WORKING PAPERS

N° 1362

September 2022

# "A Note on Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects"

Anaïs Fabre

Toulouse School of Economics

# A Note on Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects*

Anaïs Fabre†

September 21, 2022

**Abstract**

I use a generalized decomposition of the Two-Way Fixed Effects (TWFE) estimator to show that it is a weighted sum of five different types of two-by-two comparisons, with positive weights. I impose the same assumptions as de Chaisemartin and d'Haultfoeuille (2020a) for their heterogeneity-robust estimator to be unbiased. I find that these restrictions are sufficient for each comparison to estimate without bias the Average Treatment Effect (ATE) of the group switching treatment status. Thus, the TWFE estimator weighs each ATE positively, even with heterogeneous treatment effects. I exploit all available comparisons to build unbiased estimators of the ATT and ATE.

*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, 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, it is well-known that 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 exploit the variation in exposure to treatment of potentially many different groups over several 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 (ATE) in each group and period, with weights that may be negative in the presence of heterogeneous treatment effects (de Chaisemartin and d'Haultfoeuille (2020a), Borusyak et al. (2022)). The TWFE estimator may then be negative even when all ATE are positive.

This paper shows that, under the same assumptions as imposed by de Chaisemartin and d'Haultfoeuille (2020a) to prove the unbiasedness of their heterogeneity-robust alternative estimator, the TWFE estimator does not weigh any ATE negatively. I find that each ATE enters proportionally to the number of comparison groups available to identify it. These results are derived without restricting ATE to be homogeneous across groups, nor over time: 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)). Building on Strezhnev (2018), 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 difference-in-differences, (ii) reverse difference-in-differences, (iii) leaver difference-in-differences, (iv) reverse leaver difference-in-differences, 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.

Second, I establish that each of these comparisons is an unbiased estimator of the ATE of the group switching treatment status. This implies that the TWFE estimator is a weighted sum of ATE in different groups and periods, all weighted positively. This result holds under the same assumptions as imposed by de Chaisemartin and d'Haultfoeuille (2020a) to prove the unbiasedness of the alternative estimator they propose in the general case. In particular, it requires that the trends in potential outcomes both when *untreated* and *treated* evolve similarly across groups.[1] Together, these common trends assumptions imply that ATE should follow the same evolution over time across groups. However, they do not restrict treatment effects to be homogeneous over time, nor across groups.

I conclude with the two main implications of this paper for applied researchers. First, it provides sufficient conditions for the TWFE estimator not to be under the threat of negative weights. Whenever using the TWFE estimator or the one provided by de Chaisemartin and d'Haultfoeuille (2020a), one should test not only for parallel trends of potential outcomes across groups when they are untreated, but also when they are treated. I summarize tests available to the practitioner.[2] Second, I take stock of the fact that the TWFE estimator does not weigh each ATE by their sample size and propose alternative estimators. In particular, I exploit the existing valid comparisons highlighted in this paper, which may be of interest in themselves, and 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 than what they impose. I further show that one can build, under adequate restrictions, unbiased estimators of the ATT 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 ATE in each group and period, with weights that may be negative when ATE 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 provide sufficient conditions for the TWFE estimator to weigh positively all ATE, without restricting them to be homogeneous across groups, nor over time. Moreover, I expand the heterogeneity-robust estimator suggested by de Chaisemartin and d'Haultfoeuille (2020a), by highlighting that it gives zero weight to valid comparisons. I show

[1] The latter assumption is not required for their estimator in the staggered difference-in-differences design.
[2] See Roth (2022) for recommended practices to perform such tests.

how to use the latter to build unbiased estimators of the ATT and ATE.

The literature has provided decompositions of the TWFE estimator in terms of two-by-two difference-in-differences comparisons (Goodman-Bacon (2021), Strezhnev (2018)) to highlight 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 under a standard common trends assumption. In the general case, Strezhnev (2018) shows that the TWFE estimator is a uniform average of difference-in-differences comparisons. This paper further decomposes this result to highlight the five types of well-defined two-by-two difference-in-differences comparisons which enter the TWFE estimator, with no negative weights. It then establishes the assumptions under which each comparison is an unbiased estimator of its corresponding ATE, allowing to decompose the TWFE estimator as a weighted sum of ATE, with positive weights.

