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. When a treatment is omitted from the regression, we obtain a new omitted variable bias formula, where bias can arise even if the treatments are not correlated with each other, but can be smaller than in the TWFE regression with all treatments. 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.

(JEL C21, C23)

1 Introduction

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

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

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

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

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 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 an analogue of the standard omitted variable bias (OVB) formula for that regression, allowing for heterogeneous effects. Allowing for heterogeneous effects leads to two perhaps surprising departures from the standard OVB logic many researchers are used to. First, omitting from a TWFE regression a treatment uncorrelated with the main treatment of interest may still lead to an OVB. Second, controlling for more treatments may lead to a more biased estimator than not controlling for them. Specifically, we show that 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. That second term, which corresponds to the OVB term in the standard formula, may differ from zero even if the two treatments are uncorrelated conditional on the group and time fixed effects. 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.

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 estimator's robustness may come at a high price in terms of external validity and statistical precision. For instance, in our application in Section 5, 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.

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.

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). Our paper is also closely related to several papers that have considered the causal interpretation of OLS regression coefficients with several treatments (see Sun and Abraham, 2021; Hull, 2018; Goldsmith-Pinkham, Hull and Kolesár, 2021). We discuss those papers in more details later in the paper (see Section 3.4), but for now we just note that when the treatments are indicators for whether group g has started receiving a binary and staggered treatment ℓ periods ago, our decomposition of the TWFE regression reduces to one of the decompositions in Sun and Abraham (2021). Accordingly, our decomposition extends their result to situations where the treatments in the regression are different policies that could non-binary, non-staggered, and non-mutually-exclusive, rather than indicators for having received a single binary and staggered policy ℓ periods ago.

The remainder of the paper is organized as follows. Section 2 presents the set up. Section 3

presents our decomposition results for TWFE regressions with several treatments. Section 4 presents our alternative estimator. Section 5 presents our empirical application.

2 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. 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 $D_{g,t} = (D^k_{g,t})_{k\in\{1,...,K\}}$ 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 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 $(D^1_{g,t},...,D^K_{g,t}) = \boldsymbol{d}$. The observed outcome is $Y_{g,t} = Y_{g,t}(D^1_{g,t},...,D^K_{g,t})$.

Importantly, our notation does not necessarily rule out dynamic effects of past treatments on the outcome. The K treatments may for instance include lags of the same treatment variables. We discuss this issue in more details after Theorem 2 below, and in Section 4.4.

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* $((Y_{g,t}(\boldsymbol{d}))_{\boldsymbol{d}\in\{0,1\}^K}, (D^k_{g,t})_{k\in\{1,...,K\}})_{t\in\{1,...,T\}}$ *are mutually independent.*

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

1. $E(Y_{g,t}(\mathbf{0}) - Y_{g,t-1}(\mathbf{0}) | D_{g,1},...,D_{g,T}) = E(Y_{g,t}(\mathbf{0}) - Y_{g,t-1}(\mathbf{0}))$.

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.[1]

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* $D_{g,t}$*, weighted by* $N_{g,t}$*.*[2]

$$Y_{g,t} = \widehat{\alpha}_g + \widehat{\gamma}_t + \widehat{\beta}_{fe} D_{g,t} + u_{g,t}, \tag{1}$$

where $u_{g,t}$ *denotes the regression residual.*

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). In Section 3.4.2 below, we elaborate on the difference between a treatment and a covariate.

Let $\boldsymbol{D}$ be the vector $(D_{g,t})_{(g,t)\in\{1,...,G\}\times\{1,...,T\}}$ collecting all the treatments in all the (g,t) cells. let $\boldsymbol{D}_g = (D_{1,g}, ..., D_{T,g})$ be the vector collecting all the treatments in group g. Let $N_1 = \sum_{g,t} N_{g,t} D^1_{g,t}$ denote the total population of cells receiving the first treatment. Let $D^{-1}_{g,t} = (D^2_{g,t}, ..., D^K_{g,t})$ denote

[1] 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.

[2] 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 in (1), 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.

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 $D^{-1}_{g,t}$:

$$D^1_{g,t} = \widehat{\alpha} + \widehat{\gamma}_g + \widehat{\nu}_t + (D^{-1}_{g,t})'\widehat{\zeta} + \varepsilon_{g,t}. \tag{2}$$

One can show that 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: $\sum_{(g,t):D^1_{g,t}=1}(N_{g,t}/N_1)\varepsilon_{g,t} \neq 0$. Then we let $w_{g,t}$ denote $\varepsilon_{g,t}$ divided by that average:

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

3 TWFE regressions with several treatments and heterogeneous effects

3.1 Decomposition results

3.1.1 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

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

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,*

$$\beta_{fe} = E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta^1_{g,t} + \sum_{(g,t):D^2_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta^2_{g,t}\right]. \tag{3}$$

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*b*, a previous version of this paper, for a formal statement and a proof).

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.

3.1.2 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

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

$\Delta^1_{g,t}$ 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,*

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

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

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

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

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

$$E\left[\Delta^{-1}_{g,t}|\boldsymbol{D}\right] = \sum_{k=2}^{K} D^k_{g,t}\delta^k_{g,t}. \tag{4}$$

Then, we obtain Decomposition (5) below. The corresponding weights can be computed using the `twowayfeweights` Stata command.

Corollary 1 *Suppose that Assumptions 1-3 and* (4) *hold. Then,*

$$\beta_{fe} = E\left[\sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta^1_{g,t} + \sum_{k=2}^{K}\sum_{(g,t):D^k_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\delta^k_{g,t}\right]. \tag{5}$$

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. The treatments in the regression may for instance be the current treatment and its first $K - 1$ lags. In that case, our potential outcome notation allows the current treatment and its first $K - 1$ lags 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 its lags.

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

3.2.1 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

$$\widehat{\beta}_{fe} = Y_{2,2} - Y_{2,1} - (Y_{1,2} - Y_{1,1}) . \tag{6}$$

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,

$$\begin{aligned}
\beta_{fe} =& E\left(Y_{2,2}(1,0) - Y_{2,1}(0,0) - (Y_{1,2}(0,0) - Y_{1,1}(0,0))\right) \\
=& E\left(Y_{2,2}(1,0) - Y_{2,2}(0,0)\right) + E\left(Y_{2,2}(0,0) - Y_{2,1}(0,0) - (Y_{1,2}(0,0) - Y_{1,1}(0,0))\right) \\
=& E\left(Y_{2,2}(1,0) - Y_{2,2}(0,0)\right),
\end{aligned} \tag{7}$$

where the second equality follows from Assumption 3. Equation (7) is a special case of Equation (3) 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

$$\widehat{\beta}_{fe} = \frac{1}{2}\left(Y_{2,2} - Y_{2,1} - (Y_{1,2} - Y_{1,1})\right) + \frac{1}{2}\left(Y_{4,2} - Y_{4,1} - (Y_{3,2} - Y_{3,1})\right). \tag{8}$$

The first DID in Equation (8) is the same as that in the right-hand side of Equation (6) 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,

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

Equations (8) and (9) imply that

$$\begin{aligned}\beta_{fe} =& \frac{1}{2}E\left(Y_{2,2}(1,0) - Y_{2,2}(0,0)\right) + \frac{1}{2}E\left(Y_{4,2}(1,1) - Y_{4,2}(0,1)\right) \\ &+ \frac{1}{2}E\left(Y_{4,2}(0,1) - Y_{4,2}(0,0)\right) - \frac{1}{2}E\left(Y_{3,2}(0,1) - Y_{3,2}(0,0)\right). \end{aligned} \tag{10}$$

Equation (10) is a special case of Equation (3) 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 5: 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,

$$\widehat{\beta}_{fe} = Y_{1,3} - Y_{1,2} - (Y_{2,3} - Y_{2,2}) . \tag{11}$$

Accordingly, one can show that

$$\begin{aligned} \beta_{fe} =& E\left(Y_{1,3}(1,0) - Y_{1,3}(0,0)\right) \\ &+ E\left(Y_{2,2}(0,1) - Y_{2,2}(0,0)\right) - E\left(Y_{2,3}(0,1) - Y_{2,3}(0,0)\right) . \end{aligned} \tag{12}$$

$\widehat{\beta}_{fe}$ 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.

