NBER WORKING PAPER SERIES

TWO-WAY FIXED EFFECTS AND DIFFERENCES-IN-DIFFERENCES ESTIMATORS WITH SEVERAL TREATMENTS

Clément de Chaisemartin
Xavier D'Haultfoeuille

Working Paper 29734
http://www.nber.org/papers/w29734

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2022

Several of this paper's ideas arose during conversations with Enrico Cantoni, Angelica Meinhofer, Vincent Pons, Jimena Rico-Straffon, Marc Sangnier, Oliver Vanden Eynde, and Liam Wren-Lewis who shared with us their interrogations, and sometimes their referees' interrogations, on two-way fixed effects regressions with several treatments. We are grateful to them for those stimulating conversations. We are grateful to Yubo Wei for his excellent work as a research assistant. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Two-way Fixed Effects and Differences-in-Differences Estimators with Several Treatments
Clément de Chaisemartin and Xavier D'Haultfoeuille
NBER Working Paper No. 29734
February 2022
JEL No. C21,C23

ABSTRACT

We study regressions with period and group fixed effects and several treatment variables. Under a parallel trends assumption, the coefficient on each treatment identifies the sum of two terms. The first term is a weighted sum of the effect of that treatment in each group and period, with weights that may be negative and sum to one. The second term is a sum of the effects of the other treatments, with weights summing to zero. Accordingly, coefficients in those regressions are not robust to heterogeneous effects, and may be contaminated by the effect of other treatments. We propose alternative differences-in-differences estimators. To estimate, say, the effect of the first treatment, our estimators compare the outcome evolution of a group whose first treatment changes while its other treatments remain unchanged, to control groups whose treatments all remain unchanged, and with the same baseline treatments or treatments' history as the switching group. Those carefully selected comparisons are robust to heterogeneous effects, and do not suffer from the contamination problem.

Clément de Chaisemartin
Department of Economics
University of California at Santa Barbara
Santa Barbara, CA 93106
and NBER
clementdechaisemartin@ucsb.edu

Xavier D'Haultfoeuille
CREST
5 avenue Henry Le Chatelier
91764 Palaiseau cedex
FRANCE
xavier.dhaultfoeuille@ensae.fr

Two-way Fixed Effects and Differences-in-Differences Estimators with Several Treatments*

Clément de Chaisemartin Xavier D'Haultfœuille†

Abstract

We study regressions with period and group fixed effects and several treatment variables. Under a parallel trends assumption, the coefficient on each treatment identifies the sum of two terms. The first term is a weighted sum of the effect of that treatment in each group and period, with weights that may be negative and sum to one. The second term is a sum of the effects of the other treatments, with weights summing to zero. Accordingly, coefficients in those regressions are not robust to heterogeneous effects, and may be contaminated by the effect of other treatments. We propose alternative differences-in-differences estimators. To estimate, say, the effect of the first treatment, our estimators compare the outcome evolution of a group whose first treatment changes while its other treatments remain unchanged, to control groups whose treatments all remain unchanged, and with the same baseline treatments or treatments' history as the switching group. Those carefully selected comparisons are robust to heterogeneous effects, and do not suffer from the contamination problem.

(JEL C21, C23)

1 Introduction

To estimate treatment effects, researchers often use panels of groups (e.g. counties, regions), and estimate two-way fixed effect (TWFE) regressions, namely regressions of the outcome variable on group and time fixed effects and the treatment. de Chaisemartin and D'Haultfœuille (2020)

*Several of this paper's ideas arose during conversations with Enrico Cantoni, Angelica Meinhofer, Vincent Pons, Jimena Rico-Straffon, Marc Sangnier, Oliver Vanden Eynde, and Liam Wren-Lewis who shared with us their interrogations, and sometimes their referees' interrogations, on two-way fixed effects regressions with several treatments. We are grateful to them for those stimulating conversations. We are grateful to Yubo Wei for his excellent work as a research assistant.

†de Chaisemartin: Sciences Po (email: clement.dechaisemartin@sciencespo.fr); D'Haultfœuille: CREST-ENSAE (email: xavier.dhaultfoeuille@ensae.fr). Xavier D'Haultfœuille thanks the hospitality of PSE where this research was conducted.

have found that almost 20% of empirical papers published by the American Economic Review (AER) from 2010 to 2012 estimate such regressions.

de Chaisemartin and D'Haultfœuille (2020) and Borusyak and Jaravel (2017) have shown that with one treatment in the regression, under a parallel trends assumption TWFE regressions identify a weighted sum of the treatment effects of treated (g,t) cells, with weights that may be negative and sum to one.[1] Because of the negative weights, the treatment coefficient in such regressions is not robust to heterogeneous treatment effects across groups and time periods: it may be, say, negative, even if the treatment effect is strictly positive in every (g,t) cell.

However, in 18% of the TWFE papers in the 2010-2012 AER survey in de Chaisemartin and D'Haultfœuille (2020), the TWFE regression has several treatment variables. By including several treatments, researchers hope to estimate the effect of each treatment holding the other treatments constant. For instance, when studying the effect of marijuana laws, as in Meinhofer et al. (2021), one may want to separate the effect of medical and recreational laws. To do so, one may estimate a regression of the outcome of interest in state g and year t on state fixed effects, year fixed effects, an indicator for whether state g has a medical law in year t, and an indicator for whether state g has a recreational law in year t. In this example, the two treatments are binary, they can switch on but never switch off, a situation referred to as a staggered adoption design, and the second treatment always comes after the first. TWFE regressions have also been used in more complicated designs. For instance, Hotz and Xiao (2011) run TWFE regressions of measures of daycare quality in state g and year t on two daycare regulations in state g and year t: the minimum number of years of schooling required to be a daycare director and the minimum staff-child ratio. Both treatments are non-binary, and can increase or decrease over time.

In this paper, we start by investigating what TWFE regressions with several treatments identify. We show that under a parallel trends assumption, the coefficient on each treatment identifies the sum of two terms. The first term is a weighted sum of the effect of that treatment in each group and period, with weights that may be negative and sum to one. This first term also appears in decompositions of TWFE regressions with only one treatment, but we show that TWFE regressions with several treatments often have more and larger negative weights than TWFE regressions with only one treatment, and are therefore less robust to heterogeneous effects. The second term is a sum of the effects of the other treatments, with weights summing to zero. Accordingly, treatment coefficients in TWFE regressions with several treatments may also be contaminated by the effect of other treatments, an issue that was not present in TWFE regressions with one treatment. As the weights sum to zero, this second term disappears if the effect of the second treatment is homogeneous, but it is often implausible that this effect

[1] When the treatment is binary and staggered, Goodman-Bacon (2021) shows that negative weights arise because the TWFE regression leverages DIDs using groups treated at both periods as control groups.

is homogeneous. In a simple example with two periods and two treatments, we show that this contamination problem may arise because the coefficient on the first treatment may leverage a difference-in-differences (DID) comparing the outcome evolution of a group going from untreated to receiving both treatments to the outcome evolution of a group going from untreated to receiving the second treatment. If the effect of the second treatment is the same in the two groups, its effect in both groups cancel each other out in this DID. But if the effects of the second treatment differ in the two groups, they do not cancel each other out, and they contaminate the coefficient on the first treatment. The weights attached to any TWFE regression with several treatments can be computed by the `twowayfeweights` Stata and R packages.

Then, we propose alternative DID estimators that rely on common trends assumptions, like TWFE coefficients, but that are robust to heterogeneous effects and do not suffer from the contamination problem, unlike TWFE coefficients. To do so, we start by assuming that the treatment does not have dynamic effects: the current outcome is only affected by the current value of the treatment, not by past treatments. Under this assumption, we propose an estimator that generalizes the DID_M estimator in de Chaisemartin and D'Haultfœuille (2020), and that can be used in applications where the treatment switches on and off, and in instances where the treatment is discrete rather than binary. To isolate the effect of the first treatment, our new estimator compares the $t-1$-to-t outcome evolution, between groups whose first treatment changes from $t-1$ to t while their other treatments do not change, and groups whose treatments all remain the same and that had the same treatments as the switching groups in period $t-1$. In the Hotz and Xiao (2011) example, to isolate the effect of the staff to child ratio treatment, our estimator compares the $t-1$ to t outcome evolution of states whose staff to child ratio changes but whose years-of-schooling requirement for daycare directors does not change, to the outcome evolution of states whose two treatments remain the same, and with the same treatments as the switching states in $t-1$. Restricting comparisons to groups whose other treatments do not change avoids the contamination problem. Restricting the control group to groups that do not experience any change in their treatments and with the same treatments as the switchers in $t-1$ ensures that our estimator is robust to heterogeneous treatment effects and only relies on parallel trends assumptions. Our new estimator is computed by the `did_multiplegt` Stata and R packages.

Then, we propose alternative estimators that allow dynamic effects: the current outcome may be affected by past values of the treatment. To do so, we restrict attention to binary treatments following a staggered adoption design, and such that the second treatment always comes after the first, as in the marijuana laws example. In such instances, one can rely on existing estimators to isolate the effect of the first treatment, by restricting the sample to (g,t) cells that have not received the second treatment yet, and computing the estimators for the one-binary-and-staggered-treatment case proposed, say, by Callaway and Sant'Anna (2021). One can also

rely on existing estimators to estimate the combined effect of the two treatments: one can define a new treatment equal to the sum of the two treatments, and compute the estimators in de Chaisemartin and D'Haultfœuille (2021), that allow for dynamic effects and that can be used with a non-binary discrete treatment.

On the other hand, existing estimators cannot be used to isolate the effect of the second treatment. We propose a novel estimator for that purpose. Our estimator compares the outcome evolutions of groups that start receiving the second treatment and groups that have not received it yet, restricting such comparisons to groups that all started receiving the first treatment at the same date. Such comparisons are valid under a parallel trends assumption, and under the assumption that the effect of the first treatment follows the same evolution over time in every group. This second assumption may be strong, but it is necessary to isolate the effect of the second treatment: under parallel trends alone, one can only identify the combined effect of the two treatments. Those two assumptions are partly testable, by comparing the outcome evolution of groups adopting and not adopting the second treatment, before adopters adopt. Our proposed estimator may, however, not always be applicable. For instance, in the marijuana laws example, it requires that for every state adopting a recreational law, one can find a non-adopting state that adopted a medical law in the same year. Accordingly, we propose two other estimators. The first relies on the assumption that in each group, the effect of the first treatment increases or decreases linearly with the duration of exposure. The second relies on the assumption that the evolution of the effect of the first treatment is the same in every group, and does not depend on the date at which the first treatment was adopted. All our proposed estimators are computed by the `did_multiplegt` Stata package. Our robust-to-dynamic-effects estimators can easily be extended to applications with two binary and staggered treatments such that neither treatment systematically comes after the other one. We briefly discuss extensions to more general designs, but mostly leave them for future work.

Finally, we use our results to revisit Hotz and Xiao (2011). We find that some of the TWFE regressions with several treatments in that paper estimate weighted sums of effects with very large negative weights attached to them, both on a given treatment's own effects, but also on the effects of the other treatments in the regression. The authors' regression implicitly rules out dynamic effects so we follow their specification and compute our heterogeneity-robust DID estimator that also rules out dynamic effects. We find that for the years-of-schooling treatment, our estimator is seven times smaller than, and significantly different from, the TWFE coefficient.[2]

Our paper is related to the recent literature showing that TWFE regressions with one treatment variable may not be robust to heterogeneous effects (see de Chaisemartin and D'Haultfœuille, 2020; Borusyak and Jaravel, 2017; Goodman-Bacon, 2021). Our paper is also related to several

[2]There are too few states changing their staff-to-child-ratio treatment without changing their years-of-schooling treatment for us to compute an heterogeneity-robust DID estimator of that second treatment.

papers that have considered the causal interpretation of OLS regression coefficients with several treatments (see Hull, 2018; Sun and Abraham, 2021; Goldsmith-Pinkham, Hull and Kolesár, 2021). We discuss those papers in more details later in the paper, but for now we just note that when the treatments are indicators for whether group g has started receiving a binary and staggered treatment ℓ periods ago, our decomposition of the TWFE regression reduces to one of the decompositions in Sun and Abraham (2021). Accordingly, our decomposition extends their result to situations where the different treatments in the regression are different policies that could be non-binary and non-staggered, rather than indicators for having received a single binary and staggered policy ℓ periods ago. Finally, the alternative estimator we propose for the case without dynamic effects builds upon the DID_{M} estimator in de Chaisemartin and D'Haultfœuille (2020), while the alternative estimator we propose for the case with dynamic effects builds upon the estimators in Callaway and Sant'Anna (2021).

The remainder of the paper is organized as follows. Section 2 presents the set up. Section 3 presents our decomposition results for TWFE regressions with several treatment. Section 4 presents our alternative estimators. Section 5 presents our empirical application.

