf{0}_{t'-1}, 1) - Y_{k,t'}(\mathbf{0}_{t'-1}, 0) + \tau_{t'}(\sum_{\ell=1}^{t'-1} D_{k,\ell})) | \mathbf{D}] \\ = & E[Y_{g,t}(\mathbf{0}_{t-1}, 1) - Y_{g,t'}(\mathbf{0}_{t'-1}, 1) - (Y_{k,t}(\mathbf{0}_{t-1}, 1) - Y_{k,t'}(\mathbf{0}_{t'-1}, 1)) \\ & - [Y_{g,t}(\mathbf{0}_{t-1}, 0) - Y_{g,t'}(\mathbf{0}_{t'-1}, 0) - (Y_{k,t}(\mathbf{0}_{t-1}, 0) - Y_{k,t'}(\mathbf{0}_{t'-1}, 0))] | \mathbf{D}] \\ = & 0 \end{aligned}$$

where the second equality follows from Assumptions 7 and 8, and the third equality from Assumption 10. The first term of the fourth equality cancels out following from Assumption 9, while the second term does following Assumptions 7 and 8.

Let us now examine the last two terms of Equation (9) to understand what does the reverse difference-

in-differences identifies under the above assumptions. In particular, we have:

$$\begin{aligned}
& E[Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t}(\mathbf{D}_{k,t-1}, 1)] - [Y_{g,t'}(\mathbf{D}_{g,t'-1}, 1) - Y_{g,t'}(\mathbf{D}_{k,t'-1}, 1)] | \mathbf{D}] \\
&= E \left[[\Delta_{g,t}(D_{g,t-1}) + Y_{g,t}(\mathbf{D}_{g,t-1}, 0) - \Delta_{g,t}(D_{k,t-1}) - Y_{g,t}(\mathbf{D}_{k,t-1}, 0)] \right. \\
&\quad \left. - [\Delta_{g,t'}(D_{g,t'-1}) + Y_{g,t'}(\mathbf{D}_{g,t'-1}, 0) - \Delta_{g,t'}(D_{k,t'-1}) - Y_{g,t'}(\mathbf{D}_{k,t'-1}, 0)] | \mathbf{D} \right] \\
&= E \left[[Y_{g,t}(\mathbf{0}_{t-1}, 1) + \tau_t \left(\sum_{\ell=1}^{t-1} D_{g,\ell} \right) - Y_{g,t}(\mathbf{0}_{t-1}, 1) - \tau_t \left(\sum_{\ell=1}^{t-1} D_{k,\ell} \right)] \right. \\
&\quad \left. - [Y_{g,t'}(\mathbf{0}_{t'-1}, 1) + \tau_{t'} \left(\sum_{\ell=1}^{t'-1} D_{g,\ell} \right) - Y_{g,t'}(\mathbf{0}_{t'-1}, 1) - \tau_{t'} \left(\sum_{\ell=1}^{t'-1} D_{k,\ell} \right)] | \mathbf{D} \right] \\
&= E \left[[\tau_t \left(\sum_{\ell=1}^{t-1} D_{g,\ell} \right) - \tau_t \left(\sum_{\ell=1}^{t-1} D_{k,\ell} \right)] - [\tau_{t'} \left(\sum_{\ell=1}^{t'-1} D_{g,\ell} \right) - \tau_{t'} \left(\sum_{\ell=1}^{t'-1} D_{k,\ell} \right)] | \mathbf{D} \right]
\end{aligned} \tag{10}$$

where the first equality uses the definition of $\Delta_{g,t}(D_{g,t-1}) = Y_{g,t}(\mathbf{D}_{g,t-1}, 1) - Y_{g,t}(\mathbf{D}_{g,t-1}, 0)$. The second equality follows from Assumptions 10, according to which $\Delta_{g,t}(D_{g,t-1}) = Y_{g,t}(\mathbf{0}_{t-1}, 1) - Y_{g,t}(\mathbf{0}_{t-1}, 0) + \tau_t(\sum_{\ell=1}^{t-1} D_{g,t-1})$ and Assumption 8. The last equality stems from Assumption 9.

This implies that if groups g and k have been exposed to treatment the same number of periods up to $t' - 1$ and up to period $t - 1$, all those terms cancel out and one can identify the ATE of group g in t' , conditional on facing history $D_{g,t}$. This would for example happen when comparing periods 2 and 4 of group g with $\mathbf{D}_{g,t} = (0, 0, 1, 1)$ and group k with $\mathbf{D}_{k,t} = (0, 1, 0, 1)$.