3.2.2 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 $(1 - I_g)$.[3] 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

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

[3]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.

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 $D_{g,t} = 1, I_g = 0$.

3.3 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.

Regression 2 *(Short TWFE regression)*

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

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

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

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

$$\beta_{fe}^s = E\left[\sum_{(g,t):D_{g,t}^1=1} \frac{N_{g,t}}{N_1} w_{g,t}^s \Delta_{g,t}^1 + \sum_{(g,t):D_{g,t}^2=1} \frac{N_{g,t}}{N_1} w_{g,t}^s \Delta_{g,t}^2\right]. \tag{13}$$

Moreover, $\sum_{(g,t):D_{g,t}^1=1}(N_{g,t}/N_1)w_{g,t}^s = 1$ and $\sum_{(g,t):D_{g,t}^2=1}(N_{g,t}/N_1)w_{g,t}^s$ may differ from zero.

Theorem 3 is similar to Theorem 1. It shows that the coefficient on $D_{g,t}^1$ 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 $D_{g,t}^1$ from 0 to 1 while keeping $D_{g,t}^2$ 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 may be seen as a version of the standard omitted variable bias (OVB) formula, for TWFE regressions and with heterogeneous treatment effects. Allowing for heterogeneous treatment effects in this OVB formula has one important consequence. With constant treatment effects, there is no OVB if the omitted variable $D^2_{g,t}$ is uncorrelated with $\varepsilon^s_{g,t}$, the residual of $D^1_{g,t}$ from a regression on group and time FEs. On the other hand, with heterogeneous treatment effects, there may be an OVB even if $D^2_{g,t}$ and $\varepsilon^s_{g,t}$ are uncorrelated, if $\sum_{(g,t)}(N_{g,t}/N_1)w^s_{g,t}D^2_{g,t}\Delta^2_{g,t} \neq 0$, namely if $D^2_{g,t}\Delta^2_{g,t}$ and $\varepsilon^s_{g,t}$ are correlated. Conversely, there is no OVB if $D^2_{g,t}\Delta^2_{g,t}$ and $\varepsilon^s_{g,t}$ are uncorrelated, but this condition is strong, and unlike the standard condition that $D^2_{g,t}$ and $\varepsilon^s_{g,t}$ be uncorrelated, it is untestable. The implication of this result is that in TWFE regressions, if treatment effects are heterogeneous, failing to control for other treatments may lead to an OVB even if those other treatments are uncorrelated with the main treatment of interest.[4]

The previous paragraph may seem to imply that in TWFE regressions, one should control for all the time-varying treatments one observes, if not for all that exist. And indeed, 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 again, this logic breaks down under heterogeneous treatment effects. Then, including the second treatment into the regression may not be preferable: $D^1_{g,t}$'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.

[4]Not all OLS regressions are subject to this issue in the presence of heterogeneous treatment effects. For instance, in an RCT, if one regresses the outcome on the treatment omitting other determinants of the outcome, there will be no OVB in the regression, because the effect of those other determinants of the outcome is by design uncorrelated with the treatment assignment. In TWFE regressions on the other hand, there is no guarantee that $\sum_{(g,t)}(N_{g,t}/N_1)w^s_{g,t}D^2_{g,t}\Delta^2_{g,t} = 0$.

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.[5] 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 5, 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.

3.4 Related literature

3.4.1 TWFE regressions with control variables

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 $X_{g,t}$.[6] It assumes that for some vector θ, $E\left(Y_{g,t}(0)-Y_{g,t-1}(0)-(X_{g,t}-X_{g,t-1})'\theta\right)$ does not vary across groups. In other words, groups may experience different trends, but those are fully accounted for by a linear model in the evolution of their covariates. Under that assumption, 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, with weights that differ from those without controls.

Theorems 1 and 2 are related to, but different from, that result. For instance, with $D^2_{g,t}$ as the only control variable in the regression, the weighted sum of treatment effects identified by β_{fe} in Theorem S4 of de Chaisemartin and D'Haultfœuille (2020) is identical to the first weighted sum in Theorem 1. On the other hand, the second weighted sum in Theorem 1, the

[5] 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.

[6] With time-invariant group-level controls, the TWFE regression is not identified.

contamination term, does not appear in Theorem S4 of de Chaisemartin and D'Haultfœuille (2020). The only case where the decompositions in Theorems 1 and 2 reduce to that in Theorem S4 of de Chaisemartin and D'Haultfœuille (2020) is when the effect of the other treatments do not vary across (g,t) cells, an often implausible assumption.

Treating a variable in Regression 1 as a covariate or as a treatment leads to different decompositions of β_{fe}, so it is important to weigh this choice carefully. In principle, any variable that has a causal effect on the outcome, in the sense that different values of that variable lead to different counterfactual outcomes, should be regarded as a treatment. Covariates, on the other hand, are variables that do not have a causal effect on the outcome, though their evolution may be correlated with the outcome's evolution.

Overall, Theorems 1 and 2 are important extensions of Theorem S4 of de Chaisemartin and D'Haultfœuille (2020). They apply to TWFE regressions where some of the variables in the regression, other than the main treatment of interest, can affect the outcome and may have heterogeneous effects. Such regressions are likely to be common in practice.

3.4.2 Linear regressions with several treatments

Theorems 1 and 2 complement the pioneering work of Sun and Abraham (2021). The authors study the so-called event-study regression, an example of a TWFE regression with several treatments, 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 in Regression 1 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.[7] Our decompositions extend their result, by showing that the contamination bias they first uncovered is much more 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. We also

[7] 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.

show that TWFE regressions with one treatment intended at measuring heterogeneous treatment effects may also be affected by some form of contamination bias. Finally, we show that omitting the other treatments from the regression may not necessarily increase the regression coefficient's bias.

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

Another related paper, released after ours, is Goldsmith-Pinkham, Hull and Kolesár (2021), who show that a contamination phenomenon similar to that in Sun and Abraham (2021) and in Theorem 1 also arises in linear regressions with several treatments, and a set of controls such that the treatments can be assumed to be independent of the potential outcomes conditional on those controls. 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 $D^1_{g,t}$ 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.

4 Alternative estimator

4.1 Identifying assumption

In this section, we start by considering the following identifying assumption.

