Two-way fixed effects estimators with heterogeneous treatment effects*

Clément de Chaisemartin[†] Xavier D'Haultfœuille[‡]

August 2, 2018

Abstract

Linear regressions with time and group fixed effects are widely used to estimate treatment effects. We show that they identify weighted sums of the average treatment effects in each group and period, with weights that can be estimated, and that may be negative. In two articles that have used those regressions, half of the weights are negative, thus suggesting that those regressions are often not robust to heterogeneous treatment effects across groups and over time. We propose another estimator that is robust to heterogeneous treatment effects. In an application, it is of a different sign than the linear regression estimator.

Keywords: fixed effects, heterogeneous treatment effects, differences-in-differences.

JEL Codes: C21, C23

*We are very grateful to Guido Imbens, Peter Kuhn, Kyle Meng, Jesse Shapiro, Dick Startz, Doug Steigerwald, members of the UCSB econometrics research group, and seminar participants at Bergen, CIREQ Econometrics conference, CREST, Stanford, and UCSB for their helpful comments. Xavier D'Haultfœuille gratefully acknowledges financial support from the research grants Otelo (ANR-17-CE26-0015-01) and the Labex Ecodec: Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

[†]University of California at Santa Barbara, clementdechaisemartin@ucsb.edu

[‡]CREST, xavier.dhaultfoeuille@ensae.fr

1 Introduction

A popular method to estimate the effect of a treatment D on an outcome Y is to compare the evolution of Y across groups experiencing different evolutions of their exposure to D over time. In practice, this idea is implemented by estimating linear regressions controlling for both group and time fixed effects, which we hereafter refer to as two-way fixed effects regressions. Let $G \in \{0, 1, ..., \overline{g}\}$ denote the group (e.g.: the county) an observation belongs to, and let $T \in \{0, 1, ..., \overline{t}\}$ denote the period when the observation is made. A commonly-used regression is the OLS regression of Y on group fixed effects, period fixed effects, and $E(D|G,T)$, the mean value of the treatment in the group $\times$ period cell the observation belongs to. Hereafter, we refer to this regression as the fixed-effects (FE) regression, and we let β_{fe} denote the coefficient of $E(D|G,T)$. The FE regression has been used both in applications where all observations in the same group $\times$ period cell face the same value of the treatment, in which case $E(D|G,T) = D$, and in applications where the treatment varies within group $\times$ period cells.[1] Another commonly-used regression is the OLS regression of the first difference of the group-level mean of the outcome on period fixed effects, and on the first difference of the group-level mean of the treatment. Hereafter, we refer to this regression as the first-difference (FD) regression, and we let β_{fd} denote the coefficient of the first difference of the group-level mean of the treatment.

We conducted a literature review, and found that nearly 20% of all empirical articles published by the American Economic Review (AER) between 2010 and 2012 use the FD or FE regression, or another similar two-way fixed effects regression, to estimate the effect of a treatment on an outcome. When the treatment effect is constant across groups and over time, it is well-known that such regressions identify the treatment effect under the standard "common trends" assumption. However, it is often implausible that the treatment effect is constant across groups and over time. For instance, the effect of minimum wage on employment may vary across US counties, and may change over time. Perhaps surprisingly, two-way fixed effects regressions have not been analyzed yet without assuming constant effect, except in the special case with two groups and two periods. This is the purpose of this paper.

We start by showing that when the treatment is binary, under the common trends assumption

$$\beta_{fe} = \sum_{g=0}^{\overline{g}} \sum_{t=0}^{\overline{t}} w_{fe,g,t}^{TR} \Delta_{g,t}^{TR}$$

$$\beta_{fd} = \sum_{g=0}^{\overline{g}} \sum_{t=0}^{\overline{t}} w_{fd,g,t}^{TR} \Delta_{g,t}^{TR},$$

[1]Then, using $E(D|G,T)$ rather than D ensures that β_{fe} is solely identified out of comparisons of groups experiencing different evolutions of their average treatment over time, and not out of comparisons of observations with different treatments within the same group $\times$ period, as such comparisons may be plagued by selection bias.

where $\Delta^{TR}_{g,t}$ is the average treatment effect on the treated (ATT) in group g and period t, and $(w^{TR}_{fe,g,t})_{g,t}$ and $(w^{TR}_{fd,g,t})_{g,t}$ are weights summing to one. The weights are not equal to the proportion that each group $\times$ period accounts for in the treated population, so β_{fe} and β_{fd} are not equal to the population ATT. Perhaps more worryingly, many of the weights may be negative. In two applications, we find that half of the weights are negative. In Sections 3.1 and 3.3, we give further intuition as to why negative weights arise, but for now let us just note that β_{fe} and β_{fd} are related to difference-in-differences (DID) estimands, and the negative weights arise from the double-differencing embedded in those estimands. When many weights attached to β_{fe} or β_{fd} are negative, this coefficient is not robust to heterogeneous treatment effects across groups and periods. Even if all the $(\Delta^{TR}_{g,t})_{g,t}$ are positive, this coefficient may be negative. Negative weights would not be an issue if the weights were uncorrelated to treatment effects, but we show that this is unlikely in the two applications we consider.

We also show that a simple function of the weights identifies the minimal value of the standard deviation of the treatment effect across groups and periods under which the ATT in the full population may actually have the opposite sign than β_{fe} or β_{fd}. This lower bound can be used as a summary measure of a coefficient's robustness to heterogeneous treatment effects: the larger it is, the more robust the coefficient is.

As β_{fe} and β_{fd} do not identify an interpretable parameter under the common trends assumption, we consider two supplementary assumptions. The first one, hereafter referred to as the treatment monotonicity assumption, requires that within each group, the treatment of all units evolve in the same direction between each pair of consecutive periods. The second one, hereafter referred to as the stable treatment effect assumption, requires that in each group, the average treatment effect of units treated in period $t-1$ remain stable between $t-1$ and t. We show that under the common trends, treatment monotonicity, and stable treatment effect assumptions, β_{fe} and β_{fd} identify weighted sums of the local average treatment effect (LATE) of the "switchers" in each group and at each period, where "switchers" are units that experience a change in their treatment between two consecutive periods. In the special case with staggered adoption, meaning that groups go from fully untreated to fully treated at heterogeneous dates, the weights attached to β_{fe} and β_{fd} under those three assumptions are all positive. But in more general cases, some of the weights attached to β_{fe} and β_{fd} may still be negative, even under those three assumption.[2]

We also propose a new estimand. This estimand identifies an easily interpretable parameter, the LATE of all switchers in the population. It does not rely on any homogeneous treatment effect assumption. Instead, it relies on a conditional common trends assumption, whose plausibility can be assessed by looking at pre-trends, as in a standard DID. This new estimand can be used when for each pair of consecutive dates, there are groups whose exposure to the treatment does not

[2] Stata dofiles estimating the weights attached to any FE or FD regression under the two sets of assumptions we consider can be found on the authors' webpages.

change between those two dates. In such instances, one can use those groups to identify the effect of time on the outcome, and then separate the effect of time from the treatment effect in groups whose exposure to treatment changes. In one of the two applications we revisit, the corresponding estimator is of a different sign than β_{fe}, and of the same sign as, but significantly different from, β_{fd}. We recommend using our new estimand rather than β_{fe} or β_{fd}, whenever there are groups whose exposure to the treatment does not change between each pair of consecutive dates: this estimand identifies an interpretable parameter, it is robust to heterogeneous treatment effects, and it relies on an assumption whose plausibility can be assessed through placebo tests.[3]

We derive our decomposition of β_{fe} and β_{fd} as weighted sums of ATTs or LATEs with a binary treatment variable and without covariates in the regression. Nevertheless, we show that this decomposition can still be obtained with a non-binary, ordered treatment, and with covariates in the regression. Interestingly, the weights remain unchanged in these two cases. Moreover, researchers have also estimated two-stage-least-squares (2SLS) versions of the FE and FD regressions we consider, and we show that our main conclusions also apply to those regressions. Therefore, our results are widely applicable in practice.

