WORKING PAPERS

N° 1362

June 2023

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

Anaïs Fabre

Toulouse School of Economics

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

Anaïs Fabre[†]

June 15, 2023

Abstract

This paper provides necessary and sufficient conditions for the Two-Way Fixed Effects (TWFE) estimator to be robust to heterogeneous treatment effects. I decompose the TWFE estimator to show that it is a weighted sum of five different types of two-by-two comparisons, with positive weights. I show that parallel trends assumptions on either the untreated or treated potential outcomes must hold for each comparison to identify the Average Treatment Effect (ATE) of the group switching treatment status, when the effect of the treatment is contemporaneous. Both parallel trends assumptions are thus necessary and sufficient for the TWFE estimator to weigh each ATE positively, when allowing treatment effects to be heterogeneous across groups and periods. I further provide sufficient conditions under which the TWFE estimator remains valid even in the presence of dynamic treatment effects. Finally, I show how to exploit all available comparisons to build unbiased estimators of the ATT and ATE.

[*] 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, Tomas Larroucau, 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.

[†] Toulouse School of Economics: anais.fabre@tse-fr.eu.

1 Introduction

Difference-in-differences 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) in each group and period, with weights that may be negative in the presence of heterogeneous treatment effects (de Chaisemartin and d'Haultfoeuille, 2020a). The TWFE estimator may then be negative even when all ATEs are positive.

This paper provides necessary and sufficient conditions for the TWFE estimator not to weigh any ATEs negatively, when they can be heterogeneous across groups and over time. I show that this requires not only a parallel trends assumption on the *untreated* potential outcomes, but also on the *treated* potential outcomes. Under these conditions, the TWFE estimator is thus heterogeneity-robust. Furthermore, I show that its weights can be corrected to build unbiased estimators of the ATT and of the ATE.

To derive these results, I consider a general framework where groups are allowed to enter and exit 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, 2020a). 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. While most researchers are only familiar with the standard difference-in-differences, these intermediary results broaden the panel of comparisons available to them. 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 help identify treatment effects of interest.

I then establish the implication of these results for the TWFE estimator: it is a weighted sum of ATEs in different groups and periods, all weighted positively. Each ATE enters proportionally to the number of comparison groups available to identify it. When allowing for heterogeneous treatment effects over time and across groups, 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.

I further investigate the robustness of the static TWFE estimator to the presence of dynamic treatment effects. On top of parallel trends assumptions, the absence of such dynamics is a sufficient, but not necessary, condition for the TWFE estimator to weigh all ATEs positively. I provide conditions, allowing for dynamics, for the TWFE estimator to be heterogeneity-robust. In particular, I consider a setting where common trends assumptions conditional on being never-treated and on joining treatment for the first time hold. I assume that the incremental effect of having been exposed to treatment for n periods is homogeneous across groups for a given time period, while ATEs of first exposure can be heterogeneous. In this framework, I show that there exist relevant settings in which all two-by-two comparisons, and thus the TWFE estimator, remain valid.

Finally, I conclude with the two main implications of this paper for applied researchers. First, the conditions for the TWFE estimator not to be under the threat of negative weights are such that testing for parallel trends on the untreated potential outcomes is not sufficient. Whenever using the TWFE estimator, one should also test for parallel trends of potential outcomes across groups when they are treated. I summarize tests available to the practitioner.[1] Second, I take stock of the fact that the TWFE estimator does not weigh each ATE by the corresponding sample size and

[1] See Roth (2022) for recommended practices to perform such tests.

propose alternative estimators. In particular, I exploit the existing valid comparisons which receive zero-weight in the heterogeneity-robust estimator suggested by de Chaisemartin and d'Haultfoeuille (2020a). I expand the latter by building an unbiased estimator of the ATT of all switching cells under less stringent assumptions. I further show that one can build, under adequate restrictions, unbiased estimators of the ATT - requiring a parallel trends assumption on the untreated potential outcomes only - and ATE.

Related Literature The difference-in-differences literature has recently put the TWFE estimator under increased scrutiny. de Chaisemartin and d'Haultfoeuille (2020a) study a general framework with several groups and time periods where groups can enter and leave treatment, without considering dynamic treatment effects. 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. (2022) reach the same conclusion in the staggered case, where groups cannot exit treatment.

This paper revisits these results and provides necessary and sufficient conditions for the TWFE estimator to be robust to heterogeneity in treatment effects. I show that it is not required for ATEs to be homogeneous across groups, nor over time for the TWFE estimator to weigh all ATEs positively.

Both Goodman-Bacon (2021) and Strezhnev (2018) provide decompositions of the TWFE estimator in terms of two-by-two difference-in-differences comparisons to shed light on the origin of negative weights. 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'. Borusyak et al. (2022) 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 of this paper builds on these decompositions. By carefully deriving the conditions which are both necessary and sufficient for the TWFE estimator to be heterogeneity-robust, 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 these papers, this does not require treatment effects to be homogeneous over time, nor across groups.

Finally, while the above papers have focused on the staggered case, I explore the properties of each of the five types of two-by-two difference-in-differences comparisons which enter the TWFE estimator. I establish the assumptions under which each comparison is an unbiased estimator of its corresponding ATE. On top of allowing to derive the main result of this paper, it also allows to expand the tool box of the applied economist. Apart from the standard difference-in-differences, only the reverse difference-in-differences has received some attention in the literature (Kim and Lee, 2019). Additionally, I highlight how the different comparisons can be combined to build unbiased estimators of the ATT and ATE, as well as to expand the heterogeneity-robust estimator suggested by de Chaisemartin and d'Haultfoeuille (2020a).

It should be noted that, following de Chaisemartin and d'Haultfoeuille (2020a), I do not consider cases with non-binary, continuous or several treatments (de Chaisemartin and d'Haultfoeuille, 2018; de Chaisemartin and d'Haultfoeuille, 2020b; Callaway et al., 2021). de Chaisemartin and D'Haultfoeuille (2022b) and Roth et al. (2023) provide rich reviews of this recent literature.

2 Framework

I use the same framework and notations as de Chaisemartin and d'Haultfoeuille (2020a). In particular, consider observations that are divided across G groups and T periods. For each group g in period t, we observe a number $N_{g,t}$ of individuals. Let us denote $D_{i,g,t}$ the binary treatment status of individual i in group g at period t, and $(Y_{i,g,t}(0), Y_{i,g,t}(1))$ the potential outcomes when untreated and when treated, respectively. The observed outcome of individual i in group g at period t is denoted $Y_{i,g,t}(D_{i,g,t})$. The following objects can be defined, for all $(g,t) \in \{1, ..., G\} \times \{1, ..., T\}$:

$$D_{g,t} = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} D_{i,g,t}, \qquad Y_{g,t}(0) = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} Y_{i,g,t}(0),$$

$$Y_{g,t}(1) = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} Y_{i,g,t}(1), \qquad \text{and} \qquad Y_{g,t} = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} Y_{i,g,t}.$$

where $D_{g,t}$, $Y_{g,t}(0)$, $Y_{g,t}(1)$, and $Y_{g,t}$ denote the average treatment, the average potential outcomes when untreated and when treated, and the average observed outcome in group g at period t, respectively. $\mathbf{\Omega}$ collects the potential outcomes and treatment status for all g and t, i.e. $\mathbf{\Omega} = \{Y_{g,t}(0), Y_{g,t}(1), D_{g,t}\}_{\forall g,t}$.

