Two-way fixed effects and differences-in-differences estimators with several treatments

Clément de Chaisemartin, Xavier D'haultfœuille

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

Clément de Chaisemartin Xavier D'Haultfœuille*

Abstract

We study two-way-fixed-effects regressions (TWFE) with several treatment variables. Under a parallel trends assumption, we show that the coefficient on each treatment identifies a weighted sum of that treatment's effect, with possibly negative weights, plus a weighted sum of the effects of the other treatments. Thus, those estimators are not robust to heterogeneous effects and may be contaminated by other treatments' effects. We further show that omitting a treatment from the regression can actually reduce the estimator's bias, unlike what would happen under constant treatment effects. We propose an alternative difference-in-differences estimator, robust to heterogeneous effects and immune to the contamination problem. In the application we consider, the TWFE regression identifies a highly non-convex combination of effects, with large contamination weights, and one of its coefficients significantly differs from our heterogeneity-robust estimator.

Keywords: differences-in-differences, two-way-fixed-effects regressions regressions, multiple treatments, heterogeneous treatment effects.

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

Under a parallel trends assumption, TWFE regressions with one treatment identify a weighted sum of the treatment effects of treated (g, t) cells, with weights that may be negative and sum

*de Chaisemartin: Sciences Po, Département d'économie, CNRS, Paris, France (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.

to one (see de Chaisemartin and D'Haultfœuille, 2020; Borusyak and Jaravel, 2017). 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 published in the AER from 2010 to 2012, 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 paper, we investigate 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. A similar weighted sum appears in decompositions of TWFE regressions with only one treatment. The second term is a sum of the effects of the other treatments, with weights summing to zero. Accordingly, with several treatments, coefficients in TWFE regressions may be contaminated by the effect of other treatments, an issue that was not present with one treatment. As the weights sum to zero, this second term disappears if the effect of the other treatments is homogeneous, but it is often implausible that those effects are homogeneous. The weights attached to any TWFE regression with several treatments can be computed by the `twowayfeweights` Stata and R packages. Estimating those weights may be useful, to assess if a TWFE coefficient is robust to heterogeneous treatment effects, and if it is contaminated by the effect of the other treatments in the regression.

We consider simple examples with two treatments, to show that TWFE regressions may not be robust to heterogeneous effects because they may leverage two types of "forbidden comparisons", borrowing the terminology coined by Borusyak and Jaravel (2017). In a first example, the coefficient on the first treatment leverages 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 "control" group going from untreated to receiving the second treatment. If the effect of the second treatment is the same in the two groups, those two effects 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. In a second example, the coefficient on the first treatment leverages a DID comparing the outcome evolution of a group going from untreated to receiving the first treatment to the outcome evolution of a

"control" group that receives the second treatment at both periods. If the control group's effect of the second treatment is the same in the pre and in the post period, those two effects cancel each other out in this DID. But if the control group's effect of the second treatment changes over time, those two effects do not cancel out, and they contaminate the coefficient on the first treatment.

We then consider a TWFE regression that would omit the other treatments, and derive another decomposition formula. In the presence of two treatments, a TWFE regression with only the first treatment also estimates a weighted sum of the effect of that treatment in each group and period, with weights that may be negative and sum to one, plus a weighted sum of the effects of the other treatments, but with weights that do not sum to zero. Then, we use our decompositions of the TWFE regressions with one and several treatments to derive the maximal bias of both regressions for the average effect of the first treatment on the treated, under the assumption that the effect of every treatment is bounded in absolute value by a (potentially large) constant in every group and period. The ratio between the maximal biases of both regressions is independent of that constant and can be estimated, thus allowing researchers to compare the maximal bias of the two regressions. The ratio of the regressions' maximal biases can either be smaller or larger than one in practice, which means that controlling for more treatments may lead to a more biased estimator than not controlling for them. Therefore, omitting a treatment from the regression can actually reduce the estimator's bias, something that cannot happen under constant treatment effects.

Finally, we propose an alternative DID estimator that relies on common trends assumptions, like TWFE regressions, but that is robust to heterogeneous effects and does not suffer from the contamination problem, unlike TWFE regressions. Our estimator generalizes the DID_M estimator in de Chaisemartin and D'Haultfœuille (2020) to instances with several treatments. To isolate the effect of the first treatment, our estimator compares the $t-1$-to-t outcome evolution, of switching groups whose first treatment switches from $t-1$ to t while their other treatments do not change, and of control groups i) whose treatments all remain the same, and ii) that had the same treatments as the switching groups in period $t-1$. i) ensures that our new estimator is robust to heterogeneous effects across groups of all treatments. ii) ensures that it is robust to heterogeneous effects over time of all treatments. Our alternative estimator is computed by the `did_multiplegt` Stata command, we refer the reader to the command's help file for details.

Our estimator's robustness may come at a high price in terms of external validity and statistical precision. For instance, in our application in Section 6, we can only match a small number of switchers to valid control groups meeting i) and ii). Then, there may be internal-external validity and bias-variance trade-offs between our new estimator and less robust estimators, such as the DID_M estimator in de Chaisemartin and D'Haultfœuille (2020) or TWFE regressions with several treatments. To account for the fact our new estimator may sometimes be estimated

on a small sample of groups, we propose, in addition to a standard confidence interval that is asymptotically valid under weak conditions, another confidence interval that has both exact coverage under a normality assumption and is asymptotically valid without such a normality requirement, in the spirit of Donald and Lang (2007).

As an illustration, we use our results to revisit Hotz and Xiao (2011), who 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-to-child ratio. Focusing on the years-of-schooling treatment, we find that the TWFE regression with several treatments estimates weighted sums of effects with very large negative weights attached to them, both on the treatment's own effects, but also on the effects of the other treatments in the regression. The TWFE regression with only the years-of-schooling treatment has much smaller weights attached to it. As a result, the maximal bias of the TWFE regression with several treatments is almost five times larger than that of the regression including only the years-of-schooling treatment. Thus, the "short" regression seems preferable, at least per our maximal-bias metric. We finally show that our heterogeneity-robust estimator is much closer to zero than, and significantly different from, the coefficient of the TWFE regression with several treatments.

The remainder of the paper is organized as follows. Section 2 discusses the related literature. Section 3 presents the set up. Section 4 presents our decomposition results for TWFE regressions with several treatments. Section 5 presents our alternative estimator. Section 6 presents our empirical application. The Appendix gathers all proofs. The Web Appendix gathers several important extensions, summarized in Section 5.4 of the paper.

2 Related literature

2.1 Connection with papers studying TWFE regressions with one treatment

Our paper is closely 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; Goodman-Bacon, 2021; Borusyak and Jaravel, 2017). In particular, Theorem S4 in the Web Appendix of de Chaisemartin and D'Haultfœuille (2020) studies TWFE regressions with one treatment and some time-varying control variables. Under a parallel trend assumption accounting for such covariates, de Chaisemartin and D'Haultfœuille (2020) show that TWFE regressions with one treatment and some controls identify a weighted sum of the treatment effects across all treated (g,t) cells. Our decomposition results are related to, but different from, that result. The weighted sum in Theorem S4 of de Chaisemartin and D'Haultfœuille

(2020) is identical to the first weighted sum in Theorem 2 below. On the other hand, the second weighted sum in Theorem 2, the contamination term, does not appear in Theorem S4 of de Chaisemartin and D'Haultfœuille (2020). This is because the parallel trend assumptions are not the same in the two results. When the other variables in the regression are treatments rather than covariates (see below for the difference between a treatment and a covariate), one can show that the parallel trends condition in Theorem S4 implicitly assumes that the effect of the other treatments is constant, which is why the contamination term disappears.

2.2 Connection with papers studying linear regressions with several treatments

Our paper is also related to three papers that also consider linear regressions with several treatments.

Firstly, our results 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, where the treatments 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 we consider are indicators for having started receiving a single binary-and-staggered treatment ℓ periods ago, and ii) 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), provided no lags are gathered together in the event-study regression they consider.[1]

Our decompositions extend their result, by showing that the contamination bias they first uncovered is very pervasive: it can arise in any TWFE regression with several treatments, rather than in event-study regressions only. In particular, our results apply to situations where the treatments are different, potentially non-mutually exclusive policies, that may not be binary or may not follow a staggered adoption design. Another difference with their work is that with non-mutually exclusive treatments, the contamination weights do not sum to zero. We also provide some novel intuition as to why contamination may arise with different, potentially non-mutually exclusive treatments in the regression. Finally, we show that omitting the other treatments from the regression may not necessarily increase the regression coefficient's bias.

Secondly, 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

[1] 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, finally leading to the result in our Theorem 2.

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, 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.

Thirdly, 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 Theorems 1 and 2 below 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. 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. The weights in their decomposition are functions of the variance-covariance matrix of the treatments conditional on the controls. An interesting difference with our results is that under their conditional independence assumption, the weights on the effect of the first treatment are all positive.

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.

3 Set up

We consider a panel of G groups observed at T periods, respectively indexed by g and t. Typically, groups are geographical entities gathering many observations, but a group could also just be a single individual or firm. For every $(g,t) \in \{1,...,G\} \times \{1,...,T\}$, let $N_{g,t}$ denote the population of cell (g,t), and let $N = \sum_{g,t} N_{g,t}$ be the total population across all cells.

We are interested in the effect of K treatments. In this paper, we follow, e.g., Holland (1986); Holland and Rubin (1987), and define as a treatment a variable that has a causal effect on the outcome, in the sense that different values of that variable lead to different counterfactual outcomes.[2] For every $(k,g,t) \in \{1,...,K\} \times \{1,...,G\} \times \{1,...,T\}$, let $D^k_{g,t}$ denote the value of treatment k for group g at period t, and let $\boldsymbol{D}_{g,t} = (D^k_{g,t})_{k\in\{1,...,K\}}$ denote a vector stacking together the K treatments of group g at period t. For every k, let $\mathcal{D}_k$ denote the values $D^k_{g,t}$ can take. For now, we assume that the treatments are binary: $\mathcal{D}_k = \{0,1\}$ for all k. This is just

[2]By contrast, a covariate may be statistically correlated to the outcome but does not have a causal effect on it.

to simplify the exposition: our results can be extended to non-binary treatments, as explained below.

For any $\boldsymbol{d} \in \{0,1\}^K$, let $Y_{g,t}(\boldsymbol{d})$ denote the potential outcome of group g at period t if $\boldsymbol{D}_{g,t} = \boldsymbol{d}$. The observed outcome is $Y_{g,t} = Y_{g,t}(\boldsymbol{D}_{g,t})$. Importantly, our notation does not necessarily rule out dynamic effects of past or future treatments (the latter in case of anticipation effects) 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 2 below, and in Web Appendix Section 3.1.

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. All expectations below are taken with respect to the distribution of those random variables. On the other hand, the populations of cells (g,t) $N_{g,t}$ are treated as non-random throughout the paper.

Throughout the paper, we maintain the following assumptions. Below, we let $\mathbf{0} = (0,...,0)$ denote the vector of K zeros.

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

Assumption 2 *(Independent groups) The vectors* $((Y_{g,t}(\boldsymbol{d}))_{\boldsymbol{d}\in\{0,1\}^K}, \boldsymbol{D}_{g,t})_{t\in\{1,...,T\}}$ *are mutually independent.*

Assumption 3 *(Strong exogeneity and common trends) For all* $(g,t) \in \{1,...,G\} \times \{2,...,T\}$,

1. $E(Y_{g,t}(\mathbf{0}) - Y_{g,t-1}(\mathbf{0})|\boldsymbol{D}_{g,1},...,\boldsymbol{D}_{g,T}) = E(Y_{g,t}(\mathbf{0}) - Y_{g,t-1}(\mathbf{0}))$.
2. $E(Y_{g,t}(\mathbf{0}) - Y_{g,t-1}(\mathbf{0}))$ *does not vary across* g.