2 Set up

We assume that there are G groups and T periods. For every $(g,t) \in \{1,...,G\} \times \{1,...,T\}$, let $N_{g,t}$ denote the number of observations in group g at period t, and let $N = \sum_{g,t} N_{g,t}$ be the total number of observations. We are interested in the effect of K treatments. We assume that the treatments are binary, though our result can be extended to any ordered treatment. Then, for every $(k,i,g,t) \in \{1,...,K\} \times \{1,...,N_{g,t}\} \times \{1,...,G\} \times \{1,...,T\}$, let $D^k_{i,g,t}$ denote the value of treatment k for observation i in group g at period t. For any $\boldsymbol{d} \in \{0,1\}^K$, let $Y_{i,g,t}(\boldsymbol{d})$ denote her potential outcome if $(D^1_{i,g,t},...,D^K_{i,g,t}) = \boldsymbol{d}$. The outcome of observation i in group g and period t is $Y_{i,g,t} = Y_{i,g,t}(D^1_{i,g,t},...,D^K_{i,g,t})$. Importantly, our notation does not necessarily rule out dynamic effects of past treatments on the outcome. The K treatments may for instance include lags of the same treatment variables. We discuss this issue in more details after Theorem 1 below.

For all (g,t), all $k \in \{0,...,K\}$, and all $\boldsymbol{d} \in \{0,1\}^K$, let

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

$D^k_{g,t}$ denotes the average of treatment k in group g at period t, while $Y_{g,t}(\boldsymbol{d})$ and $Y_{g,t}$ respectively denote the average potential outcomes and the average observed outcome in group g at period t. Let also $D_{g,t} = (D^k_{g,t})_{k \in \{1,...,K\}}$ denote a vector stacking together the K average treatments of group g at period t.

We consider the treatments and potential outcomes of each (g,t) cell as random variables. For instance, aggregate random shocks may affect the potential outcomes of group g at period t, and that cell's treatments may also be random. The expectations below are taken with respect to the distribution of those random variables.

Throughout the paper, we maintain the following assumptions.

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* $k \in \{1,...,K\}$, *all* $(g,t) \in \{1,...,G\} \times \{1,...,T\}$, *and all* $i \in \{1,...,N_{g,t}\}$, $D^k_{i,g,t} = D^k_{g,t}$.

Assumption 3 *(Independent groups) The vectors* $((Y_{g,t}(\boldsymbol{d}))_{\boldsymbol{d}\in\{0,1\}^K}, (D^k_{g,t})_{k\in\{1,...,K\}})$ *are mutually independent.*

Let $\mathbf{0} = (0,...,0)$ denote the vector of K zeros.

Assumption 4 *(Strong exogeneity and common trends) For all* $(g,t) \in \{1,...,G\} \times \{2,...,T\}$, $E(Y_{g,t}(\mathbf{0}) - Y_{g,t-1}(\mathbf{0})|D_{g,1},...,D_{g,T})$ *does not vary across* g.

Assumption 2 holds when $N_{g,t} = 1$ or if the treatments are group-level variables, for instance county- or state-laws. Assumption 3 requires that potential outcomes and treatments of different groups be independent, but it allows these variables to be correlated over time within each group. This is a commonly-made assumption in DID analysis, where standard errors are usually clustered at the group level (see Bertrand, Duflo and Mullainathan, 2004). To understand better Assumption 4, let us state two conditions that, together, are sufficient for it to hold, and that are easily interpretable:

1. $E(Y_{g,t}(\mathbf{0}) - Y_{g,t-1}(\mathbf{0})|D_{g,1},...,D_{g,T}) = E(Y_{g,t}(\mathbf{0}) - Y_{g,t-1}(\mathbf{0}))$.
2. $\forall t \geq 2$, $E(Y_{g,t}(\mathbf{0}) - Y_{g,t-1}(\mathbf{0}))$ does not vary across g.

Point 1 is related to the strong exogeneity condition in panel data models. It requires that the shocks affecting group g's never-treated outcome be mean independent of group g's treatments. For instance, this rules out cases where a group gets treated because it experiences negative shocks, the so-called Ashenfelter's dip (see Ashenfelter, 1978). Point 2 requires that in every group, the expectation of the never-treated outcome follow the same evolution over time. It is a generalization of the standard common trends assumption in DID models (see, e.g., Abadie, 2005).

We now define the FE regression described in the introduction.[3]

[3] Throughout the paper, we assume that the treatments $D^k_{g,t}$ in Regression 1 are not collinear with the other independent variables in those regressions, so $\widehat{\beta}_{fe}$ is well-defined.

Regression 1 *(Fixed-effects regression with K treatments)*

Let $\beta_{fe} = E\left[\widehat{\beta}_{fe}\right]$, where $\widehat{\beta}_{fe}$ is the coefficient of $D^1_{g,t}$ in an OLS regression of $Y_{i,g,t}$ on group fixed effects, period fixed effects, and the vector $D_{g,t}$.

Let $\boldsymbol{D}$ be the vector $(D_{g,t})_{(g,t)\in\{1,...,G\}\times\{1,...,T\}}$ collecting all the treatments in all the (g,t) cells. let $\boldsymbol{D}_g = (D_{1,g},...,D_{T,g})$ be the vector collecting all the treatments in group g. Let $N_1 = \sum_{i,g,t} D^1_{i,g,t}$ denote the number of units receiving the first treatment. Let $D^{-1}_{g,t} = (D^2_{g,t},...,D^K_{g,t})$ denote a vector stacking together the treatments of cell (g,t), excluding treatment 1. Let $\varepsilon_{g,t}$ denote the residual of observations in cell (g,t) in the regression of $D^1_{g,t}$ on group and period fixed effects and $D^{-1}_{g,t}$:

$$D^1_{g,t} = \widehat{\alpha} + \widehat{\gamma}_g + \widehat{\nu}_t + (D^{-1}_{g,t})'\widehat{\zeta} + \varepsilon_{g,t}. \tag{1}$$

One can show that if the regressors in Regression 1 are not collinear, the average value of $\varepsilon_{g,t}$ across all treated (g,t) cells differs from 0: $\sum_{(g,t):D^1_{g,t}=1}(N_{g,t}/N_1)\varepsilon_{g,t} \neq 0$. Then we let $w_{g,t}$ denote $\varepsilon_{g,t}$ divided by that average:

$$w_{g,t} = \frac{\varepsilon_{g,t}}{\sum_{(g,t):D^1_{g,t}=1}(N_{g,t}/N_1)\varepsilon_{g,t}}.$$

3 Decomposition results

3.1 Two treatment variables

For expositional purposes, we begin by considering the case with two treatments. This excludes the case where the TWFE regression includes two treatments and their interaction, but that case is covered by our results in the next subsection, where we allow for three (or more) treatments in the regression, one of which could be the interaction of two treatments. For any $(g,t) \in \{1,...,G\}\times\{1,...,T\}$, and for any $(d_1,d_2)\in\{0,1\}^2$, let

$$\Delta^{d_1,d_2}_{g,t} = \frac{1}{N_{g,t}}\sum_{i=1}^{N_{g,t}}\left[Y_{i,g,t}(d_1,d_2) - Y_{i,g,t}(0,0)\right]$$

denote the average effect, in cell (g,t), of moving the first treatment from zero to d_1 and the second treatment from zero to d_2. Let also

$$\Delta^1_{g,t}(D^2_{g,t}) = \frac{1}{N_{g,t}}\sum_{i=1}^{N_{g,t}}\left[Y_{i,g,t}(1,D^2_{g,t}) - Y_{i,g,t}(0,D^2_{g,t})\right]$$

denote the average effect, in cell (g,t), of moving the first treatment from zero to one while keeping the second treatment at its observed value.

Theorem 1 *Suppose that Assumptions 1-4 hold and $K = 2$. Then,*

$$\beta_{fe} = E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} \Delta^1_{g,t}(D^2_{g,t}) + \sum_{(g,t):D^2_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} \Delta^{0,1}_{g,t}\right].$$

Moreover, $\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} = 1$ *and* $\sum_{(g,t):D^2_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} = 0$.

Theorem 1 shows that the coefficient of $D^1_{g,t}$ identifies the sum of two terms. The first term is a weighted sum of the average effect of moving $D^1_{g,t}$ from 0 to 1 while keeping $D^2_{g,t}$ at its observed value, across all (g,t) such that $D^1_{g,t} = 1$, and with weights summing to 1. The second term is a weighted sum of the effect of moving $D^2_{g,t}$ from 0 to 1 while keeping $D^1_{g,t}$ at 0, across all the (g,t) such that $D^2_{g,t} = 1$, and with weights summing to 0. If the effect of the second treatment is constant, meaning that there is a real number δ^2 such that for all (g,t), $\Delta^{0,1}_{g,t} = \delta^2$, this second term disappears. Then, Theorem 1 becomes equivalent to Theorem S4 in the Web Appendix of de Chaisemartin and D'Haultfœuille (2020), with $D^2_{g,t}$ playing the role of a control variable in the two-way fixed effects regression. Theorem 1 implies that on top of not being robust to heterogeneous treatment effects, β_{fe} may also be contaminated by the effect of the second treatment on the outcome. On the other hand, if the effect of the first and second treatments are both constant ($\Delta^1_{g,t}(D^2_{g,t}) = \delta^1$ and $\Delta^{0,1}_{g,t} = \delta^2$ for some real numbers δ^1 and δ^2), then $\widehat{\beta}_{fe}$ is unbiased for δ^1. Importantly, Theorem 1 and Theorem 2 below can easily be extended to the case where the treatments are non-binary. Then, the causal effects $\Delta^1_{g,t}(D^2_{g,t})$ and $\Delta^{0,1}_{g,t}$ just need to be replaced by slopes of the potential outcome function, as in Section 3.2 of the Web Appendix of de Chaisemartin and D'Haultfœuille (2020).

The contamination bias appears in Theorem 1, because $\widehat{\beta}_{fe}$ may leverage "forbidden comparisons", using the terminology coined by Borusyak and Jaravel (2017). Here, forbidden comparisons are differences-in-differences (DIDs) comparing the outcome evolution of a group that starts receiving the first and the second treatment to the outcome evolution of a group that starts receiving the second treatment only. For instance, assume that there are four groups and two time periods. No group is treated at period 1, and at period 2 groups 2 and 4 receive the first treatment while 3 and 4 receive the second treatment. Then, using the fact that a TWFE regression with two periods is equivalent to a first-difference regression, and the fact that the first differences of the two treatments are orthogonal, it is easy to show that

$$\widehat{\beta}_{fe} = \frac{1}{2}\left(Y_{2,2} - Y_{2,1} - (Y_{1,2} - Y_{1,1})\right) + \frac{1}{2}\left(Y_{4,2} - Y_{4,1} - (Y_{3,2} - Y_{3,1})\right).$$

The first DID in the previous display compares a group that starts receiving the first treatment at period 2 to a group that does not receive any treatment. Under a parallel trends assumption on the untreated outcome $Y_{g,t}(0,0)$, that DID identifies the effect of the first treatment in group 2 at period 2. On the other hand, the second DID compares a group that starts receiving the first and

second treatment at period 2 to a group that only starts receiving the second treatment. Under a parallel trends assumption on the untreated outcome $Y_{g,t}(0,0)$, that second DID identifies the sum of three terms. The first is $E(Y_{4,2}(0,1) - Y_{4,2}(0,0))$, the effect of receiving the second treatment versus nothing in group 4 at period 2. The second is $E(Y_{4,2}(1,1) - Y_{4,2}(0,1))$, the effect of receiving the first and second treatments versus the second only in group 4 at period 2. The last term is minus $E(Y_{3,2}(0,1) - Y_{3,2}(0,0))$, the effect of receiving the second treatment versus nothing in group 3 at period 2. Accordingly, $\widehat{\beta}_{fe}$ is contaminated by the effects of receiving the second treatment versus nothing, in groups 4 and 3 at period 2.

Importantly, Theorem 1 does not necessarily rule out dynamic effects of past treatments on the outcome. The two treatments in the regression may for instance be the current treatment and its first lag. In that case, our potential outcome notation allows the current and lagged treatment to affect the outcome. Theorem 2 below generalizes Theorem 1, by considering TWFE regressions with K treatments. It thus applies to cases where potential outcomes depend on up to $K-1$ lags of the treatment.[4]

In this sense, Theorems 1 and 2 complement the pioneering work of Sun and Abraham (2021). The authors study the so-called event-study regression, an example of a TWFE regression with several treatments that is often used in staggered adoption designs, and where the treatment variables in the regression are indicators for having started receiving a single binary-and-staggered treatment ℓ periods ago. In those regressions, the authors show that effects of being treated for ℓ' periods may contaminate the coefficient supposed to measure the effect of ℓ periods of treatment in the regression, and they provide a decomposition formula one can use to quantify the extent of the phenomenon. If i) the K treatments in Regression 1 are indicators for having started receiving a single binary-and-staggered treatment ℓ periods ago, ii) no lags were gathered together in the event-study regression considered by Sun and Abraham (2021), and iii) the treatment no longer has an effect after $K+1$ periods of exposure, then our Theorem 2 reduces to Proposition 3 in Sun and Abraham (2021).[5] Though they coincide under conditions i)-iii) above, Theorems 1 and 2 and their results are non-nested. Our results apply to situations where the treatment variables in the regression are different, non necessarily mutually exclusive policies, that may not be binary or may not follow a staggered adoption design. On the other hand, their results also apply to situations where some of the treatment variables may be gathered together or omitted from the regression. They also obtain a decomposition without imposing common

[4]If the treatments have dynamic effects beyond the number of lags specified by the researcher, the regression is misspecified in the sense that it does not include all the explanatory variables it should have. Theorems 1 and 2 do not consider the consequences of such misspecification. Rather, our goal is to highlight issues that may arise even when the TWFE regression is correctly specified.