It should be noted that, following de Chaisemartin and d'Haultfoeuille (2020a), this paper rules out dynamic treatment effects. It thus does not relate to the fast-growing literature studying event-study regressions, or dynamic TWFE regressions (Callaway and Sant'Anna (2021), Gardner (2021), Sun and Abraham (2021), Wooldridge (2021), Borusyak et al. (2022), de Chaisemartin and D'Haultfoeuille (2022a)). Similarly, 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. (2022) provide rich reviews of this recent literature.

# 2 Framework

I use the same set-up 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 without treatment and with treatment, and the average observed outcome in group $g$ at period $t$, respectively.

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

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

**Assumption 2** *(Sharp Design): For all* $(g,t) \in \{1,...,G\} \times \{1,...,T\}$ *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 for Y(0)):* $\forall (g,t) \in \{1,...,G\} \times \{2,...,T\}$,
$E(Y_{g,t}(0) - Y_{g,t-1}(0) | D_{g,1},...,D_{g,T}) = E(Y_{g,t}(0) - Y_{g,t-1}(0))$.

**Assumption 5** *(Strong Exogeneity for Y(1)):* $\forall (g,t) \in \{1,...,G\} \times \{2,...,T\}$,
$E(Y_{g,t}(1) - Y_{g,t-1}(1) | D_{g,1},...,D_{g,T}) = E(Y_{g,t}(1) - Y_{g,t-1}(1))$.

**Assumption 6** *(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* $g$.

**Assumption 7** *(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* $g$.

Assumption 5 is equivalent to Assumption 4 for the potential outcome under treatment. Assumption 7 imposes that the potential outcome with treatment follows the same evolution over time in every group. They correspond to Assumptions 9 and 10 in de Chaisemartin and d'Haultfoeuille (2020a), and are imposed to prove the unbiasedness of their estimator in the general framework. Importantly, Assumptions 6 and 7 together rule out dynamic treatment effects, i.e. it rules out that the treatment effects depend on the number of periods during which the group has been exposed to treatment. I follow de Chaisemartin and d'Haultfoeuille (2020a) in letting such considerations outside the scope of this paper.

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, $\delta^{TR}$ 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 $D_{g,t}$, with $\beta_{fe} = E[\hat{\beta}_{fe}]$. de Chaisemartin and d'Haultfoeuille (2020a) show that, under Assumptions 1-4 and 6, $\beta_{fe}$ is equal to the expectation of a weighted sum of $\Delta_{g,t}$, with potentially negative weights.[3] This result implies that $\hat{\beta}_{fe}$ may be negative even if all ATE are positive. The general intuition is that negative weights arise because the TWFE estimator includes so-called 'forbidden comparisons' (Borusyak et al. (2022), de Chaisemartin and D'Haultfoeuille (2022b)), comparing the outcome evolution of some groups to invalid control groups, such as always-treated units (Borusyak and Jaravel (2017)). In the next sections, I revisit these results by showing that under the additional Assumptions 5 and 7 the TWFE estimator does not weigh negatively any ATE.

# 3 A General Decomposition of the TWFE Estimator

Under which assumptions does the TWFE estimator weigh some ATE negatively? 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:

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

[3] 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 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).

I now decompose this object further, in the spirit of Goodman-Bacon (2021), in order to clarify which comparisons enter the TWFE estimator.[4] 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:

**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.[5] 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

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

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

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, $\hat{\beta}^{RL}_{g,k,t,t'}$:**

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

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, $\hat{\beta}^{S}_{g,k,t,t'}$:**

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

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 of Theorem 1 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 what assumptions.

# 4 Identification

Section 4.1 studies separately each of the five two-by-two comparisons positively weighted by the TWFE estimator. It shows that, under either common trends Assumption 6 or 7, each comparison is an unbiased estimator of the ATE of the group switching treatment status. Section 4.2 highlights the implication of this result for the TWFE estimator. Under both common trends assumptions, it is a weighted sum of ATE, with positive weights, even in the presence of heterogeneous treatment effects across groups or over time.