I also impose the same assumptions as de Chaisemartin and d'Haultfoeuille (2020a). Discussions of these assumptions can be found in their paper.

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

Assumption 2 *(Sharp Design):* $\forall (g,t) \in \{1,...,G\} \times \{1,...,T\}$ *and* $i \in \{1,...,N_{g,t}\}, D_{i,g,t} = D_{g,t}.$

Assumption 3 *(Independent Groups): The vectors* $(Y_{g,t}(0), Y_{g,t}(1), D_{g,t})_{1\leq t\leq T}$ *are mutually independent.*

Assumption 4 *(Strong Exogeneity):* $\forall (g,t,d) \in \{1,...,G\} \times \{2,...,T\} \times \{0,1\},$
$E(Y_{g,t}(d) - Y_{g,t-1}(d)|D_{g,1},...,D_{g,T}) = E(Y_{g,t}(d) - Y_{g,t-1}(d)).$

These assumptions impose very few restrictions on $\mathbf{\Omega}$, 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, let us define some objects of interest. The ATE of group g in period t, $\Delta_{g,t}$, writes:

$$\Delta_{g,t} = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} [Y_{i,g,t}(1) - Y_{i,g,t}(0)]$$

Defining $N^{(1)} = \sum_{i,g,t} D_{i,g,t}$ as the number of treated units, the expected average treatment on the treated, δ^{TR}, writes:

$$\delta^{TR} = E\left[\sum_{g,t:D_{g,t}=1} \frac{N_{g,t}}{N^{(1)}} \Delta_{g,t}\right]$$

We consider the following TWFE regression:

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

We let $\hat{\beta}_{fe}$ denote the OLS estimator of the coefficient of $D_{g,t}$, with $\beta_{fe} = E[\hat{\beta}_{fe}]$. de Chaisemartin and d'Haultfoeuille (2020a) show that, under Assumptions 1-4 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.[2] This

[2] Results in de Chaisemartin and d'Haultfoeuille (2020a) are derived when considering the regression of $Y_{i,g,t}$ on group fixed effects, period fixed effects and $D_{g,t}$, i.e. using more disaggregated outcome data. de Chaisemartin and

result implies that $\hat{\beta}_{fe}$ 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 (de Chaisemartin and D’Haultfoeuille, 2022b). In the next sections, I revisit these results and provide sufficient and necessary conditions for the TWFE 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. Importantly, Strezhnev (2018) provides a decomposition of the TWFE estimator in the general case. He shows that the TWFE estimator can be written as a uniform average of difference-in-differences comparisons:

$$\hat{\beta}_{fe} = \frac{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [(Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'})]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]} \tag{2}$$

I now decompose this object further in order to clarify which comparisons enter the TWFE estimator.[3] In particular, I show that $\hat{\beta}_{fe}$ is a weighted sum of five different types of two-by-two comparisons. $\forall g \in \{1,..,G\}, \forall k \in \{1,..,G\} \setminus g$, $\forall t \in \{1,..,T\}, \forall t' \in \{1,..,T\} \setminus t$, we let $\hat{\beta}^{DD}_{g,k,t,t'}$ denote the two-by-two comparison of the outcomes of group g and k between periods t and t':

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

We can now define the five objects entering the TWFE estimator:

d’Haultfoeuille (2022b) extend them to the aggregated version considered in this paper. The latter is equivalent to the one using individual-data, up to re-weighing by the population in each group. Re-weighing only matters for the variance of the estimator, not its unbiasedness and consistency: we can thus abstract from re-weighing, following de Chaisemartin and D’Haultfoeuille (2022b).

[3] Strezhnev (2018) interprets Equation (2) informally, and focusing on the staggered setting.

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

$$\hat{\beta}^S_{g,k,t,t'} = \hat{\beta}^{DD}_{g,k,t,t'} \times \underbrace{\mathbb{1}\{D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 0, t > t'\}}_{\equiv \omega^S_{g,k,t,t'}} \tag{3}$$

Equation (3) 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, $\hat{\beta}^R_{g,k,t,t'}$:

$$\hat{\beta}^R_{g,k,t,t'} = \hat{\beta}^{DD}_{g,k,t,t'} \times \underbrace{\mathbb{1}\{D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 1, D_{k,t'} = 1, t > t'\}}_{\equiv \omega^R_{g,k,t,t'}} \tag{4}$$

Equation (4) describes the reverse difference-in-differences comparison.[4] 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.

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

$$\hat{\beta}^L_{g,k,t,t'} = \hat{\beta}^{DD}_{g,k,t,t'} \times \underbrace{\mathbb{1}\{D_{g,t} = 1, D_{g,t'} = 1, D_{k,t} = 0, D_{k,t'} = 1, t > t'\}}_{\equiv \omega^L_{g,k,t,t'}} \tag{5}$$

Equation (5) 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, $\hat{\beta}^{RL}_{g,k,t,t'}$:

$$\hat{\beta}^{RL}_{g,k,t,t'} = \hat{\beta}^{DD}_{g,k,t,t'} \times \underbrace{\mathbb{1}\{D_{g,t} = 0, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 1, t > t'\}}_{\equiv \omega^{RL}_{g,k,t,t'}} \tag{6}$$

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

Equation (6) 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, $\hat{\beta}^{DS}_{g,k,t,t'}$:

$$\hat{\beta}^{DS}_{g,k,t,t'} = \hat{\beta}^{DD}_{g,k,t,t'} \times \underbrace{\mathbb{1}\{D_{g,t} = 1, D_{g,t'} = 0, D_{k,t} = 0, D_{k,t'} = 1, t > t'\}}_{\equiv \omega^{DS}_{g,k,t,t'}} \tag{7}$$

Equation (7) 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:*

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

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

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

Details of the proof are provided in Appendix. 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 this result for the TWFE estimator. When allowing for heterogeneous treatment effects over time and across groups, 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 allows for dynamic treatment effects, and provides conditions under which the static TWFE estimator remains valid.

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 $\mathbf{\Omega}$ such that Assumptions 1-4 hold.

Standard Difference-in-Differences, $\hat{\beta}^S_{g,k,t,t'}$: 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 identify the ATT, i.e. the ATE of group g in time t.