Assumption 1 requires that no group appears or disappears over time. Assumption 2 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). Point 1 of Assumption 3 is related to the strong exogeneity condition in panel data models. It requires that the shocks affecting group g's untreated 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 untreated 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 TWFE regression described in the introduction, as well as our estimand of interest β_{fe}, the expectation of the treatment coefficient in the regression.[3]

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

Regression 1 *(TWFE regression with K treatments)*

Let $\beta_{fe} = E\left[\widehat{\beta}_{fe}\right]$, where $\widehat{\beta}_{fe}$ denotes the coefficient on $D^1_{g,t}$ in a sample OLS regression of $Y_{g,t}$ on group fixed effects, period fixed effects, and the vector $\boldsymbol{D}_{g,t}$, weighted by $N_{g,t}$.[4]

On top of the K treatments, the regression may also include some covariates. The decompositions below can easily be extended to this case, following the same steps as those used by de Chaisemartin and D'Haultfœuille (2020) to extend their decomposition of TWFE regressions with one treatment to TWFE regressions with one treatment and some covariates (see Theorem S4 therein).

Let $\boldsymbol{D}$ be the vector $(\boldsymbol{D}_{g,t})_{(g,t)\in\{1,...,G\}\times\{1,...,T\}}$ 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 $N_1 = \sum_{g,t} N_{g,t} D^1_{g,t}$ denote the total population of cells receiving the first treatment. Let $\boldsymbol{D}^{-1}_{g,t} = (D^2_{g,t},...,D^K_{g,t})$ denote a vector stacking together the treatments of cell (g,t), excluding treatment 1. Let $\varepsilon_{g,t}$ denote the residual of cell (g,t) in the sample regression of $D^1_{g,t}$ on group and period fixed effects and $\boldsymbol{D}^{-1}_{g,t}$:

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

If the regressors in Regression 1 are not collinear, the average value of $\varepsilon_{g,t}$ across all (g,t) cells with $D^1_{g,t} = 1$ differs from 0: $\sum_{(g,t):D^1_{g,t}=1}(N_{g,t}/N_1)\varepsilon_{g,t} = \sum_{(g,t)}(N_{g,t}/N_1)\varepsilon^2_{g,t} > 0$. Then we let $w_{g,t}$ denote $\varepsilon_{g,t}$ divided by that average:

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

4 TWFE regressions with several treatments

4.1 Decomposition results

4.1.1 Two treatment variables

For expositional purposes, we begin by considering the case with two treatments. For any $(g,t) \in \{1,...,G\}\times\{1,...,T\}$, let

$$\Delta^2_{g,t} = Y_{g,t}(0,1) - Y_{g,t}(0,0)$$

[4] The regression could also be estimated using more disaggregated outcome data. For instance, groups may be US counties, and one may estimate the regression using individual-level outcome measures. This disaggregated regression is equivalent to the aggregated regression, provided $Y_{g,t}$ is defined as the average outcome of individuals in cell (g,t), and the aggregated regression is weighted by the number of individuals in cell (g,t). Accordingly, the results below also apply to disaggregated regressions.

denote the effect, in cell (g,t), of moving the second treatment from zero to 1 while keeping the first treatment at zero. Let also

$$\Delta^1_{g,t} = Y_{g,t}(1, D^2_{g,t}) - Y_{g,t}(0, D^2_{g,t})$$

denote the effect, in cell (g,t), of moving the first treatment from zero to one while keeping the second treatment at its observed value. When one estimates a TWFE regression with two treatments, a natural target parameter for β_{fe}, the coefficient on the first treatment, is

$$\delta_{ATT} = E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1}\Delta^1_{g,t}\right],$$

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)s such that $D^1_{g,t} = 1$. δ_{ATT} is the ATT of $D^1_{g,t}$ controlling for $D^2_{g,t}$. We now show that β_{fe} does not identify δ_{ATT} in general.

Theorem 1 *Suppose that Assumptions 1-3 hold and $K = 2$. Then,*

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

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

Theorem 1 shows that the coefficient on $D^1_{g,t}$ 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 (g,t) such that $D^2_{g,t} = 1$, and with weights summing to 0. If the effect of $D^2_{g,t}$ is constant ($\Delta^2_{g,t} = \delta^2$ for all (g,t)), this second term is equal to zero, but it may differ from zero if the effect of $D^2_{g,t}$ is heterogeneous.

Theorem 1 implies that there are two reasons why β_{fe} may differ from δ_{ATT}. First, some of the weights $w_{g,t}$ may differ from one. When the weights $w_{g,t}$ differ from one, one may have that

$$E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta^1_{g,t}\right] \neq \delta_{ATT},$$

if the effect of $D^1_{g,t}$ is heterogeneous across (g,t) cells. Some of the weights $w_{g,t}$ could even be negative, in which case $E\left[\sum_{(g,t):D^1_{g,t}=1}(N_{g,t}/N_1)w_{g,t}\Delta^1_{g,t}\right]$ does not satisfy the no-sign reversal property: this quantity could for instance be negative, even if $\Delta^1_{g,t} \geq 0$ for all (g,t). With two treatments, negative weights can occur even in very simple designs, where there would not be any negative weights in the absence of the second treatment. For instance, consider a standard

DID set-up without variation in treatment timing but 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. In the absence of the second treatment, one can show that the coefficient on $D^1_{g,t}$ in the regression of $Y_{g,t}$ on group fixed effects, period fixed effects, and $D^1_{g,t}$ identifies the ATT of $D^1_{g,t}$ and does not have negative weights attached to it. On the other hand, in the presence of the second treatment, one can show that β_{fe} no longer identifies the ATT of $D^1_{g,t}$ and may have negative weights attached to it (see Corollary 1 in de Chaisemartin and d'Haultfoeuille, 2021, a previous version of this paper, for a formal statement and a proof of these results).

The second reason why β_{fe} may differ from δ_{ATT} is that β_{fe} may also be contaminated by the effect of $D^2_{g,t}$: if that effect is heterogeneous across (g,t) cells, $E\left[\sum_{(g,t):D^2_{g,t}=1}(N_{g,t}/N_1)w_{g,t}\Delta^2_{g,t}\right]$ may differ from zero. Such a contamination phenomenon is not present in the presence of one treatment only (see de Chaisemartin and D'Haultfœuille, 2020). Below, we give some intuition as to why it arises.

Theorem 1 can be extended to non-binary ordered treatments, that may be continuous or discrete. When $D^1_{g,t} \neq 0$, let $S^1_{g,t} = (Y_{g,t}(D^1_{g,t}, D^2_{g,t}) - Y_{g,t}(0, D^2_{g,t}))/D^1_{g,t}$ be the slope of cell (g,t)'s potential outcome function, when moving its first treatment from 0 to $D^1_{g,t}$, while keeping its second treatment at its observed value. Similarly, when $D^2_{g,t} \neq 0$, let $S^2_{g,t} = (Y_{g,t}(0, D^2_{g,t}) - Y_{g,t}(0,0))/D^2_{g,t}$. Finally, let

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

for $k = 1, 2$. If $D^1_{g,t}$ and $D^2_{g,t}$ are non-binary, one can show, following similar steps as in the proof of Theorem 1, that

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

Moreover, $\sum_{(g,t):D^1_{g,t}\neq 0}(N_{g,t}/N_1)w^1_{g,t} = 1$ and $\sum_{(g,t):D^2_{g,t}\neq 0}(N_{g,t}/N_1)w^2_{g,t} = 0$. Essentially, Theorem 1 extends to non-binary treatments, replacing the average treatment effects $\Delta^1_{g,t}$ and $\Delta^2_{g,t}$ by slopes of (g,t)-cells' potential outcome functions, from a treatment of zero to their actual treatment. The decomposition in the previous display does not assume a linear treatment effect.

4.1.2 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

$$\Delta^1_{g,t} = Y_{g,t}(1, \boldsymbol{D}^{-1}_{g,t}) - Y_{g,t}(0, \boldsymbol{D}^{-1}_{g,t}),$$
$$\Delta^{-1}_{g,t} = Y_{g,t}(0, \boldsymbol{D}^{-1}_{g,t}) - Y_{g,t}(0, \mathbf{0}^{-1}).$$

$\Delta^1_{g,t}$ is the effect, in cell (g,t), of moving the first treatment from zero to one while keeping the other treatments at their observed values. $\Delta^{-1}_{g,t}$ is the effect, in cell (g,t), of moving the other treatments from zero to their actual values, while keeping the first treatment at zero.

Theorem 2 below generalizes Theorem 1.

Theorem 2 *Suppose that Assumptions 1-3 hold. Then,*

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

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

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

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

The contamination 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 first treatment's effect.

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

$$E\left[\Delta^{-1}_{g,t}|\boldsymbol{D}\right] = \sum_{k=2}^{K} D^k_{g,t}\delta^k_{g,t}. \tag{3}$$

Then, we obtain Decomposition (4) below. The corresponding weights can be computed using the `twowayfeweights` Stata command.

Corollary 1 *Suppose that Assumptions 1-3 and* (3) *hold. Then,*

$$\beta_{fe} = E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta^1_{g,t} + \sum_{k=2}^{K} \sum_{(g,t):D^k_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\delta^k_{g,t}\right]. \tag{4}$$

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

On the other hand, when the treatments are not mutually exclusive and may be complementary or substitutable $\widehat{\beta}_{fe}$ could be biased even under constant treatment effects. This is because in that case, Regression 1 is misspecified, and should include the interactions of the treatments.

Importantly, Theorem 2 does not necessarily rule out dynamic effects of past treatments on the outcome; similarly, it can allow for anticipation effects. The treatments in the regression may for instance be the current treatment, and some of its lags and leads. In that case, our potential outcome notation allows the current treatment and its lags and leads included in the regression to affect the outcome. Accordingly, the `twowayfeweights` Stata command can also be used to compute the weights attached to distributed-lags regressions of an outcome on the current treatment, and some of its lags and leads.

4.2 Intuition for, and a perhaps surprising implication of, the contamination bias

4.2.1 Intuition for the contamination bias

The reasons why TWFE regressions are not robust to heterogeneous treatment effects are now well understood (see de Chaisemartin and D'Haultfœuille, 2018; de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Borusyak and Jaravel, 2017). In this section, we give intuition as to why β_{fe} may be affected by contamination bias. To do so, we start by considering two very simple examples, one where contamination bias is absent, and the other where it is present.

First, assume that there are three groups and two time periods. With probability one, no group is treated at period 1, and at period 2 group 2 receives the first treatment while group 3 receives the second treatment. Then, it is easy to show that

$$\widehat{\beta}_{fe} = Y_{2,2} - Y_{2,1} - \left(Y_{1,2} - Y_{1,1}\right). \tag{5}$$

The right-hand side of the previous display is a DID comparing the period-one-to-two outcome evolution of group 2, that starts receiving the first treatment at period 2, to that of group 1, that is untreated at both dates. Therefore,

$$\begin{aligned}\beta_{fe} =& E\left(Y_{2,2}(1,0) - Y_{2,1}(0,0) - \left(Y_{1,2}(0,0) - Y_{1,1}(0,0)\right)\right)\\ =& E\left(Y_{2,2}(1,0) - Y_{2,2}(0,0)\right) + E\left(Y_{2,2}(0,0) - Y_{2,1}(0,0) - \left(Y_{1,2}(0,0) - Y_{1,1}(0,0)\right)\right)\\ =& E\left(Y_{2,2}(1,0) - Y_{2,2}(0,0)\right), \end{aligned} \tag{6}$$

where the second equality follows from Assumption 3. Equation (6) is a special case of Equation (2) in Theorem 1. In this simple example, β_{fe} is not contaminated by the effect of the second treatment. It identifies the effect, in group 2 and at period 2, of moving the first treatment from zero to one while keeping the second treatment at its observed value (zero). Because only

group 2 at period 2 receives the first treatment, this effect is equal to δ_{ATT}, the ATT of the first treatment controlling for the second treatment.