## 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.

**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, under the above assumptions, the standard difference-in-differences estimator corresponds to the ATT, i.e. the ATE of group $g$ in period $t$. We thus have:

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

where $\mathbf{D}$ is the vector of treatment status history for every group.

**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, under the standard common trends assumption (Goodman-Bacon (2021)). In particular, 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 $\mathrm{E}[\Delta_{k,t}|\mathbf{D}] \neq E[\Delta_{k,t'}|\mathbf{D}]$, i.e. in the presence of heterogeneous treatment effects over time, these comparisons are biased estimators of $E[\Delta_{g,t}]$.

Yet, Kim and Lee (2019) show that, under the common trends assumption on the potential outcomes under treatment, Assumption 7, these two-by-two comparisons are unbiased estimators of the ATE of group $g$ in the pre-treatment period, $t'$. 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 find:

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

Thus, under Assumption 7, we have:

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

**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 can show that the two-by-two comparison identifies the ATE of group $k$ in period $t$ under Assumption 7:

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

We thus have, under Assumption 7:

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

**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 $D_{k,t'}=1$. We are in a case where 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 the two-by-two comparison

identifies the ATE of group $k$ in period $t'$ under Assumption 6:

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

We thus have, under Assumption 6:

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

**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 under the standard common trends assumption on potential outcomes when untreated, Assumption 6:

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

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 group, 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 6. 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

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

the sum of $\Delta_{2,2}$ and $\Delta_{1,3}$. Subtracting the first from the second comparison would thus allow to identify separately the ATE of the first group in period 3.

Overall, under Assumption 6, we thus have:

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

## 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** *Suppose Assumptions 1-7 hold. Then,*

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

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, under Assumptions 1-7, the TWFE estimator does not weigh any ATE negatively. Each of the ATE enters proportionally to the number of relevant comparison groups that can be used to identify it. The weights are thus 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.

# 5 Implications

## 5.1 When is the TWFE estimator heterogeneity-robust?

To the extent that the TWFE estimator remains broadly used in empirical studies, it is crucial to establish under which assumptions all ATE it includes are weighted positively. As shown in Section 4, this requires an additional common trends assumption, Assumption 7. Together, Assumptions 6 and 7 do allow for heterogeneous treatment effects over time and across groups. They however restrict the evolution of the treatment effects: they should have the same trends across groups. This same assumption has to hold for the heterogeneity-robust estimator provided by de Chaisemartin and d'Haultfoeuille (2020a) to be unbiased.

Overall, in a setting with several groups and time periods, whether one considers using the TWFE estimator or the one suggested by de Chaisemartin and d'Haultfoeuille (2020a), the assumption of common trends on the potential outcomes when treated should be tested. While Assumption 7 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:

$$E\left(Y_{g,t} - Y_{g,t-1} \mid D_{g,t} = 1, D_{g,t-1} = 1\right) = E\left(Y_{k,t} - Y_{k,t-1} \mid D_{k,t} = 1, D_{k,t-1} = 1\right)$$

More generally, one can use the placebo estimator provided by de Chaisemartin and d'Haultfoeuille (2020a), $DID_M^{pl}$. This estimator compares the changes in outcomes between periods $t-2$ and $t-1$ for two groups sharing the same treatment status in these periods, but differing in $t$: one group switches treatment status between period $t-1$ and $t$, while the other's treatment status stays the same. They provide a Stata package to implement such a test.

## 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 $\delta^{TR}$. 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.[7] 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, $\delta^{TR}$, and of the ATE.

### 5.2.1 Unbiased Estimator of the ATE of Switching Cells

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

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

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

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

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 $\delta^S$ can be built while relaxing these requirements. This estimator exploits the additional comparisons highlighted above.

**Assumption 8** *(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 9** *(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 8. However, I relax the assumptions they impose with respect to the existence of 'stable groups', which correspond to Assumption 9(i)1) and Assumption 9(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 $\delta^S$.