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

1. $E(Y_{g,t}(d_{t-1}) - Y_{g,t-1}(d_{t-1})|D_{g,1},...,D_{g,t-2},D_{g,t-1} = d_{t-1},D_{g,t},...,D_{g,T}) = E(Y_{g,t}(d_{t-1}) - Y_{g,t-1}(d_{t-1})|D_{g,t-1} = d_{t-1})$.

2. $E(Y_{g,t}(d_{t-1}) - Y_{g,t-1}(d_{t-1})|D_{g,t-1} = d_{t-1})$ *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.[8] Moreover, parallel trends may be more likely to hold over consecutive time periods than over the panel's entire duration.

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 $Y_{g,t}(d_{t-1})$ for (g,t) cells such that $D_{g,t-1} = d_{t-1}$. 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

[8]Because it imposes parallel trends conditional on $D_{g,t-1}$, 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.

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 5, there are two non-binary treatments, and while there are (g,t) cells whose two treatments are equal to 0, there is no group that does not receive any of the two treatments at two dates at least. Accordingly, we cannot construct an heterogeneity-robust DID estimator relying on Assumption 3, while we can construct one relying on Assumption 4.

We also consider a second identifying assumption.

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

1. $E(Y_{g,t}(d_t) - Y_{g,t-1}(d_t) | D_{g,1},...,D_{g,t-1}, D_{g,t} = d_t, D_{g,t+1},...,D_{g,T}) = E(Y_{g,t}(d_t) - Y_{g,t-1}(d_t) | D_{g,t} = d_t)$.

2. $E(Y_{g,t}(d_t) - Y_{g,t-1}(d_t) | D_{g,t} = d_t)$ *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. Assumptions 4 and 5 are both parallel trends assumptions over two consecutive periods, among groups with the same treatments at one of the two periods. Accordingly, in instances where Assumption 4 is plausible, Assumption 5 may be plausible too. Imposing jointly Assumptions 4 and 5 may imply that the treatment effects follow the same evolution over time in some groups.[9]

4.2 Target parameters

Let

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

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

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

[9]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.

the average effect of moving the first treatment from 0 to 1 while keeping all other treatments at their observed value, across all switchers.[10]

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 $\mathcal{D}_1 = \{(g,t) : D^1_{g,t} = 1\}$. 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 $\mathcal{D}_1$.

(g,t) cells belonging to $\mathcal{D}_1$ but not to $\mathcal{S}_1$ can be divided into five mutually exclusive subgroups, detailed in Section 1 of the Web Appendix. Identifying the effect of the first treatment in each of those subgroups would require imposing stronger assumptions than Assumption 4. For instance, the first subgroup belonging to $\mathcal{D}_1$ but not to $\mathcal{S}_1$ are all (g,t)s such that $D^1_{g,t} = 1$ for all t. As those cells' first treatment never changes, their first-treatment's effect cannot be identified under a parallel trends assumption. The second and third subgroups are cells whose first-treatment's effect could only be identified under a stronger parallel trends assumption than Assumption 4, which only imposes parallel trends over consecutive periods and conditional on cells' period-$t-1$ treatments. The fourth subgroup are cells whose first treatment changes while at least one of their other treatment also changes. Identifying their first-treatment's effect would require assuming that the effect of the other treatments is constant between groups, as discussed in Section 3.2.1 (see Equation (10) therein). The fifth subgroup are cells whose first treatment changes while their other treatments do not change, but such that all potential control cells experiencing no treatment change have different baseline treatments. Identifying their first-treatment's effect would require assuming that the effects of the other treatments are constant over time, as discussed in Section 3.2.1 (see Equation (12) therein). Therefore, $\mathcal{S}_1$ is the maximal set of (g,t) cells for which the effect of the first treatment can be identified under a minimal parallel trends assumption and without restricting treatment effect heterogeneity.

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

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

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

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

where

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

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

4.3 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 $d^{-1} \in \mathcal{D}_2 \times ... \times \mathcal{D}_K$, let

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

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 d^{-1} at both dates. We then let $N_{d,d',d^{-1},t} = \sum_{g\in\mathcal{G}_{d,d',d^{-1},t}} N_{g,t}$ denote the total population of groups in $\mathcal{G}_{d,d',d^{-1},t}$. Let also

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

$$\text{DID}^f_{-,d^{-1},t} = \sum_{g\in\mathcal{G}_{1,1,d^{-1},t}} \frac{N_{g,t}}{N_{1,1,d^{-1},t}} \left(Y_{g,t} - Y_{g,t-1}\right) - \sum_{g\in\mathcal{G}_{0,1,d^{-1},t}} \frac{N_{g,t}}{N_{0,1,d^{-1},t}} \left(Y_{g,t} - Y_{g,t-1}\right). \tag{15}$$

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

$\text{DID}^f_{+,d^{-1},t}$ compares the $t-1$-to-t outcome evolution of groups whose first treatment goes from 0 to 1 from $t-1$ to t while their other treatments are equal to d^{-1} at both dates, to the outcome evolution of groups whose first and other treatments are respectively equal to 0 and d^{-1} at both dates. Under Assumption 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. $\text{DID}^f_{-,d^{-1},t}$'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

$$\text{DID}^f_{\text{M}} = \sum_{t=2}^{T} \sum_{d^{-1} \in \{0,1\}^{K-1}} \left(\frac{N_{1,0,d^{-1},t}}{N_{\mathcal{S}_1}} \text{DID}^f_{+,d^{-1},t} + \frac{N_{0,1,d^{-1},t}}{N_{\mathcal{S}_1}} \text{DID}^f_{-,d^{-1},t} \right) \tag{16}$$

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 $\text{DID}^f_{+,d^{-1},t}$ and $\text{DID}^f_{-,d^{-1},t}$ estimators, across values of the other treatments 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, even if switchers and non-switchers are equally likely to experience a change in their other treatments, the analogue of having that the treatments are uncorrelated conditional on the group and time fixed effects in the TWFE regression. 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.

Like in de Chaisemartin and D'Haultfœuille (2020), it is straightforward to propose a placebo version of the DID^f_{M} 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 (14) and (15) 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 DID^f_{M} 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 Section 4 of the Web Appendix 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 $d^{-1} \in \mathcal{D}_2 \times ... \times \mathcal{D}_K$, let

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

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 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 DID^f_{M} estimator is a weighted average of the $\text{DID}^f_{d,d',d^{-1},t}$ estimators, across d, d', 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 $d^{-1} \in \mathcal{D}_2 \times ... \times \mathcal{D}_K$, let $N_{d,d',d^{-1},t+1,t} = \sum_{g \in \mathcal{G}_{d,d',d^{-1},t+1}} N_{g,t}$ denote the total population, at period t, of groups in $\mathcal{G}_{d,d',d^{-1},t+1}$. Then, let

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

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

In contrast to $\text{DID}^f_{+,d^{-1},t}$, which is a "forward" DID, $\text{DID}^b_{+,d^{-1},t}$ 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 d^{-1} at both dates, to the outcome evolution of groups whose first and other treatments are respectively equal to 0 and d^{-1} at both dates. $\text{DID}^b_{-,d^{-1},t}$ 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

$$\text{DID}^b_{\text{M}} = \sum_{t=1}^{T-1} \sum_{d^{-1} \in \{0,1\}^{K-1}} \left(\frac{N_{0,1,d^{-1},t+1,t}}{N_{\mathcal{S}_2}} \text{DID}^b_{+,d^{-1},t} + \frac{N_{1,0,d^{-1},t+1,t}}{N_{\mathcal{S}_2}} \text{DID}^b_{-,d^{-1},t} \right) \tag{17}$$

if $N_{\mathcal{S}_2} > 0$, and $\text{DID}^b_{\text{M}} = 0$ if $N_{\mathcal{S}_2} = 0$. One can show that if Assumptions 1-2 and 5 hold, $E\left[\text{DID}^b_{\text{M}}\right] = \delta_2$.