Assumption 5 *(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:

$$\textit{Assumption 5 holds if and only if } \forall g, k \neq g, t, t' \neq t, \forall \mathbf{\Omega}, E[\hat{\beta}^S_{g,k,t,t'}|\mathbf{D}] = E[\Delta_{g,t}|\mathbf{D}] \times \omega^S_{g,k,t,t'}$$

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

Reverse Difference-in-Differences, $\hat{\beta}^R_{g,k,t,t'}$: 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 5, we have:

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

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 are biased estimators of $E[\Delta_{g,t}|\mathbf{D}]$.[5] 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 6 *(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:

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

In a framework allowing for heterogeneous treatment effects, Assumption 6 is thus both necessary and sufficient for the reverse difference-in-differences to identify the ATE of the switching group in period t'. We can write:

$$\textit{Assumption 6 holds if and only if } \forall g, k \neq g, t, t' \neq t, \forall \mathbf{\Omega}, E[\hat{\beta}^R_{g,k,t,t'}|\mathbf{D}] = E[\Delta_{g,t'}|\mathbf{D}] \times \omega^R_{g,k,t,t'}$$

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.

[5] Note 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:

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

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.

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

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

That is:

$$\textit{Assumption 6 holds if and only if } \forall g, k \neq g, t, t' \neq t, \forall \mathbf{\Omega}, E[\hat{\beta}^{L}_{g,k,t,t'}|\mathbf{D}] = E[\Delta_{k,t}|\mathbf{D}] \times \omega^{L}_{g,k,t,t'}$$

Reverse Leaver Difference-in-Differences, $\hat{\beta}^{RL}_{g,k,t,t'}$: 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 5:

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

We thus have:

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

Double-Switcher Difference-in-Differences, $\hat{\beta}^{DS}_{g,k,t,t'}$: 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 identified if the standard common trends assumption on potential outcomes when untreated holds:

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

Overall, we thus have:[6]

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

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 ATE 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.[7]

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 5. 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 $\Delta_{1,3}$.

[6] Note that, alternatively, we can show:

$$\textit{Assumption 6 holds if and only if } \forall g, k \neq g, t, t' \neq t, \forall \mathbf{\Omega}, E[\hat{\beta}^{DS}_{g,k,t,t'}|\mathbf{D}] = E[\Delta_{g,t'} + \Delta_{k,t}|\mathbf{D}] \times \omega^{DS}_{g,k,t,t'}$$

[7] Note that assuming that ATE are homogeneous will not be necessary to show that under Assumptions 5 and 6 none of the ATE that enter the TWFE estimator are weighted negatively.

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:

Theorem 2

$$E\left[\hat{\beta}_{fe}|\mathbf{D}\right] = \frac{\sum_t \sum_g \left[E\left[\Delta_{g,t}|\mathbf{D}\right] \sum_{t' \neq t} \sum_{k \neq g} \left[\omega^S_{g,k,t,t'} + \omega^R_{g,k,t',t} + \omega^L_{k,g,t,t'} + \omega^{RL}_{k,g,t',t} + 2\omega^{DS}_{g,k,t,t'} + 2\omega^{DS}_{k,g,t',t}\right]\right]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t}\left[1 - D_{g,t'} + D_{k,t'}\right]}$$

$\forall \boldsymbol{\Omega}$ *such that Assumptions 1-4 hold, if and only if, Assumptions 5 and 6 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. 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.[8] In particular, the TWFE estimator will never be negative when all ATEs are positive. Note that Assumptions 5 and 6 are also both imposed by de Chaisemartin and d'Haultfoeuille (2020a) for the estimator they propose as an alternative to the TWFE estimator to be unbiased.

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.

Note that if we consider 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.

[8] Note indeed that $\frac{\sum_t \sum_g \sum_{t' \neq t} \sum_{k \neq g} \left[\omega^S_{g,k,t,t'} + \omega^R_{g,k,t',t} + \omega^L_{k,g,t,t'} + \omega^{RL}_{k,g,t',t} + 2\omega^{DS}_{g,k,t,t'} + 2\omega^{DS}_{k,g,t',t}\right]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t}\left[1 - D_{g,t'} + D_{k,t'}\right]} = 1$.

4.3 Extension: 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'Haultfoeuille (2022a). 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 $Y_{g,t} = Y_{g,t}(D_{g,1}, D_{g,2}, ..., D_{g,T}) = Y_{g,t}(\mathbf{D}_{g,T})$, where $\mathbf{D}_{g,t}$ is the vector of treatment status for group g up to period t. Now, $\mathbf{\Omega}$ is the vector $\{(Y_{g,t}(\mathbf{d}))_{d\in\{0,1\}^T}, \mathbf{D}_{g,t}\}_{\forall g,t}$. We only maintain Assumptions 1 and 2. Let us also follow de Chaisemartin and D'Haultfoeuille (2022a) 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 7 *For all g, for all $\mathbf{d} \in \{0,1\}^T$, $Y_{g,t}(\mathbf{d}) = Y_{g,t}(d_1, ..., d_t)$.*

Assumption 8 $\forall t \geq 2, \forall g, E(Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})|\mathbf{D})$*, 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 $\mathbf{D}_{t-1} \neq \mathbf{D}'_{t-1}$. Can the static TWFE estimator (Equation (1)) be robust to dynamic treatment effects? I now use the decomposition provided in Theorem 1[9] to show that it is the case in some relevant scenarios. Let us assume that untreated potential outcomes do not depend on the treatment history of a group, that is:

Assumption 9 *(No Dynamics for Untreated Potential Outcomes):*

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

A General Decomposition Result in the Dynamic Framework Given Assumptions 8 and 9, 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

[9] Note indeed that this decomposition does not rely on any assumptions about the true data generating process for $Y_{g,t}$, and thus remains valid in the presence of dynamic treatment effects.

$E[Y_{g,t'}(\mathbf{D}_{g,t'-1},1)|\mathbf{D}]$, we have:

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

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 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

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

$\forall \mathbf{\Omega}$ *such that Assumptions 1-2, 7-9 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. Note that, together with Assumption 6, the absence of dynamic treatment effects is thus a sufficient condition for the TWFE estimator not to weigh any ATEs negatively. However, it is not necessary, and I now show that the condition in Theorem 3 can be satisfied even in the presence of dynamic treatment effects.

Sufficient Conditions in the Presence of Dynamic Treatment Effects Let us now derive sufficient conditions for the TWFE estimator to be heterogeneity-robust even in the presence of dynamics. I impose the following assumption:

Assumption 10

1. *(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*.*

2. *(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 10.1. 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.