Our paper is related to the DID literature. Our main result generalizes Theorem 1 in de Chaisemartin and D'Haultfœuille (2018). When the data only bears two groups and two periods, the Wald-DID estimand considered therein is equal to β_{fe} and β_{fd}. Our main result is thus an extension of that theorem to multiple periods and groups.[4] Moreover, the Wald-TC estimand we propose in this paper is closely related to, but improves upon, the Wald-TC estimand with many groups and periods proposed in de Chaisemartin and D'Haultfœuille (2018). In Section 4, we give more details on the differences between those two Wald-TC estimands.

More recently, several papers study the special case with staggered adoption, and derive some interesting results specific to that design that we do not consider here. Still, some of their results are related to ours. Borusyak and Jaravel (2017), Abraham and Sun (2018), and Goodman-Bacon (2018) show that if the common trends assumption holds, β_{fe} identifies a weighted sum of treatment effects, with possibly some negative weights. These results can be obtained from the first point of Theorem 1 below. Abraham and Sun (2018) and Callaway and Sant'Anna (2018) propose estimands that identify the instantaneous effect of the treatment, as well as its dynamic effects. Actually, with staggered adoption, our Wald-TC estimand is equal to the estimand of the instantaneous effect proposed by Abraham and Sun (2018), and closely related to that proposed by Callaway and Sant'Anna (2018). In Section 5, we also discuss how the Wald-TC estimand can be used to identify dynamic effects.

With respect to these papers, we make several contributions. For instance, we study both the

[3]The corresponding estimator is computed by the Stata package `fuzzydid`, see de Chaisemartin et al. (2018).

[4]In fact, a preliminary version of our main result appeared in a working paper version of de Chaisemartin and D'Haultfœuille (2018) (see Theorems S1 and S2 in de Chaisemartin and D'Haultfoeuille, 2015).

fixed effects and the first difference regressions, while they focus on the former. In addition, we derive a useful summary statistic of a coefficient's robustness to treatment effect heterogeneity: the minimal value of the standard deviation of the treatment effect across groups and periods under which the ATT may have a different sign than that coefficient. But most importantly, our results apply to all fixed effects and first difference regressions, not only to those with staggered adoption. Less than 10% of the papers estimating two-way fixed effects regressions and published in the AER between 2010 and 2012 have a staggered adoption design. This suggests that while staggered adoptions are an important research design, they may account for a relatively small minority of the applications where two-way fixed effects regressions have been used. Thus, our results illuminate the implications of treatment-effect heterogeneity between groups or over time in a broader class of applications.

The paper is organized as follows. Section 2 presents the regressions we consider, our assumptions, and our parameters of interest. Section 3 presents our main results, considering first the special case with two groups and two periods to build intuition. Section 4 presents our alternative estimand. Section 5 extends our results to non-binary treatments, regressions with covariates, 2SLS regressions, and models allowing for dynamic treatment effects. Section 6 discusses inference. Section 7 presents our literature review of the articles published in the AER, and the results of our two empirical applications.

2 Two-way FE regressions, assumptions, and parameters of interest

Assume that one is interested in measuring the effect of a treatment D on some outcome Y. We start by assuming that D is binary, but our results apply to any ordered treatment, as we show in Subsection 5.1. Then, let $Y(0)$ and $Y(1)$ denote the two potential outcomes of the same observation without and with the treatment. The observed outcome is $Y = Y(D)$. We assume that two variables $G \in \{0, 1, ..., \overline{g}\}$ and $T \in \{0, 1, ..., \overline{t}\}$ are attached to each observation. G denotes the group (e.g. the county) the observed unit belongs to, and T denotes the time period when the observation is made. Each group $\times$ period cell may bear several or a single observation: the case where each group bears a single unit is a special case of our framework.

Our first assumption requires that all group $\times$ period cells are non-empty, meaning that no group appears or disappears over time. This assumption is often satisfied. Without it, our results still hold but the notation becomes more complicated as some of the objects considered below are no longer well-defined. Therefore, we just make this assumption to simplify our notation.

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

We now introduce some notation we use throughout the paper. For any random variable R and for every $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, let $R_{g,.}$, $R_{.,t}$, and $R_{g,t}$ respectively be random variables such that $R_{g,.} \sim R|G=g$, $R_{.,t} \sim R|T=t$, and $R_{g,t} \sim R|G=g, T=t$, where $\sim$ denotes equality in distribution. Finally, let $FD_R(g,t) = E(R_{g,t}) - E(R_{g,t-1})$.

We now formally define the FE and FD regressions described in the introduction.

Regression 1 *(Fixed-effects regression)*

Let β_{fe} denote the coefficient of $E(D|G,T)$ in an OLS regression of Y on a constant, $(1\{G=g\})_{1\leq g\leq \overline{g}}$, $(1\{T=t\})_{1\leq t\leq \overline{t}}$, and $E(D|G,T)$.

Regression 2 *(First-difference regression)*

Let β_{fd} denote the coefficient of $FD_D(G,T)$ in an OLS regression of $FD_Y(G,T)$ on a constant, $(1\{T=t\})_{2\leq t\leq \overline{t}}$, and $FD_D(G,T)$, conditional on $T \geq 1$.

The FE regression is very pervasive: 13 articles published in the AER between 2010 and 2012 have estimated it. Other articles have estimated regressions similar to it, e.g. with two treatment variables instead of one. The FD regression is also very pervasive: six articles published in the AER between 2010 and 2012 have estimated it, and other articles have estimated regressions similar to it. When $\overline{t} = 1$ and $T \perp\!\!\!\perp G$, meaning that groups' distribution remains stable over time, one can show using the Frisch-Waugh theorem that $\beta_{fe} = \beta_{fd}$. β_{fe} differs from β_{fd} if $\overline{t} > 1$ or T is not independent from G.

We now introduce the main assumptions we consider.

Assumption 2 *(Common trends) For all $t \in \{1,...,\overline{t}\}$, $E(Y(0)|G,T=t) - E(Y(0)|G,T=t-1)$ does not depend on G.*

The common trends assumption requires that the mean of $Y(0)$ follow the same evolution over time in every group. This assumption is sufficient to have that the DID estimand identifies the ATT in standard two groups two periods sharp designs (see, e.g., Abadie, 2005). However, this assumption is not sufficient to have that β_{fe} and β_{fd} identify an interpretable treatment effect parameter in more general designs. We therefore consider two supplementary assumptions.

Assumption 3 *(Treatment monotonicity) There exist random variables $D(0),...,D(\overline{t})$ such that:*

1. *$D = D(T)$;*
2. *For all $t \in \{1,...,\overline{t}\}$, $D(t) \perp\!\!\!\perp T|G$;*
3. *For all $t \in \{1,...,\overline{t}\}$, $P(D(t) \geq D(t-1)|G) = 1$ or $P(D(t) \leq D(t-1)|G) = 1$.*

The treatment monotonicity assumption requires that each unit has $\bar{t}+1$ variables $D(0)$, $D(1)$, ..., $D(\bar{t})$ attached to her, which respectively denote her treatment at T =0, 1, ..., $\bar{t}$. It also requires that the distribution of those variables be stable across periods in each group. Finally, it implies that between each pair of consecutive periods, in a given group there cannot be both units whose treatment increases and units whose treatment decreases.