4.4 Dynamic treatment effects

With a single treatment $D^s_{g,t}$, DID^f_{M} can be used to estimate the effect of the current value of $D^s_{g,t}$, allowing for dynamic effects. Assume that $(D^1_{g,t}, ..., D^K_{g,t}) = (D^s_{g,t}, ..., D^s_{g,t-(K-1)})$. Then,

our potential outcome notation allows the current treatment and its first $K-1$ lags to affect the outcome, so DID^f_{M} is an estimator of the effect of the current value of $D^s_{g,t}$ robust to dynamic effects up to $K-1$ lags. This is an improvement over the DID_{M} estimator in de Chaisemartin and D'Haultfœuille (2020), which is not robust to dynamic effects, except with a binary and staggered treatment. To achieve some robustness to dynamic effects, DID^f_{M} restricts the estimation to groups that did not experience a treatment change from $t-K$ to $t-1$. For instance, with $K=2$ and $(D^1_{g,t}, D^2_{g,t}) = (D^s_{g,t}, D^s_{g,t-1})$, the DID^f_{M} estimator compares groups with $(D^s_{g,t-2}, D^s_{g,t-1}, D^s_{g,t}) = (0,0,1)$ to groups with $(D^s_{g,t-2}, D^s_{g,t-1}, D^s_{g,t}) = (0,0,0)$, and groups with $(D^s_{g,t-2}, D^s_{g,t-1}, D^s_{g,t}) = (1,1,0)$ to groups with $(D^s_{g,t-2}, D^s_{g,t-1}, D^s_{g,t}) = (1,1,1)$. On the other hand, the DID^f_{M} estimator may not be used to estimate the effect of past treatments on the outcome. For instance, with $K=2$ and $(D^1_{g,t}, D^2_{g,t}) = (D^s_{g,t-1}, D^s_{g,t})$, $\mathcal{S}_1$ is empty: for any group g such that $(D^s_{g,t-2} \neq D^s_{g,t-1} = D^s_{g,t})$, there cannot exist another group g' such that $(D^s_{g',t-2} = D^s_{g',t-1} = D^s_{g',t})$, and $(D^s_{g',t-2}, D^s_{g',t-1}) = (D^s_{g,t-2}, D^s_{g,t-1})$.

The opposite applies to DID^b_{M}: it may not be used to estimate the effect of the current treatment allowing for dynamic effects, but it may be used to estimate the effect of past treatments on the outcome. For instance, with $K=2$ and $(D^1_{g,t}, D^2_{g,t}) = (D^s_{g,t-1}, D^s_{g,t})$, $\mathcal{S}_2$ is not empty: it contains all (g,t) cells such that $D^s_{g,t-1} \neq D^s_{g,t} = D^s_{g,t+1}$, for which there exists another group g' such that $D^s_{g',t-1} = D^s_{g',t} = D^s_{g',t+1} = D^s_{g,t+1}$. Then, DID^b_{M} is a weighted average of two types of DIDs. DIDs of the first type compare the $t+1$ to t outcome evolution between groups such that $D^s_{g,t-1} = 1, D^s_{g,t} = 0, D^s_{g,t+1} = 0$ and groups such that $D^s_{g,t-1} = 0, D^s_{g,t} = 0, D^s_{g,t+1} = 0$. DIDs of the second type compare the $t+1$ to t outcome evolution between groups such that $D^s_{g,t-1} = 1, D^s_{g,t} = 1, D^s_{g,t+1} = 1$ and groups such that $D^s_{g,t-1} = 0, D^s_{g,t} = 1, D^s_{g,t+1} = 1$. If the current outcome only depends on the current treatment and its first lag, and if Assumption 5 holds for $(D^1_{g,t}, D^2_{g,t}) = (D^s_{g,t-1}, D^s_{g,t})$, then DID^b_{M} is unbiased for the average effect of switching the treatment's first lag from 0 to 1 holding the current treatment fixed, across all (g,t)s in $\mathcal{S}_2$.

Of course, assuming that the current outcome only depends on the current treatment and its first lag is restrictive. One could instead assume, say, that the current outcome only depends on the current treatment and its first two lags. Then, with $K=3$ and $(D^1_{g,t}, D^2_{g,t}, D^3_{g,t}) = (D^s_{g,t-2}, D^s_{g,t-1}, D^s_{g,t})$, DID^b_{M} is unbiased for the average effect of switching the treatment's second lag from 0 to 1, holding the current treatment and its first lag fixed, across all (g,t) cells in $\mathcal{S}_2$. $\mathcal{S}_2$ now becomes the set of all (g,t) cells such that $D^s_{g,t-2} \neq D^s_{g,t-1} = D^s_{g,t} = D^s_{g,t+1}$ and for which there exists another group g' such that $D^s_{g',t-2} = D^s_{g',t-1} = D^s_{g',t} = D^s_{g',t+1} = D^s_{g,t+1}$. $\mathcal{S}_2$ contains fewer cells with $K=3$ and $(D^1_{g,t}, D^2_{g,t}, D^3_{g,t}) = (D^s_{g,t-2}, D^s_{g,t-1}, D^s_{g,t})$ than with $K=2$ and $(D^1_{g,t}, D^2_{g,t}) = (D^s_{g,t-1}, D^s_{g,t})$: allowing more treatment lags to affect the outcome may be more plausible, but it may also result in less precise estimators, that apply to a smaller population. Note also that with dynamic effects up to $K-1$ treatment lags, DID^b_{M} can be used to estimate the effect of the $K-1$th lag, but it cannot be used to estimate the effect of earlier lags.

Overall, $\text{DID}_{\text{M}}^{f}$ and $\text{DID}_{\text{M}}^{b}$ can be used for some but not for all purposes in the presence of a single treatment with dynamic effects. Assuming constant treatment effects, one can use a TWFE regression of the outcome on the treatment and its lags, the so-called distributed lag regression, to separately estimate the effect of the current and past treatments on the outcome. Separately estimating each of those effects while allowing for heterogeneous treatment effects is inherently difficult. This may be the reason why a substantial branch of the heterogeneity-robust DID literature has instead proposed to estimate the total effect of current and past treatments on the outcome (see Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; de Chaisemartin and D'Haultfœuille, 2021*a*; Borusyak, Jaravel and Spiess, 2021). This literature has focused on the case with one treatment. In Section 2 of the Web Appendix, we extend that literature, and propose estimators of instantaneous and dynamic effects when there are several binary and staggered treatments. Those estimators can also be used in the presence of a single treatment that can change multiple times, to isolate the effect of each treatment change. For instance, with a treatment that can switch on and then off, one may be interested in separately estimating the effect of switching the treatment on/off.