[5]In their decomposition, Sun and Abraham (2021) gather groups that started receiving the treatment at the same period into cohorts. Their decomposition can then be further decomposed, thus finally yielding the result in our Theorem 2.

trends.

Theorem 1 is also related to the pioneering work of Hull (2018). In his Section 2.2, the author studies TWFE regressions where indicators for each value that a multinomial treatment may take are included in the regression, an example of a TWFE regression with several treatments. Equation (15) therein is, to our knowledge, the first instance where a contamination phenomenon was shown. However, the paper does not discuss this phenomenon. It also does not give a decomposition formula like Theorem 1, so one cannot use the paper's results to compute the contamination weights, and assess whether they are important in a given regression. Finally, the paper's result applies when the data has two periods, and in instances were the treatments in the regression are indicators for each value that a multinomial treatment may take. Accordingly, the paper does not cover the case with more two time periods, and non-exclusive treatments.

Another related paper, released after ours, is Goldsmith-Pinkham, Hull and Kolesár (2021), who show that a contamination phenomenon similar to that in Sun and Abraham (2021) and in Theorem 1 also arises in linear regressions with several treatments, and a set of controls such that the treatments can be assumed to be independent of the potential outcomes conditional on those controls. Again, their result is not nested within and does not nest the results of Sun and Abraham (2021) nor ours: both Sun and Abraham (2021) and us assume parallel trends rather than conditional independence. Interestingly, under their conditional independence assumption, the weights on the effect of $D^1_{g,t}$ are all positive. Thus, their result shows that the contamination phenomenon can also arise in instances where the negative weighting phenomenon put forward by de Chaisemartin and D'Haultfœuille (2020) in the context of TWFE regressions is absent.

Overall, our four papers complement each other, and show that the contamination phenomenon is very pervasive, as it arises under several identifying assumptions (parallel trends and conditional independence), and irrespective of the nature of the treatments included in the regression.

Theorem 1 has an important consequence for TWFE regressions estimating heterogeneous treatment effects. Often times, researchers run a TWFE regression with a treatment variable $D_{g,t}$ interacted with a group-level binary variable I_g, and with $(1-I_g)$. For instance, to study if the treatment effect differs in poor and rich counties, one interacts the treatment with an indicator for counties above the median income, and with an indicator for counties below the median income. Theorem 1 also applies to those regressions. Specifically, one has

$$\beta_{fe}^{I=1} = E\left[\sum_{(g,t):D_{g,t}=1,I_g=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta_{g,t} + \sum_{(g,t):D_{g,t}=1,I_g=0} \frac{N_{g,t}}{N_1} w_{g,t}\Delta_{g,t}\right].$$

where $\beta_{fe}^{I=1}$ is the coefficient of $D_{g,t}\times I_g$, and $\Delta_{g,t} = \frac{1}{N_{g,t}}\sum_{i=1}^{N_{g,t}}\left[Y_{i,g,t}(1) - Y_{i,g,t}(0)\right]$. The previous display implies that the coefficient of $D_{g,t}\times I_g$ is contaminated by the treatment effect in (g,t) cells such that $I_g = 0$. In the example, the coefficient of the treatment interacted with the

indicator for rich counties is contaminated by the treatment effect in poor counties.[6] This calls into question the use of such TWFE regressions to estimate heterogeneous effects.

Theorem 1 shows that TWFE regressions with several treatments may be affected by a contamination phenomenon that does not affect TWFE regressions with one treatment. However, TWFE regressions with several treatments also tend to have a larger number of negative weights and a larger absolute value of the sum of negative weights than TWFE regressions with only one treatment. Thus TWFE regressions with several treatments may be less robust to heterogeneous effects. We start by showing this formally, in the following, simple design.

Assumption 5 *(Standard DID with two treatments) For all* $(g,t) \in \{1,...,G\} \times \{1,...,T\}$ *and* $k \in \{1,2\}$, $D^k_{g,t} = 1\{g \geq G^k\}1\{t \geq T^k\}$ *for some* $1 < G^1 < G^2 \leq G$ *and* $1 < T^1 < T^2 \leq T$.

Assumption 5 corresponds to a standard DID set-up, with two treatments: some groups start receiving the first treatment at a date T^1, and a subset of those groups then start receiving the second treatment at a later date T^2. These conditions are typically satisfied when the second treatment is a reinforcement of the first. Note that there is no variation in treatment timing: all treated groups start receiving the first (resp. second) treatment at the same date.

In this set-up, we compare the TWFE regression with two treatments considered above to a TWFE regression with only the first treatment. Specifically, we consider the regression of $Y_{i,g,t}$ on group fixed effects, period fixed effects, and $D^1_{g,t}$, estimated on all periods before T^2. We let β^1_{fe} denote the expectation of the coefficient on $D^1_{g,t}$ in that regression. We also let $N'_1 = \sum_{g,t:t<T_2,D^1_{g,t}=1} N_{g,t}$ denote the number of units receiving the first treatment before period T_2.

Corollary 1 *Suppose that Assumptions 1-5 hold,* $K = 2$ *and for all* $t \geq 2$, $N_{g,t}/N_{g,t-1}$ *does not vary across* g. *Then,*

$$\beta^1_{fe} = E\left[\sum_{(g,t):t<T_2,D^1_{g,t}=1} \frac{N_{g,t}}{N'_1}\Delta^1_{g,t}(0)\right],$$

$$\beta_{fe} = E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta^1_{g,t}(D^2_{g,t})\right].$$

If $\sum_{g,t} N_{g,t}D^2_{g,t} > \sum_{g,t} N_{g,t}1\{g < G^2\}1\{t < T^2\}$, *the weights* $w_{g,t}$ *are negative for all* (g,t) *satisfying* $g \in \{G^1,...,G^2-1\}$ *and* $t \in \{T^1,...,T^2-1\}$.

The first result shows that β^1_{fe} identifies the average treatment effect on the treated from period 1 to $T_2 - 1$. On the other hand, the second result shows that once we include the subsequent

[6] This contamination phenomenon disappears if the time fixed effects are interacted with I_g in the regression.

periods and the second treatment in the regression, the coefficient on $D^1_{g,t}$ may now identify a weighted sum of the effect of $D^1_{g,t}$ across the treated (g,t) cells, with some negative weights. The condition $\sum_{g,t} N_{g,t} D^2_{g,t} > \sum_{g,t} N_{g,t} 1\{g < G^2\} 1\{t < T^2\}$ requires that there are more units that receive the second treatment than units in "control groups" ($g < G^2$) during periods before the second treatment starts ($t < T^2$). Interestingly, the contamination term in Theorem 1 vanishes under Assumption 5: adding the second treatment to the regression generates some negative weights but does not lead to a contamination bias.

Beyond the simple design considered in Assumption 5, TWFE regressions with several treatments seem to often have more and larger negative weights than TWFE regressions with only one treatment. At least, this is the case in the application we revisit in Section 5, where Assumption 5 fails. In that application, the TWFE regression with several treatments has much larger negative weights attached to it than the TWFE regressions with only one treatment.

3.2 More than two treatment variables

We now go back to the general case where K may be greater than 2. We let $\mathbf{0}^{-1} = (0, ..., 0)$ be the vector of $K-1$ zeros. We also define

$$\Delta^1_{g,t}(D^{-1}_{g,t}) = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} \left[Y_{i,g,t}(1, D^{-1}_{g,t}) - Y_{i,g,t}(0, D^{-1}_{g,t}) \right],$$
$$\Delta^{-1}_{g,t} = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} \left[Y_{i,g,t}(0, D^{-1}_{g,t}) - Y_{i,g,t}(0, \mathbf{0}^{-1}) \right].$$

Theorem 2 below generalizes Theorem 1.

Theorem 2 *Suppose that Assumptions 1-4 hold. Then,*

$$\beta_{fe} = E\left[\sum_{(g,t): D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} \Delta^1_{g,t}(D^{-1}_{g,t}) + \sum_{(g,t): D^{-1}_{g,t} \neq \mathbf{0}^{-1}} \frac{N_{g,t}}{N_1} w_{g,t} \Delta^{-1}_{g,t} \right].$$

Moreover, $\sum_{(g,t): D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} = 1$*, and if* $K = 2$ *or the treatments* $D^2_{g,t}, ..., D^K_{g,t}$ *are mutually exclusive,* $\sum_{(g,t): D^{-1}_{g,t} \neq \mathbf{0}^{-1}} \frac{N_{g,t}}{N_1} w_{g,t} = 0$.

Theorem 2 is similar to Theorem 1, except that when $K > 2$, we do not always have

$$\sum_{(g,t): D^{-1}_{g,t} \neq \mathbf{0}^{-1}} \frac{N_{g,t}}{N_1} w_{g,t} = 0.$$

The weights on the effects of the other treatments may not sum to 0. Accordingly, even if the effects of all treatments are constant, $\widehat{\beta}_{fe}$ may still be biased for the effect of the first treatment.

There are three special cases where the weights on the effects of the other treatments sum to 0. The first one is when $K = 2$, as shown in Theorem 1. The second one is when the treatments $D^2_{g,t}, ..., D^K_{g,t}$ are mutually exclusive, as stated in Theorem 2. The third one is when there is no complementarity or substitutability between the treatments $D^2_{g,t}, ..., D^K_{g,t}$. Specifically, assume that for all (g, t), there exists $(\delta^k_{g,t})_{k=2,...,K}$ such that

$$E\left[\Delta^{-1}_{g,t}|\boldsymbol{D}\right] = \sum_{k=2}^{K} D^k_{g,t}\delta^k_{g,t}. \tag{2}$$

Then, we have the following decomposition:

Corollary 2 *Suppose that Assumptions 1-4 and* (2) *hold. Then,*

$$\beta_{fe} = E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta^1_{g,t}(D^{-1}_{g,t}) + \sum_{k=2}^{K}\sum_{(g,t):D^k_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\delta^k_{g,t}\right].$$

Moreover, $\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} = 1$, *and* $\sum_{(g,t):D^k_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} = 0$ *for every* $k \in \{2, ..., K\}$.

Accordingly, it is only when the treatments are not mutually exclusive and may be complementary or substitutable that $\widehat{\beta}_{fe}$ could be biased even under constant treatment effects. This is because in that case, Regression 1 is misspecified, and one should include the interactions of the treatments. The weights in Corollary 2 can be computed using the `twowayfeweights` Stata command.

In the special case where $K = 3$ and the three treatment variables in the regression are two treatments and their interaction, we have $\Delta^{-1}_{g,t} = \Delta^{0,1}_{g,t}$: when the first treatment is equal to 0, as in $\Delta^{-1}_{g,t}$, the interaction of the two treatments must be equal to 0. Accordingly, the decomposition of β_{fe} in Theorem 2 involves the same causal effects as the decomposition of the coefficient on the first treatment in the regression without the interaction term in Theorem 1. The weights, on the other hand, differ. In the decomposition in Theorem 1, they involve residuals of a TWFE regression of $D^1_{g,t}$ on $D^2_{g,t}$, while in the decomposition in Theorem 2, they involve residuals of a TWFE regression of $D^1_{g,t}$ on $D^2_{g,t}$ and $D^1_{g,t}D^2_{g,t}$. In the special case with only two time periods and where groups do not receive any of the two treatments in the first period, one can show that the coefficient on $D^1_{g,t}$ in the regression with the interaction estimates a weighted average of the effect of the first treatment and is not contaminated by the effect of the second treatment. On the other hand, in the regression without the interaction, the coefficient on $D^1_{g,t}$ estimates a weighted average of the effects of the first treatment plus a weighted sum of the effects of the second treatment with non-zero weights. In that special case, the regression with the interaction term is preferable, as it makes the contamination problem disappear. This result does not, however, translate to more general designs with more than two time periods and where groups may receive the treatments at every period. It is easy to find examples where adding the

interaction to the regression actually increases the contamination weights. This is the case for instance in the application we consider below: in the regression without control variables and with the two main treatments (the minimum staff-child ratio and the minimum number of years of schooling required for daycare directors), adding the interaction between the two treatments actually increases the absolute value of the contamination weights.