Assumption 10.2. 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 8 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.[10] Under these assumptions, we can derive the following theorem:

Theorem 4 *Suppose Assumptions 1-2, 7-10 hold. If:*

1. *$\nexists g, k \neq g, t, t' \neq t$ such that $\omega^R_{g,k,t',t} = 1$, or if $g, k \neq g, t, t' \neq t$ for which $\omega^R_{g,k,t',t} = 1$ 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 $\omega^L_{k,g,t,t'} = 1$, or if $g, k \neq g, t, t' \neq t$ for which $\omega^L_{k,g,t,t'} = 1$ are such that $\sum_{\ell=1}^{t'-1} D_{g,\ell} = \sum_{\ell=1}^{t'-1} D_{k,\ell}$,*

then,

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

Details can be found in Appendix. 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

[10] Details can be found in Appendix.

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.

Overall, the absence of dynamic treatment effects over time is a sufficient condition for the TWFE estimator to be heterogeneity-robust. This may be a plausible assumption 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.

However, I also show that the TWFE estimator remains valid in relevant settings, even in the presence of dynamic treatment effects. The framework provided should help researchers evaluate whether their setting contains comparisons threatening the validity of the static TWFE estimator. Under the dynamics described in Assumption 10 and when reverse and leaver difference-in-differences exist, the potential for the TWFE estimator to weigh biased estimators of ATEs is null under two conditions. First, the existing reverse difference-in-differences should be such that the cumulative number of treated periods is the same in the two groups prior to the last period of the comparison. Second, the existing leaver difference-in-differences should be such that the cumulative number of treated periods is the same in the two groups prior to the first period of the comparison.

5 Implications for Empirical Analyses

I now derive the implications of Section 4.2 for researchers. First, when using the TWFE estimator, researchers should test for both *untreated* and *treated* parallel trends. Second, I provide alternative estimators exploiting the two-by-two comparisons highlighted above.

5.1 Testing for Common Trends on *Treated* Potential Outcomes

To the extent that the static TWFE estimator remains broadly used in empirical studies, it is crucial to establish under which assumptions all ATEs it includes are weighted positively. Suppose the researcher is in a framework where dynamic treatment effects can safely be discarded. As shown in Section 4.2, the TWFE estimator will then be heterogeneity-robust under two common trends assumption, Assumptions 5 and 6.

Overall, in a setting with several groups and time periods, when using the TWFE estimator, one should test not only for common trends on the potential outcomes when untreated, but also when treated.[11] While Assumption 6 is not directly testable, it has a natural testable counterpart, suggested by Kim and Lee (2019) who study the reverse difference-in-differences. In particular, one can use groups which stay treated over at least two periods in order to compare the evolution of their outcomes. In particular, one should test whether the following relation holds:

$$E\left(Y_{g,t} - Y_{g,t'} \mid D_{g,t} = 1, D_{g,t'} = 1\right) = E\left(Y_{k,t} - Y_{k,t'} \mid D_{k,t} = 1, D_{k,t'} = 1\right)$$

5.2 Alternative Estimators

Even if non-negative, the weights derived in Section 4.2 do not correspond to each group's sample size. This might make the TWFE estimator difficult to interpret, and implies that it is generally a biased estimator of δ^{TR}. This may motivate the use of other heterogeneity-robust estimators, such as the one suggested by de Chaisemartin and d'Haultfoeuille (2020a), which estimates a quantity which may be of interest: the ATE of switching cells.[12] This paper highlights additional comparisons which can be exploited to construct alternative estimators. I first show that one can build an augmented, unbiased estimator of the ATE of switching cells, while relaxing some of the assumptions de Chaisemartin and d'Haultfoeuille (2020a) impose. Second, I provide an unbiased estimator of the ATT, δ^{TR}, and of the ATE.

[11] Note that one should perform a similar test when using the heterogeneity-robust estimator provided by de Chaisemartin and d'Haultfoeuille (2020a).

[12] Note, however, that it is not an unbiased estimator of the ATT.

5.2.1 Unbiased Estimator of the ATE of Switching Cells

Let us consider the following object, the ATE of all switching cells:

$$\delta^S = E\left[\frac{1}{N_S}\sum_{(i,g,t):t\geq 2, D_{g,t}\neq D_{g,t-1}}[Y_{i,g,t}(1) - Y_{i,g,t}(0)]\right]$$

where $N_S = \sum_{(g,t):t\geq 2, D_{g,t}\neq D_{g,t-1}} N_{g,t}$.

de Chaisemartin and d'Haultfoeuille (2020a) provide an unbiased estimator of this object, under assumptions ensuring the existence of relevant comparison groups. I now show that an unbiased estimator of δ^S can be built while relaxing these requirements. This estimator exploits the additional comparisons highlighted above.

Assumption 11 *(Mean Independence between a Group's Outcome and Other Group Treatments): For all g and t, $E[Y_{g,t}(0)|\mathbf{D}] = E[Y_{g,t}(0)|\mathbf{D}_g]$ and $E[Y_{g,t}(1)|\mathbf{D}] = E[Y_{g,t}(1)|\mathbf{D}_g]$.*

Assumption 12 *(Existence of Stable Groups): For all $t \geq 2$:*

(i) If there is at least one $g \in \{1,...,G\}$ such that $D_{g,t-1} = 0$, $D_{g,t} = 1$, then 1) there exists at least one $g' \neq g$, $g' \in \{1,...,G\}$ such that either $D_{g',t-1} = D_{g',t} = 0$ or 2) g is such that $D_{g,t+1} = 0$ and there exists at least one g' such that $D_{g',t} = D_{g',t+1} = 0$.

(ii) If there is at least one $g \in \{1,...,G\}$ such that $D_{g,t-1} = 1$, $D_{g,t} = 0$, then 1) there exists at least one $g' \neq g$, $g' \in \{1,...,G\}$ such that either $D_{g',t-1} = D_{g',t} = 1$ or 2) g is such that $D_{g,t+1} = 1$ and there exists at least one g' such that $D_{g',t} = D_{g',t+1} = 1$.

I follow de Chaisemartin and d'Haultfoeuille (2020a) in imposing Assumption 11. However, I relax the assumptions they impose with respect to the existence of 'stable groups', which correspond to Assumption 12(i)1) and Assumption 12(ii)1). They require the existence of a group which stays untreated (treated, respectively) over two consecutive periods if there exists a group which joins (leaves, respectively) treatment over these periods. When such assumptions hold, one can identify the ATE in period t for each group switching treatment status between $t-1$ and t, and hence δ^S.

Yet, if these assumptions fail to hold, I now show that one can exploit other comparisons to identify the ATE in period t for each group switching treatment status between $t-1$ and t. In particular, while the estimator suggested by de Chaisemartin and d'Haultfoeuille (2020a) comprises only two kinds of comparisons, at least four two-by-two comparisons could be included.[13] First, one

[13] For simplicity, I do not consider using the double-switcher difference-in-differences comparison.

could use the two-by-two comparisons initially thought of as 'forbidden comparisons', the reverse difference-in-differences. Second, one could include the reverse leaver difference-in-differences.