The treatment monotonicity assumption automatically holds in applications where the treatment is constant within each group $\times$ period, as is for instance the case in Gentzkow et al. (2011), where the treatment is the number of newspapers in county g and election year t. Then, all units of the same group have the same treatment in period t, so $D(t)$ is constant within group and $D(t) \perp\!\!\!\perp T|G$. Moreover, $P(D(t) \geq D(t-1)|G) = 1$ in groups where the treatment increases between $t-1$ and t, and $P(D(t) \leq D(t-1)|G) = 1$ in groups where it decreases. The treatment monotonicity assumption also holds in cases where being treated is an absorbing state.

On the other hand, the treatment monotonicity assumption may fail in applications where the treatment varies within each group $\times$ period. For instance, Point 3 will fail to hold if there are groups where the treatment of some units diminishes between $t-1$ and t, while the treatment of other units increases. When the FE or the FD regression is estimated with individual-level panel data, the variables $(D(t))_{0\leq t\leq \bar{t}}$ are observed: they are just the treatments of each unit at each period. Then, one can assess from the data whether Point 3 holds or not.[5] On the other hand, when the FE or the FD regression is estimated with individual-level repeated cross-sections, only $D(T)$ is observed so Point 3 is not testable.

Assumption 4 *(Stable treatment effect) For all* $(g,t) \in \{0,...,\overline{g}\} \times \{1,...,\bar{t}\}$,
$E(Y(1)-Y(0)|G=g,T=t,D(t-1)=1) = E(Y(1)-Y(0)|G=g,T=t-1,D(t-1)=1)$.

The stable treatment effect assumption requires that in every group, the average treatment effect among units treated in period $t-1$ does not change between $t-1$ and t. This assumption restricts treatment effect heterogeneity over time, but not across groups. It was first introduced in de Chaisemartin and D'Haultfœuille (2018). It may be implausible, but it is necessary for some of our identification results concerning β_{fe} and β_{fd}.

Our two parameters of interest are

$$\Delta^{TR} = E(Y(1)-Y(0)|D=1),$$
$$\Delta^{S} = E\left(Y(1)-Y(0)|S\right),$$

where $S = \{D(T-1) \neq D(T), T \geq 1\}$ denotes units whose treatment status switches between $T-1$ and T. Δ^{TR} is the Average Treatment effect on the Treated (ATT), and Δ^{S} is the Local

[5]This test may reveal that Point 3 fails. However, our results still hold if the treatment variables satisfy the threshold crossing Equation (3.2) in de Chaisemartin and D'Haultfœuille (2018), which is weaker than Point 3.

Average Treatment Effect (LATE) of the switchers. For any $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, let $\Delta^{TR}_{g,t} = E(Y(1)-Y(0)|D=1, G=g, T=t)$ denote the ATT in group g and at period t. For any $(g,t) \in \{0,...,\overline{g}\} \times \{1,...,\overline{t}\}$, let $\Delta^{S}_{g,t} = E(Y(1)-Y(0)|S, G=g, T=t)$ denote the LATE of switchers in group g and at period t. Then $\Delta^{TR} = E[\Delta^{TR}_{G,T}|D=1]$ and $\Delta^{S} = E[\Delta^{S}_{G,T}|S]$: Δ^{TR} and Δ^{S} are respectively equal to weighted averages of the $(\Delta^{TR}_{g,t})_{g,t}$ and $(\Delta^{S}_{g,t})_{g,t}$.

Throughout the paper, we assume that groups experience common trends, but that the effect of the treatment may be heterogeneous between groups and / or over time. We discuss in Appendix B.1 two examples where this may happen, and we argue that the mechanisms behind these examples are fairly general.

Also, when the treatment is constant within each group $\times$ period, $\Delta^{S}_{g,t} = E(Y(1)-Y(0)|G = g, T=t)$ in groups whose treatment changes between $t-1$ and t, because everybody's treatment switches between $t-1$ and t in those groups. But even then, Δ^S is still a "local" measure of the treatment effect, because it only factors in the period t average treatment effect of groups switching treatment between $t-1$ and t, not that of groups whose treatment does not change.

3 Main results

3.1 The special case with two groups and periods

To introduce our main result, let us start by considering the special case where the population only bears two groups and two periods, which we studied in de Chaisemartin and D'Haultfœuille (2018). In that case, we have, provided that $E(D_{1,1}) - E(D_{1,0}) \neq E(D_{0,1}) - E(D_{0,0})$,[6]

$$\beta_{fe} = \beta_{fd} = \frac{E(Y_{1,1}) - E(Y_{1,0}) - E(Y_{0,1}) + E(Y_{0,0})}{E(D_{1,1}) - E(D_{1,0}) - E(D_{0,1}) + E(D_{0,0})}. \tag{1}$$

The right-hand side is the DID comparing the evolution of the mean outcome from period 0 to 1 in groups 0 and 1, divided by the DID comparing the evolution of the mean treatment in those two groups. We let DID_D denote the denominator of this ratio.

Let us first assume that $D = G \times T$: all observations in group 1 and period 1 receive the treatment and no other observation receives it, a case we hereafter refer to as a "sharp" DID. Then, $\beta_{fe} = \beta_{fd} = E(Y_{1,1}) - E(Y_{1,0}) - E(Y_{0,1}) + E(Y_{0,0})$, and one can show that if the common trends assumption holds,

$$\beta_{fe} = \beta_{fd} = \Delta^{TR}_{1,1}.$$

The common trends assumption requires that if all observations had remained untreated, the mean outcome would have followed parallel trends in groups 0 and 1. In a sharp DID, the only

[6] If this condition does not hold, the regressors in the FE and FD regressions are collinear.

departure from the scenario where nobody is treated is that observations in group 1 and period 1 receive the treatment. Therefore, any discrepancy between the trends of the mean outcome in the two groups must come from the effect of the treatment in group 1 and period 1, so β_{fe} and β_{fd} identify the ATT in group 1 and period 1.

Now, let us assume that $D \neq G \times T$ but $DID_D > 0$,[7] a case we hereafter refer to as a "fuzzy" DID. Then, there may be treated observations in each of the four group $\times$ period cells. One can show, using (1) and Point 1 of Lemma 2 in the appendix, that under the common trends assumption, β_{fe} and β_{fd} identify a weighted sum of the four ATTs in each group and time periods, where two ATTs enter with negative weights:

$$\beta_{fe} = \beta_{fd} = \frac{E(D_{1,1})}{DID_D}\Delta^{TR}_{1,1} - \frac{E(D_{1,0})}{DID_D}\Delta^{TR}_{1,0} - \frac{E(D_{0,1})}{DID_D}\Delta^{TR}_{0,1} + \frac{E(D_{0,0})}{DID_D}\Delta^{TR}_{0,0}. \tag{2}$$

The intuition for Equation (2) goes as follows. In a fuzzy DID there are potentially four departures from the scenario where nobody is treated: some observations may be treated in group 1 and period 1, in group 1 and period 0, etc. Therefore, the discrepancy between the outcome trends in the two groups can come from the treatment effect in any group and time period. The negative weights arise from the double-differencing embedded in the DID estimand: the average treatment effect on the treated in group 1 $\times$ period 0 and group 0 $\times$ period 0 enter with a negative sign in (2) because the average outcomes in these cells enter with a negative sign in (1).