4 Alternative estimators of the effect of a treatment controlling for other treatments

4.1 Alternative estimators when the treatments do not have dynamic effects

Let us first introduce

$$\mathcal{S} = \Big\{ (g,t) : t \geq 2, D^1_{g,t} \neq D^1_{g,t-1}, D^{-1}_{g,t} = D^{-1}_{g,t-1}, \exists g' : D^1_{g',t} = D^1_{g',t-1} = D^1_{g,t-1}, \\ D^{-1}_{g',t} = D^{-1}_{g',t-1} = D^{-1}_{g,t-1} \Big\}$$

and $N_{\mathcal{S}} = \sum_{(g,t)\in\mathcal{S}} N_{g,t}$. $\mathcal{S}$ is the set of cells (g,t) whose first treatment changes between $t-1$ and t while their other treatments do not change, and such that there is another group g' whose treatments do not change between $t-1$ and t, and with the same treatments as g in $t-1$. Then, let

$$\delta^{\mathcal{S}} = E\left[\frac{1}{N_{\mathcal{S}}} \sum_{(g,t)\in\mathcal{S}} \sum_{i=1}^{N_{g,t}} \left[Y_{i,g,t}(1, D^{-1}_{g,t}) - Y_{i,g,t}(0, D^{-1}_{g,t}) \right] \right]$$

denote the average effect of moving the first treatment from 0 to 1 while keeping all other treatments at their observed value, across the units in $\mathcal{S}$.[7] One may be interested in estimating other average treatment effects, such as the effect of moving several treatments at the same time. Such parameters can be estimated following a similar strategy as that we follow to estimate $\delta^{\mathcal{S}}$.

We now show that $\delta^{\mathcal{S}}$ can be unbiasedly estimated by a weighted average of DID estimators. This result holds under the following condition.

Assumption 6 *(Strong exogeneity and common trends, v2) For all* $(g,t) \in \{1,...,G\}\times\{2,...,T\}$ *and* $\boldsymbol{d} = (d^1_1, d^{-1}_1, ..., d^1_T, d^{-1}_T)$ *such that* $d^{-1}_t = d^{-1}_{t-1}$*, we have*

$$E\left[Y_{g,t}(d^1_{t-1}, d^{-1}_{t-1}) - Y_{g,t-1}(d^1_{t-1}, d^{-1}_{t-1}) | \boldsymbol{D}_g = \boldsymbol{d} \right] \\ = E\left[Y_{g,t}(d^1_{t-1}, d^{-1}_{t-1}) - Y_{g,t-1}(d^1_{t-1}, d^{-1}_{t-1}) | D^1_{g,t-1} = d^1_{t-1}, D^{-1}_{g,t-1} = d^{-1}_{t-1} \right].$$

Moreover, these conditional expectations do not depend on g*.*

[7] When $N_{\mathcal{S}} = 0$, we simply let the term inside brackets be equal to 0.

Assumption 6 imposes both a strong exogeneity and a parallel trends condition. The strong exogeneity condition requires that groups' $t-1$ to t outcome evolution, in the counterfactual scenario where all their treatments are at their $t-1$ value at period t, be mean independent of their treatments at every period other than $t-1$. The parallel trends assumption requires that groups with the same period-$t-1$ treatments have the same counterfactual trends. Note that Assumption 6 does not restrict treatment effect heterogeneity. It also does not imply parallel trends on the treatment effect, because for a given value of $\boldsymbol{D}_g$ it only imposes parallel trends on one potential outcome. Below we compare Assumption 6 to the more standard Assumption 4.

We can now define our estimator. For all $t \in \{2,...,T\}$, for all $(d,d') \in \{0,1\}^2$, and for all $d^{-1} \in \{0,1\}^{K-1}$, let

$$\mathcal{G}_{d,d',d^{-1},t} = \left\{ g : \ D^1_{g,t} = d, D^1_{g,t-1} = d', D^{-1}_{g,t} = D^{-1}_{g,t-1} = d^{-1} \right\}.$$

We then let $N_{d,d',d^{-1},t} = \sum_{g \in \mathcal{G}_{d,d',d^{-1},t}} N_{g,t}$ denote the number of observations with treatment 1 equal to d' at period $t-1$ and d at period t, and with other treatments equal to d^{-1} at both dates. Let also

$$\begin{aligned}
\text{DID}_{+,d^{-1},t} &= \sum_{g \in \mathcal{G}_{1,0,d^{-1},t}} \frac{N_{g,t}}{N_{1,0,d^{-1},t}} \left(Y_{g,t} - Y_{g,t-1}\right) - \sum_{g \in \mathcal{G}_{0,0,d^{-1},t}} \frac{N_{g,t}}{N_{0,0,d^{-1},t}} \left(Y_{g,t} - Y_{g,t-1}\right), \\
\text{DID}_{-,d^{-1},t} &= \sum_{g \in \mathcal{G}_{1,1,d^{-1},t}} \frac{N_{g,t}}{N_{1,1,d^{-1},t}} \left(Y_{g,t} - Y_{g,t-1}\right) - \sum_{g \in \mathcal{G}_{0,1,d^{-1},t}} \frac{N_{g,t}}{N_{0,1,d^{-1},t}} \left(Y_{g,t} - Y_{g,t-1}\right).
\end{aligned}$$

Note that $\text{DID}_{+,d^{-1},t}$ is not defined when $N_{1,0,d^{-1},t} = 0$ or $N_{0,0,d^{-1},t} = 0$. In such instances, we let $\text{DID}_{+,d^{-1},t} = 0$. Similarly, we let $\text{DID}_{-,d^{-1},t} = 0$ when $N_{1,1,d^{-1},t} = 0$ or $N_{0,1,d^{-1},t} = 0$. $\text{DID}_{+,d^{-1},t}$ compares the $t-1$-to-t outcome evolution of groups whose first treatment goes from 0 to 1 from $t-1$ to t while their other treatments are equal to d^{-1} at both dates, to the outcome evolution of groups whose first and other treatments are respectively equal to 0 and d^{-1} at both dates. Under Assumption 6, the latter evolution is a valid counterfactual of the outcome evolution that the first groups would have experienced if their first treatment had remained equal to 0 at period t. $\text{DID}_{-,d^{-1},t}$ has a similar interpretation.

Finally, let

$$\text{DID}_{\text{M}} = \sum_{t=2}^{T} \sum_{d^{-1} \in \{0,1\}^{K-1}} \left(\frac{N_{1,0,d^{-1},t}}{N_{\mathcal{S}}} \text{DID}_{+,d^{-1},t} + \frac{N_{0,1,d^{-1},t}}{N_{\mathcal{S}}} \text{DID}_{-,d^{-1},t} \right) \tag{3}$$

if $N_{\mathcal{S}} > 0$, and $\text{DID}_{\text{M}} = 0$ if $N_{\mathcal{S}} = 0$.

Theorem 3 *If Assumptions 1-3 and 6 hold, $E\left[DID_M\right] = \delta^{\mathcal{S}}$.*

DID_M extends the DID_M estimator in de Chaisemartin and D'Haultfœuille (2020) to settings with several treatments. Relative to the estimator in our previous paper, the estimator in this paper does not estimate the effect of the first treatment in (g,t) cells such that at least one of g's other treatments changes between $t-1$ and t. Similarly, it drops control groups whose first treatment does not change but such that at least one of their other treatments changes between $t-1$ and t. These two modifications ensure that our estimator is not contaminated by the effects of other treatments. Another modification is that our new estimator compares switchers and non-switchers with the same baseline values of their other treatments. This ensures it does not require that the effect of the other treatments be constant over time. The asymptotic normality of the DID_M estimator, when the number of groups goes to infinity, could be established under similar assumptions and using similar arguments as those used to show Theorem S6 in the Web Appendix of de Chaisemartin and D'Haultfœuille (2020). DID_M can be computed by the `did_multiplegt` Stata command, see the command's help file for more details.

Our estimators rely on a new assumption, Assumption 6, instead of the more standard Assumption 4. Though the two assumptions are non-nested, Assumption 6 may be more plausible, because it imposes parallel trends conditional on groups' treatments in the baseline period, rather than unconditionally. Groups with the same treatments in the baseline period may be more similar, and may then be more likely to experience parallel trends. Moreover, by imposing parallel trends in the counterfactual scenario where groups' treatments are at their $t-1$ value at period t rather than in the counterfactual where they do not receive any treatment, Assumption 6 may lead to more precise estimators than Assumption 4, especially when the number of treatments is large or when the treatments are non binary. Under Assumption 4, an heterogeneity-robust DID estimator can only use as controls groups that do not receive any of the treatments at two dates at least. Moreover, treatment effects can only be estimated for groups that do not receive any of the treatments at one date at least. Those two sets of groups may be small. In our empirical application in Section 5, there are two non-binary treatments, and while there are (g,t) cells whose two treatments are equal to 0, there is no group that does not receive any of the two treatments at two dates at least. Accordingly, one cannot construct an heterogeneity-robust DID estimator based on Assumption 4 in this application. Finally, note that Assumption 6 can be tested using placebo estimators similar to those proposed in de Chaisemartin and D'Haultfœuille (2020), and adapted to the case with several treatments.

The DID_M estimator in this paper can be extended to accommodate non-binary treatments, just as the DID_M estimator in de Chaisemartin and D'Haultfœuille (2020) can also be extended to accommodate a non-binary treatment (see Section 4 of the Web Appendix of de Chaisemartin and D'Haultfœuille, 2020). With a non-binary treatment, the DID_M estimator starts by computing a weighted average of DIDs comparing the $t-1$ to t outcome evolution in groups whose first treatment changes and whose other treatments do not change, to the same outcome evolution

in groups whose treatments do not change and with the same treatments in $t-1$. Then, the estimator normalizes that weighted average by the average treatment change among switchers.

4.2 Alternative estimators when the treatments have dynamic effects

In the previous subsection, we have implicitly assumed that the treatments do not have dynamic effects, since the outcome of a unit at period t only depended on her period-t treatment, not on her previous treatments.[8] When treatments can have dynamic effects, estimating the effect of a treatment controlling for other treatments is difficult. We propose an estimation strategy when there are two binary treatments, which both follow a staggered adoption design. For any $g \in \{1,...,G\}$, let $F_g^1 = \min\{t : D_{g,t}^1 = 1\}$ denote the first date at which group g receives the first treatment, with the convention that $F_g^1 = T+1$ if group g never receives that treatment. Similarly, let $F_g^2 = \min\{t : D_{g,t}^2 = 1\}$ denote the first date at which group g receives the second treatment, with the convention that $F_g^2 = T+1$ if group g never receives that treatment.

Assumption 7 *(Staggered design with two binary treatments) For all* $(g,t) \in \{1,...,G\} \times \{2,...,T\}$, $D_{g,t}^1 \in \{0,1\}$, $D_{g,t}^2 \in \{0,1\}$, $D_{g,t-1}^1 \leq D_{g,t}^1$, $D_{g,t-1}^2 \leq D_{g,t}^2$, *and* $F_g^2 \geq F_g^1$.

Assumption 7 requires that both treatments weakly increase over time, which means that once a group has switched from untreated to treated, it cannot switch back to being untreated. Assumption 7 also requires that groups start receiving the second treatment after the first. This is typically satisfied when the second treatment is a reinforcement of the first. Our running example will be that of a researcher seeking to separately estimate the effects of medical and recreational marijuana laws in the US: so far, states have passed the former before the latter, and none of the medical and recreational laws passed since the late 1990s have been reverted. Another example where Assumption 7 holds include voter ID laws in the US, where non-strict laws are typically passed before strict ones (see Cantoni and Pons, 2021). Another example are anti-deforestation policies, where plots of lands are typically put into a concession, and then some concessions get certified (see Panlasigui et al., 2018).[9]

To allow for dynamic effects, we need to modify our potential outcome notation. For all $(\boldsymbol{d}^1, \boldsymbol{d}^2) \in \{0,1\}^{2T}$, let $Y_{i,g,t}(\boldsymbol{d}^1; \boldsymbol{d}^2)$ denote the potential outcome of observation i in group g at period t, if her two treatments from period 1 to T are equal to $\boldsymbol{d}^1, \boldsymbol{d}^2$, and let $Y_{g,t}(\boldsymbol{d}^1; \boldsymbol{d}^2)$ be the average outcome in group g and at period t under that scenario. This dynamic potential outcome framework is similar to that in Robins (1986). It allows for the possibility that observations' outcome at time t be affected by their past and future treatments.

[8] On the other hand and as discussed above, Theorems 1 and 2 do apply to dynamic effect cases, when the other treatment variables in the regression are lags of the treatment.

[9] Assumption 5 is a special case of Assumption 7, without variation in treatment timing.

Our estimator relies on the following assumptions.