4.5 Inference

In Section 3 of the Web Appendix, we prove the asymptotic normality of $\text{DID}_{\text{M}}^{f}$ when the number of groups goes to infinity, and we propose confidence intervals. Those 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 of this approach, though, 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. The larger the number of treatments, the more likely it is that $\text{DID}_{\text{M}}^{f}$ uses data from a small number of groups. In this section, 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). The exactness of those confidence intervals relies on strong conditions, but they remain asymptotically valid under much weaker assumptions. The main price to pay for using them, rather than those described in Section 3 of the Web Appendix, is that doing so may result in an adjustment of the definition of δ_1, as explained below.

To ease the exposition, in this section we condition on $\boldsymbol{D}$. Accordingly, functions of $\boldsymbol{D}$ can be treated as non-stochastic terms. For simplicity, we also assume that $N_{g,t} = 1$ for all (g,t). Let $s = 1, ..., S$ index "switches", that is to say a $K+2$-uple (d, d', d^{-1}, t) with $d \neq d'$ for which there exists (g, g') satisfying $D_{g,t}^{1} = d$, $D_{g,t-1}^{1} = D_{g',t-1}^{1} = D_{g',t}^{1} = d'$ and $D_{g,t}^{-1} = D_{g,t-1}^{-1} = D_{g',t}^{-1} =$

$D^{-1}_{g',t-1} = d^{-1}$. Then, note that

$$\text{DID}_\text{M} = \sum_{s=1}^{S} \alpha_s \text{DID}_s,$$

for some non-stochastic weights $(\alpha_s)_{s=1,...,S}$, where DID_s is the DID corresponding to switch s. Let $\mathcal{G}_s$ denote the set of groups intervening in DID_s, either as a "switcher" or as a "control". The following assumption is needed to ensure the validity of our approach.

Assumption 6 (Non-overlapping groups) *for any* $(s, s') \in \{1, ..., S\}^2$, $s \neq s'$, $\mathcal{G}_s \cap \mathcal{G}_{s'} = \emptyset$.

Note that Assumption 6 automatically holds with $T = 2$. Otherwise, it is more likely to hold if T is small. When Assumption 6 fails, we can ensure it holds on a modified sample, by removing groups belonging to several sets $\mathcal{G}_s$ from all those sets except one. This sample modification will modify the estimator DID^f_M. It may also lead to removing a switching group from a set $\mathcal{G}_s$. This would change the target parameter, which would become the average treatment effect across all switching cells in the modified sample, in lieu of δ_1, the average treatment effect across all switching cells in the original sample. For simplicity, we still denote the parameter and its estimator on the modified sample δ_1 and DID^f_M.

Our confidence interval relies on the following variance estimator:

$$\widehat{V} = \sum_{s=1}^{S} \alpha_s^2 \left[\frac{1}{n_{1s}(n_{1s}-1)} \sum_{g \in \mathcal{G}_{1s}} (\Delta Y_{g,t_s} - \overline{\Delta Y}_{1s})^2 + \frac{1}{n_{0s}(n_{0s}-1)} \sum_{g \in \mathcal{G}_{0s}} (\Delta Y_{g,t_s} - \overline{\Delta Y}_{0s})^2 \right],$$

where $\mathcal{G}_{1s}$ (resp. $\mathcal{G}_{0s}$) is the subset of switching (resp. control) cells in $\mathcal{G}_s$, n_{ks} =card($\mathcal{G}_{ks}$), and $\overline{\Delta Y}_{ks}$ is the average of $\Delta Y_{g,t_s}$ over $g \in \mathcal{G}_{ks}$. Our definition of $\widehat{V}$ uses the convention that 0/0=0.

Next, let $q_{1-\alpha}$ denote the quantile of order $1-\alpha$ of $|T|$, defined as

$$T = \left(\frac{\sum_{s=1}^{S} \alpha_s^2 \left(1/n_{1s} + 1/n_{0s}\right)}{\sum_{s=1}^{S} \alpha_s^2 \left[W_{1s}/[n_{1s}(n_{1s}-1)] + W_{0s}/[n_{0s}(n_{0s}-1)]\right]} \right)^{1/2} \times Z, \tag{18}$$

where $(Z, W_{01}, W_{11}, ..., W_{0S}, W_{1S})$ are independent of the data, mutually independent and satisfy $Z \sim \mathcal{N}(0,1)$ and $W_{ks} \sim \chi^2(n_{ks}-1)$. Note that $q_{1-\alpha}$ does not have a closed-form expression, but it can be approximated by simulations.

Then, the confidence interval of order $1-\alpha$ we consider is

$$\text{CI}^{\text{ex}}_{1-\alpha} = \left[\text{DID}_\text{M} \pm q_{1-\alpha} \sqrt{\widehat{V}} \right].$$

Below, we introduce two assumptions under which $\text{CI}^{\text{ex}}_{1-\alpha}$ is valid: under Assumption 7, $\text{CI}^{\text{ex}}_{1-\alpha}$ is valid in finite samples; under Assumption 8, $\text{CI}^{\text{ex}}_{1-\alpha}$ is valid asymptotically.

Assumption 7 (Restrictions for finite-sample validity of $\text{CI}^{\text{ex}}_{1-\alpha}$)

1. *For all* $s = 1, ..., S$ *and* $g \in \mathcal{G}_s$, $Y_{g,t_s}(d_s^1, d_s^{-1}) - Y_{g,t_s}(d_s^{1\prime}, d_s^{-1}) = \delta_{1s}$ *where* $(t_s, d_s^1, d_s^{1\prime}, d_s^{-1})$ *are the* $(t, d^1, d^{1\prime}, d^{-1})$ *associated with* s *and* δ_{1s} *is non-stochastic.*

2. *For all* s *and* $g \in \mathcal{G}_s$, $\Delta Y_{g,t_s}(0, d_s^{-1}) \sim \mathcal{N}(\mu_s, \sigma^2)$.

Point 1 of Assumption 7 assumes that the first treatment's effect is homogeneous within each set of groups s, but may vary across s. Point 2 of Assumption 7 assumes that $\Delta Y_{g,t_s}(0, d_s^{-1})$ is normally distributed and homoskedastic: the variance of $\Delta Y_{g,t_s}(0, d_s^{-1})$ should not depend on s.

Assumption 8 (Restrictions for asymptotic validity of $\mathbf{CI}_{1-\alpha}^{\mathbf{ex}}$)

1. *There exists* G_0 *such that for all* $G \geq G_0$, $\mathscr{S}_G := \{(d, d', d^{-1}, t) : N_{d,d',d^{-1},t} > 0,\ N_{d',d',d^{-1},t} > 0\}$ *does not vary across* G *and is finite. We denote by* $\overline{S}$ *its cardinal.*[11]

2. *For all* $(k, s) \in \{0, 1\} \times \{1, ..., \overline{S}\}$, *the* $(\Delta Y_{g,t_s})_{g \in \mathcal{G}_{ks}}$ *are i.i.d. and for* $g \in \mathcal{G}_{ks}$, $E[\Delta Y_{g,t_s}^2] < \infty$ *and* $V(\Delta Y_{g,t_s}) > 0$.

3. *For all* $(k, s) \in \{0, 1\} \times \{1, ..., \overline{S}\}$, $\liminf_{G \to \infty} n_{ks}/G > 0$.