Assumptions 12(i)2) and 12(ii)2) define the alternative stable groups which are needed to derive an unbiased estimator of δ^S when Assumptions 12(i)1) or 12(ii)1) are not satisfied. Let us now define the augmented estimator. For all $t \in \{2, ..., T\}$ and for all $(d, d') \in \{0,1\}^2$, let

$$N_{d,d',t} = \sum_{g: D_{g,t}=d, D_{g,t-1}=d'} N_{g,t}$$

denote the number of observations with treatment d' at period $t-1$ and d at period t. The following objects will be included in the augmented estimator:

$$DID_{+,g,t} = \mathbb{1}\{D_{g,t} = 1, D_{g,t-1} = 0\}\left[(Y_{g,t} - Y_{g,t-1}) - \sum_{k: D_{k,t}=D_{k,t-1}=0} \frac{N_{k,t}}{N_{0,0,t}}(Y_{k,t} - Y_{k,t-1})\right]$$

$$DID_{-,g,t} = \mathbb{1}\{D_{g,t} = 0, D_{g,t-1} = 1\}\left[\sum_{k: D_{k,t}=1, D_{k,t-1}=1} \frac{N_{k,t}}{N_{1,1,t}}(Y_{k,t} - Y_{k,t-1}) - (Y_{g,t} - Y_{g,t-1})\right]$$

$$DID^R_{+,g,t} = \mathbb{1}\{D_{g,t} = 1, D_{g,t-1} = 0\}\left[(Y_{g,t} - Y_{g,t-1}) - \sum_{k: D_{k,t}=D_{k,t-1}=1} \frac{N_{k,t}}{N_{1,1,t}}(Y_{k,t} - Y_{k,t-1})\right]$$

$$DID^R_{-,g,t} = \mathbb{1}\{D_{g,t} = 0, D_{g,t-1} = 1\}\left[\sum_{k: D_{k,t}=0, D_{k,t-1}=0} \frac{N_{k,t}}{N_{0,0,t}}(Y_{k,t} - Y_{k,t-1}) - (Y_{g,t} - Y_{g,t-1})\right]$$

$DID_{+,g,t}$ and $DID_{-,g,t}$ are defined in a similar fashion as in de Chaisemartin and d'Haultfoeuille (2020a). Following them, we let $DID_{+,g,t} = 0$ if there is no group such that $D_{g,t} = 1$ and $D_{g,t-1} = 0$ or no group such that $D_{g,t} = D_{g,t-1} = 0$. Similarly, we let $DID_{-,g,t} = 0$ if there is no group such that $D_{g,t} = 0$ and $D_{g,t-1} = 1$ or no group such that $D_{g,t} = D_{g,t-1} = 1$. We follow the same rule for the two additional objects. We let $DID^R_{+,g,t} = 0$ if there is no group such that $D_{g,t} = 1$ and $D_{g,t-1} = 0$ or no group such that $D_{g,t} = D_{g,t-1} = 1$. Similarly, we let $DID_{-,g,t} = 0$ if there is no group such that $D_{g,t} = 0$ and $D_{g,t-1} = 1$ or no group such that $D_{g,t} = D_{g,t-1} = 0$.

Finally, let us define the augmented estimator of δ^S:

$$
\begin{aligned}
DID_M^A = \sum_{t=2}^{T} \Bigg[& \frac{1}{N_S} \sum_{g: D_{g,t}=1, D_{g,t-1}=0} N_{g,t} \big(\mathbb{1}\{\exists k \neq g, D_{k,t} = D_{k,t-1} = 0\} DID_{+,g,t} \\
& \qquad + \mathbb{1}\{\nexists k \neq g, D_{k,t} = D_{k,t-1} = 0\} DID^R_{-,g,t+1} \big) \\
& + \frac{1}{N_S} \sum_{g: D_{g,t}=0, D_{g,t-1}=1} N_{g,t} \big(\mathbb{1}\{\exists k \neq g, D_{k,t} = D_{k,t-1} = 1\} DID_{-,g,t} \\
& \qquad + \mathbb{1}\{\nexists k \neq g, D_{k,t} = D_{k,t-1} = 1\} DID^R_{+,g,t+1} \big) \Bigg]
\end{aligned}
$$

Theorem 5 *If Assumption 1, 2, 4-6, 11-12 hold, then* $E[DID_M^A] = \delta^S$.

The proof, following closely the steps of de Chaisemartin and d'Haultfoeuille (2020a), is provided in Appendix. This estimator uses the same comparisons as the one suggested by de Chaisemartin and d'Haultfoeuille (2020a). However, when a given comparison is equal to zero due to the absence of a stable group, it augments it by using either the reverse or reverse leaver difference-in-differences. Note that these two comparisons contrast outcomes between $t+1$ and t, as they identify the ATE of the switching group in the period before it switches treatment status.

5.2.2 Unbiased Estimators of the ATT and ATE

Each two-by-two comparison highlighted in Section 3 is an unbiased estimator of the ATE of the group switching treatment status. Yet, most of them receive a weight equal to zero in the heterogeneity-robust estimator provided by de Chaisemartin and d'Haultfoeuille (2020a). Moreover, the latter is not an unbiased estimator of the ATT, δ^{TR}. I now show that one can exploit the comparisons identifying the ATE of the switching group when it is treated, the standard and reverse leaver difference-in-differences, to build an estimator of δ^{TR}.

Assumption 13 *(Existence of Stable Groups): For all* g, t *such that* $D_{g,t} = 1$, *there exists at least one group* k *and time period* t' *such that* $\omega^S_{g,k,t,t'} = 1$ *or* $\omega^{RL}_{k,g,t',t} = 1$.

Assumption 13 simply specifies that, for each group g treated in period t, there exists at least one group and time period such that a comparison allowing to identify its ATE can be performed. We can then re-weigh appropriately the existing standard and reverse leaver difference-in-differences identifying the ATE corresponding to group g in period t when $D_{g,t} = 1$ in order to build an estimator of δ^{TR}. We obtain the following estimator:

$$DID^{TR} = \frac{1}{N^{(1)}} \sum_{g,t:D_{g,t}=1} N_{g,t} \left[\frac{\sum_{k \neq g} \sum_{t' \neq t} \left[\hat{\beta}^{S}_{g,k,t,t'} + \hat{\beta}^{RL}_{k,g,t',t} \right]}{\sum_{k \neq g} \sum_{t' \neq t} \left[\omega^{S}_{g,k,t,t'} + \omega^{RL}_{k,g,t',t} \right]} \right]$$

Theorem 6 *If Assumption 1, 2, 4-6, 11 and 13 hold, then* $E[DID^{TR}] = \delta^{TR}$.