Now let us consider another example, very similar to that above, but with a fourth group that receives both treatments at period 2. Then, using the equivalence between TWFE regressions and first-difference regressions with two periods and the fact that the first difference of the two treatments are uncorrelated, we obtain

$$\widehat{\beta}_{fe} = \frac{1}{2}\left(Y_{2,2} - Y_{2,1} - (Y_{1,2} - Y_{1,1})\right) + \frac{1}{2}\left(Y_{4,2} - Y_{4,1} - (Y_{3,2} - Y_{3,1})\right). \tag{7}$$

The first DID in Equation (7) is the same as that in the right-hand side of Equation (5) and it is unbiased for $E\left(Y_{2,2}(1,0) - Y_{2,2}(0,0)\right)$. The second DID compares the period-one-to-two outcome evolution of group 4, that starts receiving the first and second treatments at period 2, to that of group 3, that only starts receiving the second treatment. Therefore,

$$\begin{aligned}
&E\left(Y_{4,2} - Y_{4,1} - (Y_{3,2} - Y_{3,1})\right)\\
=&E\left(Y_{4,2}(1,1) - Y_{4,1}(0,0) - (Y_{3,2}(0,1) - Y_{3,1}(0,0))\right)\\
=&E\left(Y_{4,2}(1,1) - Y_{4,2}(0,1)\right) + E\left(Y_{4,2}(0,1) - Y_{4,2}(0,0)\right) - E\left(Y_{3,2}(0,1) - Y_{3,2}(0,0)\right)\\
+&E\left(Y_{4,2}(0,0) - Y_{4,1}(0,0) - (Y_{3,2}(0,0) - Y_{3,1}(0,0))\right)\\
=&E\left(Y_{4,2}(1,1) - Y_{4,2}(0,1)\right) + E\left(Y_{4,2}(0,1) - Y_{4,2}(0,0)\right) - E\left(Y_{3,2}(0,1) - Y_{3,2}(0,0)\right).
\end{aligned} \tag{8}$$

Equations (7) and (8) imply that

$$\begin{aligned}
\beta_{fe} =&\frac{1}{2}E\left(Y_{2,2}(1,0) - Y_{2,2}(0,0)\right) + \frac{1}{2}E\left(Y_{4,2}(1,1) - Y_{4,2}(0,1)\right)\\
&+\frac{1}{2}E\left(Y_{4,2}(0,1) - Y_{4,2}(0,0)\right) - \frac{1}{2}E\left(Y_{3,2}(0,1) - Y_{3,2}(0,0)\right).
\end{aligned} \tag{9}$$

Equation (9) is a special case of Equation (2) in Theorem 1. β_{fe} identifies the sum of two terms. The term on the first line is the average effect, in groups two and four and at period two, of moving the first treatment from zero to one while keeping the second treatment at its observed value (zero in group 2, one in group 4). The term on the second line is a contamination bias term, equal to the difference, between groups 4 and 3, of the effect of moving the second treatment from zero to one while keeping the first treatment at zero.

The contamination bias appears in the second example because $\widehat{\beta}_{fe}$ leverages a DID comparing a group that starts receiving the first and the second treatments to a group that starts receiving the second treatment only. With heterogeneous treatment effects, this comparison is contaminated by the effect of the second treatment. On the other hand, if the effect of the second treatment does not vary across groups, this contamination bias disappears. To our knowledge, our paper is the first to show that TWFE regressions with several treatments leverage this type of "forbidden comparisons", using the terminology coined by Borusyak and Jaravel (2017).

In the example with four groups, a simple solution to eliminate the contamination bias is to add the interaction of the two treatments to the regression. One can in fact show the following, slightly more general result. With only two time periods, and groups that do not receive any of the two treatments in the first period, the coefficient on $D^1_{g,t}$ in the regression of $Y_{g,t}$ on $D^1_{g,t}$, $D^2_{g,t}$, and $D^1_{g,t}D^2_{g,t}$ is not contaminated by the effect of the second treatment. In such cases, 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 in Section 6: in the regression without control variables and with the two main treatments (the minimum staff-to-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.

In the first example, the two treatments are mutually exclusive so $\widehat{\beta}_{fe}$ cannot leverage a "forbidden" DID comparing a group that starts receiving the first and the second treatments to a group that starts receiving the second treatment only, which is why there is no contamination bias in this example. This does not mean contamination bias never arises with mutually exclusive treatments. To illustrate this point, let us consider a third example with two groups and three periods. Group 1 receives the first treatment at period 3, and Group 2 receives the second treatment at periods 2 and 3. Then, because this regression is equivalent to a regression of $Y_{2,t} - Y_{1,t}$ on a constant, $D^1_{2,t} - D^1_{1,t}$ and $D^2_{2,t} - D^2_{1,t}$, we obtain, after some algebra,

$$\widehat{\beta}_{fe} = Y_{1,3} - Y_{1,2} - (Y_{2,3} - Y_{2,2})\,. \tag{10}$$

Accordingly, one can show that

$$\begin{aligned}\beta_{fe} =& E\left(Y_{1,3}(1,0) - Y_{1,3}(0,0)\right)\\ &+E\left(Y_{2,2}(0,1) - Y_{2,2}(0,0)\right) - E\left(Y_{2,3}(0,1) - Y_{2,3}(0,0)\right).\end{aligned} \tag{11}$$

$\widehat{\beta}_{fe}$ is contaminated by the effect of the second treatment, because it leverages a DID where the control group receives the second treatment at both dates. This second type of "forbidden" DID is very similar to the late- versus early-treated DIDs due to which TWFE regressions with one treatment are not robust to heterogeneous treatment effects (see de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Borusyak and Jaravel, 2017). Note that if the effect of the second treatment is constant over time, the contamination bias term disappears. However, constant effects over time is often an implausible assumption.

Overall, TWFE regressions with several treatments are not affected by contamination bias in very simple designs with two time periods, where groups are only treated in the second period, and where the treatments are mutually exclusive. In designs with non-mutually exclusive treatments,

contamination bias may appear because $\widehat{\beta}_{fe}$ may leverage DIDs comparing a group that starts receiving, say, the first and the second treatments to a group that starts receiving the second treatment only. With more than two time periods, even if the treatments are mutually exclusive, $\widehat{\beta}_{fe}$ may leverage DIDs comparing a group that starts receiving, say, the first treatment, to a group receiving the second treatment at both dates.

4.2.2 A perhaps surprising implication of the contamination bias

Theorem 1 has an important and perhaps surprising consequence for TWFE regressions with one treatment where one seeks to estimate heterogeneous treatment effects. Oftentimes, researchers run a TWFE regression with a treatment variable $D_{g,t}$ interacted with a group-level binary variable I_g, and with $(1 - I_g)$.[5] 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 on $D_{g,t} \times I_g$, and $\Delta_{g,t} = Y_{g,t}(1) - Y_{g,t}(0)$. The previous display implies that the coefficient on $D_{g,t} \times I_g$ is contaminated by the treatment effect in (g, t) cells such that $I_g = 0$. In the example, the coefficient on the treatment interacted with the indicator for rich counties is contaminated by the treatment effect in poor counties. This calls into question the use of such TWFE regressions to estimate heterogeneous effects.

This contamination phenomenon disappears if the time fixed effects are interacted with I_g in the regression. Then, the coefficient on $D_{g,t} \times I_g$ becomes equivalent to that one would obtain by running a TWFE regression restricting the sample to groups such that $I_g = 1$. It follows from de Chaisemartin and D'Haultfœuille (2020) that this coefficient identifies a weighted sum of the treatment effects across (g, t) cells such that $D_{g,t} = 1, I_g = 1$: it is not contaminated by the treatment effect in (g, t) cells such that $D_{g,t} = 1, I_g = 0$.

4.3 Should one control for other treatments?

In this section, we derive a decomposition similar to that in Theorem 1, when there are two treatments but the second treatment is omitted from the regression.

[5] Researchers may instead have $D_{g,t}$ and $D_{g,t}I_g$ in the regression. The coefficient on $D_{g,t}$ in this regression is equal to that on $D_{g,t}(1 - I_g)$ in the regression described in the text. The coefficient on $D_{g,t}I_g$ is equal to the difference between that on $D_{g,t}I_g$ and that on $D_{g,t}(1 - I_g)$ in the regression described in the text. Accordingly, the discussion in this section also applies to those regressions.

Regression 2 *(Short TWFE regression)*

Let $\beta^s_{fe} = E\left[\widehat{\beta}^s_{fe}\right]$*, where* $\widehat{\beta}^s_{fe}$ *denotes the coefficient on* $D^1_{g,t}$ *in a sample OLS regression of* $Y_{g,t}$ *on group fixed effects, period fixed effects, and* $D^1_{g,t}$*, weighted by* $N_{g,t}$*.*

Let $\varepsilon^s_{g,t}$ denote the residual of cell (g,t) in the sample regression of $D^1_{g,t}$ on group and period fixed effects. If the regressors in Regression 2 are not collinear, the average value of $\varepsilon^s_{g,t}$ across all (g,t) cells with $D^1_{g,t}=1$ differs from 0: $\sum_{(g,t):D^1_{g,t}=1}(N_{g,t}/N_1)\varepsilon^s_{g,t} \neq 0$. Then we let $w^s_{g,t}$ denote $\varepsilon^s_{g,t}$ divided by that average:

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

Theorem 3 *Suppose that Assumptions 1-3 hold and* $K=2$*. Then,*

$$\beta^s_{fe} = E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w^s_{g,t}\Delta^1_{g,t} + \sum_{(g,t):D^2_{g,t}=1} \frac{N_{g,t}}{N_1} w^s_{g,t}\Delta^2_{g,t}\right]. \tag{12}$$

Moreover, $\sum_{(g,t):D^1_{g,t}=1}(N_{g,t}/N_1)w^s_{g,t} = 1$ *and* $\sum_{(g,t):D^2_{g,t}=1}(N_{g,t}/N_1)w^s_{g,t}$ *may differ from zero.*

Theorem 3 is similar to Theorem 1. It shows that the coefficient on $D^1_{g,t}$ in the short regression identifies the sum of two terms. The first term in Theorem 3 is similar to that in Theorem 1, namely a weighted sum of the effect of moving $D^1_{g,t}$ from 0 to 1 while keeping $D^2_{g,t}$ at its observed value, but with different weights that still sum to one. The second term in Theorem 3 is also similar to that in Theorem 1, namely a weighted sum of the effect of moving $D^2_{g,t}$ from 0 to 1 while keeping $D^1_{g,t}$ at 0, but with different weights that no longer sum to zero. Note that Theorem 3 can easily be extended to instances with more than two treatments.

Theorem 3 is also similar to Theorem 1 in de Chaisemartin and D'Haultfœuille (2020). With the notation of this paper, Theorem 1 in de Chaisemartin and D'Haultfœuille (2020) provides a decomposition of β^s_{fe} under a parallel trends assumption on $Y_{g,t}(0, D^{-1}_{g,t})$, the potential outcome of g at t with the first treatment set at 0 and the other treatments set at their actual values. The weighted sum of effects in Theorem 1 of de Chaisemartin and D'Haultfœuille (2020) is identical to the first weighted sum in Theorem 3. On the other hand, the contamination term in Theorem 3 does not appear in Theorem 1 of de Chaisemartin and D'Haultfœuille (2020), because the parallel trend assumptions underlying the two results are not the same. There may be instances where parallel trends on $Y_{g,t}(0, D^{-1}_{g,t})$ is plausible, in which case the decomposition in Theorem 1 in de Chaisemartin and D'Haultfœuille (2020) is applicable. In this paper, as the researcher estimates a TWFE regression with several treatments, it is natural to consider instead a parallel trends assumption on $Y_{g,t}(\mathbf{0})$.