Assumption 8 does not make any treatment-effect homogeneity or homoscedasticity assumption. Note that we impose that the $(\Delta Y_{g,t_s})_{g \in \mathcal{G}_{ks}}$ are identically distributed for simplicity. If we instead assumed that the $(\Delta Y_{g,t_s})_{g \in \mathcal{G}_{ks}}$ are independent but not identically distributed, one could still show that $\mathrm{CI}_{1-\alpha}^{\mathrm{ex}}$ is asymptotically conservative under appropriate regularity conditions, as in, e.g., Theorem S6 in the Web Appendix of de Chaisemartin and D'Haultfœuille (2020).

The following theorem shows that $\mathrm{CI}_{1-\alpha}^{\mathrm{ex}}$ is exact under Assumption 7, and asymptotically valid under Assumption 8.

Theorem 5 *If Assumptions 1-2, 4, and 6 hold, then, for any* $\alpha \in (0, 1)$*:*

1. *if Assumption 7 further holds,* $P\left(\delta_1 \in CI_{1-\alpha}^{ex}\right) = 1 - \alpha$.

2. *if Assumption 8 further holds,* $\lim_{G \to \infty} P\left(\delta_1 \in CI_{1-\alpha}^{ex}\right) = 1 - \alpha$.

Our approach in this section generalizes that in Donald and Lang (2007) to designs with more than two time periods and/or several treatments. A difference with that paper is that our confidence intervals use critical values from a non-standard distribution, instead of critical values from a t-distribution.

[11] Accordingly, we keep the same indexation for switches $s \in \{1, ..., \overline{S}\}$ for all $G \geq G_0$.

5 Application

In this section, we revisit Hotz and Xiao (2011).[12] 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.

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.

[12] 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 DID^{f}_{M} estimator).

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

Min. years of schooling	# of (g,t) cells
0	26
12	36
12.5	5
13	4
14	61
16	21
Min. staff-to-child ratio	**# of (g,t) cells**
0	5
1/8	2
1/7	4
1/6	30
1/5	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=$[-0.735, -0.155]$),[13] 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 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 DID^f_{M} estimator we compute below: while the DID^f_{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.

[13]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).

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 4, 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 5 (g,t) cells our estimator applies to are (Kentucky,1992), (Minnesota,1992), (Utah,1992), (Vermont,1992), and (Rhode Island,1997).[14] Of the 122 (g,t) cells we lose when focusing on $\mathcal{S}_1$, 93 belong to the first subgroup in Appendix 1, 19 belong to the second or third subgroup, and 10 belong to the fourth or fifth subgroup. Therefore, the vast majority of the cells we lose do not experience any change of their minimum years of schooling, so their treatment effect cannot be identified under a parallel trends assumption. We may seem to lose 19 cells by imposing only a minimal parallel trends assumption. In reality, estimating the treatment effects of 14 of the 19 cells in the second or third subgroup would also require assuming that the effect of the minimum staff-to-child ratio is homogeneous: either their minimum staff-to-child ratio also changes when their minimum years of schooling changes, or they cannot be matched to a control state with the same baseline treatments.

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 Section 3 of the Web Appendix may not be very reliable for that estimator. Instead, we compute the confidence interval $\text{CI}^{\text{ex}}_{1-\alpha}$ for $\alpha = 0.95$ and find that it is equal to $[-0.821, 0.807]$.[15] 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} 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

[14] 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.

[15] $\text{CI}^{\text{ex}}_{1-\alpha}$ relies on Assumption 6, 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 Section 4.5) changes very slightly the value of DID^f_{M} (-0.0072 in lieu of -0.029).

in Section 3 of the Web Appendix may be more reasonable for that estimator,[16] 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.

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 heterogeneous in this application.

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	$[-0.852, -0.280]$
$\widehat{\beta}_s$	-0.022	$[-0.114, 0.074]$
DID^f_{M}	-0.029	$[-0.821, 0.807]$
$1/2(\text{DID}^f_{\text{M}} + \text{DID}^b_{\text{M}})$	-0.016	$[-0.126, 0.094]$

[16] 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).

6 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.

In most instances where TWFE and DID estimators are used, it is likely that besides the main treatment of interest, many other determinants of the outcome change over the study period. We show that in the presence of heterogeneous treatment effects, failing to control for those other treatments, be it in a TWFE regression or using an heterogeneity-robust DID estimator, may lead to a biased estimator, even if those other treatments are uncorrelated with the main treatment of interest. Accordingly, all those other treatments should be controlled for, but our results also show that a non-parametric DID estimator robust to the heterogeneous effects of many treatments will often be subject to a curse of dimensionality. Data-driven methods to select the treatments that should be controlled for, as well as more parametric methods to control for them, would be useful additions to the econometrics literature. In the meantime, applied researchers could discuss more systematically whether other treatments than the one under consideration have changed over their study period. If so, they could assess if their estimates are robust to controlling for at least some of those other treatments, using the tools provided in this paper.

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.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2021. "Revisiting event study designs: Robust and efficient estimation." *arXiv preprint arXiv:2108.12419*.

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

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'Haultfœuille. 2021*a*. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." arXiv preprint arXiv:2007.04267.

de Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2021*b*. "Two-way fixed effects regressions with several treatments." *arXiv preprint arXiv:2012.10077, v4*.

D'Haultfœuille, Xavier, and Purevdorj Tuvaandorj. 2022. "A Robust Permutation Test for Subvector Inference in Linear Regressions." arXiv preprint 2205.06713.

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 $K = 2$, $D_{g,t}^{-1} = D_{g,t}^{2}$. Then, $D_{g,t}^{-1} \neq \mathbf{0}^{-1}$ if and only if $D_{g,t}^{2} = 1$, and one then has $D_{g,t}^{2}\Delta_{g,t}^{-1} = D_{g,t}^{2}\Delta_{g,t}^{2}$.

A.2 Theorem 2

We first establish the following lemma.

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

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

Proof of Lemma 1

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

$$\begin{aligned}
E\left(Y_{g,t}|\boldsymbol{D}\right) =&E\Bigg(Y_{g,t}(0, \mathbf{0}^{-1}) + D_{g,t}^{1}(Y_{g,t}(1, D_{g,t}^{-1}) - Y_{g,t}(0, D_{g,t}^{-1}) + Y_{g,t}(0, D_{g,t}^{-1}) - Y_{g,t}(0, \mathbf{0}^{-1})) \\
&+ (1 - D_{g,t}^{1})(Y_{g,t}(0, D_{g,t}^{-1}) - Y_{g,t}(0, \mathbf{0}^{-1}))\Big|\boldsymbol{D}\Bigg) \\
=&E\left(Y_{g,t}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}\right) + D_{g,t}^{1}E\left(\Delta_{g,t}^{1}\middle|\boldsymbol{D}\right) + E\left(\Delta_{g,t}^{-1}\middle|\boldsymbol{D}\right) \\
=&E\left(Y_{g,t}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}_g\right) + D_{g,t}^{1}E\left(\Delta_{g,t}^{1}\middle|\boldsymbol{D}\right) + E\left(\Delta_{g,t}^{-1}\middle|\boldsymbol{D}\right), \qquad (19)
\end{aligned}$$

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

$$\begin{aligned}
&E\left(Y_{g,t}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}_g\right) - E\left(Y_{g,t'}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}_g\right) - E\left(Y_{g',t}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}_g\right) + E\left(Y_{g',t'}(0, \mathbf{0}^{-1})\middle|\boldsymbol{D}_g\right) \\
=&0. \qquad (20)
\end{aligned}$$

The result follows by combining (19) and (20).