Because of the negative weights, β_{fe} and β_{fd} could for instance be negative while all the $(\Delta^{TR}_{g,t})_{g,t}$ are strictly positive. For instance, if $E(D_{1,0}) = 0.3$, $E(D_{1,1}) = 0.6$, $E(D_{0,1}) = E(D_{0,0})$, $\Delta^{TR}_{1,0} = 3$, $\Delta^{TR}_{1,1} = 1$, and $\Delta^{TR}_{0,1} = \Delta^{TR}_{0,0} > 0$, then it follows from Equation (2) that $\beta_{fe} = \beta_{fd} = -1$.

If we further impose the treatment monotonicity and stable treatment effect assumptions, one can show, using (1) and Point 2 of Lemma 2, that β_{fe} and β_{fd} identify a weighted sum of switchers' LATEs in groups 0 and 1:

$$\beta_{fe} = \beta_{fd} = \frac{E(D_{1,1}) - E(D_{1,0})}{DID_D}\Delta^{S}_{1,1} - \frac{E(D_{0,1}) - E(D_{0,0})}{DID_D}\Delta^{S}_{0,1}. \tag{3}$$

Let us give the intuition underlying Equation (3) in the case where in both groups the treatment rate increases from period 0 to 1. Under the treatment monotonicity assumption, for $g \in \{0, 1\}$ $\Delta^{TR}_{g,1}$ can be decomposed into the period 1 LATE of units already treated in period 0, and the period 1 LATE of units switching into treatment. Under the stable treatment effect assumption, the period 1 LATE of units already treated in period 0 is equal to their period 0 LATE, $\Delta^{TR}_{g,0}$. Thus, those terms cancel out in Equation (2), and we are left with a weighted sum of switchers' LATEs in both groups. When the share of treated observations increases in one group but decreases in the other, then both LATEs in Equation (3) enter with a positive weight, thus

[7] The results below hold whenever $DID_D \neq 0$. But assuming $DID_D > 0$ simplifies the discussion.

illustrating that imposing Assumption 4 can eliminate the negative weights. On the other hand, when the share of treated observations increases or decreases in both groups, then one of the two LATEs in Equation (3) enters with a negative weight. Then again, because of the negative weights, β_{fe} and β_{fd} could for instance be negative while $\Delta^S_{1,1}$ and $\Delta^S_{0,1}$ are both strictly positive.

3.2 Two-way fixed effects estimands as weighted sums of average treatment effects

Our main result extends Equations (2) and (3) to the case where the population bears more than two groups and periods. Then, we show that under the common trends assumption (resp. the common trends, treatment monotonicity, and stable treatment effect assumptions), β_{fe} and β_{fd} still identify weighted sums of $(\Delta^{TR}_{g,t})_{g,t}$ (resp. $(\Delta^S_{g,t})_{g,t}$), with weights that can be negative, as in the two-groups and two-periods case.

We start by defining the weights attached to $(\Delta^{TR}_{g,t})_{g,t}$ in the FE and FD regressions under the common trends assumption. First, for any $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, let $\varepsilon_{fe,g,t}$ denote the residual of observations in group g and at period t in the regression of $E(D|G,T)$ on a constant, $(1\{G=g\})_{1\leq g\leq \overline{g}}$, and $(1\{T=t\})_{1\leq t\leq \overline{t}}$. Then, provided that $E[E(D|G,T)\varepsilon_{fe,G,T}] \neq 0$, let

$$v^{TR}_{fe,g,t} = \frac{E(D)\varepsilon_{fe,g,t}}{E[E(D|G,T)\varepsilon_{fe,G,T}]}, \quad w^{TR}_{fe,g,t} = \frac{E(D1\{G=g,T=t\})}{E(D)} v^{TR}_{fe,g,t}.$$

Second, for any $(g,t) \in \{0,...,\overline{g}\} \times \{1,...,\overline{t}\}$, let $\varepsilon_{fd,g,t}$ denote the residual of observations in group g and at period t in the regression of $FD_D(G,T)$ on a constant and $(1\{T=t\})_{2\leq t\leq \overline{t}}$, conditional on $T \geq 1$. For any $g \in \{0,...,\overline{g}\}$, let also $\varepsilon_{fd,g,0} = 0$, $\varepsilon_{fd,g,\overline{t}+1} = 0$. Then, for any $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, let

$$\widetilde{v}^{TR}_{fd,g,t} = \varepsilon_{fd,g,t} - \frac{P(G=g,T=t+1)}{P(G=g,T=t)}\varepsilon_{fd,g,t+1}.$$

Finally, provided that $E[\widetilde{v}^{TR}_{fd,G,T}E(D|G,T)] \neq 0$, let

$$v^{TR}_{fd,g,t} = \frac{E(D)\widetilde{v}^{TR}_{fd,g,t}}{E[E(D|G,T)\widetilde{v}^{TR}_{fd,G,T}]}, \quad w^{TR}_{fd,g,t} = \frac{E(D1\{G=g,T=t\})}{E(D)} v^{TR}_{fd,g,t}.$$

$w^{TR}_{fe,g,t}$ and $w^{TR}_{fd,g,t}$ are the weights attached to $\Delta^{TR}_{g,t}$ in the FE and FD regressions under the common trends assumption.

We then define the weights attached to $(\Delta^S_{g,t})_{g,t}$ in the two regressions under the common trends, treatment monotonicity, and stable treatment effect assumptions. First, for all $(g,t) \in \{0,...,\overline{g}\}\times \{1,...,\overline{t}\}$, let

$$s_{g,t} = \text{sgn}\left(E(D_{g,t}) - E(D_{g,t-1})\right),$$

where for any real number x, $\text{sgn}(x) = 1\{x > 0\} - 1\{x < 0\}$. $s_{g,t}$ is equal to 1 (resp. -1, 0) for groups where the share of treated observations increases (resp. decreases, does not change) between $t-1$ and t. Let also, for all $(g,t) \in \{0,...,\overline{g}\} \times \{1,...,\overline{t}\}$,

$$\widetilde{v}^S_{fe,g,t} = \frac{s_{g,t}E[1\{G=g, T \geq t\}\varepsilon_{fe,G,T}]}{P(G=g,T=t)}.$$

Finally, provided that $E[\widetilde{v}^S_{fe,G,T}|S] \neq 0$, let

$$v^S_{fe,g,t} = \frac{\widetilde{v}^S_{fe,g,t}}{E[\widetilde{v}^S_{fe,G,T}|S]}, \quad w^S_{fe,g,t} = P(G=g,T=t|S)v^S_{fe,g,t}.$$

Second, provided that $E[s_{G,T}\varepsilon_{k,G,T}|S] \neq 0$, let us define, for all $(g,t) \in \{0,...,\overline{g}\} \times \{1,...,\overline{t}\}$,

$$v^S_{fd,g,t} = \frac{s_{g,t}\varepsilon_{fd,g,t}}{E[s_{G,T}\varepsilon_{fd,G,T}|S]}, \quad w^S_{fd,g,t} = P(G=g,T=t|S)v^S_{fd,g,t}.$$

$w^S_{fe,g,t}$ and $w^S_{fd,g,t}$ are the weights attached to $\Delta^S_{g,t}$ in the FE and FD regressions under the common trends, treatment monotonicity, and stable treatment effect assumptions.