Under constant effects and a parallel trends assumption on $Y_{g,t}(0,0)$, omitting the second treatment from the regression leads to an omitted variable bias, and including the second treatment

into the regression is always preferable. But this may not be the case with heterogeneous treatment effects: because the weights associated with the two regressions differ, $D^1_{g,t}$'s coefficient in the long regression may be more biased for δ_{ATT} than $D^1_{g,t}$'s coefficient in the short regression. The following corollary formalizes this idea.

Corollary 2 *Suppose that Assumptions 1-3 hold, $K = 2$, and there is a real number B such that $|\Delta^1_{g,t}| \leq B$ and $|\Delta^2_{g,t}| \leq B$ for all (g,t). Then,*

$$|\beta_{fe} - \delta_{ATT}| \leq B \times E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1}|w_{g,t} - 1| + \sum_{(g,t):D^2_{g,t}=1} \frac{N_{g,t}}{N_1}|w_{g,t}|\right],$$

$$|\beta^s_{fe} - \delta_{ATT}| \leq B \times E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1}|w^s_{g,t} - 1| + \sum_{(g,t):D^2_{g,t}=1} \frac{N_{g,t}}{N_1}|w^s_{g,t}|\right].$$

Moreover, both upper bounds are sharp.

Corollary 2 assumes that the effects of the first and second treatments are both bounded in every (g,t) cell by a constant B. Under that assumption, it gives the maximal biases of $\widehat{\beta}_{fe}$ and $\widehat{\beta}^s_{fe}$ as estimators of δ_{ATT}, the ATT of $D^1_{g,t}$ controlling for $D^2_{g,t}$. One can compare those maximal biases by comparing (estimates of)

$$E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1}|w_{g,t} - 1| + \sum_{(g,t):D^2_{g,t}=1} \frac{N_{g,t}}{N_1}|w_{g,t}|\right]$$

and

$$E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1}|w^s_{g,t} - 1| + \sum_{(g,t):D^2_{g,t}=1} \frac{N_{g,t}}{N_1}|w^s_{g,t}|\right],$$

which does not require specifying B.[6] The maximal bias of $\widehat{\beta}_{fe}$ could be larger than that of $\widehat{\beta}^s_{fe}$, if for (g,t)s such that $D^1_{g,t} = 1$ the weights $w_{g,t}$ are on average further away from one than the weights $w^s_{g,t}$, and/or if for (g,t)s such that $D^2_{g,t} = 1$ the contamination weights $w_{g,t}$ are on average further away from zero than the weights $w^s_{g,t}$. In our application in Section 6, we find that the estimated maximal bias of the long regression is almost five times larger than that of the short regression. Then, the short regression is preferable, at least per our maximal-bias metric.

[6] A similar result holds if we consider distinct bounds B_1 and B_2 for $|\Delta^1_{g,t}|$ and $|\Delta^2_{g,t}|$. Then, one has to multiply $\sum_{(g,t):D^2_{g,t}=1}(N_{g,t}/N_1)|w_{g,t}|$ by (B_2/B_1) when performing the comparison of the maximal biases. Hence, in this case, one needs to take a stand on the ratio B_2/B_1.

5 Alternative estimator

5.1 Identifying assumption

In this section, we start by considering the following identifying assumption. Recall that $Y_{g,t}(\boldsymbol{d})$ denotes the potential outcome of g at t, if the treatment vector is equal to $\boldsymbol{d}$.

Assumption 4 *(Strong exogeneity and common trends from $t-1$ to t, conditional on $\boldsymbol{D}_{g,t-1}$) For all $(g,t) \in \{1,...,G\} \times \{2,...,T\}$ and all $\boldsymbol{d} \in \{0,1\}^K$,*

1. $E(Y_{g,t}(\boldsymbol{d}) - Y_{g,t-1}(\boldsymbol{d}) | \boldsymbol{D}_{g,1},...,\boldsymbol{D}_{g,t-2},\boldsymbol{D}_{g,t-1}=\boldsymbol{d},\boldsymbol{D}_{g,t},...,\boldsymbol{D}_{g,T}) = E(Y_{g,t}(\boldsymbol{d}) - Y_{g,t-1}(\boldsymbol{d}) | \boldsymbol{D}_{g,t-1}=\boldsymbol{d})$.

2. $E(Y_{g,t}(\boldsymbol{d}) - Y_{g,t-1}(\boldsymbol{d}) | \boldsymbol{D}_{g,t-1}=\boldsymbol{d})$ *does not vary across g.*

Like Assumption 3, Assumption 4 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 their period-t treatments all remain at their $t-1$ value, 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. Then, consider a group whose first treatment changes between $t-1$ and t, but whose other treatments remain constant. Under Assumption 4, the $t-1$-to-t evolution of its outcome had its first treatment not changed is identified by the outcome evolution of groups whose treatments all remain constant and with the same period-$t-1$ treatments.

We now compare our new assumption, Assumption 4, to the more standard Assumption 3. The two assumptions are non-nested, and there are two main differences between them. First, Assumption 3 requires that all groups be on parallel trends, over the entire duration of the panel. Assumption 4, on the other hand, only requires that groups with the same period-$t-1$ treatments be on parallel trends, from $t-1$ to t. Assumption 4 may then be more plausible: groups with the same treatments in the baseline period may be more similar, and may be more likely to experience parallel trends.[7] Moreover, parallel trends may be more likely to hold over consecutive time periods than over the panel's entire duration.

[7]Because it imposes parallel trends conditional on $\boldsymbol{D}_{g,t-1}$, Assumption 4 may be seen as "in-between" a standard parallel trends assumption and the sequential ignorability assumption, another commonly-used identifying assumption in panel data models (see, e.g., Robins, 1986; Bojinov, Rambachan and Shephard, 2021). Sequential ignorability requires that treatment be uncounfounded conditional on prior treatment and outcome, which implies parallel trends conditional on prior treatment and outcome. Because Assumption 4 does not condition on groups' $t-1$ outcomes, it may be less plausible than sequential ignorability. At the same time, estimators relying on sequential ignorability need to compare groups with the same prior treatments and outcomes. This may lead to a curse of dimensionality.

Second, Assumption 3 is a parallel trends assumption in the counterfactual where groups do not receive any treatment, while Assumption 4 is a parallel trends assumption in the counterfactual where groups' treatments do not change from $t-1$ to t. Accordingly, Assumption 3 only restricts one potential outcome, the one without any treatment, while Assumption 4 imposes restrictions on many potential outcomes. Still, Assumption 4 does not impose any restriction on treatment effect heterogeneity, because it restricts only one potential outcome per (g,t) cell, namely $Y_{g,t}(\boldsymbol{d})$ for (g,t) cells such that $\boldsymbol{D}_{g,t-1}=\boldsymbol{d}$. In particular, Assumption 4 does not require that all groups experience the same evolution of their treatment effect. Moreover, in complicated designs where the number of treatments is large and/or when the treatments are non binary, Assumption 4 may have considerably more identifying power than Assumption 3. Under Assumption 3, an heterogeneity-robust DID estimator can only use as controls groups that do not receive any treatment at two dates at least. Moreover, treatment effects can only be estimated for groups that do not receive any treatment at one date at least. With many treatments and/or when the treatments are non binary, those two sets of groups may be small. In our empirical application in Section 6, 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, we cannot construct an heterogeneity-robust DID estimator relying on Assumption 3, while we can construct one relying on Assumption 4.

Importantly, there are special cases where: i) the two assumptions are equivalent and ii) our decomposition of $\widehat{\beta}_{fe}$ under Assumption 3 has contamination weights attached to it. This shows that decomposing $\widehat{\beta}_{fe}$ under Assumption 4 can also lead to contamination weights. For instance, with two periods and $\boldsymbol{D}_{g,1}=0$ almost surely, Assumption 3 and 4 are equivalent. As a result, contamination weights may arise under Assumption 4, as the example we give p.13 demonstrates. In designs where Assumptions 4 and 3 are not equivalent, under Assumption 4 we cannot in general write β_{fe} as a function of the design and treatment effects only, and replacing Assumption 3 by Assumption 4 may actually exacerbate the problems of TWFE regressions: in addition to not being robust to heterogeneous treatment effects, TWFE regressions may now be biased even with homogenous treatment effects. We provide an example in Web Appendix Section 1.

We also consider a second identifying assumption.

Assumption 5 *(Strong exogeneity and common trends from $t-1$ to t, conditional on $\boldsymbol{D}_{g,t}$) For all $(g,t)\in\{1,...,G\}\times\{2,...,T\}$ and all $\boldsymbol{d}_t\in\{0,1\}^K$,*

1. $E(Y_{g,t}(\boldsymbol{d}_t)-Y_{g,t-1}(\boldsymbol{d}_t)|\boldsymbol{D}_{g,1},...,\boldsymbol{D}_{g,t-1},\boldsymbol{D}_{g,t}=\boldsymbol{d}_t,\boldsymbol{D}_{g,t+1},...,\boldsymbol{D}_{g,T})=E(Y_{g,t}(\boldsymbol{d}_t)-Y_{g,t-1}(\boldsymbol{d}_t)|\boldsymbol{D}_{g,t}=\boldsymbol{d}_t)$.
2. $E(Y_{g,t}(\boldsymbol{d}_t)-Y_{g,t-1}(\boldsymbol{d}_t)|\boldsymbol{D}_{g,t}=\boldsymbol{d}_t)$ *does not vary across g.*

Assumption 5 is similar to Assumption 4, except that it assumes parallel trends from $t-1$ to t, in the counterfactual where groups keep their period-t rather than their period-$t-1$ treatments.

Imposing jointly Assumptions 4 and 5 may imply that the treatment effects follow the same evolution over time in some groups.[8]

5.2 Target parameters

Let us define

$$\mathcal{S}_1 = \left\{ (g,t) : t \geq 2, D^1_{g,t} \neq D^1_{g,t-1}, \boldsymbol{D}^{-1}_{g,t} = \boldsymbol{D}^{-1}_{g,t-1}, \exists g' : \boldsymbol{D}_{g',t} = \boldsymbol{D}_{g',t-1} = \boldsymbol{D}_{g,t-1} \right\}$$

and let $N_{\mathcal{S}_1} = \sum_{(g,t)\in\mathcal{S}_1} N_{g,t}$. $\mathcal{S}_1$ 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 $t-1$. Hereafter, those cells are referred to as switchers. We show below that under Assumption 4, one can unbiasedly estimate

$$\delta_1 = E\left[\sum_{(g,t)\in\mathcal{S}_1} \frac{N_{g,t}}{N_{\mathcal{S}_1}} \Delta^1_{g,t} \right],$$

the average effect of moving the first treatment from 0 to 1 while keeping all other treatments at their observed value, across all switchers.[9]

δ_1 may differ from δ_{ATT}, arguably a more natural target parameter. The two parameters apply to different and non-nested sets of (g,t) cells. Let $\mathcal{T}_1 = \{(g,t) : D^1_{g,t} = 1\}$. δ_1 is the average of $\Delta^1_{g,t}$ across all cells in $\mathcal{S}_1$. δ_{ATT} is the average effect of $\Delta^1_{g,t}$ across all cells in $\mathcal{T}_1$. The following proposition shows that in our set-up, we cannot identify treatment effects on cells outside $\mathcal{S}_1$.