The proof is immediate, as Section 4.1 establishes that each of the two-by-two comparisons incorporated in DID^{TR} identifies the ATE of group g in t. It thus suffices to re-weigh each available comparison appropriately to form an unbiased estimator of the ATT.[14]

Finally, one could additionally use the reverse and leaver difference-in-differences which identify the ATE of the group switching status at the time where it is untreated in order to identify the ATE, δ^T:

$$\delta^T = E\left[\frac{1}{N} \sum_{(i,g,t)} [Y_{i,g,t}(1) - Y_{i,g,t}(0)] \right]$$

Building an unbiased estimator of this object requires to assume that relevant comparisons can be performed.

Assumption 14 *(Existence of Stable Groups): For all g, t, there exists at least one group k and time period t' such that $\omega^{S}_{g,k,t,t'} = 1$ or $\omega^{R}_{g,k,t',t} = 1$ or $\omega^{RL}_{k,g,t',t} = 1$ or $\omega^{L}_{k,g,t,t'} = 1$.*

We can then define:

$$DID^{T} = \frac{1}{N} \sum_{g,t} N_{g,t} \left[\frac{\sum_{k \neq g} \sum_{t' \neq t} \left[\hat{\beta}^{S}_{g,k,t,t'} + \hat{\beta}^{RL}_{k,g,t',t} + \hat{\beta}^{L}_{k,g,t,t'} + \hat{\beta}^{R}_{g,k,t',t} \right]}{\sum_{k \neq g} \sum_{t' \neq t} \left[\omega^{S}_{g,k,t,t'} + \omega^{RL}_{k,g,t',t} + \omega^{L}_{k,g,t,t'} + \omega^{R}_{g,k,t',t} \right]} \right]$$

where N is the total number of observations in the sample.

Theorem 7 *If Assumption 1, 2, 4-6, 11 and 14 hold, then* $E[DID^{T}] = \delta^{T}$.

Again, the proof is immediate and stems from the fact that each of the comparisons included in DID^T identifies the ATE of group g in period t, irrespective of its treatment status.[15]

[14] Note that the comparisons included in this estimator are also unbiased in the dynamic framework considered in Section 4.3.

[15] Note that I derive this estimator within a static framework. It could be easily adapted to the framework of Section 4.3 if one is interested in the sample ATE even in the presence of dynamic treatment effects. It would require to consider either frameworks in which all comparisons are unbiased estimators of their corresponding ATE, or by including only valid comparisons in the estimator.

6 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 in the presence of heterogeneous treatment effects.

This paper provides necessary and sufficient assumptions for the TWFE estimator to weigh all the ATEs it includes positively. 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 consider the presence of dynamic treatment effects, and show that the TWFE estimator remains valid under plausible conditions.

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. Finally, I show how to combine the highlighted comparisons in order to construct unbiased estimators of the ATT and ATE. I also use them to build an unbiased estimator of the ATE of all switching cells under less stringent assumptions than de Chaisemartin and d'Haultfoeuille (2020a).

As noted by de Chaisemartin and D'Haultfoeuille (2022b), '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 be very similar in practice. 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):

$$\hat{\beta}_{fe} = \frac{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [(Y_{g,t} - Y_{g,t'}) - (Y_{k,t} - Y_{k,t'})]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]} \tag{8}$$

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:

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

Similarly, we can decompose B:

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

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:

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

Hence, we can write:

$$
\hat{\beta}_{fe} = \frac{\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} [\hat{\beta}^S_{g,k,t,t'} + \hat{\beta}^R_{k,g,t',t} + \hat{\beta}^L_{g,k,t,t'} + \hat{\beta}^{RL}_{k,g,t',t} + 2\hat{\beta}^{DS}_{g,k,t,t'}]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]} \tag{9}
$$

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 5 holds if and only if, $\forall g, k \neq g, t, t' \neq t, \forall \mathbf{\Omega}, E[\hat{\beta}^S_{g,k,t,t'} | \mathbf{D}] = E[\Delta_{g,t} | \mathbf{D}] \times \omega^S_{g,k,t,t'}$.

First, it is well known that Assumption 5 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 5 does not hold, it is not true that $E[\hat{\beta}^S_{g,k,t,t'} | \mathbf{D}] = E[\Delta_{g,t} | \mathbf{D}] \times \omega^S_{g,k,t,t'}$, $\forall g, k \neq g, t, t' \neq t$ and $\forall \mathbf{\Omega}$.

Let us assume that Assumption 5 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:

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

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 1-6. Taking the expectation of $\hat{\beta}_{fe}$ conditional on $\mathbf{D}$, we have:

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

where $\beta^C_{g,k,t,t'} \equiv E\left[\hat{\beta}^C_{g,k,t,t'}|\mathbf{D}\right]$,

We can rewrite the numerator:

$$\begin{aligned} &\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} \left[\beta^S_{g,k,t,t'} + \beta^R_{k,g,t',t} + \beta^L_{g,k,t,t'} + \beta^{RL}_{k,g,t',t} + 2\beta^{DS}_{g,k,t,t'}\right] \\ =&\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} \left[E\left[\Delta_{g,t}|\mathbf{D}\right] \times \left[\omega^S_{g,k,t,t'} + \omega^{RL}_{k,g,t',t} + 2\omega^{DS}_{g,k,t,t'}\right]\right. \\ &\left.+E\left[\Delta_{k,t}|\mathbf{D}\right] \times \left[\omega^R_{k,g,t',t} + \omega^L_{g,k,t,t'}\right] + E\left[\Delta_{k,t'}|\mathbf{D}\right] \times \left[2\omega^{DS}_{g,k,t,t'}\right]\right] \end{aligned}$$

The fist term of the sum rewrites:

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

Let us focus on the term $\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} \left[E\left[\Delta_{k,t}|\mathbf{D}\right] \times \left[\omega^R_{k,g,t',t} + \omega^L_{g,k,t,t'}\right]\right]$:

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

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

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

The numerator thus writes:

$$\begin{aligned}&\sum_t \sum_{t' \neq t} \sum_g \sum_{k \neq g} \left[\beta^S_{g,k,t,t'} + \beta^R_{k,g,t',t} + \beta^L_{g,k,t,t'} + \beta^{RL}_{k,g,t',t} + 2\beta^{DS}_{g,k,t,t'}\right]\\ &= \sum_t \sum_g \left[E\left[\Delta_{g,t}|\mathbf{D}\right] \sum_{t' \neq t} \sum_{k \neq g} \left[\omega^S_{g,k,t,t'} + \omega^R_{g,k,t',t} + \omega^L_{k,g,t,t'} + \omega^{RL}_{k,g,t',t} + 2\omega^{DS}_{g,k,t,t'} + 2\omega^{DS}_{k,g,t',t}\right]\right]\end{aligned}$$