Importantly, for $k \in \{fe, fd\}$, $(v^{TR}_{k,g,t})_{g,t}$ and $(v^S_{k,g,t})_{g,t}$ are identified and can easily be estimated. This is obvious for $(v^{TR}_{k,g,t})_{g,t}$, but less so for $(v^S_{k,g,t})_{g,t}$, as their denominators involve expectations over switchers, a population that is generally not identified. Nonetheless, we show in the appendix that under the treatment monotonicity assumption, for any function f, $E(f(G,T)|S)$ is identified by

$$E(f(G,T)|S) = \frac{\sum_{(g,t):t\geq 1} P(G=g,T=t)\,|E(D_{g,t}) - E(D_{g,t-1})|\, f(g,t)}{\sum_{(g,t):t\geq 1} P(G=g,T=t)\,|E(D_{g,t}) - E(D_{g,t-1})|}. \tag{4}$$

We can now state our main result. We say below that β_k (resp. the $(v^{TR}_{k,g,t})_{g,t}$, the $(v^S_{k,g,t})_{g,t}$) is well-defined whenever there is a unique solution to the linear system corresponding to the regression attached to β_k (resp. when the denominator in their definition is not zero).

Theorem 1 *Suppose that Assumption 1 holds, D is binary, $k \in \{fe, fd\}$ and β_k is well defined.*

1. *If Assumption 2 holds, then the $(v^{TR}_{k,g,t})_{g,t}$ are well defined and*

$$\beta_k = E\left[v^{TR}_{k,G,T}\Delta^{TR}_{G,T}|D=1\right].$$

2. *If Assumptions 2-4 hold, then the $(v^S_{k,g,t})_{g,t}$ are well defined and*

$$\beta_k = E\left[v^S_{k,G,T}\Delta^S_{G,T}|S\right].$$

In Theorem 1, we show that under Assumption 2 (resp. Assumptions 2-4), β_{fe} and β_{fd} identify weighted sums of the ATTs (resp. of switchers' LATEs) in each group and at each period. Indeed, the displayed equations in the first and second statements of the theorem are respectively equivalent to

$$\beta_k = \sum_{g=0}^{\overline{g}} \sum_{t=0}^{\overline{t}} w_{k,g,t}^{TR} \Delta_{g,t}^{TR} \tag{5}$$

$$\beta_k = \sum_{g=0}^{\overline{g}} \sum_{t=1}^{\overline{t}} w_{k,g,t}^{S} \Delta_{g,t}^{S}. \tag{6}$$

The weights $(w_{k,g,t}^{TR})_{g,t}$ differ from $(P(G=g,T=t|D=1))_{g,t}$, so under Assumption 2, β_k does not identify the ATT in the full population. Similarly, under Assumptions 2-4 β_k does not identify the LATE of all switchers. Perhaps more worryingly, some of the weights $(w_{k,g,t}^{TR})_{g,t}$ may be strictly negative. Then, β_k does not satisfy the no-sign-reversal property under the common trends assumption: β_k may for instance be negative while the treatment effect is positive for everybody in the population. Similarly, β_k may not satisfy the no-sign-reversal property under the common trends, treatment monotonicity, and stable treatment effect assumptions.

3.3 Conditions under which some weights are strictly negative

We first present a necessary and sufficient condition to have that some of the weights $(w_{fe,g,t}^{TR})_{g,t}$ are strictly negative. One can show that $v_{fe,g,t}^{TR} = \varepsilon_{fe,g,t}E(D)/V(\varepsilon_{fe,G,T})$. Because $\varepsilon_{fe,g,t}$ is a regression residual, $E(\varepsilon_{fe,G,T}) = 0$. Therefore, there must be some values of (g,t) for which $\varepsilon_{fe,g,t} < 0$.[8] Then, some of the $(v_{fe,g,t}^{TR})_{g,t}$ must be negative. Accordingly, the only instance where all the weights $(w_{fe,g,t}^{TR})_{g,t}$ are positive is when all the (g,t) such that $\varepsilon_{fe,g,t} < 0$, meaning that $E(D_{g,t})$ is lower than its predicted value in a linear regression with group and time fixed effects, are also such that $E(D_{g,t}) = 0$, thus implying that $P(G=g,T=t|D=1) = 0$. This condition is for instance satisfied in sharp DIDs with two groups and two periods. But it is unlikely to hold more generally. For instance, when $E(D_{g,t}) > 0$ for all (g,t), some of the weights $(w_{fe,g,t}^{TR})_{g,t}$ must be negative. Similarly, one can show, using $E(\tilde{w}_{2,G,T}) = 0$, that some of the $(v_{fd,g,t}^{TR})_{g,t}$ must be negative. Here as well, the only instance where all the weights $(w_{fd,g,t}^{TR})_{g,t}$ are positive is when all the (g,t) such that $v_{fd,g,t}^{TR} < 0$ are also such that $E(D_{g,t}) = 0$.

We then present a necessary and sufficient condition to have that some of the weights $(w_{fd,g,t}^{S})_{g,t}$ are strictly negative. First, one can show that for any $(g,t) \in \{0,...,\overline{g}\} \times \{1,...,\overline{t}\}$, $\varepsilon_{fd,g,t} = E(D_{g,t}) - E(D_{g,t-1}) - (E(D_{.,t}) - E(D_{.,t-1}))$. Moreover, under Assumptions 2-4, β_{fd} well defined implies $E(s_{G,T}\varepsilon_{fd,G,T}|S) > 0$. Therefore, one can show that the $(w_{fd,g,t}^{S})_{g,t}$ are all positive if and only if for all $t \geq 1$:

[8] Otherwise we would have $\varepsilon_{fe,g,t} = 0$ for all (g,t) and β_{fe} would not be well defined.

1. in all the groups where the share of treated observations strictly increases between $t-1$ and t (i.e. $E(D_{g,t}) - E(D_{g,t-1}) > 0$), that increase is higher than $E(D_{.,t}) - E(D_{.,t-1})$, the evolution of the treatment rate in the full population between $t-1$ and t.

2. in all the groups where the share of treated observations strictly decreases between $t-1$ and t (i.e. $E(D_{g,t}) - E(D_{g,t-1}) < 0$), that decrease is lower than $E(D_{.,t}) - E(D_{.,t-1})$.

Those two conditions will for instance hold when the treatment is constant within each group $\times$ period cell. Then, groups where the treatment increases have $E(D_{g,t}) - E(D_{g,t-1}) = 1$, while groups where the treatment decreases have $E(D_{g,t}) - E(D_{g,t-1}) = -1$. On the other hand, that condition will for instance fail if for some t $E(D_{g,t}) - E(D_{g,t-1})$ is strictly positive for all g, but not constant across g.

We now consider two pervasive research designs in which the necessary and sufficient conditions to have that some of the weights are strictly negative have a simpler expression. First, we consider the staggered adoption design.

Assumption 5 *(Staggered adoption)* $D = 1\{T \geq a_G\}$ *for some* $a_G \in \{0, ..., \overline{t}, \overline{t}+1\}$.

Assumption 6 *(Group $\times$ period regression with a balanced panel of groups) For every* $(g,t) \in \{0, ..., \overline{g}\} \times \{0, ..., \overline{t}\}$, $P(G = g, T = t) = \frac{1}{\overline{g}+1}\frac{1}{\overline{t}+1}$.

Assumption 5 is satisfied when each group is either fully untreated at each period, fully treated at each period, or fully untreated until a period $a_g - 1$ and fully treated from period a_g onwards. Groups g with $a_g = \overline{t}+1$ never adopt the treatment. This type of staggered adoption design is often met in practice, see e.g. Athey and Stern (2002). Hereafter, we refer to $g(e) = \text{argmin}_{g \in \{0,...,\overline{g}\}} a_g$ as the earliest adopting group. Assumption 6 is for instance satisfied when the FE or FD regression is estimated with only one observation per group and period, and when no group appears or disappears over time. It is necessary to obtain some, but not all the results in Proposition 1 below.