Proposition 1 *Suppose that Assumptions 1-2 and 4 hold. Then, for any subset $\mathcal{V}$ of $\mathcal{T}_1 \backslash \mathcal{S}_1$, and letting $N_{\mathcal{V}} = \sum_{(g,t)\in\mathcal{V}} N_{g,t}$, $E[\sum_{(g,t)\in\mathcal{V}}(N_{g,t}/N_{\mathcal{V}})\Delta^1_{g,t}]$ is not identified.*

Proposition 1 shows that $\mathcal{S}_1$ is the maximal set of cells for which treatment effects can be identified under Assumptions 1-2 and 4. The (g,t) cells belonging to $\mathcal{T}_1$ but not to $\mathcal{S}_1$ can be divided into five mutually exclusive subgroups, detailed in Web Appendix Section 2. Identifying the effect of the first treatment in each of those subgroups would either require restricting treatment effect heterogeneity, or making parallel trend restrictions different from those in Assumption 4.

While we expect $\mathcal{S}_1$ to be often smaller than $\mathcal{T}_1$, there are also (g,t) cells that belong to $\mathcal{S}_1$ but not to $\mathcal{T}_1$. Those are the switching-out cells, such that $D^1_{g,t} = 0, D^1_{g,t-1} = 1, \boldsymbol{D}^{-1}_{g,t} = \boldsymbol{D}^{-1}_{g,t-1}, \exists g' : \boldsymbol{D}_{g',t} = \boldsymbol{D}_{g',t-1} = \boldsymbol{D}_{g,t-1}$.

[8] For instance, if $K = 1$, $G = 4$, $T = 2$, $D^1_{1,1} = D^1_{1,2} = 0$, $D^1_{2,1} = D^1_{2,2} = 1$, $D^1_{3,1} = 0, D^1_{3,2} = 1$, and $D^1_{4,1} = 1, D^1_{4,2} = 0$, one can show that together, Assumptions 4 and 5 imply that the treatment effect follows the same evolution in groups 3 and 4.

[9] When $N_{\mathcal{S}_1} = 0$, we simply let the term inside brackets be equal to 0.

As δ_1 and δ_{ATT} apply to different, non-nested subpopulations, a significant difference between $\widehat{\beta}_{fe}$ and the estimator of δ_1 we propose below cannot be interpreted as evidence that $\widehat{\beta}_{fe}$ is biased for δ_{ATT}. It could also be the case that $\widehat{\beta}_{fe}$ is unbiased for δ_{ATT} and δ_1 and δ_{ATT} differ. On the other hand, under Assumptions 3 and 4, a significant difference between $\widehat{\beta}_{fe}$ and the estimator of δ_1 implies that the effect of at least one treatment is not constant.

Similarly, we show below that under Assumption 5, one can unbiasedly estimate

$$\delta_2 = E\left[\sum_{(g,t)\in\mathcal{S}_2} \frac{N_{g,t}}{N_{\mathcal{S}_2}} \Delta^1_{g,t}\right],$$

where

$$\mathcal{S}_2 = \left\{(g,t) : t \leq T-1, D^1_{g,t} \neq D^1_{g,t+1}, \boldsymbol{D}^{-1}_{g,t} = \boldsymbol{D}^{-1}_{g,t+1}, \exists g' : \boldsymbol{D}_{g',t} = \boldsymbol{D}_{g',t+1} = \boldsymbol{D}_{g,t+1}\right\},$$

and $N_{\mathcal{S}_2} = \sum_{(g,t)\in\mathcal{S}_2} N_{g,t}$. $\mathcal{S}_2$ is the set of cells (g,t) whose first treatment changes between t and $t+1$ while their other treatments do not change, and such that there is another group g' whose treatments do not change between t and $t+1$, and with the same treatments as g in $t+1$. $\mathcal{S}_1$ and $\mathcal{S}_2$ are not necessarily disjoints: a (g,t) cell experiencing two consecutive changes of its first treatment ($D^1_{g,t-1} \neq D^1_{g,t}$ and $D^1_{g,t} \neq D^1_{g,t+1}$) may belong both to δ_1 and to δ_2. On the other hand, a (g,t) cell that does not experience two consecutive changes of its first treatment ($D^1_{g,t-1} = D^1_{g,t}$ or $D^1_{g,t} = D^1_{g,t+1}$) may belong to δ_1 or to δ_2 but cannot belong to both sets.

Finally, under Assumptions 4 and 5, one can unbiasedly estimate

$$\delta = E\left[\sum_{(g,t)\in\mathcal{S}_1\cup\mathcal{S}_2} \frac{N_{g,t}}{N_{\mathcal{S}_1\cup\mathcal{S}_2}} \Delta^1_{g,t}\right],$$

where $N_{\mathcal{S}_1\cup\mathcal{S}_2} = \sum_{(g,t)\in\mathcal{S}_1\cup\mathcal{S}_2} N_{g,t}$.

5.3 Estimation

We now show that under Assumption 4, δ_1 can be unbiasedly estimated by a weighted average of DIDs. For all $t \in \{2,...,T\}$, for all $(d,d') \in (\mathcal{D}_1)^2$, and for all $\boldsymbol{d}^{-1} \in \mathcal{D}_2 \times ... \times \mathcal{D}_K$, let

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

be the set of groups whose first treatment goes from d' to d from $t-1$ to t while their other treatments are equal to $\boldsymbol{d}^{-1}$ at both dates. We then let $N_{d,d',\boldsymbol{d}^{-1},t} = \sum_{g\in\mathcal{G}_{d,d',\boldsymbol{d}^{-1},t}} N_{g,t}$ denote the total population of groups in $\mathcal{G}_{d,d',\boldsymbol{d}^{-1},t}$. Let also

$$\text{DID}^f_{+,\boldsymbol{d}^{-1},t} = \sum_{g\in\mathcal{G}_{1,0,\boldsymbol{d}^{-1},t}} \frac{N_{g,t}}{N_{1,0,\boldsymbol{d}^{-1},t}} \left(Y_{g,t} - Y_{g,t-1}\right) - \sum_{g\in\mathcal{G}_{0,0,\boldsymbol{d}^{-1},t}} \frac{N_{g,t}}{N_{0,0,\boldsymbol{d}^{-1},t}} \left(Y_{g,t} - Y_{g,t-1}\right), \tag{13}$$

$$\text{DID}^f_{-,\boldsymbol{d}^{-1},t} = \sum_{g\in\mathcal{G}_{1,1,\boldsymbol{d}^{-1},t}} \frac{N_{g,t}}{N_{1,1,\boldsymbol{d}^{-1},t}} \left(Y_{g,t} - Y_{g,t-1}\right) - \sum_{g\in\mathcal{G}_{0,1,\boldsymbol{d}^{-1},t}} \frac{N_{g,t}}{N_{0,1,\boldsymbol{d}^{-1},t}} \left(Y_{g,t} - Y_{g,t-1}\right). \tag{14}$$

Note that $\text{DID}^f_{+,\boldsymbol{d}^{-1},t}$ is not defined when $N_{1,0,\boldsymbol{d}^{-1},t}=0$ or $N_{0,0,\boldsymbol{d}^{-1},t}=0$. In such instances, we let $\text{DID}^f_{+,\boldsymbol{d}^{-1},t}=0$. Similarly, we let $\text{DID}^f_{-,\boldsymbol{d}^{-1},t}=0$ when $N_{1,1,\boldsymbol{d}^{-1},t}=0$ or $N_{0,1,\boldsymbol{d}^{-1},t}=0$.

$\text{DID}^f_{+,\boldsymbol{d}^{-1},t}$ 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 $\boldsymbol{d}^{-1}$ at both dates, to the outcome evolution of groups whose first and other treatments are respectively equal to 0 and $\boldsymbol{d}^{-1}$ at both dates. Under Assumption 4, 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. $\text{DID}^f_{-,\boldsymbol{d}^{-1},t}$'s interpretation is similar, except that it compares groups whose first treatment is equal to 1 at both dates to groups whose first treatment goes from 1 to 0.

Finally, let

$$\text{DID}^f_{\text{M}}=\sum_{t=2}^{T}\sum_{\boldsymbol{d}^{-1}\in\{0,1\}^{K-1}}\left(\frac{N_{1,0,\boldsymbol{d}^{-1},t}}{N_{\mathcal{S}_1}}\text{DID}^f_{+,\boldsymbol{d}^{-1},t}+\frac{N_{0,1,\boldsymbol{d}^{-1},t}}{N_{\mathcal{S}_1}}\text{DID}^f_{-,\boldsymbol{d}^{-1},t}\right) \tag{15}$$

if $N_{\mathcal{S}_1}>0$, and $\text{DID}^f_{\text{M}}=0$ if $N_{\mathcal{S}_1}=0$. DID^f_{M} is just a weighted average of the $\text{DID}^f_{+,\boldsymbol{d}^{-1},t}$ and $\text{DID}^f_{-,\boldsymbol{d}^{-1},t}$ estimators, across values of the other treatments $\boldsymbol{d}^{-1}$ and across time periods t.

Theorem 4 *If Assumptions 1-2 and 4 hold,* $E\left[DID^f_M\right]=\delta_1$.

DID^f_{M} extends the DID_{M} estimator in de Chaisemartin and D'Haultfœuille (2020) to settings with several treatments. With several treatments, one could show the analogue of Theorem 3 for the DID_{M} estimator in de Chaisemartin and D'Haultfœuille (2020): the fact that this estimator does not control for the other treatments may lead to a bias. To avoid that, the DID^f_{M} and DID_{M} estimators differ on three important dimensions: DID^f_{M} 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; 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; and it compares switchers and non-switchers with the same baseline values of their other treatments. All those modifications ensure that our new estimator is not biased in the presence of other treatments with potentially heterogeneous treatment effects, but they may also come at a cost in terms of precision: the DID^f_{M} estimator in this paper discards several cells from the estimation. Accordingly, there may be a bias-variance trade-off between the two estimators. DID^f_{M} is computed by the `did_multiplegt` Stata command, we refer the reader to the command's help file for details.[10]

[10] DID^f_{M} estimates the effect of the first treatment, holding other treatments fixed. One may instead be interested in interaction effects. DID^f_{M} can easily be used for that purpose. For instance, with two binary treatments, to estimate their interaction effect, one can compute DID^f_{M} separately for $D^1_{g,t}$ switchers with $D^2_{g,t}=1$ and $D^2_{g,t}=0$, and then take the difference between the two estimated effects. A significant difference between the two estimated effects is only suggestive evidence of an interaction effect: the effect of $D^1_{g,t}$ could also just differ in the two subsamples.

Like in de Chaisemartin and D'Haultfœuille (2020), it is straightforward to propose a placebo version of the $\text{DID}_{\text{M}}^{f}$ estimator that one can use to test Assumption 4. To do so, one just needs to replace $Y_{g,t}-Y_{g,t-1}$ by $Y_{g,t-1}-Y_{g,t-2}$ in Equations (13) and (14) above, and exclude from the estimation groups experiencing a change in any of their treatments from $t-2$ to $t-1$. The resulting placebo estimator compares the outcome evolution of switchers and non-switchers, before switchers switch.