We thus have, $\forall \mathbf{\Omega}$:

$$E\left[\hat{\beta}_{fe}|\mathbf{D}\right] = \frac{\sum_t \sum_g \left[E\left[\Delta_{g,t}|\mathbf{D}\right] \sum_{t' \neq t} \sum_{k \neq g} \left[\omega^S_{g,k,t,t'} + \omega^R_{g,k,t',t} + \omega^L_{k,g,t,t'} + \omega^{RL}_{k,g,t',t} + 2\omega^{DS}_{g,k,t,t'} + 2\omega^{DS}_{k,g,t',t}\right]\right]}{\sum_t \sum_{g:D_{g,t}=1} \sum_{k:D_{k,t}=0} \sum_{t' \neq t} [1 - D_{g,t'} + D_{k,t'}]}$$

Following Section 4.1, if Assumption 5 or 6 does not hold, then we would be able to find an $\mathbf{\Omega}$ 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 5 and 6 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 8 and 9 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 8 and 9, 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 8 and 9:

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

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.

Sufficient Conditions in the Presence of Dynamic Treatment Effects 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:

$$
\begin{aligned}
&E[Y_{g,t}(\mathbf{D}_{g,t-1},1)-Y_{g,t'}(\mathbf{D}_{g,t'-1},1)-(Y_{k,t}(\mathbf{D}_{k,t-1},1)-Y_{k,t'}(\mathbf{D}_{k,t'-1},1))|\mathbf{D}]\\
=&E[Y_{g,t}(\mathbf{D}_{k,t-1},1)-Y_{g,t'}(\mathbf{D}_{k,t'-1},1)-(Y_{k,t}(\mathbf{D}_{k,t-1},1)-Y_{k,t'}(\mathbf{D}_{k,t'-1},1))\\
&+[Y_{g,t}(\mathbf{D}_{g,t-1},1)-Y_{g,t}(\mathbf{D}_{k,t-1},1)]-[Y_{g,t'}(\mathbf{D}_{g,t'-1},1)-Y_{g,t'}(\mathbf{D}_{k,t'-1},1)]\,|\mathbf{D}]
\end{aligned}
\tag{10}
$$

Let us first show that, under Assumptions 8-10 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 8 and 9, and the third equality from Assumption 10.2. The first term of the fourth equality cancels out following from Assumption 10.1, while the second term does following Assumptions 8 and 9.

Let us now examine the last two terms of Equation (10) to understand what does the reverse difference-in-differences identifies under Assumption 10. In particular, we have:

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

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 10.2, 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 9. The last equality stems from Assumption 10.1.

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:

$$\begin{aligned}
&E[\Delta_{g,t'}(\mathbf{D}_{g,t'-1})-[\tau_{t'}(\sum_{\ell=1}^{t'-1}D_{g,\ell})-\tau_{t'}(\sum_{\ell=1}^{t'-1}D_{k,\ell})]|\mathbf{D}]\\
&=E[\Delta_{g,t'}(\mathbf{0}_{t'-1})+\tau_{t'}(\sum_{\ell=1}^{t'-1}D_{g,\ell})-[\tau_{t'}(\sum_{\ell=1}^{t'-1}D_{g,\ell})-\tau_{t'}(\sum_{\ell=1}^{t'-1}D_{k,\ell})]|\mathbf{D}]\\
&=E[\Delta_{g,t'}(\mathbf{0}_{t'-1})+\tau_{t'}(\sum_{\ell=1}^{t'-1}D_{k,\ell})]|\mathbf{D}]\\
&=E[\Delta_{g,t'}(\mathbf{D}_{k,t'-1})]|\mathbf{D}]
\end{aligned}\tag{12}$$

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.

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

$$E[\Delta_{g,t'}(\mathbf{0}_{t'-1}) + \tau_{t'}(\sum_{\ell=1}^{t'-1} D_{g,\ell}) + [\tau_t(\sum_{\ell=1}^{t-1} D_{g,\ell}) - \tau_t(\sum_{\ell=1}^{t-1} D_{k,\ell})]|\mathbf{D}] \tag{13}$$

Thus, in this case, the reverse difference-in-differences is a biased estimator of the ATE of group g in period t' conditional on history $\mathbf{D}_{g,t'-1}$. In particular, this is the case of staggered difference-in-differences, when groups are not allowed to leave treatment.

In a similar fashion, one can show that the leaver 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.

A.3 Proof of Theorem 5

We want to prove that DID_M^A is an unbiased estimator of δ^S. Let us write the expectation of DID_M^A:

$$\begin{aligned}
E\left[DID_M^A\right] ==&\sum_{t=1}^{T} E\left[\frac{1}{N_S}\sum_{g:D_{g,t}=1,D_{g,t-1}=0}\left[N_{g,t}\mathbb{1}\{\exists k \neq g, D_{k,t}=D_{k,t-1}=0\}E\left[DID_{+,g,t}|\mathbf{D}\right]\right.\right.\\
&+N_{g,t}\mathbb{1}\{\nexists k \neq g, D_{k,t}=D_{k,t-1}=0\}E\left[DID^R_{-,g,t+1}|\mathbf{D}\right]\Big]\\
&+\frac{1}{N_S}\sum_{g:D_{g,t}=0,D_{g,t-1}=1}\left[N_{g,t}\mathbb{1}\{\exists k \neq g, D_{k,t}=D_{k,t-1}=1\}E\left[DID_{-,g,t}|\mathbf{D}\right]\right.\\
&+\left.\left. N_{g,t}\mathbb{1}\{\nexists k \neq g, D_{k,t}=D_{k,t-1}=1\}E\left[DID^R_{+,g,t+1}|\mathbf{D}\right]\right]\right]
\end{aligned} \tag{14}$$

Let us look separately at each conditional expectation:

$$\begin{aligned}
E\left(DID_{+,g,t}|\mathbf{D}\right) =&E\left(\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}\left[\left(Y_{g,t}-Y_{g,t-1}\right)-\sum_{k:D_{k,t}=D_{k,t-1}=0}\frac{N_{k,t}}{N_{0,0,t}}\left(Y_{k,t}-Y_{k,t-1}\right)\right]|\mathbf{D}\right)\\
=&\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}\left[E\left(Y_{g,t}-Y_{g,t-1}|\mathbf{D}\right)-\sum_{k:D_{k,t}=D_{k,t-1}=0}\frac{N_{k,t}}{N_{0,0,t}}E\left(Y_{k,t}-Y_{k,t-1}|\mathbf{D}\right)\right]
\end{aligned}$$

For every g that $D_{g,t-1}=0$ and $D_{g,t}=1$, we have:

$$E\left(Y_{g,t}-Y_{g,t-1}|\mathbf{D}\right)=E\left(\Delta_{g,t}|\mathbf{D}\right)+E\left(Y_{g,t}(0)-Y_{g,t-1}(0)|\mathbf{D}\right) \tag{15}$$