Proposition 1 *Suppose that Assumption 1 and 5 hold, and that β_{fe} and β_{fd} are well defined. Then,*

1. *If Assumption 6 also holds,* $P(v^{TR}_{fe,G,T} \geq 0 | D = 1) = 1$ *if and only if*

$$\frac{a_{g(e)}}{\overline{t}+1} \geq E(D_{.,\overline{t}}) - E(D).$$

2. *If Assumption 6 also holds,* $P(v^{TR}_{fd,G,T} \geq 0 | D = 1) = 1$ *if and only if* $a_g \geq \overline{t}$ *for every* $g \in \{0, ..., \overline{g}\}$.

3. *If Assumption 6 also holds,* $P(v^{S}_{fe,G,T} \geq 0 | S) = 1$.

4. $P(v^S_{fd,G,T} \geq 0|S) = 1$.

The first point of Proposition 1 shows that in the staggered-adoption research design, the weights $(w^{TR}_{fe,g,t})_{g,t}$ are all positive if and only if the proportion of periods during which the earliest adopting group remains untreated is larger than the difference between the proportion of groups treated at $\overline{t}$ and the average of that proportion across all periods. This condition is rarely met. For instance, it fails when at least one group is treated at $T = 0$,[9] or when all groups are treated in the last period.[10] The second point of Proposition 1 then shows that in this research design, the weights $(w^{TR}_{fd,g,t})_{g,t}$ are all positive if and only if every group either remains untreated or becomes treated at the last period. This condition is also rarely met. The reason why negative weights arise in the staggered adoption design is that β_{fe} and β_{fd} are weighted averages of DID estimands, some of which use treated groups as controls. Following Equation (2), the treatment effect in those treated control groups gets differenced out in those DID estimands, hence the negative weights.

On the other hand, the third and fourth points of Proposition 1 show that in the staggered adoption research design, the weights $(w^S_{fe,g,t})_{g,t}$ and $(w^S_{fd,g,t})_{g,t}$ are always all positive. Contrary to the third, the fourth point of the proposition does not rely on Assumption 6.

Overall, in staggered adoption designs, the common trends assumption is not sufficient for β_{fe} and β_{fd} to identify convex combinations of the treatment effect in each group and period. On the other hand, if the common trends and stable treatment effect assumptions hold, these estimands identify a convex combination of the LATEs of switchers in each group and period.[11] Still, that convex combination is not equal to the LATE of all switchers.

The second design we consider is the heterogeneous adoption design.

Assumption 7 *(Heterogenous adoption)* $\overline{t} = 1$*, D is binary, and for every* $g \in \{0, ..., \overline{g}\}$*,* $E(D_{g,0}) = 0$ *and* $E(D_{g,1}) > 0$*.*

Assumption 7 is satisfied in applications with two time periods, where all groups are fully untreated at $T = 0$, and where all groups become at least partly treated at $T = 1$. This type of heterogeneous adoption design is often met in practice, for instance when an innovation is introduced and heterogeneously adopted by various groups (see, e.g., Enikolopov et al., 2011).

Proposition 2 *Suppose that Assumptions 1 and 7 hold, and that for every* $k \in \{fe, fd\}$*,* β_k *is well defined. Then,* $P(v^{TR}_{k,G,T} < 0|D = 1) > 0$ *and* $P(v^S_{k,G,T} < 0|S) > 0$*.*

[9] We then have $a_{g(e)}/(\overline{t} + 1) = 0$, while $E(D_{.,\overline{t}}) - E(D) > 0$, otherwise β_{fe} would not be well defined.

[10] One can show that $E(D) < E(D_{.,\overline{t}})\left(1 - \frac{a_{g(e)}}{\overline{t}+1}\right)$, so the condition in point 1 is violated when $E(D_{.,\overline{t}}) = 1$.

[11] Under Assumption 5, Assumption 3 automatically holds.

Proposition 2 shows that in the heterogeneous adoption design, for every $k \in \{fe, fd\}$ some of the weights $(w^{TR}_{k,g,t})_{g,t}$ and $(w^{S}_{k,g,t})_{g,t}$ are strictly negative. One can show that in this design, β_{fe} is a weighted average of Wald-DID estimands comparing the evolution of the outcome in group g to that in group $g-1$ (see point 1 of Theorem S1 in de Chaisemartin and D'Haultfoeuille, 2015). Following Equation (2), the treatment effect in group $g-1$ and period 1 gets differenced out in those DID estimands, hence the negative weights.

3.4 Robustness of double fixed effects estimands to heterogeneous treatment effects

When the treatment effect is heterogeneous across groups and periods, β_k may be a misleading measure of Δ^{TR} (resp. Δ^S). In the corollary below, we derive the minimum amount of heterogeneity of the $(\Delta^{TR}_{g,t})_{g,t}$ (resp. $(\Delta^{S}_{g,t})_{g,t}$) that could actually lead β_k to be of a different sign than Δ^{TR} (resp. Δ^S). Let $\sigma^{TR} = V\left(\Delta^{TR}_{G,T}|D=1\right)^{1/2}$ (resp. $\sigma^{S} = V\left(\Delta^{S}_{G,T}|S\right)^{1/2}$) denote the standard deviation of the ATTs (resp. of switchers' LATEs) across groups and periods.

Corollary 1 *Suppose that Assumption 1 holds, D is binary, $k \in \{fe, fd\}$ and β_k is well defined.*

1. *Suppose that Assumption 2 holds and $V(v^{TR}_{k,G,T}|D=1) > 0$. Then, the minimal value of σ^{TR} compatible with β_k and $\Delta^{TR} = 0$ is*

$$\underline{\sigma}^{TR}_k = \frac{|\beta_k|}{V\left(v^{TR}_{k,G,T}|D=1\right)^{1/2}}.$$

2. *Suppose that Assumptions 2-4 hold and $V(v^{S}_{k,G,T}|S) > 0$. Then, the minimal value of σ^S compatible with β_k and $\Delta^S = 0$ is*

$$\underline{\sigma}^{S}_k = \frac{|\beta_k|}{V\left(v^{S}_{k,G,T}|S\right)^{1/2}}.$$

Estimators of $\underline{\sigma}^{TR}_k$ and $\underline{\sigma}^{S}_k$ can be used to assess the sensitivity of β_k to treatment effect heterogeneity across groups and periods. Assume for instance that β_{fe} is large and positive, while $\underline{\sigma}^{TR}_{fe}$ is close to 0. Then, even under a small amount of treatment effect heterogeneity, β_{fe} could be of a different sign than Δ^{TR}. Indeed, the data is compatible with β_{fe} large and positive, $\Delta^{TR} = 0$, and a small dispersion of the $(\Delta^{TR}_{g,t})_{g,t}$. Estimators of $\underline{\sigma}^{TR}_{fe}$ and $\underline{\sigma}^{TR}_{fd}$ (resp. $\underline{\sigma}^{S}_{fe}$ and $\underline{\sigma}^{S}_{fd}$) can also be used to determine which of β_{fe} or β_{fd} is more robust to heterogeneous treatment effects in a given application. Finally, note that Corollary 1 applies beyond two-way fixed effects regressions. For any estimand that identifies a weighted sum of conditional average treatment effects,[12] one can follow the same steps as those we use in our proof to derive a lower bound on

[12]Many estimands satisfy that property. For instance, Angrist (1998) shows that the coefficient of D in an OLS regression of Y on D and some covariates X identifies a weighted average of conditional average treatment effects under the conditional independence assumption, if the regression is saturated in X.

the amount of treatment effect heterogeneity under which that estimand could be of a different sign than the unconditional average treatment effect.