The $\text{DID}_{\text{M}}^{f}$ estimator can be extended to accommodate discrete non-binary treatments taking values in $\mathcal{D}_1=\{0,...,\overline{d}\}$, like the DID_{M} estimator in de Chaisemartin and D'Haultfœuille (2020) (see Web Appendix Section 4 of de Chaisemartin and D'Haultfœuille, 2020). For all $t\in\{2,...,T\}$, for all $(d,d')\in(\mathcal{D}_1)^2$, and for all $\boldsymbol{d}^{-1}\in\mathcal{D}_2\times...\times\mathcal{D}_K$, let

$$\text{DID}^{f}_{d,d',\boldsymbol{d}^{-1},t}=[1\{d'<d\}-1\{d<d'\}]\Bigg[\sum_{g\in\mathcal{G}_{d,d',\boldsymbol{d}^{-1},t}}\frac{N_{g,t}}{N_{d,d',\boldsymbol{d}^{-1},t}}[Y_{g,t}-Y_{g,t-1}]$$
$$-\sum_{g\in\mathcal{G}_{d',d',\boldsymbol{d}^{-1},t}}\frac{N_{g,t}}{N_{d',d',\boldsymbol{d}^{-1},t}}[Y_{g,t}-Y_{g,t-1}]\Bigg]$$

be a DID estimator comparing the $t-1$-to-t outcome evolution in groups whose first treatment changes from d' to d and whose other treatments are equal to $\boldsymbol{d}^{-1}$ at both dates, to the same outcome evolution in groups whose treatments do not change and with the same treatments in $t-1$. With a non-binary treatment, the $\text{DID}_{\text{M}}^{f}$ estimator is a weighted average of the $\text{DID}^{f}_{d,d',\boldsymbol{d}^{-1},t}$ estimators, across d, d', $\boldsymbol{d}^{-1}$, and t, normalized by the average change of the first treatment among switchers, to ensure the estimator can be interpreted as an effect produced by a one-unit increase of the first treatment.

Similarly, under Assumption 5, and getting back to the binary treatment case, δ_2 can be unbiasedly estimated by a weighted average of DIDs. For all $t\in\{1,...,T-1\}$, for all $(d,d')\in(\mathcal{D}_1)^2$, and for all $\boldsymbol{d}^{-1}\in\mathcal{D}_2\times...\times\mathcal{D}_K$, let $N_{d,d',\boldsymbol{d}^{-1},t+1,t}=\sum_{g\in\mathcal{G}_{d,d',\boldsymbol{d}^{-1},t+1}}N_{g,t}$ denote the total population, at period t, of groups in $\mathcal{G}_{d,d',\boldsymbol{d}^{-1},t+1}$. Then, let

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

In contrast to $\text{DID}^{f}_{+,\boldsymbol{d}^{-1},t}$, which is a "forward" DID, $\text{DID}^{b}_{+,\boldsymbol{d}^{-1},t}$ is a "backward" DID, from the future to the past. It 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 $\boldsymbol{d}^{-1}$ at both dates, to the outcome evolution of groups whose first and other treatments are respectively equal to 0 and $\boldsymbol{d}^{-1}$ at both dates. $\text{DID}^{b}_{-,\boldsymbol{d}^{-1},t}$ has a similar interpretation, except that it compares groups

whose first treatment is equal to 1 at both dates to groups whose first treatment goes from 1 to 0 from $t+1$ to t. Let

$$\text{DID}_{\text{M}}^{b}=\sum_{t=1}^{T-1}\sum_{\boldsymbol{d}^{-1}\in\{0,1\}^{K-1}}\left(\frac{N_{0,1,\boldsymbol{d}^{-1},t+1,t}}{N_{\mathcal{S}_2}}\text{DID}_{+,\boldsymbol{d}^{-1},t}^{b}+\frac{N_{1,0,\boldsymbol{d}^{-1},t+1,t}}{N_{\mathcal{S}_2}}\text{DID}_{-,\boldsymbol{d}^{-1},t}^{b}\right) \tag{16}$$

if $N_{\mathcal{S}_2}>0$, and $\text{DID}_{\text{M}}^{b}=0$ if $N_{\mathcal{S}_2}=0$. By the exact same reaoning as in the proof of Theorem 4, we obtain, under Assumptions 1-2 and 5, $E\left[\text{DID}_{\text{M}}^{b}\right]=\delta_2$.

5.4 Additional results

In Web Appendix Section 3, we show that with a single treatment, $\text{DID}_{\text{M}}^{f}$ can be used to estimate the effect of the contemporaneous value of the treatment, controlling for some lags of that treatment. Similarly, $\text{DID}_{\text{M}}^{b}$ can be used to estimate the effect of a lag of the treatment, controlling for more recent lags and the treatment's contemporaneous value. We also show how to estimate dynamic effects with several binary and staggered treatments.

In Web Appendix Section 4, we consider both asymptotic and finite-sample inference. The asymptotic results are established under similar assumptions and arguments as those used to show the asymptotic normality of the DID_{M} estimator in de Chaisemartin and D'Haultfœuille (2020) (see Theorem S6 in the Web Appendix therein), without any important conceptual difference. One limitation is that the asymptotic approximation may not be accurate. $\text{DID}_{\text{M}}^{f}$ compares carefully selected treatment and control groups, and it could be the case that only a small number of groups can be included in those comparisons. We deal with this issue by proposing confidence intervals that are exact in a finite sample of groups under a normality assumption, in the spirit of Donald and Lang (2007). Though the exactness of those confidence intervals relies on strong conditions, we show that they remain asymptotically valid under much weaker assumptions.

6 Application

In this section, we revisit Hotz and Xiao (2011).[11] Unfortunately, many tables in this paper rely on proprietary data. The only table with TWFE regressions with several treatments that we can replicate is Table 11. Therefore, we focus on this table in our replication, though it is not the paper's main table.

[11] This paper is the only one, in the census of TWFE papers published by the AER from 2010 to 2012 that we conducted in de Chaisemartin and D'Haultfœuille (2020), that has several treatments in the regression, relies at least partially on non-proprietary data, and for which the treatments are not continuous (thus making it possible to compute the $\text{DID}_{\text{M}}^{f}$ estimator).

Hotz and Xiao (2011) use a panel of the 50 US states and the District of Columbia, in 1987, 1992, and 1997, to estimate the effect of state center-based daycare regulations, namely the minimum years of schooling required to be the director of a center-based care and the minimum staff-to-child ratio, 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. The distributions of these regulations are shown in Table 1. The minimum years of schooling is a discrete treatment taking six values included between 0 (no minimum) and 16, with 14 (associate degree) being the most frequent value. The minimum staff-to-child ratio is a also discrete treatment variable, taking seven values included between 0 (no minimum) and 1/3 (one professional per three children), with 1/4 being the most frequent value.

Table 1: Distribution of the two treatments in Hotz and Xiao (2011)

Min. years of schooling	# of (g,t) cells	Min. staff-to-child ratio	# of (g,t) cells
0	26	0	5
12	36	1/8	2
12.5	5	1/7	4
13	4	1/6	30
14	61	1/5	21
16	21	1/4	82
		1/3	9

Hotz and Xiao (2011) regress the revenue of family home day cares in state g and year t on state fixed effects, year fixed effects, 12 control variables, the minimum years of schooling required to be the director of a center-based care, the minimum staff-to-child ratio, and two indicators for whether there is no such minima, to allow for potentially non-linear effects. In Column (3) of their Table 11, the coefficient on the minimum years of schooling treatment, $\widehat{\beta}^X_{fe}$, is equal to -0.445 and is highly significant (95% confidence interval=$[-0.735, -0.155]$),[12] 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.

Dropping the 12 control variables from the regression does not affect that conclusion very much: the coefficient on the minimum years of schooling treatment, $\widehat{\beta}_{fe}$, is now equal to -0.566 and is

[12]This confidence interval is slightly larger than that in Hotz and Xiao (2011), because we cluster standard errors 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).

still highly significant (95% confidence interval=$[-0.852, -0.280]$). Below, we study $\widehat{\beta}_{fe}$, rather than $\widehat{\beta}^{X}_{fe}$, the coefficient estimated by Hotz and Xiao (2011). This is to ensure that the TWFE estimator we study is comparable to the $\text{DID}^{f}_{\text{M}}$ estimator we compute below: while the $\text{DID}^{f}_{\text{M}}$ estimator can be extended to allow for control variables, the sample on which it is computed in this application is not large enough to include 12 control variables.

We now show that $\widehat{\beta}_{fe}$ 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 1, 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 44 effects receive a positive weight and 83 receive a negative weight, and where the positive and negative weights respectively sum to 7.897 and -6.897. 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 0.148 and -0.148. The third term is a sum of the effects of increasing by one the staff to child ratio in 148 state×year cells, where 51 effects receive a positive weight and 97 receive a negative weight, and where the positive and negative weights respectively sum to 0.160 and -0.160. 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 0.055 and -0.055. 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 246.222 and -246.222.

When the other three treatment variables are dropped from the regression, the coefficient on the minimum years of schooling becomes small (-0.020) and insignificant (95% confidence interval=$[-0.114, 0.074]$). We follow Theorem 3 to decompose the coefficient in this "short" regression, and compare it to the coefficient in the "long" regression with the four treatments. The short regression's 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 56 cells receive a positive weight and 71 receive a negative weight, and where the positive and negative weights respectively sum to 1.759 and -0.759. Thus, the short regression has considerably smaller negative weights in this first term than the long regression. 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 5 effects receive a positive weight and 21 receive a negative weight, and where the positive and negative weights respectively sum to 0.008 and -0.077. The third term is a sum of the effects of increasing by one the staff to child ratio in 148 state×year cells, where 61 effects receive a positive weight and 87 receive a negative weight, and where the

positive and negative weights respectively sum to 0.030 and -0.022. 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 all effects receive a negative weight, and where the negative weights sum to -0.035. Thus, the short regression also has considerably less contamination weights than the long regression. Accordingly, the estimated maximal bias in Corollary 2 is almost five times lower for the short than for the long regression ($4.233 \times B$ versus $20.741 \times B$), so the short regression is preferable per this maximal-bias metric.

Finally, we compute the estimator proposed in Section 5, for the minimum years of schooling treatment, controlling for the staff-to-child ratio treatment. Our estimators do not assume linear treatment effects, so we do not need to control for the indicators for whether there is no such minima.

There are 127 (g,t) cells with a non-zero minimum years of schooling. On the other hand, there are only five (g,t) cells in $\mathcal{S}_1$, all of which have a non-zero minimum years of schooling. The five (g,t) cells our estimator applies to are (Kentucky,1992), (Minnesota,1992), (Utah,1992), (Vermont,1992), and (Rhode Island,1997).[13] Of the 122 (g,t) cells we lose when focusing on $\mathcal{S}_1$, 93 belong to states that do not experience any change of their minimum years of schooling, so their treatment effect cannot be identified under a parallel trends assumption. 24 (g,t) cells either also experience a change of their minimum staff-to-child ratio when their minimum years of schooling changes, or cannot be matched to a control state with the same baseline treatments. Then, estimating their treatment effect would require making constant effects assumptions. Finally, estimating the treatment of the remaining 5 (g,t) cells would require assuming parallel trends over a longer horizon than over consecutive time periods.

We find that $\text{DID}^f_{\text{M}} = -0.029$. DID^f_{M} uses data from 5 switching and 19 control (g,t) cells, so the asymptotic approximation in Web Appendix Section 4.1 may not be very reliable for that estimator. Instead, we compute the exact confidence interval developed in Web Appendix Section 4.2 and find that it is equal to $[-0.821, 0.807]$.[14] In this application, the assumption that the first-differenced outcome is normally distributed is not rejected. We conduct a Shapiro-Wilk test separately for the 1987 to 1992 and for the 1992 to 1997 first differences, as the test assumes independent observations. None of the two tests is rejected (p-value= 0.98 and 0.46, respectively).

To gain precision, one may further impose Assumption 5. Doing so allows us to use DID^b_{M}

[13]For the staff-to-child ratio treatment, the set $\mathcal{S}_1$ is even smaller as it only contains two (g,t) cells. This is why we focus on the minimum-years-of-schooling treatment.