Proof of Theorem 2

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

$$E\left(\widehat{\beta}_{fe}\middle|\boldsymbol{D}\right) = \frac{\sum_{g,t} N_{g,t}\varepsilon_{g,t}E\left(Y_{g,t}|\boldsymbol{D}\right)}{\sum_{g,t} N_{g,t}\varepsilon_{g,t}D_{g,t}^{1}}. \qquad (21)$$

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

$$\sum_{t=1}^{T} N_{g,t}\varepsilon_{g,t} = 0 \text{ for all } g \in \{1, ..., G\}, \tag{22}$$

$$\sum_{g=1}^{G} N_{g,t}\varepsilon_{g,t} = 0 \text{ for all } t \in \{1, ..., T\}, . \tag{23}$$

Then,

$$\begin{aligned}
&\sum_{g,t} N_{g,t}\varepsilon_{g,t} E\left(Y_{g,t}|\boldsymbol{D}\right) \\
=&\sum_{g,t} N_{g,t}\varepsilon_{g,t}\left(E\left(Y_{g,t}|\boldsymbol{D}\right) - E\left(Y_{g,1}|\boldsymbol{D}\right) - E\left(Y_{1,t}|\boldsymbol{D}\right) + E\left(Y_{1,1}|\boldsymbol{D}\right)\right) \\
=&\sum_{g,t} N_{g,t}\varepsilon_{g,t}\left(D^1_{g,t}E\left(\Delta^1_{g,t}\middle|\boldsymbol{D}\right) + E\left(\Delta^{-1}_{g,t}\middle|\boldsymbol{D}\right) - D^1_{1,t}E\left(\Delta^1_{1,t}(D^{-1}_{1,t})\middle|\boldsymbol{D}\right) - E\left(\Delta^{-1}_{1,t})\middle|\boldsymbol{D}\right)\right. \\
&\left. - D^1_{g,1}E\left(\Delta^1_{g,1}(D^{-1}_{g,1})\middle|\boldsymbol{D}\right) - E\left(\Delta^{-1}_{g,1}\middle|\boldsymbol{D}\right) + D^1_{1,1}E\left(\Delta^1_{1,1}(D^{-1}_{1,1})\middle|\boldsymbol{D}\right) + E\left(\Delta^{-1}_{1,1})\middle|\boldsymbol{D}\right)\right) \\
=&\sum_{g,t} N_{g,t}\varepsilon_{g,t}\left(D^1_{g,t}E\left(\Delta^1_{g,t}\middle|\boldsymbol{D}\right) + E\left(\Delta^{-1}_{g,t}\middle|\boldsymbol{D}\right)\right) \\
=&\sum_{(g,t):D^1_{g,t}=1} N_{g,t}\varepsilon_{g,t}E\left(\Delta^1_{g,t}\middle|\boldsymbol{D}\right) + \sum_{(g,t):D^{-1}_{g,t}\neq\mathbf{0}^{-1}} N_{g,t}\varepsilon_{g,t}E\left(\Delta^{-1}_{g,t}\middle|\boldsymbol{D}\right).
\end{aligned} \tag{24}$$

The first and third equalities follow from Equations (22) and (23). The second equality follows from Lemma 1. The fourth equality follows from the fact that $\Delta^0_{g,t}(\mathbf{0}^{-1}) = 0$. Finally,

$$\sum_{g,t} N_{g,t}\varepsilon_{g,t}D^1_{g,t} = \sum_{(g,t):D^1_{g,t}=1} N_{g,t}\varepsilon_{g,t}. \tag{25}$$

Combining (21), (24), (25) yields

$$E\left(\widehat{\beta}_{fe}\middle|\mathbf{D}\right) = \sum_{(g,t):D^1_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}E\left(\Delta^1_{g,t}\middle|\boldsymbol{D}\right) + \sum_{(g,t):D^{-1}_{g,t}\neq\mathbf{0}^{-1}} \frac{N_{g,t}}{N_1} w_{g,t}E\left(\Delta^{-1}_{g,t}\middle|\boldsymbol{D}\right). \tag{26}$$

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

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

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.

Proof of 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.

Proof of 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 $\Delta^2_{g,t} = B(2 \times 1\{w^s_{g,t} \geq 0\} - 1)$.

Theorem 4

First, by definition of DID^f_{M},

$$\text{DID}^f_{\text{M}} = \sum_{t=2}^{T} \sum_{d^{-1} \in \{0,1\}^{K-1}} \frac{N_{1,0,d^{-1},t}}{N_{\mathcal{S}_1}} \text{DID}^f_{+,d^{-1},t} + \frac{N_{0,1,d^{-1},t}}{N_{\mathcal{S}_1}} \text{DID}^f_{-,d^{-1},t}, \tag{27}$$

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

$$E\left(Y_{g,t} - Y_{g,t-1} | \boldsymbol{D}\right) = E\left(\Delta^1_{g,t} \middle| \boldsymbol{D}\right) + E\left(Y_{g,t}(0,d^{-1}) - Y_{g,t-1}(0,d^{-1}) \middle| \boldsymbol{D}\right). \tag{28}$$

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

$$\begin{aligned} E\left(Y_{g,t}(0,d^{-1}) - Y_{g,t-1}(0,d^{-1}) \middle| \boldsymbol{D}\right) &= E\left(Y_{g,t}(0,d^{-1}) - Y_{g,t-1}(0,d^{-1}) \middle| \boldsymbol{D}_g\right) \\ &= E\left(Y_{g,t}(0,d^{-1}) - Y_{g,t-1}(0,d^{-1}) \middle| D^1_{g,t-1} = 0, D^{-1}_{g,t-1} = d^{-1}\right) \\ &= \psi_{0,d^{-1},t}. \end{aligned} \tag{29}$$

As a result,

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

The first equality follows by (28), the second by (29), and the third after some algebra. Given that $\text{DID}^f_{+,d^{-1},t} = 0$ if $N_{1,0,d^{-1},t} = 0$ or $N_{0,0,d^{-1},t} = 0$, we obtain, by definition of $\mathcal{S}_1$ and with the

convention that sums over empty sets are 0,

$$E\left(N_{1,0,d^{-1},t}\text{DID}^f_{+,d^{-1},t}\middle|\boldsymbol{D}\right)=E\Bigg(\sum_{\substack{g:D^1_{g,t}=1,D^{-1}_{g,t}=d^{-1}\\(g,t)\in\mathcal{S}_1}}N_{g,t}\Delta^1_{g,t}\middle|\boldsymbol{D}\Bigg). \tag{30}$$

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

$$E\left(N_{0,1,d^{-1},t}\text{DID}^f_{-,d^{-1},t}\middle|\boldsymbol{D}\right)=E\Bigg(\sum_{\substack{g:D^1_{g,t}=0,D^{-1}_{g,t}=d^{-1}\\(g,t)\in\mathcal{S}_1}}N_{g,t}\Delta^1_{g,t}\middle|\boldsymbol{D}\Bigg). \tag{31}$$

Plugging (30) and (31) into (27) yields

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

A.3 Theorem 5

As in Subsection 4.5, the proof is conditional on $\boldsymbol{D}$.