Assumption 8 *(No Anticipation) For all g, for all $(\boldsymbol{d}^1, \boldsymbol{d}^2) \in \{0,1\}^{2T}$,*

$$Y_{g,t}(\boldsymbol{d}^1; \boldsymbol{d}^2) = Y_{g,t}(d_1^1, ..., d_t^1; d_1^2, ..., d_t^2).$$

Assumption 8 requires that a group's current outcome do not depend on her future treatments, the so-called no-anticipation hypothesis. Abbring and Van den Berg (2003) have discussed that assumption in the context of duration models, and Malani and Reif (2015), Botosaru and Gutierrez (2018), and Sun and Abraham (2021) have discussed it in the context of DID models.

For any $j \in \{1, ..., T\}$, let $\mathbf{0}_j$ and $\mathbf{1}_j$ denote vectors of j zeros and ones, respectively. We also adopt the convention that $\mathbf{0}_0$ and $\mathbf{1}_0$ denote empty vectors. Hereafter, we refer to $Y_{g,t}(\mathbf{0}_T; \mathbf{0}_T)$ as group g's -treated potential outcome at period t, her outcome if she never receives either of the two treatments. Our estimators rely on the following assumption on $Y_{g,t}(\mathbf{0}_T; \mathbf{0}_T)$.

Assumption 9 *(Independent groups, strong exogeneity, and common trends for the never-treated outcome) For all $t \geq 2$ and $g \in \{1, ..., G\}$, $E(Y_{g,t}(\mathbf{0}_T; \mathbf{0}_T) - Y_{g,t-1}(\mathbf{0}_T; \mathbf{0}_T)|\boldsymbol{D})$ does not vary across g.*

Assumption 9 is an adaptation of Assumptions 3-4 to the set-up we consider in this section, and where we allow for dynamic effects.

Under Assumption 7, the estimators of instantaneous and dynamic treatment effects proposed in de Chaisemartin and D'Haultfœuille (2021) can still be used with two treatments, redefining the treatment as $\tilde{D}_{g,t} = D_{g,t}^1 + D_{g,t}^2$. However, those estimators will average together effects of the first and of the second treatment. Estimating separately the effect of the first treatment is straightforward: one can just compute the estimators in Callaway and Sant'Anna (2021) or de Chaisemartin and D'Haultfœuille (2021), restricting the sample to all (g,t)s such that $D_{g,t}^2 = 0$. In the marijuana laws example, to estimate the effect of medical marijuana laws, one can just restrict the sample to all state×year (g,t) such that state g has not passed a recreational law yet in year t. The horizon until which dynamic effects can be estimated will just be truncated by the second treatment.

Estimating separately the effect of the second treatment is more challenging but can still be achieved, under the following, supplementary assumption.

Assumption 10 *(Restriction on the effect of the first treatment) For all $g \in \{1, ..., G\}$, $j \in \{1, ..., T\}$, and $t > j$, there exists $\lambda_{j,g}(\boldsymbol{D})$ and $\mu_{j,t}(\boldsymbol{D})$ such that*

$$E(Y_{g,t}((\mathbf{0}_{j-1}, \mathbf{1}_{T-(j-1)}); \mathbf{0}_T) - Y_{g,t}(\mathbf{0}_T; \mathbf{0}_T)|\boldsymbol{D}) = \lambda_{j,g}(\boldsymbol{D}) + \mu_{j,t}(\boldsymbol{D}).$$

Assumption 10 requires that the effect of the first treatment evolves over time in the same way in every group: for any $t > j+1$,

$$E(Y_{g,t}((\mathbf{0}_{j-1}, \mathbf{1}_{T-(j-1)}); \mathbf{0}_T) - Y_{g,t}(\mathbf{0}_T; \mathbf{0}_T)|\boldsymbol{D}) - E(Y_{g,t-1}((\mathbf{0}_{j-1}, \mathbf{1}_{T-(j-1)}); \mathbf{0}_T) - Y_{g,t-1}(\mathbf{0}_T; \mathbf{0}_T)|\boldsymbol{D}),$$

the difference between group g's effect of being treated for $t-(j-1)$ and $t-1-(j-1)$ periods, should be the same in every group. Still, Assumption 10 allows such treatment effects to vary in an unrestricted way with the number of time periods over which a group has been treated, and to vary with the time period at which the treatment was adopted. It also allows, to some extent, the treatment effect to vary across groups: groups' treatment effects can be arbitrarily heterogeneous at the first period where they start receiving the treatment, but the period to period evolution of that effect should then be the same in every group.

To understand why that assumption is needed, let us go back to the marijuana law example. Without Assumption 10, a state passing a recreational marijuana law may start experiencing a different outcome trend than other states that have only passed a medical law, either because of the recreational law, or because its evolution of the effect of the medical law differs from that in other states. In other words, that assumption is key to disentangle the effects of the two treatments, which is often of interest. Under the standard parallel trends assumption on the never-treated outcome (Assumption 9), one could only estimate the combined effects of the two treatments.

Though it is arguably strong, this assumption is partly testable, as we explain in more details below: it implies that groups that start receiving the treatment at the same time should then have the same outcome evolution until they adopt the second treatment. A violation of this assumption would lead our estimators to be upward (resp. downward biased) if the effect of the first treatment increases less (resp. more) in groups that adopt the second treatment than in groups that do not adopt it.

Assumption 11 *(Non-pathological design) There exists* $(g, g') \in \{1, ..., G\}^2$ *such that* $F_g^1 = F_{g'}^1$ *and* $1 < F_g^2 < F_{g'}^2$.

For any $f \in \{1, ..., T\}$, let

$$\mathcal{G}_f = \{g \in \{1, ..., G\} : F_g^1 = f\}$$

denote the set of groups that started receiving the first treatment at date f. Let

$$\mathcal{F} = \{f \in \{1, ..., T\} : \exists (g, g') \in \mathcal{G}_f^2 : 1 < F_g^2 < F_{g'}^2\}$$

be the set of dates such that at least two groups start receiving the first treatment at that date and start receiving the second treatment at different dates. Assumption 11 ensures that $\mathcal{F}$ is

not empty. For any $f \in \mathcal{F}$, let

$$NT_f = \max_{g \in \mathcal{G}_f} F_g^2 - 1$$

be the last period at which at least one group that started receiving the first treatment at period f has still not received the second treatment. Then, we let

$$L_{nt,f} = NT_f - \min_{g \in \mathcal{G}_f : F_g^2 \geq 2} F_g^2$$

denote the number of time periods between the first date at which a group that started receiving the first treatment at date f starts receiving the second treatment, and the last date at which a group that started receiving the first treatment at date f has not received the second treatment yet. Note that $L_{nt,f} \geq 0$ for all $f \in \mathcal{F}$. Let also

$$L_{nt} = \max_{f \in \mathcal{F}} L_{nt,f}.$$

For any $\ell \in \{0, ..., L_{nt}\}$, $f \in \mathcal{F}$ such that $NT_f \geq \ell + f + 1$, and $t \in \{\ell + f + 1, ..., NT_f\}$, let

$$N_{t,\ell}^f = \sum_{g \in \mathcal{G}_f : F_g^2 = t - \ell} N_{g,t}$$

denote the number of units in groups that started receiving the first treatment at date f and the second treatment ℓ periods ago at t, and such that at least one group also started receiving the first treatment at date f and has not started receiving the second treatment yet at t. Let

$$N_\ell = \sum_{f \in \mathcal{F} : NT_f \geq \ell + f + 1} \sum_{t = \ell + f + 1}^{NT_f} N_{t,\ell}^f$$

be the number of units reaching ℓ periods after they started receiving the second treatment at a date where there is still a group that started receiving the first treatment at the same date as their group and that has not received the second treatment yet. Across those units, the average cumulative effect of having received the second treatment for $\ell + 1$ periods while fixing the first treatment at its observed value is

$$\delta_\ell = E\left[\frac{1}{N_\ell} \sum_{f \in \mathcal{F} : NT_f \geq \ell + f + 1} \sum_{t = \ell + f + 1}^{NT_f} \sum_{(i,g) : g \in \mathcal{G}_f, F_g^2 = t - \ell} Y_{i,g,t}(\boldsymbol{D}_g^1; (\mathbf{0}_{t-\ell-1}, \mathbf{1}_{\ell+1})) - Y_{i,g,t}(\boldsymbol{D}_g^1; \mathbf{0}_t) \right].$$

Remark that by construction, $N_\ell > 0$ for all $\ell \in \{0, ..., L_{nt}\}$, so δ_ℓ is well-defined for such ℓ. Note also that δ_ℓ does not include the effect of the second treatment for groups that start receiving the two treatments at the same time. For those groups, it is impossible to separately estimate the effects of the first and second treatments, using our DID estimation strategy at least.

We now define an estimator of δ_ℓ. For any $f \in \mathcal{F}$ and t such that $NT_f \geq t$, let

$$N_t^{nt,f} = \sum_{g \in \mathcal{G}_f : F_g^2 > t} N_{g,t}.$$

Then, for any $\ell \in \{0, ..., L_{nt}\}$, $f \in \mathcal{F}$ such that $NT_f \geq \ell + f + 1$, and $t \in \{\ell + f + 1, ..., NT_f\}$, we define

$$\text{DID}^f_{t,\ell} = \sum_{g \in \mathcal{G}_f : F^2_g = t-\ell} \frac{N_{g,t}}{N^f_{t,\ell}} (Y_{g,t} - Y_{g,t-\ell-1}) - \sum_{g \in \mathcal{G}_f : F^2_g > t} \frac{N_{g,t}}{N^{nt,f}_t} (Y_{g,t} - Y_{g,t-\ell-1})$$

if $N^f_{t,\ell} > 0$ and $N^{nt,f}_t > 0$, and we let $\text{DID}^f_{t,\ell} = 0$ if $N^f_{t,\ell} = 0$ or $N^{nt,f}_t = 0$. Then, for all $\ell \in \{0, ..., L_{nt}\}$, we let

$$\text{DID}_\ell = \sum_{f \in \mathcal{F} : NT_f \geq \ell + f + 1} \sum_{t = \ell + f + 1}^{NT_f} \frac{N^f_{t,\ell}}{N_\ell} \text{DID}^f_{t,\ell}.$$

Theorem 4 *Suppose that Assumptions 1-2 and 7-11 hold. Then,* $E\left[DID_\ell\right] = \delta_\ell$ *for all* $\ell \in \{0, ..., L_{nt}\}$.

DID_ℓ can be computed by the `did_multiplegt` Stata command, restricting the sample to the (g,t)s such that $D^1_{g,t} = 1$, and including F^1_g in the `trends_nonparam` option. The asymptotic normality of the DID_ℓ estimators, when the number of groups goes to infinity, could be established under similar assumptions and using similar arguments as those used to show Theorem 4 in de Chaisemartin and D'Haultfœuille (2021).

Beyond the somewhat complicated notation above, the idea underlying DID_ℓ is actually quite simple: it amounts to comparing the outcome evolution of groups that adopt/do not adopt the second treatment, and that adopted the first treatment at the same date. This ensures that the "treatment" and "control" groups involved in this comparison have been exposed to the first treatment for the same number of periods. Under Assumptions 9 and 10, this in turn ensures that their outcome evolution would have been the same if the "treatment groups" had not adopted the second treatment. The estimation procedure we propose can easily be extended to more than two treatments. For instance, if there was a third treatment following a staggered adoption design and always adopted after the second one, one could estimate its effect using the `did_multiplegt` Stata command, restricting the sample to the (g,t)s such that $D^2_{g,t} = 1$, and including the interaction of F^1_g and F^2_g in the `trends_nonparam` option.

Theorem 4 complements the pioneering work of Callaway and Sant'Anna (2021) and Sun and Abraham (2021), who provide DID estimators of the effect of a single treatment following a staggered adoption design. To our knowledge, our paper is the first to consider the case with several treatments following consecutive staggered designs, which arises relatively often, as the examples given above show. Our main insight is to show that one can obtain unbiased estimators of the effect of the second treatment, under the restriction on the effect of the first treatment stated in Assumption 10, and provided one controls for the first treatment's adoption date.

The assumptions underlying Theorem 4 are refutable. They imply that groups that start receiving the treatment at the same time should have the same outcome evolution until they adopt the

second treatment, see Equation (19) in the Appendix. This can be tested, using similar placebo estimators as in de Chaisemartin and D'Haultfœuille (2021), the main difference being that one should compare groups with the same value of F_g^1. The placebo estimators one can use to test the assumptions underlying Theorem 4 can also be computed by the `did_multiplegt` command, restricting the sample to the (g,t)s such that $D_{g,t}^1 = 1$, including F_g^1 in the `trends_nonparam` option, and requesting the `placebo` option. One should still keep in mind that such pre-trends tests come with some caveats, as shown by Roth (2019): they may be underpowered and could fail to detect violations of the assumptions, and they may lead to pre-testing issues. However, our placebo estimators can be used to conduct the sensitivity analysis proposed by Manski and Pepper (2018) or Rambachan and Roth (2019).