[14]This confidence interval relies on Assumption 16 in the Web Appendix, which does not hold in our data: Rhode Island and Washington are used twice in DID^f_{M}. Removing these two states in one of the two s they belong to (using the notation in Web Appendix Section 4.2) changes very slightly the value of DID^f_{M} (-0.0072 in lieu of -0.029).

to estimate the treatment effect in five (g,t) cells in $\mathcal{S}_2$. $\mathcal{S}_1$ and $\mathcal{S}_2$ do not overlap and have the same numbers of cells, so we can also use $1/2(\text{DID}^f_{\text{M}} + \text{DID}^b_{\text{M}})$ to estimate δ, the average treatment effect in $\mathcal{S}_1 \cup \mathcal{S}_2$. We find that $1/2(\text{DID}^f_{\text{M}} + \text{DID}^b_{\text{M}}) = -0.016$. $1/2(\text{DID}^f_{\text{M}} + \text{DID}^b_{\text{M}})$ uses data from 50 (g,t) cells, coming from 30 different states. The asymptotic approximation in Web Appendix Section 4.1 may be more reasonable for that estimator,[15] so we follow Theorem 7 therein to compute a 95% confidence interval for δ. We find that this confidence interval is equal to $[-0.126, 0.094]$. We also test the equality between δ and β_{fe}, and reject the null hypothesis at all conventional levels (p-value$=4 \times 10^{-4}$). Hence, as discussed above, we can reject the hypothesis that the effects of the minimum years of schooling and staff-to-child ratio treatments are homogenous.

Table 2: Estimators of the effect of the minimum years of schooling treatment

	Estimate	95% Confidence Interval
$\widehat{\beta}^X_{fe}$	-0.445	$[-0.735, -0.155]$
$\widehat{\beta}_{fe}$	-0.566	$[-0.860, -0.272]$
$\widehat{\beta}_s$	-0.022	$[-0.117, 0.077]$
DID^f_{M}	-0.029	$[-0.821, 0.807]$
$1/2(\text{DID}^f_{\text{M}} + \text{DID}^b_{\text{M}})$	-0.016	$[-0.126, 0.094]$

Let us summarize our results. Using a TWFE regression with several treatments, Hotz and Xiao (2011) find that increasing the years of schooling required for directors of center-based daycare significantly decreases the revenue of family home daycare. We show that in the presence of heterogeneous treatment effects, their regression estimates a highly-non-convex combination of the effects of the years of schooling treatment, and is contaminated by the effects of the other treatments. Therefore, their finding may not be robust to heterogeneous treatment effects. Then, we use our robust estimators to assess if, in the presence of heterogeneous effects, one can conclude, for at least a subset of (g,t) cells, that increasing the years of schooling requirement significantly decreases the revenue of family home daycare. The answer is negative, as our estimators are insignificant. Moreover, one of our estimators is significantly different from the TWFE estimator, thus allowing us to reject the null hypothesis that the effects of all treatments are constant in this application. Overall, there is no evidence that the finding in Hotz and Xiao (2011) is robust to heterogeneous effects, while there is evidence that treatment effects are

[15] To verify that, we considered simulations with the same design as in the application but with no effects of the treatments, and $(\Delta Y_{g,2}(\mathbf{0}), \Delta Y_{g,3}(\mathbf{0}))$ drawn either from a normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$, with Σ equal to the estimated variance matrix on the sample, or from the empirical distribution of $(\Delta Y_{g,2}, \Delta Y_{g,3})$. In both cases, the coverage of our confidence interval was higher than 95% (95.4% and 99.3%, respectively).

heterogeneous in this application.

7 Conclusion

In this paper, we show that treatment coefficients in TWFE regressions with several treatments may not be 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.

Acknowledgements

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. We are also grateful to the editor, associate editor, and two anonymous referees for their very helpful comments. Clément de Chaisemartin was funded by the European Union (ERC, REALLYCREDIBLE,GA No 101043899). Views and opinions expressed are those of the authors and do not reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

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

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.

Bojinov, Iavor, Ashesh Rambachan, and Neil Shephard. 2021. "Panel experiments and dynamic causal effects: A finite population perspective." *Quantitative Economics*, 12(4): 1171–1196.

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

de Chaisemartin, C, and X D'Haultfœuille. 2018. "Fuzzy Differences-in-Differences." *The Review of Economic Studies*, 85(2): 999–1028.

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'Haultfoeuille. 2021. "Two-way fixed effects regressions with several treatments." *arXiv preprint arXiv:2012.10077, v4*.

Donald, Stephen G, and Kevin Lang. 2007. "Inference with difference-in-differences and other panel data." *The review of Economics and Statistics*, 89(2): 221–233.

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.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American statistical Association*, 81(396): 945–960.

Holland, Paul W, and Donald B Rubin. 1987. "Causal inference in retrospective studies." *ETS Research Report Series*, 1987(1): 203–231.

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.

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

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.

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

A Proofs

A.1 Theorem 1

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

A.2 Theorem 2

We first establish the following lemma.

Lemma 1 *If Assumptions 1-3 hold, for all* $(g, g', t, t') \in \{1, ..., G\}^2 \times \{1, ..., T\}^2$,

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

Proof of Lemma 1

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

$$\begin{aligned}
E\left(Y_{g,t}|\boldsymbol{D}\right) =& E\Bigg(Y_{g,t}(0, \mathbf{0}^{-1}) + D_{g,t}^1 (Y_{g,t}(1, \boldsymbol{D}_{g,t}^{-1}) - Y_{g,t}(0, \boldsymbol{D}_{g,t}^{-1}) + Y_{g,t}(0, \boldsymbol{D}_{g,t}^{-1}) - Y_{g,t}(0, \mathbf{0}^{-1})) \\
&+ (1 - D_{g,t}^1)(Y_{g,t}(0, \boldsymbol{D}_{g,t}^{-1}) - Y_{g,t}(0, \mathbf{0}^{-1})) \Big| \boldsymbol{D}\Bigg) \\
=& E\left(Y_{g,t}(0, \mathbf{0}^{-1}) \middle| \boldsymbol{D}\right) + D_{g,t}^1 E\left(\Delta_{g,t}^1 \middle| \boldsymbol{D}\right) + E\left(\Delta_{g,t}^{-1} \middle| \boldsymbol{D}\right) \\
=& E\left(Y_{g,t}(0, \mathbf{0}^{-1}) \middle| \boldsymbol{D}_g\right) + D_{g,t}^1 E\left(\Delta_{g,t}^1 \middle| \boldsymbol{D}\right) + E\left(\Delta_{g,t}^{-1} \middle| \boldsymbol{D}\right), \qquad (17)
\end{aligned}$$

where the last equality follows from Assumption 2. Moreover, by Assumption 3

$$\begin{aligned}
&E\left(Y_{g,t}(0, \mathbf{0}^{-1}) \middle| \boldsymbol{D}_g\right) - E\left(Y_{g,t'}(0, \mathbf{0}^{-1}) \middle| \boldsymbol{D}_g\right) - E\left(Y_{g',t}(0, \mathbf{0}^{-1}) \middle| \boldsymbol{D}_g\right) + E\left(Y_{g',t'}(0, \mathbf{0}^{-1}) \middle| \boldsymbol{D}_g\right) \\
=& 0. \qquad (18)
\end{aligned}$$

The result follows by combining (17) and (18).

Proof of Theorem 2

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

$$E\left(\widehat{\beta}_{fe} \middle| \boldsymbol{D}\right) = \frac{\sum_{g,t} N_{g,t} \varepsilon_{g,t} E\left(Y_{g,t}|\boldsymbol{D}\right)}{\sum_{g,t} N_{g,t} \varepsilon_{g,t} D_{g,t}^1}. \qquad (19)$$

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

$$\sum_{t=1}^{T} N_{g,t}\varepsilon_{g,t} = 0 \text{ for all } g \in \{1, ..., G\}, \tag{20}$$

$$\sum_{g=1}^{G} N_{g,t}\varepsilon_{g,t} = 0 \text{ for all } t \in \{1, ..., T\}, . \tag{21}$$

Then,

$$\begin{aligned}
&\sum_{g,t} N_{g,t}\varepsilon_{g,t} E\left(Y_{g,t}|\boldsymbol{D}\right) \\
=&\sum_{g,t} N_{g,t}\varepsilon_{g,t} \left(E\left(Y_{g,t}|\boldsymbol{D}\right) - E\left(Y_{g,1}|\boldsymbol{D}\right) - E\left(Y_{1,t}|\boldsymbol{D}\right) + E\left(Y_{1,1}|\boldsymbol{D}\right)\right) \\
=&\sum_{g,t} N_{g,t}\varepsilon_{g,t} \left(D^1_{g,t} E\left(\Delta^1_{g,t}\middle|\boldsymbol{D}\right) + E\left(\Delta^{-1}_{g,t}\middle|\boldsymbol{D}\right) - D^1_{1,t} E\left(\Delta^1_{1,t}(\boldsymbol{D}^{-1}_{1,t})\middle|\boldsymbol{D}\right) - E\left(\Delta^{-1}_{1,t})\middle|\boldsymbol{D}\right)\right. \\
&\left. - D^1_{g,1} E\left(\Delta^1_{g,1}(\boldsymbol{D}^{-1}_{g,1})\middle|\boldsymbol{D}\right) - E\left(\Delta^{-1}_{g,1}\right)\middle|\boldsymbol{D}\right) + D^1_{1,1} E\left(\Delta^1_{1,1}(\boldsymbol{D}^{-1}_{1,1})\middle|\boldsymbol{D}\right) + E\left(\Delta^{-1}_{1,1})\middle|\boldsymbol{D}\right)\right) \\
=&\sum_{g,t} N_{g,t}\varepsilon_{g,t} \left(D^1_{g,t} E\left(\Delta^1_{g,t}\middle|\boldsymbol{D}\right) + E\left(\Delta^{-1}_{g,t}\middle|\boldsymbol{D}\right)\right) \\
=&\sum_{(g,t):D^1_{g,t}=1} N_{g,t}\varepsilon_{g,t} E\left(\Delta^1_{g,t}\middle|\boldsymbol{D}\right) + \sum_{(g,t):\boldsymbol{D}^{-1}_{g,t}\neq \mathbf{0}^{-1}} N_{g,t}\varepsilon_{g,t} E\left(\Delta^{-1}_{g,t}\middle|\boldsymbol{D}\right).
\end{aligned} \tag{22}$$

The first and third equalities follow from Equations (20) and (21). The second equality follows from Lemma 1. The fourth equality follows from the fact that $\Delta^0_{g,t}(\mathbf{0}^{-1}) = 0$. Finally,

$$\sum_{g,t} N_{g,t}\varepsilon_{g,t} D^1_{g,t} = \sum_{(g,t):D^1_{g,t}=1} N_{g,t}\varepsilon_{g,t}. \tag{23}$$

Combining (19), (22), (23) yields

$$E\left(\widehat{\beta}_{fe}\middle|\mathbf{D}\right) = \sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} E\left(\Delta^1_{g,t}\middle|\boldsymbol{D}\right) + \sum_{(g,t):\boldsymbol{D}^{-1}_{g,t}\neq \mathbf{0}^{-1}} \frac{N_{g,t}}{N_1} w_{g,t} E\left(\Delta^{-1}_{g,t}\middle|\boldsymbol{D}\right). \tag{24}$$

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

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

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

A.3 Theorem 3

The proof is the same as that of Theorem 1, with just one difference: we do not have $\sum_{(g,t):D^2_{g,t}=1} N_{g,t} \times \varepsilon^s_{g,t} = 0$, since $\varepsilon^s_{g,t}$ is not orthogonal to $D^2_{g,t}$ in general.

A.4 Corollary 2

The result directly follows from Theorems 1 and 3, the triangle inequality, and the fact there is a real number B such that $|\Delta^1_{g,t}| \leq B$ and $|\Delta^2_{g,t}| \leq B$ for all (g,t). The first bound is reached when $\Delta^1_{g,t} = B \times (2 \times 1\{w_{g,t} \geq 1\} - 1)$ and $\Delta^2_{g,t} = B(2 \times 1\{w_{g,t} \geq 0\} - 1)$, the second bound is reached when $\Delta^1_{g,t} = B \times (2 \times 1\{w^s_{g,t} \geq 1\} - 1)$ and $\Delta^2_{g,t} = B(2 \times 1\{w^s_{g,t} \geq 0\} - 1)$.