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.[8] First, one 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 9(i)2) and 9(ii)2) define the alternative stable groups which are needed to derive an unbiased estimator of $\delta^S$ when Assumptions 9(i)1) or 9(ii)1) are not satisfied. Let

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

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 $\delta^S$:

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

**Theorem 3** *If Assumption 1, 2, 4-9 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, $\delta^{TR}$. 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 $\delta^{TR}$.

**Assumption 10** *(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 10 simply specifies that, for each group $g$ treated in period $t$, there exist 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 $\delta^{TR}$. 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 4** *If Assumption 1, 2, 4-8 and 10 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.

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, $\delta^{T}$:

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

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

**Assumption 11** *(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* $\omega^{L}_{k,g,t,t'} = 1$.

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

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

**Theorem 5** *If Assumption 1, 2, 4-8 and 11 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.

# 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 ATE in the presence of heterogeneous treatment effects.

This paper shows that the assumptions imposed by de Chaisemartin and d'Haultfoeuille (2020a) to prove the unbiasdeness of their heterogeneity-robust alternative estimator are sufficient for the TWFE estimator to weigh all the ATE it includes positively. This guarantees that the TWFE estimator is heterogeneity-robust.

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 may be a useful first step in explaining 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.

# References


**Borusyak, Kirill and Xavier Jaravel**, "Revisiting event study designs," *Available at SSRN 2826228*, 2017.

**_ , _ , and Jann Spiess**, "Revisiting event study designs: Robust and efficient estimation," *arXiv preprint arXiv:2108.12419*, 2022.

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

**_ , 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.

**Gardner, John**, "Two-stage differences in differences," *Unpublished working paper*, 2021.

**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.

**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," *arXiv preprint arXiv:2201.01194*, 2022.

**Strezhnev, Anton**, "Semiparametric weighting estimators for multi-period differencein-differences designs," 2018, *30.*

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

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

# 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 Proof of Theorem 2

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 we can rewrite the numerator:

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

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:

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

## A.3 Proof of Theorem 3

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

$$\begin{aligned} E\left[DID_M^A\right] == \sum_{t=1}^{T} E\Bigg[ & \frac{1}{N_S} \sum_{g:D_{g,t}=1,D_{g,t-1}=0} [N_{g,t}\mathbb{1}\{\exists k \neq g, D_{k,t} = D_{k,t-1} = 0\} E\left[DID_{+,g,t}|\mathbf{D}\right] \\ & + N_{g,t}\mathbb{1}\{\nexists k \neq g, D_{k,t} = D_{k,t-1} = 0\} E\left[DID^R_{-,g,t+1}|\mathbf{D}\right]] \\ & + \frac{1}{N_S} \sum_{g:D_{g,t}=0,D_{g,t-1}=1} [N_{g,t}\mathbb{1}\{\exists k \neq g, D_{k,t} = D_{k,t-1} = 1\} E\left[DID_{-,g,t}|\mathbf{D}\right] \\ & + N_{g,t}\mathbb{1}\{\nexists k \neq g, D_{k,t} = D_{k,t-1} = 1\} E\left[DID^R_{+,g,t+1}|\mathbf{D}\right]]\Bigg] \end{aligned} \tag{10}$$

Let us look separately at each conditional expectation:

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

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

$$E\left(Y_{g,t} - Y_{g,t-1}|\mathbf{D}\right) = E\left(\Delta_{g,t}|\mathbf{D}\right) + E\left(Y_{g,t}(0) - Y_{g,t-1}(0)|\mathbf{D}\right) \tag{11}$$

Following de Chaisemartin and d'Haultfoeuille (2020a), under Assumptions 4, 6 and 8, there exists a real number $\psi_{0,t}$ such that for all $g$,

$$\begin{aligned} E\left(Y_{g,t}(0) - Y_{g,t-1}(0)|\mathbf{D}\right) &= E\left(Y_{g,t}(0) - Y_{g,t-1}(0)|\mathbf{D}_g\right) \\ &= E\left(Y_{g,t}(0) - Y_{g,t-1}(0)\right) \\ &= \psi_{0,t} \end{aligned} \tag{12}$$