The estimation strategy proposed in Theorem 4 requires that there is at least one pair of groups that receive the first treatment at the same date, and such that the first group receives the second treatment strictly before the second group. When the number of groups is relatively low (e.g.: the 50 US states), there may not be any pair of groups receiving the first treatment at the same time period. Then, two alternative estimation strategies can be proposed. First, instead of Assumption 10, one may assume that the effect of the first treatment evolves linearly with the number of periods of exposure, with a slope that differs across groups:

$$E(Y_{g,t}((\mathbf{0}_{j-1}, \mathbf{1}_{T-(j-1)}); \mathbf{0}_T) - Y_{g,t}(\mathbf{0}_T; \mathbf{0}_T) | \boldsymbol{D}) = \lambda_{j,g}(\boldsymbol{D}) + \mu_g(\boldsymbol{D})(t-j).$$

Then, one can recover the counterfactual outcome that a group adopting the second treatment would have obtained without it by extrapolating a linear estimate of its outcome evolution prior to adoption. The resulting estimator can be computed by the `did_multiplegt` command, restricting the sample to the (g,t)s such that $D_{g,t}^1 = 1$, and including the group indicator in the `trends_lin` option. Second, one could also strengthen Assumption 10, by assuming that the effect of the first treatment evolves potentially non-linearly with the number of periods of exposure, but that this evolution is the same in every group and at every time period:

$$E(Y_{g,t}((\mathbf{0}_{j-1}, \mathbf{1}_{T-(j-1)}); \mathbf{0}_T) - Y_{g,t}(\mathbf{0}_T; \mathbf{0}_T) | \boldsymbol{D}) = \lambda_{j,g}(\boldsymbol{D}) + \mu_{t-j}(\boldsymbol{D}).$$

Then, one can recover the counterfactual outcome that a group adopting the second treatment would have obtained without it, by extrapolating the outcome evolution experienced by groups that reached a similar number of periods of exposure to the first treatment without adopting the second one. The resulting estimator can be computed by the `did_multiplegt` command, restricting the sample to the (g,t)s such that $D_{g,t}^1 = 1$, and including indicators for reaching 1, 2, etc. periods of exposure to the first treatment in the `controls` option.

The approach in Theorem 4 can easily be extended to some instances where the assumptions of Theorem 4 fail. For example, there may applications with two binary treatments following a staggered adoption design, but such that some groups receive treatment 1 before treatment 2,

other groups receive treatment 2 first, and other groups receive both treatments at the same time. Then, one can start by restricting attention to the subsample of groups such that $D^1_{g,t} \geq D^2_{g,t}$ for all t and $F^1_g < F^2_g$. This subsample includes groups that only receive treatment 1, groups that receive both treatments but receive the second one strictly after the first, and groups that do not receive any treatment. In that subsample, one can estimate the instantaneous and dynamic effects of receiving only the first treatment, using the `did_multiplegt` command and restricting the sample to (g,t)s such that $D^2_{g,t} = 0$. One can then estimate the effect of receiving the second treatment when one has already received the first one, using the `did_multiplegt` command, restricting the sample to the (g,t)s such that $D^1_{g,t} = 1$ and including F^1_g in the `trends_nonparam` option. Second, one can restrict attention to the subsample of groups such that $D^2_{g,t} \geq D^1_{g,t}$ for all t and $F^2_g < F^1_g$. In that subsample, one can estimate the effect of receiving only the second treatment, and the effect of receiving the first treatment when one has already received the second one, using the same steps as above but reverting the roles of the first of second treatments. Finally, one can restrict attention to the subsample of groups that either receive both treatments at the same time or that do not receive any treatment, and estimate the effect of receiving both treatments at the same time using the `did_multiplegt` command. Comparing these five sets of estimates may be indicative of whether the treatments are complements or substitutes, even though differences could also be driven by heterogeneous effects across the various subsamples.

The approach outlined above can also be used when a single treatment changes several times over the duration of the panel, to isolate the effect of each change. To simplify, take the example of a binary treatment that can switch at most once from 0 to 1, and then once from 1 to 0. To estimate the effect of switching from 0 to 1, one can just use the `did_multiplegt` command in the subsample of (g,t)s such that group g has never switched from 1 to 0 at or before t. To estimate the effect of switching from 1 to 0, one can just use the `did_multiplegt` command in the subsample of (g,t)s such that group g has switched from 0 to 1 at or before t, including the date of that switch in the `trends_nonparam` option, and defining the treatment as an indicator for switching from 1 to 0. More generally, assume one is interested in the effect of a single treatment, that may not be binary and that can change multiple times, and one is interested in separately estimating the effect of each treatment change. Following similar steps as those used in the proof of Theorem 4, one can show that the estimators computed by the `did_multiplegt` command, restricting the sample to the (g,t)s such that g has experienced a first treatment change at or before t and has not experienced a third treatment change at or before t, and including the date of the first treatment change interacted with groups' treatment values before and after that change in the `trends_nonparam` option, are unbiased for the instantaneous and dynamic effects of a second treatment change, under assumptions similar to Assumptions 9 and 10. One can proceed similarly to estimate effects of a third, fourth, etc. treatment change,

but the corresponding estimators may soon become noisy, especially so when the treatment is not binary. In such instances, the approach proposed in de Chaisemartin and D'Haultfœuille (2021) of estimating the total cumulative effects of all treatment changes, rather than trying to separately estimate their effects, may be more feasible in practice.

5 Application

In this section, we revisit Table 11 in Hotz and Xiao (2011).[10] The authors use a 1987, 1992, and 1997 US state-level panel data set to estimate the effect of state center-based daycare regulations on the demand for family home daycare. Family home day cares are not subject to those regulations. More stringent regulations may increase the cost of center-based establishments, but may also increase their safety and quality. Accordingly, the effects of those regulations on the demand for family home daycare is ambiguous. In Column (3) of Table 11, the authors regress the revenue of family home day cares in state g and year t on state fixed effects, year fixed effects, some control variables, and four state center-based daycare regulations: the minimum staff-child ratio, the minimum years of schooling required to be the director of a center-based care, and two indicators for whether there is no such minima to allow for potentially non-linear effects.

The coefficient on the minimum years of schooling treatment is equal to -0.445 and is highly significant (s.e.=0.167),[11] thus suggesting that increasing by one the years of schooling required for directors of center-based daycare decreases the revenue of family home daycare by 0.44 million USD. However, this coefficient may not be robust to heterogeneous effects across state and years, and may also be contaminated by the effects of the other treatments in the regression. Following Corollary 2, this coefficient can be decomposed into the sum of four terms. The first term is a weighted sum of the effects of increasing by one the years of schooling required in 127 state×year cells, where 63 effects receive a positive weight and 64 receive a negative weight, and where the positive and negative weights respectively sum to 10.022 and -9.022. The second term is a sum of the effects of not having a requirement on directors' years of schooling in 26 state×year cells, where 11 effects receive a positive weight and 15 receive a negative weight, and where the positive and negative weights respectively sum to 0.175 and -0.175. The third term is a sum of the effects of increasing by one the staff to child ratio in 148 state×year cells, where 70 effects receive a

[10] That table is not the main one in the paper, but it is the only one that can be replicated using the publicly available data set. Several other tables report results from TWFE regressions with several treatments, but all make use of proprietary data.

[11] This standard error is slightly larger than that in Hotz and Xiao (2011), because we cluster it at the state rather than at the state×year level, which is more in line with the standard practice in empirical work (see Bertrand, Duflo and Mullainathan, 2004). We also use the bootstrap to compute it, to ensure that it is comparable to the standard error of the DID_{M} estimator below.

positive weight and 78 receive a negative weight, and where the positive and negative weights respectively sum to 0.199 and -0.199. The last term is a sum of the effects of not having a requirement on staff to child ratio in 5 state×year cells, where 4 effects receive a positive weight and 1 receive a negative weight, and where the positive and negative weights respectively sum to 0.056 and -0.056. Results are similar for the other three treatment coefficients in the regression, except that the contamination weights attached to them are even larger. For instance, for the coefficient on the staff to child ratio treatment, the weighted sum of the effects of the minimum years of schooling treatment has positive and negative weights summing to 334.916 and -334.916.

When the other three treatment variables are dropped from the regression, the coefficient on the minimum years of schooling becomes small (-0.022) and insignificant (s.e.=0.035). We follow Theorem 1 in de Chaisemartin and D'Haultfœuille (2020) to decompose this coefficient, and find that it estimates a weighted sum of the effects of increasing by one the years of schooling required in 127 state×year cells, where 64 cells receive a positive weight and 63 receive a negative weight, and where the positive and negative weights respectively sum to 1.856 and -0.856. Thus, the regression with only one treatment has considerably less negative weights attached to it than the regression with several treatments. The same holds for the three other treatment variables.

Finally, we compute the estimator proposed in Section 4.1, for the minimum years of schooling treatment, controlling for the staff-to-child ratio treatment. Our estimators do not assume linear treatment effects, so unlike the authors we do not need to control for the indicators for whether there is no such minima. There are 8 state×year cells (g,t) such that the years of schooling treatment has changed from $t-1$ to t in g, while the staff to child ratio treatment has not changed. Our estimator estimates the average effect of increasing the years of schooling requirement by one in those 8 (g,t) cells.[12] As the data only has three time periods, accounting for dynamic effects may not be a first-order concern here. Moreover, the authors' TWFE regression implicitly rules out such dynamic effects so we follow their specification. There are only 2 state×year cells (g,t) such that the staff to child ratio treatment changes, while the years of schooling treatment does not change, thus making it challenging to estimate the effect of the staff to child ratio treatment separately from that of the years of schooling treatment.

We find that DID_M, computed with the same controls as in the TWFE regression estimated by the authors, is equal to -0.066 and is insignificant (s.e.=0.136).[13] DID_M is almost seven times smaller than the coefficient on the years of schooling treatment in the TWFE regression. The

[12]The estimator in Section 4.1 can easily be extended to non-binary treatments. In Section 4 of their Web Appendix, de Chaisemartin and D'Haultfœuille (2020) cover that extension in the case with only one treatment. With several treatments, the extension to non-binary treatments is similar.

[13]We use the bootstrap, clustered at the state level, to compute the standard error of the DID_M estimator. In this application, that standard error is lower than the standard error of the TWFE coefficient. Though in practice, TWFE coefficients tend to have lower standard errors than heterogeneity-robust DID estimators, the opposite can also happen, as this example demonstrates.

two estimators are significantly different (t-stat=2.253). Under parallel trends, this means that the constant treatment effect assumption is rejected: if the effects of the years of schooling and staff to child ratio treatments were both constant over time and between groups, DID_M and the TWFE coefficient would both estimate the constant effect of the years of schooling treatment.

Overall, the conclusion that increasing the years of schooling required for directors of center-based daycare decreases the revenue of family home daycare may not be robust. First, the significant difference between the DID_M and TWFE estimators suggests that effects are indeed heterogeneous in this application. Then, owing to the large negative weights attached to it, in the presence of heterogeneous effects the TWFE coefficient may be biased and contaminated by other treatments' effects. Finally, the DID_M estimator, which is robust to heterogeneous effects and not contaminated by other treatments' effects, is 7 times smaller than the TWFE coefficient and insignificant.

6 Conclusion

In this paper, we show that treatment coefficients in TWFE regressions with several treatments may not robust to heterogeneous effects, and could be contaminated by the effects of other treatments in the regression. We propose alternative DID estimators that are robust to heterogeneous effects and do not suffer from this contamination problem.

7 Proofs

7.1 Theorem 1

The result directly follows from Theorem 2. If $K = 2$, $D_{g,t}^{-1} = D_{g,t}^2$. Then, $D_{g,t}^{-1} \neq \mathbf{0}^{-1}$ if and only if $D_{g,t}^2 = 1$, and one then has $D_{g,t}^2 \Delta_{g,t}^{-1} = D_{g,t}^2 \Delta_{g,t}^{0,1}$.

7.2 Corollary 1

The first result is a direct consequence of Theorem 1 in de Chaisemartin and D'Haultfœuille (2020): by symmetry all the weights of the treated units are identical. Turning to the second result, let us introduce, to simplify notation, $I_g^k = 1\{g > G^k\}$ and $J_t^k = 1\{t > T^k\}$, for $(k, g, t) \in \{1, 2\} \times \{1, ..., G\} \times \{1, ..., T\}$. Because $N_{g,t}/N_{g,t-1}$ does not vary with g, we can also write $N_{g,t} = A a_g b_t$ with $A = \sum_{g,t} N_{g,t}$, $a_g = \sum_t N_{g,t}/A$ and $b_t = \sum_g N_{g,t}/A$. Then, $\sum_g a_g = \sum_t b_t = 1$. Finally, we let, for $k \in \{1, 2\}$, $p_G^k = \sum_g a_g I_g^k$ and $p_T^k = \sum_t b_t J_t^k$.

