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

δ_1 may di er from δ_{ATT}, arguably a more natural target parameter. The two parameters apply to di erent 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 e ect 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 ve mutually exclusive subgroups, detailed in Section 1 of the Web Appendix. Identifying the e ect of the rst treatment in each of those subgroups would require imposing stronger assumptions than Assumption 4. For instance, the rst 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' rst treatment never changes, their rst-treatment's e ect cannot be identi ed under a parallel trends assumption. The second and third subgroups are cells whose rst-treatment's e ect could only be identi ed 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 rst treatment changes while at least one of their other treatment also changes. Identifying their rst-treatment's e ect would require assuming that the e ect of the other treatments is constant between groups, as discussed in Section 3.2.1 (see Equation (10) therein). The fth subgroup are cells whose rst treatment changes while their other treatments do not change, but such that all potential control cells experiencing no treatment change have di erent baseline treatments. Identifying their rst-treatment's e ect would require assuming that the e ects 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 e ect of the rst treatment can be identi ed under a minimal parallel trends assumption and without restricting treatment e ect 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 di erent, non-nested subpopulations, a signi cant di erence 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} di er. On the other hand, under Assumptions 3 and 4, a signi cant di erence between $\widehat{\beta}_{fe}$ and the estimator of δ_1 implies that the e ect 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 $N_{S_1} = 0$, 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.