1. First, under Assumption 7, we have $\delta_1=\sum_{s=1}^{S}\alpha_s\delta_{1s}$. Thus, using again Assumption 7 but also Assumption 6,

$$\text{DID}_{\text{M}}-\delta_1=\sum_{s=1}^{S}\alpha_s(\text{DID}_s-\delta_{1s})\sim\mathcal{N}\left(0,\sigma^2\sum_{s=1}^{S}\alpha_s^2\left(\frac{1}{n_{1s}}+\frac{1}{n_{0s}}\right)\right). \tag{32}$$

For the same reasons, we have

$$\frac{\widehat{V}}{\sigma^2}\sim\sum_{s=1}^{S}\alpha_s^2\left[\frac{W_{1s}}{n_{1s}(n_{1s}-1)}+\frac{W_{0s}}{n_{0s}(n_{0s}-1)}\right],$$

where the $(W_{01},W_{11},...,W_{0S},W_{1S})$ are mutually independent, independent of DID_{M} and $W_{ks}\sim\chi^2(n_{ks}-1)$, by Cochran's theorem. By definition of T (see (18)), this implies that

$$\frac{\text{DID}_{\text{M}}-\delta_1}{\sqrt{\widehat{V}}}\sim T.$$

The result follows.

2. Without loss of generality, we assume hereafter that $G \geq G_0$, so that $\mathscr{S}_G$ does not vary across G and has cardinal $\overline{S}$.

Consider the ratio $R := (\text{DID}_\text{M} - \delta_1)/\widehat{V}^{1/2}$. We first show that as $G \to \infty$, $R \xrightarrow{d} \mathcal{N}(0,1)$. For any $(k,s) \in \{0,1\} \times \{1,...,\overline{S}\}$, let $\mu_{ks} := E[\Delta Y_{g,t_s}]$ and $\sigma^2_{ks} := V(\Delta Y_{g,t_s})$. Given that $n_{ks} \to \infty$, we have, by the central limit theorem,

$$\frac{\overline{\Delta Y}_{ks} - \mu_{ks}}{\sqrt{\sigma^2_{ks}/n_{ks}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Remark that $\text{DID}_\text{M} = \sum_{s=1}^{\overline{S}} \alpha_s \left(\overline{\Delta Y}_{1s} - \overline{\Delta Y}_{0s}\right)$ and $\delta_1 = \sum_{s=1}^{\overline{S}} \alpha_s \left(\mu_{1s} - \mu_{0s}\right)$. Moreover, $(\overline{\Delta Y}_{01}$, $\overline{\Delta Y}_{11}, ..., \overline{\Delta Y}_{0\overline{S}}, \overline{\Delta Y}_{1\overline{S}})$ are mutually independent by Assumptions 2 and 6. Then, by, e.g., Lemma C.5 of D'Haultfœuille and Tuvaandorj (2022), we obtain

$$\frac{\text{DID}_\text{M} - \delta_1}{\sqrt{\sum_{s=1}^{\overline{S}} \alpha_s^2 \left(\sigma^2_{1s}/n_{1s} + \sigma^2_{0s}/n_{0s}\right)}} \xrightarrow{d} \mathcal{N}\left(0,1\right). \tag{33}$$

Moreover, by the law of large numbers,

$$\frac{1}{n_{1s}-1} \sum_{g \in \mathcal{G}_{1s}} \left(\Delta Y_{g,t_s} - \overline{\Delta Y}_{1s}\right)^2 \xrightarrow{P} \sigma^2_{ks}.$$

Then, $\alpha_s^2 \leq 1$ and $\min_{k,s} \liminf_G n_{ks}/G > 0$ implies that

$$G\left[\widehat{V} - \sum_{s=1}^{\overline{S}} \alpha_s^2 \left(\sigma^2_{1s}/n_{1s} + \sigma^2_{0s}/n_{0s}\right)\right] \xrightarrow{P} 0. \tag{34}$$

Next,

$$G \sum_{s=1}^{\overline{S}} \alpha_s^2 \left(\sigma^2_{1s}/n_{1s} + \sigma^2_{0s}/n_{0s}\right) \geq \left(\min_{k,s} \sigma^2_{ks}\right) \sum_{s=1}^{\overline{S}} \alpha_s^2 \geq \frac{\min_{k,s} \sigma^2_{ks}}{\overline{S}},$$

where the latter holds by convexity of $x \mapsto x^2$ and $\sum_{s=1}^{\overline{S}} \alpha_s = 1$. Hence, by Assumption 8, we obtain

$$\liminf_G G \sum_{s=1}^{\overline{S}} \alpha_s^2 \left(\sigma^2_{1s}/n_{1s} + \sigma^2_{0s}/n_{0s}\right) > 0.$$

Combined with (34), this yields

$$\frac{\widehat{V}}{\sum_{s=1}^{\overline{S}} \alpha_s^2 \left(\sigma^2_{1s}/n_{1s} + \sigma^2_{0s}/n_{0s}\right)} \xrightarrow{P} 1. \tag{35}$$

Taken together, (33) and (35) imply that $R \xrightarrow{d} \mathcal{N}(0,1)$.

Next, we prove that $T \xrightarrow{d} \mathcal{N}(0,1)$. First, for all (k,s), $n_{ks} \to \infty$ by Assumption 8. Thus, by the law of large numbers, $W_{ks}/(n_{ks}-1) \xrightarrow{P} 1$. In turn, using $\liminf_G n_{ks}/G > 0$, we obtain

$$G\left[\sum_{s=1}^{\overline{S}} \alpha_s^2 \left(\frac{W_{1s}}{n_{1s}(n_{1s}-1)} + \frac{W_{0s}}{n_{0s}(n_{0s}-1)}\right) - \sum_{s=1}^{\overline{S}} \alpha_s^2 \left(\frac{1}{n_{1s}} + \frac{1}{n_{0s}}\right)\right] \xrightarrow{P} 0.$$

Moreover, since $G/n_{ks} \geq 1$ for all k, s, we have

$$G \sum_{s=1}^{\overline{S}} \alpha_s^2 \left(\frac{1}{n_{1s}} + \frac{1}{n_{0s}} \right) \geq 2 \sum_{s=1}^{\overline{S}} \alpha_s^2 \geq 1/\overline{S}.$$

As a result, $\liminf_G G \sum_{s=1}^{\overline{S}} \alpha_s^2 \left(1/n_{1s} + 1/n_{0s}\right) > 0$. Hence,

$$\frac{\sum_{s=1}^{S} \alpha_s^2 \left[W_{1s}/[n_{1s}(n_{1s}-1)] + W_{0s}/[n_{0s}(n_{0s}-1)]\right]}{\sum_{s=1}^{S} \alpha_s^2 \left(1/n_{1s} + 1/n_{0s}\right)} \xrightarrow{P} 1.$$

Thus, by definition of T, $T \xrightarrow{d} \mathcal{N}(0,1)$.

By continuity of the normal distribution, this implies that $q_{1-\alpha} \to \Phi^{-1}(1-\alpha/2)$. Now, note that

$$P\left(\delta_1 \in \mathrm{CI}_{1-\alpha}^{\mathrm{ex}}\right) = F_{|R|}(q_{1-\alpha}),$$

where $F_{|R|}$ denotes the cumulative distribution function of R, which converges to $x \mapsto \max(0, 2\Phi(x)-1)$ by what precedes. Moreover, by Pólya's theorem, the convergence is uniform. The result follows.