Now, consider Regression (1) of the first treatment on other regressors. We can assume therein and without loss of generality that $\sum_g a_g \gamma_g = \sum_t b_g \nu_t = 0$. Combining this with Conditions (8) and (9) below, we obtain

$$\begin{aligned}
\alpha &= p_G^1 p_T^1 - p_G^2 p_T^2 \zeta, \\
\gamma_g &= (I_g^1 - p_G^1) p_T^1 - (I_g^2 - p_G^2) p_T^2 \zeta, \\
\nu_t &= p_G^1 (J_t^1 - p_T^1) - p_G^2 (J_t^2 - p_T^2) \zeta.
\end{aligned}$$

By definition of the residual $\varepsilon_{g,t}$ of (1), we also have $\sum_{g,t} N_{g,t} D_{g,t}^2 \varepsilon_{g,t} = 0$. Then, using the fact that $G^1 < G^2$ and $T^1 < T^2$, we obtain

$$p_G^2 p_T^2 = \alpha p_G^2 p_T^2 + \left(\sum_g a_g \gamma_g I_g^2 \right) p_T^2 + p_G^2 \left(\sum_t b_t \nu_t J_t^2 \right) p_T^2 + p_G^2 p_T^2 \zeta.$$

Plugging the expressions of α, γ_g and ν_t in this equation, we obtain, after some algebra,

$$\zeta = \frac{(1 - p_G^1)(1 - p_T^1)}{(1 - p_G^2)(1 - p_T^2)}.$$

Using again the expressions of α, γ_g and ν_t above, we get

$$\varepsilon_{g,t} = (I_g^1 - p_G^1)(J_t^1 - p_T^1) - \frac{(1 - p_G^1)(1 - p_T^1)}{(1 - p_G^2)(1 - p_T^2)} (I_g^2 - p_G^2)(J_t^2 - p_T^2). \tag{4}$$

Now, if $D_{g,t}^2 = 1$, then $I_g^2 = J_t^2 = 1$ and also $I_g^1 = J_t^1 = 1$. As a result, $\varepsilon_{g,t} = 0$. This shows that there is no contamination weights. Finally, suppose that $(g, t) \in \{G^1, ..., G^2 - 1\} \times \{T^1, ..., T^2 - 1\}$.

Then $D^1_{g,t} = 1$ but $I^2_g = J^2_t = 0$ and we get, using (4),

$$\varepsilon_{g,t} = (1 - p^1_G)(1 - p^1_T)\left[1 - \frac{p^2_G p^2_T}{(1 - p^2_G)(1 - p^2_T)}\right].$$

Thus, $\varepsilon_{g,t} < 0$ if $p^2_G p^2_T > (1 - p^2_G)(1 - p^2_T)$. The last result follows by observing that by definition of p^2_G, p^2_T and using $N_{g,t} = Aa_g b_t$, we have

$$Ap^2_G p^2_T = \sum_{g,t} N_{g,t} D^2_{g,t},$$
$$A(1 - p^2_G)(1 - p^2_T) = \sum_{g,t} N_{g,t} 1\{g < G^2\} 1\{t < T^2\}.$$

7.3 Theorem 2

We first establish the following lemma.

Lemma 1 *If Assumptions 1-4 hold, for all* $(g, g', t, t') \in \{1, ..., G\}^2 \times \{1, ..., T\}^2$,

$$\begin{aligned}
&E\left(Y_{g,t}|\boldsymbol{D}\right) - E\left(Y_{g,t'}|\boldsymbol{D}\right) - \left(E\left(Y_{g',t}|\boldsymbol{D}\right) - E\left(Y_{g',t'}|\boldsymbol{D}\right)\right) \\
=&D^1_{g,t} E\left(\Delta^1_{g,t}(D^{-1}_{g,t})\middle|\boldsymbol{D}\right) + E\left(\Delta^{-1}_{g,t}\middle|\boldsymbol{D}\right) - D^1_{g',t} E\left(\Delta^1_{g',t}(D^{-1}_{g',t})\middle|\boldsymbol{D}\right) - E\left(\Delta^{-1}_{g',t}\middle|\boldsymbol{D}\right) \\
-&D^1_{g,t'} E\left(\Delta^1_{g,t'}(D^{-1}_{g,t'})\middle|\boldsymbol{D}\right) - E\left(\Delta^{-1}_{g,t'}\middle|\boldsymbol{D}\right) + D^1_{g',t'} E\left(\Delta^1_{g',t'}(D^{-1}_{g',t'})\middle|\boldsymbol{D}\right) + E\left(\Delta^{-1}_{g',t'}\middle|\boldsymbol{D}\right).
\end{aligned}$$

Proof of Lemma 1

For all $(g, t) \in \{1, ..., G\} \times \{1, ..., T\}$,

$$\begin{aligned}
E\left(Y_{g,t}|\boldsymbol{D}\right) =&E\left(\frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} Y_{i,g,t}\middle|\boldsymbol{D}\right) \\
=&E\Bigg(\frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} \Big[Y_{i,g,t}(0, \mathbf{0}^{-1}) + D^1_{i,g,t}(Y_{i,g,t}(1, D^{-1}_{g,t}) - Y_{i,g,t}(0, D^{-1}_{g,t}) \\
&+Y_{i,g,t}(0, D^{-1}_{g,t}) - Y_{i,g,t}(0, \mathbf{0}^{-1})) + (1 - D^1_{i,g,t})(Y_{i,g,t}(0, D^{-1}_{g,t}) - Y_{i,g,t}(0, \mathbf{0}^{-1}))\Big]\Bigg|\boldsymbol{D}\Bigg) \\
=&E\left(Y_{g,t}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}\right) + D^1_{g,t} E\left(\Delta^1_{g,t}(D^{-1}_{g,t})\middle|\boldsymbol{D}\right) + E\left(\Delta^{-1}_{g,t}\middle|\boldsymbol{D}\right) \\
=&E\left(Y_{g,t}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}_g\right) + D^1_{g,t} E\left(\Delta^1_{g,t}(D^{-1}_{g,t})\middle|\boldsymbol{D}\right) + E\left(\Delta^{-1}_{g,t}\middle|\boldsymbol{D}\right), \qquad (5)
\end{aligned}$$

where the third equality follows from Assumption 2, and the fourth from Assumption 3. Moreover, by Assumption 4

$$\begin{aligned}
&E\left(Y_{g,t}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}_g\right) - E\left(Y_{g,t'}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}_g\right) - E\left(Y_{g',t}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}_g\right) + E\left(Y_{g',t'}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}_g\right) \\
=&0. \qquad (6)
\end{aligned}$$

The result follows by combining (5) and (6).

Proof of Theorem 2

It follows from the Frisch-Waugh theorem and the definition of $\varepsilon_{g,t}$ that

$$E\left(\widehat{\beta}_{fe}\middle|\boldsymbol{D}\right)=\frac{\sum_{g,t}N_{g,t}\varepsilon_{g,t}E\left(Y_{g,t}|\boldsymbol{D}\right)}{\sum_{g,t}N_{g,t}\varepsilon_{g,t}D_{g,t}^{1}}. \tag{7}$$

Now, by definition of $\varepsilon_{g,t}$ again,

$$\sum_{t=1}^{T}N_{g,t}\varepsilon_{g,t}=0 \text{ for all } g\in\{1,...,G\}, \tag{8}$$

$$\sum_{g=1}^{G}N_{g,t}\varepsilon_{g,t}=0 \text{ for all } t\in\{1,...,T\},. \tag{9}$$

Then,

$$\begin{aligned}
&\sum_{g,t}N_{g,t}\varepsilon_{g,t}E\left(Y_{g,t}|\boldsymbol{D}\right)\\
=&\sum_{g,t}N_{g,t}\varepsilon_{g,t}\left(E\left(Y_{g,t}|\boldsymbol{D}\right)-E\left(Y_{g,1}|\boldsymbol{D}\right)-E\left(Y_{1,t}|\boldsymbol{D}\right)+E\left(Y_{1,1}|\boldsymbol{D}\right)\right)\\
=&\sum_{g,t}N_{g,t}\varepsilon_{g,t}\left(D_{g,t}^{1}E\left(\Delta_{g,t}^{1}(D_{g,t}^{-1})\middle|\boldsymbol{D}\right)+E\left(\Delta_{g,t}^{-1}\middle|\boldsymbol{D}\right)-D_{1,t}^{1}E\left(\Delta_{1,t}^{1}(D_{1,t}^{-1})\middle|\boldsymbol{D}\right)-E\left(\Delta_{1,t}^{-1}\right)\middle|\boldsymbol{D}\right)\\
-&D_{g,1}^{1}E\left(\Delta_{g,1}^{1}(D_{g,1}^{-1})\middle|\boldsymbol{D}\right)-E\left(\Delta_{g,1}^{-1}\right)\middle|\boldsymbol{D}\right)+D_{1,1}^{1}E\left(\Delta_{1,1}^{1}(D_{1,1}^{-1})\middle|\boldsymbol{D}\right)+E\left(\Delta_{1,1}^{-1}\right)\middle|\boldsymbol{D}\right)\Big)\\
=&\sum_{g,t}N_{g,t}\varepsilon_{g,t}\left(D_{g,t}^{1}E\left(\Delta_{g,t}^{1}(D_{g,t}^{-1})\middle|\boldsymbol{D}\right)+E\left(\Delta_{g,t}^{-1}\middle|\boldsymbol{D}\right)\right)\\
=&\sum_{(g,t):D_{g,t}^{1}=1}N_{g,t}\varepsilon_{g,t}E\left(\Delta_{g,t}^{1}(D_{g,t}^{2})\middle|\boldsymbol{D}\right)+\sum_{(g,t):D_{g,t}^{-1}\neq\mathbf{0}^{-1}}N_{g,t}\varepsilon_{g,t}E\left(\Delta_{g,t}^{-1}\middle|\boldsymbol{D}\right).
\end{aligned} \tag{10}$$

The first and third equalities follow from Equations (8) and (9). The second equality follows from Lemma 1. The fourth equality follows from Assumption 2 and the fact that $\Delta_{g,t}^{0}(\mathbf{0}^{-1})=0$. Finally, Assumption 2 also implies that

$$\sum_{g,t}N_{g,t}\varepsilon_{g,t}D_{g,t}^{1}=\sum_{(g,t):D_{g,t}^{1}=1}N_{g,t}\varepsilon_{g,t}. \tag{11}$$

Combining (7), (10), (11) yields

$$E\left(\widehat{\beta}_{fe}\middle|\mathbf{D}\right)=\sum_{(g,t):D_{g,t}^{1}=1}\frac{N_{g,t}}{N_{1}}w_{g,t}E\left(\Delta_{g,t}^{1}(D_{g,t}^{2})\middle|\boldsymbol{D}\right)+\sum_{(g,t):D_{g,t}^{-1}\neq\mathbf{0}^{-1}}\frac{N_{g,t}}{N_{1}}w_{g,t}E\left(\Delta_{g,t}^{-1}\middle|\boldsymbol{D}\right). \tag{12}$$

Then, the first result follows from the law of iterated expectations. Finally, if $K=2$ or the treatments are mutually exclusive,

$$\sum_{(g,t):D_{g,t}^{-1}\neq\mathbf{0}^{-1}}N_{g,t}\varepsilon_{g,t}E\left(\Delta_{g,t}^{-1}\middle|\boldsymbol{D}\right)=\sum_{k=2}^{K}\sum_{(g,t):D_{g,t}^{k}=1}N_{g,t}\varepsilon_{g,t}E\left(\Delta_{g,t}^{-1}\middle|\boldsymbol{D}\right).$$

Moreover, by definition of $\varepsilon_{g,t}$, $\sum_{(g,t):D_{g,t}^{k}=1}N_{g,t}\varepsilon_{g,t}=0$ for all $k=2,...,K-1$. The second result follows.