A.5 Proposition 1

The set $\mathcal{T}_1 \backslash \mathcal{S}_1$ can be partitioned into three groups, $\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3$, defined by:

$$\begin{aligned}
\mathcal{V}_1 &= \left\{(g,t) : D^1_{g,t} = 1 \text{ and either } D^1_{g,t-1} = 1 \text{ or } t = 1\right\}, \\
\mathcal{V}_2 &= \left\{(g,t) : D^1_{g,t} = 1,\ D^1_{g,t-1} = 0 \text{ and } \boldsymbol{D}^{-1}_{g,t} \neq \boldsymbol{D}^{-1}_{g,t-1}\right\}, \\
\mathcal{V}_3 &= \Big\{(g,t) : D^1_{g,t} = 1,\ D^1_{g,t-1} = 0,\ \boldsymbol{D}^{-1}_{g,t} = \boldsymbol{D}^{-1}_{g,t-1},\ \forall g' \neq g, \text{ either } \boldsymbol{D}_{g',t} \neq \boldsymbol{D}_{g',t-1} \\
&\qquad \text{or } \boldsymbol{D}_{g',t-1} \neq \boldsymbol{D}_{g,t-1}\Big\}.
\end{aligned}$$

Since the assumptions impose no joint restrictions on these three groups, it suffices to prove that $\delta_{\mathcal{V}} = E[\sum_{(g,t)\in\mathcal{V}}(N_{g,t}/N_{\mathcal{V}})\Delta^1_{g,t}]$ is not identified if $\mathcal{V} \subset \mathcal{V}_k$ $(k = 1, ..., 3)$. To this end, fix $c \neq 0$ and for all $(g,t) \in \mathcal{V}$, let $\widetilde{Y}_{g,t}(0, \boldsymbol{D}^{-1}_{g,t}) = Y_{g,t}(0, \boldsymbol{D}^{-1}_{g,t}) + c$, $\widetilde{Y}_{g,t}(\boldsymbol{d}) = Y_{g,t}(\boldsymbol{d})$ for all $\boldsymbol{d} \neq (0, \boldsymbol{D}^{-1}_{g,t})$ and $\widetilde{Y}_{g',t'}(\boldsymbol{d}) = Y_{g',t'}(\boldsymbol{d})$ for all $\boldsymbol{d}$ and $(g',t') \notin \mathcal{V}$.

If $\mathcal{V} \subset \mathcal{V}_1$, Assumption 4 does not impose any restriction on $Y_{g,t}(0, \boldsymbol{D}^{-1}_{g,t})$ for $(g,t) \in \mathcal{V}_1$, since either $D^1_{g,t-1} = 1$ or $t = 1$. The potential outcomes $(\widetilde{Y}_{g,t}(\boldsymbol{d}))_{(g,t,\boldsymbol{d})}$ are thus compatible with the data and Assumptions 1-2 and 4. Since they lead to $\widetilde{\delta}_{\mathcal{V}} = \delta_{\mathcal{V}} - c$, $\delta_{\mathcal{V}}$ is not identified.

Now, if $\mathcal{V} \subset \mathcal{V}_2$, Assumption 4 imposes a restriction on $Y_{g,t}(0, \boldsymbol{D}^{-1}_{g,t-1}) - Y_{g,t-1}(0, \boldsymbol{D}^{-1}_{g,t-1})$ for $(g,t) \in \mathcal{V}$. But since $\boldsymbol{D}^{-1}_{g,t} \neq \boldsymbol{D}^{-1}_{g,t-1}$ for such cells, we have

$$\widetilde{Y}_{g,t}(0, \boldsymbol{D}^{-1}_{g,t-1}) - \widetilde{Y}_{g,t-1}(0, \boldsymbol{D}^{-1}_{g,t-1}) = Y_{g,t}(0, \boldsymbol{D}^{-1}_{g,t-1}) - Y_{g,t-1}(0, \boldsymbol{D}^{-1}_{g,t-1}),$$

using here $(g, t-1) \notin \mathcal{V}_2$. Therefore, $\widetilde{Y}_{g,t}(0, \boldsymbol{D}^{-1}_{g,t-1}) - \widetilde{Y}_{g,t-1}(0, \boldsymbol{D}^{-1}_{g,t-1})$ satisfies the restriction of Assumption 4. So again, $(\widetilde{Y}_{g,t}(\boldsymbol{d}))_{(g,t,\boldsymbol{d})}$ are compatible with the data and Assumptions 1-2 and 4, and $\delta_{\mathcal{V}}$ is not identified.

Finally, assume $\mathcal{V} \subset \mathcal{V}_3$. In this case, $(\widetilde{Y}_{g,t}(\boldsymbol{d}))_{(g,t,\boldsymbol{d})}$ may violate Assumption 4: if $\boldsymbol{D}_{g',t-1} = \boldsymbol{D}_{g,t-1} = (0, \boldsymbol{d}^{-1})$ for some $(g,t) \in \mathcal{V}$ and $(g',t) \notin \mathcal{V}_3$, we have

$$\begin{aligned}
&E[\widetilde{Y}_{g',t}(0, \boldsymbol{d}^{-1}) - \widetilde{Y}_{g',t-1}(0, \boldsymbol{d}^{-1}) | \boldsymbol{D}_{g',t-1} = (0, \boldsymbol{d}^{-1})] \\
=&E[Y_{g',t}(0, \boldsymbol{d}^{-1}) - Y_{g',t-1}(0, \boldsymbol{d}^{-1}) | \boldsymbol{D}_{g',t-1} = (0, \boldsymbol{d}^{-1})] \\
=&E[Y_{g,t}(0, \boldsymbol{d}^{-1}) - Y_{g,t-1}(0, \boldsymbol{d}^{-1}) | \boldsymbol{D}_{g,t-1} = (0, \boldsymbol{d}^{-1})] \\
=&E[\widetilde{Y}_{g,t}(0, \boldsymbol{d}^{-1}) - \widetilde{Y}_{g,t-1}(0, \boldsymbol{d}^{-1}) | \boldsymbol{D}_{g,t-1} = (0, \boldsymbol{d}^{-1})] - c.
\end{aligned}$$

To fix this, we simply let $\widetilde{Y}_{g',t}(0,\boldsymbol{d}^{-1}) = Y_{g',t}(0,\boldsymbol{d}^{-1}) + c$ for such $(g',t) \notin \mathcal{V}_3$. This change is still compatible with the data: by definition of $\mathcal{V}_3$ and because $\boldsymbol{D}_{g',t-1} = \boldsymbol{D}_{g,t-1}$, we have $\boldsymbol{D}_{g',t} \neq (0,\boldsymbol{d}^{-1})$. Hence, with this modification, $(\widetilde{Y}_{g,t}(\boldsymbol{d}))_{(g,t,\boldsymbol{d})}$ are compatible with the data and Assumptions 1-2 and 4. This shows that again, $\delta_{\mathcal{V}}$ is not identified.

A.6 Theorem 4

First, by definition of $\mathrm{DID}^f_{\mathrm{M}}$,

$$\mathrm{DID}^f_{\mathrm{M}} = \sum_{t=2}^{T} \sum_{\boldsymbol{d}^{-1}\in\{0,1\}^{K-1}} \frac{N_{1,0,\boldsymbol{d}^{-1},t}}{N_{\mathcal{S}_1}} \mathrm{DID}^f_{+,\boldsymbol{d}^{-1},t} + \frac{N_{0,1,\boldsymbol{d}^{-1},t}}{N_{\mathcal{S}_1}} \mathrm{DID}^f_{-,\boldsymbol{d}^{-1},t}, \tag{25}$$

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

$$E\left(Y_{g,t} - Y_{g,t-1} | \boldsymbol{D}\right) = E\left(\Delta^1_{g,t} \middle| \boldsymbol{D}\right) + E\left(Y_{g,t}(0,\boldsymbol{d}^{-1}) - Y_{g,t-1}(0,\boldsymbol{d}^{-1}) \middle| \boldsymbol{D}\right). \tag{26}$$

Under Assumptions 2 and 4, for all $t \geq 2$, there exists $\psi_{0,\boldsymbol{d}^{-1},t} \in \mathbb{R}$ such that for all $g \in \mathcal{G}_{0,0,\boldsymbol{d}^{-1},t} \cup \mathcal{G}_{1,0,\boldsymbol{d}^{-1},t}$,

$$\begin{aligned} E\left(Y_{g,t}(0,\boldsymbol{d}^{-1}) - Y_{g,t-1}(0,\boldsymbol{d}^{-1}) \middle| \boldsymbol{D}\right) &= E\left(Y_{g,t}(0,\boldsymbol{d}^{-1}) - Y_{g,t-1}(0,\boldsymbol{d}^{-1}) \middle| \boldsymbol{D}_g\right) \\ &= E\left(Y_{g,t}(0,\boldsymbol{d}^{-1}) - Y_{g,t-1}(0,\boldsymbol{d}^{-1}) \middle| D^1_{g,t-1} = 0, \boldsymbol{D}^{-1}_{g,t-1} = \boldsymbol{d}^{-1}\right) \\ &= \psi_{0,\boldsymbol{d}^{-1},t}. \end{aligned} \tag{27}$$

As a result,

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

The first equality follows by (26), the second by (27), and the third after some algebra. Given that $\mathrm{DID}^f_{+,\boldsymbol{d}^{-1},t} = 0$ if $N_{1,0,\boldsymbol{d}^{-1},t} = 0$ or $N_{0,0,\boldsymbol{d}^{-1},t} = 0$, we obtain, by definition of $\mathcal{S}_1$ and with the

convention that sums over empty sets are 0,

$$E\left(N_{1,0,\boldsymbol{d}^{-1},t}\text{DID}^{f}_{+,\boldsymbol{d}^{-1},t}\middle|\boldsymbol{D}\right)=E\left(\sum_{\substack{g:D^1_{g,t}=1,\boldsymbol{D}^{-1}_{g,t}=\boldsymbol{d}^{-1}\\(g,t)\in\mathcal{S}_1}}N_{g,t}\Delta^1_{g,t}\middle|\boldsymbol{D}\right). \tag{28}$$

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

$$E\left(N_{0,1,\boldsymbol{d}^{-1},t}\text{DID}^{f}_{-,\boldsymbol{d}^{-1},t}\middle|\boldsymbol{D}\right)=E\left(\sum_{\substack{g:D^1_{g,t}=0,\boldsymbol{D}^{-1}_{g,t}=\boldsymbol{d}^{-1}\\(g,t)\in\mathcal{S}_1}}N_{g,t}\Delta^1_{g,t}\middle|\boldsymbol{D}\right). \tag{29}$$

Plugging (28) and (29) into (25) yields

$$\begin{aligned}
E(\text{DID}^{f}_{\text{M}}) =&E\left(E\left(\sum_{t=2}^{T}\sum_{\boldsymbol{d}^{-1}\in\{0,1\}^{K-1}}\sum_{\substack{g:\boldsymbol{D}^{-1}_{g,t}=\boldsymbol{d}^{-1}\\(g,t)\in\mathcal{S}_1}}N_{g,t}\Delta^1_{g,t}\middle|\boldsymbol{D}\right)\right)\\
=&E\left(E\left(\sum_{(g,t)\in\mathcal{S}_1}N_{g,t}\Delta^1_{g,t}\middle|\boldsymbol{D}\right)\right)\\
=&\delta_1.
\end{aligned}$$