Following de Chaisemartin and d'Haultfoeuille (2020a), under Assumptions 4, 5 and 11, there exists a real number $\psi_{0,t}$ such that for all g,

$$\begin{aligned}
E\left(Y_{g,t}(0)-Y_{g,t-1}(0)|\mathbf{D}\right) &= E\left(Y_{g,t}(0)-Y_{g,t-1}(0)|\mathbf{D}_g\right)\\
&= E\left(Y_{g,t}(0)-Y_{g,t-1}(0)\right)\\
&= \psi_{0,t}
\end{aligned} \tag{16}$$

where $\mathbf{D}_g$ is the vector collecting treatment status of group g over time. Then, we have:

$$\begin{aligned}
&E\left(DID_{+,g,t}|\mathbf{D}\right)\\
=&\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}\left[E\left(\Delta_{g,t}|\mathbf{D}\right)+E\left(Y_{g,t}(0)-Y_{g,t-1}(0)|\mathbf{D}\right)-\sum_{k:D_{k,t}=D_{k,t-1}=0}\frac{N_{k,t}}{N_{0,0,t}}E\left(Y_{k,t}(0)-Y_{k,t-1}(0)|\mathbf{D}\right)\right]\\
=&\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}\left[E\left(\Delta_{g,t}|\mathbf{D}\right)+\psi_{o,t}(1-\sum_{k:D_{k,t}=D_{k,t-1}=0}\frac{N_{k,t}}{N_{0,0,t}})\right]\\
=&\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}\left[E\left(\Delta_{g,t}|\mathbf{D}\right)+\psi_{o,t}(1-\frac{1}{N_{0,0,t}}\underbrace{\sum_{k:D_{k,t}=D_{k,t-1}=0}N_{k,t})}_{=N_{0,0,,t}}\right]\\
=&\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}E\left[\Delta_{g,t}|\mathbf{D}\right]
\end{aligned} \tag{17}$$

where the first equality follows from (15), the second equality from (16) and the third one uses the definition of $N_{0,0,t}$.

A similar reasoning yields:

$$E\left(DID_{-,g,t}|\mathbf{D}\right)=\mathbb{1}\{D_{g,t}=0,D_{g,t-1}=1\}E\left[\Delta_{g,t}|\mathbf{D}\right] \tag{18}$$

$$E\left(DID^{R}_{+,g,t}|\mathbf{D}\right)=\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}E\left[\Delta_{g,t-1}|\mathbf{D}\right] \tag{19}$$

$$E\left(DID^{R}_{-,g,t}|\mathbf{D}\right)=\mathbb{1}\{D_{g,t}=0,D_{g,t-1}=1\}E\left[\Delta_{g,t-1}|\mathbf{D}\right] \tag{20}$$

Plugging (17), (18), (19) and (20) in (14), we have:

$$\begin{aligned}
E\left[DID^{A}_{M}\right]=&E\left[\sum_{t=2}^{T}\frac{1}{N_S}\left[\sum_{g:D_{g,t}=1,D_{g,t-1}=0}[N_{g,t}\mathbb{1}\{\exists k\neq g,D_{k,t}=D_{k,t-1}=0\}E\left[\Delta_{g,t}|\mathbf{D}\right]\right.\right.\\
&\left.+N_{g,t}\mathbb{1}\{\nexists k\neq g,D_{k,t}=D_{k,t-1}=0\}\mathbb{1}\{D_{g,t+1}=0,D_{g,t}=1\}E\left[\Delta_{g,t}|\mathbf{D}\right]]\right]\\
&+\frac{1}{N_S}\left[\sum_{g:D_{g,t}=0,D_{g,t-1}=1}[N_{g,t}\mathbb{1}\{\exists k\neq g,D_{k,t}=D_{k,t-1}=1\}E\left[\Delta_{g,t}|\mathbf{D}\right]\right.\\
&\left.\left.+N_{g,t}\mathbb{1}\{\nexists k\neq g,D_{k,t}=D_{k,t-1}=1\}\mathbb{1}\{D_{g,t+1}=1,D_{g,t}=0\}E\left[\Delta_{g,t}|\mathbf{D}\right]]\right]\right]
\end{aligned} \tag{21}$$

Under Assumption 12, we have that if there is a group g such that $D_{g,t}=1$ and $D_{g,t-1}=0$ then there either exists at least one comparison group which is untreated, or g is such that $D_{g,t+1} = 0$ and there exists a group k which is untreated. Then, this implies that, for a given g such that $D_{g,t} = 1$ and $D_{g,t-1} = 0$:

$$\mathbb{1}\{\exists k \neq g, D_{k,t} = D_{k,t-1} = 0\} + \mathbb{1}\{\nexists k \neq g, D_{k,t} = D_{k,t-1} = 0\}\mathbb{1}\{D_{g,t+1} = 0, D_{g,t} = 1\} = 1$$

Similarly, for g such that $D_{g,t} = 0$ and $D_{g,t-1} = 1$:

$$\mathbb{1}\{\exists k \neq g, D_{k,t} = D_{k,t-1} = 1\} + \mathbb{1}\{\nexists k \neq g, D_{k,t} = D_{k,t-1} = 1\}\mathbb{1}\{D_{g,t+1} = 1, D_{g,t} = 0\} = 1$$

Thus, Equation (21) writes:

$$\begin{aligned} E\left[DID_M^A\right] &= \sum_{t=2}^{T} E\left[E\left[\frac{1}{N_S}\left(\sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t}\Delta_{g,t} + \sum_{g:D_{g,t}=0,D_{g,t-1}=1} N_{g,t}\Delta_{g,t}\right)|\mathbf{D}\right]\right] \\ &= \delta^S \end{aligned} \tag{22}$$

References

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting event study designs: Robust and efficient estimation," *arXiv preprint arXiv:2108.12419*, 2022.

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," 2021.

de Chaisemartin, Clément and Xavier d'Haultfoeuille, "Two-way fixed effects estimators with heterogeneous treatment effects," *American Economic Review*, 2020, *110* (9), 2964–96.

_ and _ , "Two-way fixed effects regressions with several treatments," *arXiv preprint arXiv:2012.10077*, 2020.

_ and Xavier D'Haultfoeuille, "Difference-in-differences estimators of intertemporal treatment effects," Technical Report, National Bureau of Economic Research 2022.

_ and _ , "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey," 2022.

_ and Xavier d'Haultfoeuille, "Fuzzy differences-in-differences," *The Review of Economic Studies*, 2018, *85* (2), 999–1028.

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, "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends," *American Economic Review: Insights*, 2022, *4* (3), 305–22.

_ **, 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," 2018, *30.*