Theorem 3

First, by definition of DID_M,

$$\text{DID}_\text{M} = \sum_{t=2}^{T} \sum_{d^{-1} \in \{0,1\}^{K-1}} \frac{N_{1,0,d^{-1},t}}{N_\mathcal{S}} \text{DID}_{+,d^{-1},t} + \frac{N_{0,1,d^{-1},t}}{N_\mathcal{S}} \text{DID}_{-,d^{-1},t}, \tag{13}$$

using here the convention that $0/0 = 0$. Let $t \geq 2$ and $d^{-1} \in \{0,1\}^{K-1}$ be such that $N_{1,0,d^{-1},t} > 0$ and $N_{0,0,d^{-1},t} > 0$. For every g such that $D^1_{g,t-1} = 0$, $D^1_{g,t} = 1$, and $D^{-1}_{g,t} = D^{-1}_{g,t-1} = d^{-1}$, we have

$$E\left(Y_{g,t} - Y_{g,t-1} | \boldsymbol{D}\right) = E\left(\Delta^1_{g,t}(D^{-1}_{g,t-1}) \middle| \boldsymbol{D}\right) + E\left(Y_{g,t}(0,d^{-1}) - Y_{g,t-1}(0,d^{-1}) \middle| \boldsymbol{D}\right). \tag{14}$$

Under Assumptions 3 and 6, for all $t \geq 2$, there exists $\psi_{0,d^{-1},t} \in \mathbb{R}$ such that for all $g \in \mathcal{G}_{0,0,d^{-1},t} \cup \mathcal{G}_{1,0,d^{-1},t}$,

$$\begin{aligned} E\left(Y_{g,t}(0,d^{-1}) - Y_{g,t-1}(0,d^{-1}) \middle| \boldsymbol{D}\right) &= E\left(Y_{g,t}(0,d^{-1}) - Y_{g,t-1}(0,d^{-1}) \middle| \boldsymbol{D}_g\right) \\ &= E\left(Y_{g,t}(0,d^{-1}) - Y_{g,t-1}(0,d^{-1}) \middle| D^1_{g,t-1} = 0, D^{-1}_{g,t-1} = d^{-1}\right) \\ &= \psi_{0,d^{-1},t}. \end{aligned} \tag{15}$$

As a result,

$$\begin{aligned} & N_{1,0,d^{-1},t} E\left(\text{DID}_{+,d^{-1},t} | \boldsymbol{D}\right) \\ &= \sum_{g \in \mathcal{G}_{1,0,d^{-1},t}} N_{g,t} E\left(\Delta^1_{g,t}(D^{-1}_{g,t-1}) \middle| \boldsymbol{D}\right) + \sum_{g \in \mathcal{G}_{1,0,d^{-1},t}} N_{g,t} E\left(Y_{g,t}(0,d^{-1}) - Y_{g,t-1}(0,d^{-1}) \middle| \boldsymbol{D}\right) \\ &\quad - \frac{N_{1,0,d^{-1},t}}{N_{0,0,d^{-1},t}} \sum_{g \in \mathcal{G}_{0,0,d^{-1},t}} N_{g,t} E\left(Y_{g,t}(0,d^{-1}) - Y_{g,t-1}(0,d^{-1}) \middle| \boldsymbol{D}\right) \\ &= \sum_{g \in \mathcal{G}_{1,0,d^{-1},t}} N_{g,t} E\left(\Delta^1_{g,t}(D^{-1}_{g,t-1}) \middle| \boldsymbol{D}\right) + \psi_{0,d^{-1},t} \left(\sum_{g \in \mathcal{G}_{1,0,d^{-1},t}} N_{g,t} - \frac{N_{1,0,d^{-1},t}}{N_{0,0,d^{-1},t}} \sum_{g \in \mathcal{G}_{0,0,d^{-1},t}} N_{g,t} \right) \\ &= \sum_{g \in \mathcal{G}_{1,0,d^{-1},t}} N_{g,t} E\left(\Delta^1_{g,t}(D^{-1}_{g,t-1}) \middle| \boldsymbol{D}\right). \end{aligned}$$

The first equality follows by (14), the second by (15), and the third after some algebra. Given that $\text{DID}_{+,d^{-1},t} = 0$ if $N_{1,0,d^{-1},t} = 0$ or $N_{0,0,d^{-1},t} = 0$, we obtain, by definition of $\mathcal{S}$ and with the convention that sums over empty sets are 0,

$$E\left(N_{1,0,d^{-1},t} \text{DID}_{+,d^{-1},t} | \boldsymbol{D}\right) = E\left(\sum_{\substack{g: D^1_{g,t}=1, D^{-1}_{g,t}=d^{-1} \\ (g,t) \in \mathcal{S}}} N_{g,t} \Delta^1_{g,t}(D^{-1}_{g,t-1}) \middle| \boldsymbol{D} \right). \tag{16}$$

A similar reasoning yields, for all $t \geq 2$ and $d^{-1} \in \{0,1\}^{K-1}$,

$$E\left(N_{0,1,d^{-1},t} \text{DID}_{-,d^{-1},t} | \boldsymbol{D}\right) = E\left(\sum_{\substack{g: D^1_{g,t}=0, D^{-1}_{g,t}=d^{-1} \\ (g,t) \in \mathcal{S}}} N_{g,t} \Delta^1_{g,t}(D^{-1}_{g,t-1}) \middle| \boldsymbol{D} \right). \tag{17}$$

Plugging (16) and (17) into (13) yields

$$\begin{aligned}E(\text{DID}_\text{M}) =&E\left(E\left(\sum_{t=2}^{T}\sum_{d^{-1}\in\{0,1\}^{K-1}}\sum_{\substack{g:D_{g,t}^{-1}=d^{-1}\\(g,t)\in\mathcal{S}}} N_{g,t}\Delta_{g,t}^{1}(D_{g,t-1}^{-1})\middle|\boldsymbol{D}\right)\right)\\ =&E\left(E\left(\sum_{(g,t)\in\mathcal{S}} N_{g,t}\Delta_{g,t}^{1}(D_{g,t-1}^{-1})\middle|\boldsymbol{D}\right)\right)\\ =&\delta^{\mathcal{S}}.\end{aligned}$$

Theorem 4

First, by Assumption 9, for all $t \geq 2$ there is a function of $\boldsymbol{D}$ $\psi_t(\boldsymbol{D})$ such that

$$\psi_t(\boldsymbol{D}) = E(Y_{g,t}(\mathbf{0}_T;\mathbf{0}_T) - Y_{g,t-1}(\mathbf{0}_T;\mathbf{0}_T)|\boldsymbol{D}). \tag{18}$$

Then, for all $1 \leq f < t \leq T$,

$$\begin{aligned}&E[Y_{g,t}((\mathbf{0}_{f-1},\mathbf{1}_{T-f+1});\mathbf{0}_T) - Y_{g,t-1}((\mathbf{0}_{f-1},\mathbf{1}_{T-f+1});\mathbf{0}_T)|\boldsymbol{D}]\\ =&E[Y_{g,t}((\mathbf{0}_{f-1},\mathbf{1}_{T-f+1});\mathbf{0}_T) - Y_{g,t}(\mathbf{0}_T;\mathbf{0}_T)|\boldsymbol{D}] - E[Y_{g,t-1}((\mathbf{0}_{f-1},\mathbf{1}_{T-f+1});\mathbf{0}_T) - Y_{g,t-1}(\mathbf{0}_T;\mathbf{0}_T)|\boldsymbol{D}]\\ &+ E[Y_{g,t}(\mathbf{0}_T;\mathbf{0}_T) - Y_{g,t-1}(\mathbf{0}_T;\mathbf{0}_T)|\boldsymbol{D}]\\ =&\mu_{f,t}(\boldsymbol{D}) - \mu_{f,t-1}(\boldsymbol{D}) + \psi_t(\boldsymbol{D});\end{aligned} \tag{19}$$

where the second equality uses (18) and Assumption 10. Then, for any $\ell \in \{0,...,L_{nt}\}$, $f \in \mathcal{F}$ such that $NT_f \geq \ell + f + 1$ and $t \in \{\ell + f + 1,...,NT_f\}$ such that $N_{t,\ell}^f > 0$ and $N_t^{nt,f} > 0$,

$$\begin{aligned}&E\left(\text{DID}_{t,\ell}^f|\boldsymbol{D}\right)\\ =&\sum_{g\in\mathcal{G}_f:F_g^2=t-\ell}\frac{N_{g,t}}{N_{t,\ell}^f}E\left(Y_{g,t}-Y_{g,t-\ell-1}|\boldsymbol{D}\right) - \sum_{g\in\mathcal{G}_f:F_g^2>t}\frac{N_{g,t}}{N_t^{nt,f}}E\left(Y_{g,t}-Y_{g,t-\ell-1}|\boldsymbol{D}\right)\\ =&\sum_{g\in\mathcal{G}_f:F_g^2=t-\ell}\frac{N_{g,t}}{N_{t,\ell}^f}E\left(Y_{g,t}(D_g^1;(\mathbf{0}_{t-\ell-1},\mathbf{1}_{\ell+1})) - Y_{g,t}(D_g^1;\mathbf{0}_T)|\boldsymbol{D}\right)\\ &+\sum_{g\in\mathcal{G}_f:F_g^2=t-\ell}\frac{N_{g,t}}{N_{t,\ell}^f}E\left(Y_{g,t}(D_g^1;\mathbf{0}_T) - Y_{g,t-\ell-1}(D_g^1;\mathbf{0}_T)|\boldsymbol{D}\right)\\ &-\sum_{g\in\mathcal{G}_f:F_g^2>t}\frac{N_{g,t}}{N_t^{nt,f}}E\left(Y_{g,t}(D_g^1;\mathbf{0}_T) - Y_{g,t-\ell-1}(D_g^1;\mathbf{0}_T)|\boldsymbol{D}\right)\\ =&\sum_{g\in\mathcal{G}_f:F_g^2=t-\ell}\frac{N_{g,t}}{N_{t,\ell}^f}E\left(Y_{g,t}(D_g^1;(\mathbf{0}_{t-\ell-1},\mathbf{1}_{\ell+1})) - Y_{g,t}(D_g^1;\mathbf{0}_T)|\boldsymbol{D}\right).\end{aligned} \tag{20}$$

The first equality follows from the definition of $\text{DID}_{t,\ell}^f$, and $N_{t,\ell}^f > 0$ and $N_t^{nt,f} > 0$. The second equality follows from Assumption 8. The third equality follows from (19).

By definition of NT_f, we have $N_t^{nt,f} > 0$ for all $f \in \mathcal{F}$ and t such that $NT_f \geq t$. We adopt the convention that a sum over an empty set is equal to 0. Then, for any $\ell \in \{0, ..., L_{nt}\}$, $f \in \mathcal{F}$ such that $NT_f \geq \ell + f + 1$ and $t \in \{\ell + f + 1, ..., NT_f\}$, Equation (20) implies that

$$\begin{aligned} & N_{t,\ell}^f E\left(\mathrm{DID}_{t,\ell}^f | \boldsymbol{D}\right) \\ = & \sum_{g \in \mathcal{G}_f : F_g^2 = t-\ell} N_{g,t} E\left(Y_{g,t}(D_g^1; (\mathbf{0}_{t-\ell-1}, \mathbf{1}_{\ell+1})) - Y_{g,t}(D_g^1; \mathbf{0}_T) | \boldsymbol{D}\right). \end{aligned}$$

We obtain the result by summing over $f \in \mathcal{F}$ and t such that $NT_f \geq \ell + f + 1$ and $t \in \{\ell + f + 1, ..., NT_f\}$, and by the law of iterated expectations.

References

Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies*, 72(1): 1–19.

Abbring, Jaap H, and Gerard J Van den Berg. 2003. "The nonparametric identification of treatment effects in duration models." *Econometrica*, 71(5): 1491–1517.

Ashenfelter, Orley. 1978. "Estimating the effect of training programs on earnings." *The Review of Economics and Statistics*, 47–57.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How much should we trust differences-in-differences estimates?" *The Quarterly Journal of Economics*, 119(1): 249–275.

Borusyak, Kirill, and Xavier Jaravel. 2017. "Revisiting event study designs." Working Paper.

Botosaru, Irene, and Federico H Gutierrez. 2018. "Difference-in-differences when the treatment status is observed in only one period." *Journal of Applied Econometrics*, 33(1): 73–90.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics*, 225: 200–230.

Cantoni, Enrico, and Vincent Pons. 2021. "Strict ID laws don't stop voters: Evidence from a US nationwide panel, 2008–2018." *The Quarterly Journal of Economics*, 136(4): 2615–2660.

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

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2021. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." arXiv preprint arXiv:2007.04267.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2021. "On Estimating Multiple Treatment Effects with Regression." arXiv preprint arXiv:2106.05024.

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

Hotz, V Joseph, and Mo Xiao. 2011. "The impact of regulations on the supply and quality of care in child care markets." *American Economic Review*, 101(5): 1775–1805.

Hull, Peter. 2018. "Estimating Treatment Effects in Mover Designs." arXiv preprint 1804.06721.

Malani, Anup, and Julian Reif. 2015. "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform." *Journal of Public Economics*, 124: 1–17.

Manski, Charles F, and John V Pepper. 2018. "How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions." *Review of Economics and Statistics*, 100(2): 232–244.

Meinhofer, Angélica, Allison Witman, Jesse Hinde, and Kosali Simon. 2021. "Marijuana liberalization policies and perinatal health." *Journal of Health Economics*, 102537.

Panlasigui, Stephanie, Jimena Rico-Straffon, Alexander Pfaff, Jennifer Swenson, and Colby Loucks. 2018. "Impacts of certification, uncertified concessions, and protected areas on forest loss in Cameroon, 2000 to 2013." *Biological conservation*, 227: 160–166.

Rambachan, Ashesh, and Jonathan Roth. 2019. "An honest approach to parallel trends." Working paper.

Robins, James. 1986. "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*, 7(9-12): 1393–1512.

Roth, Jonathan. 2019. "Pre-test with caution: Event-study estimates after testing for parallel trends." *Department of Economics, Harvard University, Unpublished manuscript*.

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