What if the number of periods of exposure to treatment between group g and k differs only in $t' - 1$, but not in $t - 1$, i.e. $\sum_{\ell=1}^{t-1} D_{g,\ell} = \sum_{\ell=1}^{t-1} D_{k,\ell}$, but $\sum_{\ell=1}^{t'-1} D_{g,\ell} \neq \sum_{\ell=1}^{t'-1} D_{k,\ell}$? Now the reverse difference-in-differences would identify the following object:

$$\begin{aligned}
& E[\Delta_{g,t'}(\mathbf{D}_{g,t'-1}) - [\tau_{t'} \left(\sum_{\ell=1}^{t'-1} D_{g,\ell} \right) - \tau_{t'} \left(\sum_{\ell=1}^{t'-1} D_{k,\ell} \right)] | \mathbf{D}] \\
&= E[\Delta_{g,t'}(\mathbf{0}_{t'-1}) + \tau_{t'} \left(\sum_{\ell=1}^{t'-1} D_{g,\ell} \right) - [\tau_{t'} \left(\sum_{\ell=1}^{t'-1} D_{g,\ell} \right) - \tau_{t'} \left(\sum_{\ell=1}^{t'-1} D_{k,\ell} \right)] | \mathbf{D}] \\
&= E[\Delta_{g,t'}(\mathbf{0}_{t'-1}) + \tau_{t'} \left(\sum_{\ell=1}^{t'-1} D_{k,\ell} \right)] | \mathbf{D}] \\
&= E[\Delta_{g,t'}(\mathbf{D}_{k,t'-1})] | \mathbf{D}]
\end{aligned} \tag{11}$$

Thus, in this case, the reverse difference-in-differences identifies the ATE of group g in period t' conditional on history $\mathbf{D}_{k,t'-1}$. For example, comparing a group g such that $\mathbf{D}_{g,t} = (1, 0, 0, 1)$ to a group k with $\mathbf{D}_{k,t} = (0, 0, 1, 1)$ in periods 3 and 4 would allow to identify the ATE of group g in period 3 when being exposed to treatment for the first time.

In a similar fashion, one can show that the lever difference-in-differences comparing periods t and t' ,

with $t > t'$ identifies the ATE of the group switching treatment status in period t under its own history if both $\sum_{\ell=1}^{t-1} D_{g,\ell} = \sum_{\ell=1}^{t-1} D_{k,\ell}$ and $\sum_{\ell=1}^{t'-1} D_{g,\ell} = \sum_{\ell=1}^{t'-1} D_{k,\ell}$. It would identify the ATE of the group switching treatment status in period t under the history of the comparison group if $\sum_{\ell=1}^{t'-1} D_{g,\ell} = \sum_{\ell=1}^{t'-1} D_{k,\ell}$ while $\sum_{\ell=1}^{t-1} D_{g,\ell} \neq \sum_{\ell=1}^{t-1} D_{k,\ell}$.

We now know under which assumptions the reverse and leaver difference-in-differences are unbiased estimators of the ATE of group g in period t , under the treatment history of the comparison group (which may be the same as the one of group g). When only such valid comparisons exist - or that there exist no reverse or leaver difference-in-differences - the TWFE estimator is thus robust to heterogeneity: it only weighs positively unbiased estimators of their corresponding ATE.

□

References

- Borusyak, Kirill, Xavier Jaravel, and Jann Spiess**, “Revisiting event-study designs: robust and efficient estimation,” *Review of Economic Studies*, 2024, *91* (6), 3253–3285.
- Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant’Anna**, “Difference-in-differences with a continuous treatment,” *arXiv preprint arXiv:2107.02637*, 2021.
- Chabé-Ferret, Sylvain and Anca Voia**, “Are grassland conservation programs a cost-effective way to fight climate change? Evidence from France,” Technical Report 2021.
- de Chaisemartin, Clément and Xavier D’Haultfoeuille**, “Fuzzy differences-in-differences,” *The Review of Economic Studies*, 2018, *85* (2), 999–1028.
- and –, “Two-way fixed effects estimators with heterogeneous treatment effects,” *American Economic Review*, 2020, *110* (9), 2964–96.
- and –, “Two-way fixed effects and differences-in-differences estimators with several treatments,” *Journal of Econometrics*, 2023, *236* (2), 105480.
- and –, “Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey,” *The Econometrics Journal*, 2023, *26* (3), C1–C30.
- and –, “Difference-in-differences estimators of intertemporal treatment effects,” *Review of Economics and Statistics*, 2024, pp. 1–45.
- Goodman-Bacon, Andrew**, “Difference-in-differences with variation in treatment timing,” *Journal of Econometrics*, 2021, *225* (2), 254–277.
- Kim, Kimin and Myoung jae Lee**, “Difference in differences in reverse,” *Empirical Economics*, 2019, *57* (3), 705–725.
- Robins, James**, “A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect,” *Mathematical modelling*, 1986, *7* (9-12), 1393–1512.
- Rossi, Pauline and Paola Villar**, “Private health investments under competing risks: evidence from malaria control in Senegal,” *Journal of Health Economics*, 2020, *73*, 102330.
- Roth, Jonathan, Pedro HC Sant’Anna, Alyssa Bilinski, and John Poe**, “What’s trending in difference-in-differences? A synthesis of the recent econometrics literature,” *Journal of Econometrics*, 2023.

Strezhnev, Anton, “Semiparametric weighting estimators for multi-period difference-in-differences designs,” in “Annual Conference of the American Political Science Association, August,” Vol. 30 2018.