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

$$
\begin{aligned}
&E\left(DID_{+,g,t}|\mathbf{D}\right)\\
=&\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}\left[E\left(\Delta_{g,t}|\mathbf{D}\right)+E\left(Y_{g,t}(0)-Y_{g,t-1}(0)|\mathbf{D}\right)-\sum_{k:D_{k,t}=\overline{D}_{k,t-1}=0}\frac{N_{k,t}}{N_{0,0,t}}E\left(Y_{k,t}(0)-Y_{k,t-1}(0)|\mathbf{D}\right)\right]\\
=&\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}\left[E\left(\Delta_{g,t}|\mathbf{D}\right)+\psi_{o,t}(1-\sum_{k:D_{k,t}=\overline{D}_{k,t-1}=0}\frac{N_{k,t}}{N_{0,0,t}})\right]\\
=&\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}\left[E\left(\Delta_{g,t}|\mathbf{D}\right)+\psi_{o,t}(1-\frac{1}{N_{0,0,t}}\underbrace{\sum_{k:D_{k,t}=\overline{D}_{k,t-1}=0}N_{k,t}}_{=N_{0,0,,t}})\right]\\
=&\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}E\left[\Delta_{g,t}|\mathbf{D}\right]
\end{aligned}
\tag{13}
$$

where the first equality follows from (11), the second equality from (12) and the third one uses the definition of $N_{0,0,t}$.

A similar reasoning yields:

$$E\left(DID_{-,g,t}|\mathbf{D}\right)=\mathbb{1}\{D_{g,t}=0,D_{g,t-1}=1\}E\left[\Delta_{g,t}|\mathbf{D}\right] \tag{14}$$

$$E\left(DID^{R}_{+,g,t}|\mathbf{D}\right)=\mathbb{1}\{D_{g,t}=1,D_{g,t-1}=0\}E\left[\Delta_{g,t-1}|\mathbf{D}\right] \tag{15}$$

$$E\left(DID^{R}_{-,g,t}|\mathbf{D}\right)=\mathbb{1}\{D_{g,t}=0,D_{g,t-1}=1\}E\left[\Delta_{g,t-1}|\mathbf{D}\right] \tag{16}$$

Plugging (13), (14), (15) and (16) in (10), we have:

$$
\begin{aligned}
E\left[DID^{A}_{M}\right]=&E\Bigg[\sum_{t=2}^{T}\frac{1}{N_S}\Bigg[\sum_{g:D_{g,t}=1,D_{g,t-1}=0}\Big[N_{g,t}\mathbb{1}\{\exists k\neq g,D_{k,t}=D_{k,t-1}=0\}E\left[\Delta_{g,t}|\mathbf{D}\right]\\
&+N_{g,t}\mathbb{1}\{\nexists k\neq g,D_{k,t}=D_{k,t-1}=0\}\mathbb{1}\{D_{g,t+1}=0,D_{g,t}=1\}E\left[\Delta_{g,t}|\mathbf{D}\right]\Big]\Bigg]\\
&+\frac{1}{N_S}\Bigg[\sum_{g:D_{g,t}=0,D_{g,t-1}=1}\Big[N_{g,t}\mathbb{1}\{\exists k\neq g,D_{k,t}=D_{k,t-1}=1\}E\left[\Delta_{g,t}|\mathbf{D}\right]\\
&+N_{g,t}\mathbb{1}\{\nexists k\neq g,D_{k,t}=D_{k,t-1}=1\}\mathbb{1}\{D_{g,t+1}=1,D_{g,t}=0\}E\left[\Delta_{g,t}|\mathbf{D}\right]\Big]\Bigg]\Bigg]
\end{aligned}
\tag{17}
$$

Under Assumption 9, 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 (17) writes:

$$\begin{aligned} E\left[DID_M^A\right] &= \sum_{t=2}^{T} E\left[E\left[\frac{1}{N_S}\left(\sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t}\Delta_{g,t} + \sum_{g:D_{g,t}=0,D_{g,t-1}=1} N_{g,t}\Delta_{g,t}\right)|\mathbf{D}\right]\right] \\ &= \delta^S \end{aligned} \tag{18}$$