The results above show that β_{fe} and β_{fd} may not be robust to heterogeneous treatment effects. We now provide conditions under which they identify Δ^{TR} or Δ^S despite such heterogeneity.

Assumption 8_k *(Random weights or homogeneous ATTs) cov* $\left(v^{TR}_{k,G,T}, \Delta^{TR}_{G,T}|D=1\right)=0.$

Assumption 9_k *(Random weights or homogeneous LATEs) cov* $\left(v^{S}_{k,G,T}, \Delta^{S}_{G,T}|S\right)=0.$

Assumptions 8_k and 9_k are indexed by $k \in \{fe, fd\}$, because the assumption one needs to invoke for identification depends on whether one considers β_{fe} or β_{fd}.

Corollary 2 *Suppose that Assumption 1 holds, D is binary, $k \in \{fe, fd\}$ and β_k is well defined.*

1. *If Assumptions 2 and 8_k hold, $\beta_k = \Delta^{TR}$.*
2. *If Assumptions 2-4 and 9_k hold, $\beta_k = \Delta^{S}$.*

Assumptions 8_{fe} and 8_{fd} (resp. 9_{fe} and 9_{fd}) hold if $\Delta^{TR}_{G,T}$ (resp. $\Delta^{S}_{G,T}$) is constant. However, it is often implausible that the treatment effect is constant across groups and time periods. Moreover, this assumption has several testable implications. For instance, it follows from Theorem 1 that under Assumption 2 and constant treatment effects, Δ is equal to β_{fe} in a regression weighted by the population of each group $\times$ period, and in the same but unweighted regression. Therefore, the corresponding estimators should not significantly differ. In the first application we revisit in Section 7, we find that this test is rejected.

When the treatment effect is not constant, Assumptions 8_{fe} and 8_{fd} can still hold if $v^{TR}_{k,G,T}$ is not systematically correlated to $\Delta^{TR}_{G,T}$. To simplify the discussion of that condition, let us momentarily assume that $\bar{t}=1$ and $G \perp\!\!\!\perp T$. Then $\beta_{fe}=\beta_{fd}$, so Assumptions 8_{fe} and 8_{fd} are equivalent and hold if and only if cov $\left(\varepsilon_{fe,G,T}, \Delta^{TR}_{G,T}|D=1\right)=0$. This is restrictive. Positive values of $\varepsilon_{fe,g,t}$ correspond to values of (g,t) for which the share of treated observations is larger than predicted by the regression of $E(D_{g,t})$ on group and period fixed effects. Those may also be the values of (g,t) with the largest treatment effect. For instance, one can show that $\varepsilon_{fe,g,t}$ and $\Delta^{TR}_{g,t}$ are positively related if selection into treatment is determined by a Roy model. Assumptions 8_{fe} and 8_{fd} are also partly testable. First, if β_{fe} and β_{fd} are statistically different, under Assumption 2, one can reject Assumption 8_k for at least one k in $\{fe, fd\}$. In the second application we revisit in Section 7, β_{fe} and β_{fd} are statistically different. Second, assume that researchers observe a variable $P_{g,t}$ that is likely to be correlated with the intensity of the treatment effect across groups and time periods. Then, they can run a suggestive test of Assumption 8_k by testing whether $\text{cov}(v^{TR}_{k,G,T}, P_{G,T}|D=1)=0$. In the first application we revisit in Section 7, this test is rejected. Finally, one can show that similar conclusions apply to Assumptions 9_{fe} and 9_{fd}: they are unlikely to be satisfied and have strong testable implications.

4 Alternative estimand

In this section, we show that Δ^S is actually identified without imposing any restriction on treatment effect heterogeneity, when Assumption 10 below, a testable condition, is satisfied. For every $t \in \{1, ..., \overline{t}\}$, let us introduce the "supergroup" variable $G_t^* = s_{G,t}$. Groups with $G_t^* = 1$ are those where the share of treated units strictly increases between $t-1$ and t, groups with $G_t^* = 0$ are those where that share remains constant, and groups with $G_t^* = -1$ are those where that share strictly decreases.

Assumption 10 *(Existence of "stable" groups) For all $t \in \{1, ..., \overline{t}\}$:*

1. *There exists $g \in \{0, ..., \overline{g}\}$ such that $E(D_{g,t-1}) = E(D_{g,t})$.*
2. *$Supp(D|G_t^* \neq 0, T = t-1) \subset Supp(D|G_t^* = 0, T = t-1)$.*

The first point of the stable groups assumption requires that between each pair of consecutive time periods, there are groups where the share of treated units does not change. In our literature review below, we find that this condition is satisfied in many articles published in the AER between 2010 and 2012 that have used two-way fixed effects regressions.

The second point requires that if there are untreated (resp. treated) units in period $t-1$ in groups where the share of treated units changes between $t-1$ and t, then there should also be untreated (resp. treated units) in period $t-1$ in groups where the share of treated units does not change. In Appendix C, we propose an alternative estimand that identifies Δ^S even if this second condition fails, but that relies on the stable treatment effect assumption.

When the stable groups assumption holds, one can use the groups whose exposure to treatment does not change between $t-1$ and t to identify the effect of time on the two potential outcomes $Y(0)$ and $Y(1)$, and then separate the effect of time from the treatment effect in groups whose exposure to treatment changes, as we show in Theorem 2 below.

We now define our estimand. For all $(d, g, g^*, t, t') \in \{0, 1\} \times \{0, ..., \overline{g}\} \times \{-1, 0, 1\} \times \{1, ..., \overline{t}\}^2$, let $r(g|g^*, t', t) = P(G = g|G_{t'}^* = g^*, T = t)$, $r_d(g|g^*, t', t) = P(G = g|G_{t'}^* = g^*, T = t, D = d)$, and let

$$Q = \frac{r(G|G_{T+1}^*, T+1, T+1)}{r(G|G_{T+1}^*, T+1, T)}, \; Q_d = \frac{r_d(G|G_{T+1}^*, T+1, T+1)}{r_d(G|G_{T+1}^*, T+1, T)}.$$

The ratio $r(g|g^*, t+1, t+1)/r(g|g^*, t+1, t)$ is the share of group g in the supergroup $G_t^* = g^*$ at $T = t+1$, divided by the share of group g in the supergroup $G_t^* = g^*$ at $T = t$. Therefore, Q is larger (resp. lower) than 1 for observations whose group's size has increased faster (resp. more slowly) than their period-$T+1$ super-group's size between $T = t$ and $T = t+1$. Q_d can be interpreted similarly, but conditional on $D = d$. Note that if $G \perp\!\!\!\perp T$, the sizes of each group remain constant over time, so $Q = Q_0 = Q_1 = 1$.

Then, let us define, for all $(d, g, t) \in \{0, 1\} \times \{-1, 1\} \times \{1, ..., \bar{t}\}$, the following quantities:

$$\begin{aligned}
FD^*_D(g^*, t) =& E(D|G^*_t = g^*, T = t) - E(QD|G^*_t = g^*, T = t-1),\\
\delta_{d,t} =& E(Y|D = d, G^*_t = 0, T = t) - E(Q_d Y|D = d, G^*_t = 0, T = t-1),\\
CFD^*(g^*, t) =& E(Y|G^*_t = g^*, T = t) - E(Q(Y + \delta_{D,t})|G^*_t = g^*, T = t-1).
\end{aligned}$$

$FD^*_D(g^*, t)$ is the first-difference of the treatment rate in the supergroup $G^*_t = g^*$ between $T = t-1$ and $T = t$, after reweighting observations in period $t-1$ by Q to ensure that groups' distribution is the same in periods $t-1$ and t. $CFD^*(g^*, t)$ stands for "corrected first difference". It compares the mean outcome in the supergroup $G^*_t = g^*$ at $T = t$ to the mean outcome in the same group at $T = t-1$, after imputing the trends on $Y(0)$ and $Y(1)$ observed in the supergroup $G^*_t = 0$ to the period $T = t-1$ mean outcome in the supergroup $G^*_t = g^*$. Here as well, the reweighting of observations by Q_d and Q in period $t-1$ ensures that groups' distribution is the same in periods $t-1$ and t.[13] Let also

$$\omega_{g^*,t} = \frac{g^* FD^*_D(g^*, t) P(G^*_t = g^*, T = t)}{\sum_{t'=1}^{\bar{t}} \sum_{g^{*\prime} \in \{-1,1\}} g^{*\prime} FD^*_D(g^{*\prime}, t') P(G^*_t = g^{*\prime}, T = t')}.$$

Note that by construction, $\omega_{g^*,t} \geq 0$. Our estimand is defined by

$$W_{TC} = \sum_{t=1}^{\bar{t}} \sum_{g^* \in \{-1,1\}} \omega_{g^*,t} \frac{CFD^*(g^*, t)}{FD^*_D(g^*, t)},$$

with the convention that $\omega_{g^*,t}/FD^*_D(g^*, t) = 0$ when $P(G^*_t = g^*, T = t) = 0$.

Theorem 2 below shows that W_{TC} identifies Δ^S under the following conditional common trends condition.

Assumption 2' *(Common trends conditional on $D(t-1)$) For all $(d, t) \in \{0, 1\} \times \{1, ..., \bar{t}\}$, $E(Y(d)|D(t-1) = d, T = t, G) - E(Y(d)|D(t-1) = d, T = t-1, G)$ does not depend on G.*

Assumption 2' imposes two conditional common trends conditions. First, it requires that for all $t \geq 1$, the evolution from $t-1$ to t of the mean $Y(0)$ of units untreated at $T = t-1$ be the same in every group. Then, it requires that for all $t \geq 1$, the evolution from $t-1$ to t of the mean $Y(1)$ of units treated at $T = t-1$ be the same in every group. This assumption may seem unusual, as it involves conditioning on the lagged treatment. However, it is compatible with selection à la Roy, as shown in Appendix B.2.[14] Moreover, its plausibility can be assessed through placebo tests, as discussed below.

[13] Note that $\delta_{d,t}$ may not be defined for $d = 0$ or $d = 1$ if $\text{Supp}(D|G^*_t = 0, T = t-1) \neq \{0, 1\}$. Still, $E(Q\delta_{D,t}|G^*_t = g^*, T = t-1)$ is well-defined under Assumption 10, since $\text{Supp}(D|G^*_t = g^*, T = t-1) \subset \text{Supp}(D|G^*_t = 0, T = t-1)$.

[14] Interestingly, the stable treatment effect condition is violated in the model we consider.

Theorem 2 *Suppose that Assumption 1 holds and D is binary. If Assumptions 2', 3, and 10 hold, then* $W_{TC} = \Delta^S$.

Theorem 2 shows that when "stable" groups exist for each pair of consecutive time periods, a weighted average of the Wald-TC estimands with two groups and two periods proposed in de Chaisemartin and D'Haultfœuille (2018) identifies Δ^S under Assumptions 2' and 3. In section 1.2 of the supplement of de Chaisemartin and D'Haultfœuille (2018), we also propose a weighted average of Wald-TC estimands in the case with multiple groups and periods. That estimand is close to the one we propose here, but the weights differ. By reweighting observations at period $t-1$ and by using slightly different weights $\omega_{g,t}$, the estimand proposed here does not rely on the assumption that $G \perp\!\!\!\perp T$, and it identifies Δ^S, instead of a less interpretable weighted average of switchers' LATEs.

Let us start by giving the intuition underlying the W_{TC} estimand when the treatment is constant within each group $\times$ period cell and $G \perp\!\!\!\perp T$. Then, let $NT_t = 1\{D_{G,t} = D_{G,t-1} = 0\}$ (resp. $AT_t = 1\{D_{G,t} = D_{G,t-1} = 1\}$) be an indicator for groups that are untreated (resp. treated) in periods $t-1$ and t, the "never treated" (resp. "always treated") groups. One can then show that

$$\begin{aligned}
CFD^*(1,t)/FD^*_D(1,t) =& E(Y|G^*_t = 1, T = t) - E(Y|G^*_t = 1, T = t-1) \\
& - \left(E(Y|NT_t = 1, T = t) - E(Y|NT_t = 1, T = t-1)\right), \\
CFD^*(-1,t)/FD^*_D(-1,t) =& E(Y|AT_t = 1, T = t) - E(Y|AT_t = 1, T = t-1) \\
& - \left(E(Y|G^*_t = -1, T = t) - E(Y|G^*_t = -1, T = t-1)\right).
\end{aligned}$$

$CFD^*(1,t)/FD^*_D(1,t)$ is a DID estimand comparing the evolution of the mean outcome between groups switching from being untreated to treated between $t-1$ and t, and groups that remain untreated between these two dates. Because these groups have the same treatment in $t-1$, under Assumption 2' this DID identifies the treatment effect in groups switching from being untreated to treated. Similarly, $CFD^*(-1,t)/FD^*_D(-1,t)$ is a DID estimand comparing the evolution of the mean outcome between groups that remain treated between $t-1$ and t, and groups switching from being treated to untreated between these two dates. Here again, under Assumption 2' this DID identifies the treatment effect in groups switching from being treated to untreated. Finally, W_{TC} is a weighted average of those estimands.

Notice that in the staggered adoption design where $D = 1\{T \geq a_G\}$, W_{TC} has an even simpler expression. In this design, there is no group whose treatment diminishes between consecutive time periods, so W_{TC} is a weighted average of $CFD^*(1,t)/FD^*_D(1,t)$, with

$$\begin{aligned}
\frac{CFD^*(1,t)}{FD^*_D(1,t)} =& E\left[Y|a_G = t, T = t\right] - E\left[Y|a_G = t, T = t-1\right] \\
& - \left(E(Y|a_G > t, T = t) - E(Q_0 Y|a_G > t, T = t-1)\right).
\end{aligned}$$

In other words, W_{TC} is just a weighted average of the DIDs comparing the evolution of the outcome in groups that become treated at t and in groups not yet treated.

When the treatment is not constant within each group $\times$ period cell, the formula of W_{TC} is more complicated, but here is the intuition underlying it. Under Assumption 2', the trend affecting the $Y(0)$ (resp. $Y(1)$) of units with $D(t-1)=0$ (resp. $D(t-1)=1$) between $t-1$ and t is the same in every group. This trend is identified by the evolution of the mean of Y of untreated (resp. treated) units between $t-1$ and t in all "stable" groups with $G_t^*=0$: under the treatment monotonicity assumption, $D(t-1)=D(t)$ in those groups. Then, one can add the trend on $Y(0)$ (resp. $Y(1)$) to the outcome of untreated (resp. treated) units in group $G_t^*=1$ in period $t-1$, and thus recover the mean outcome we would have observed in this group in period t if switchers had not changed treatment between the two periods. This is what $Y+\delta_{D,t}$ does. Therefore, $CFD^*(1,t)$ compares the mean outcome in group $G_t^*=1$ at $T=t$ to the counterfactual mean we would have observed in that group at $T=t$ if switchers had remained untreated. $FD_D^*(1,t)$ identifies the proportion of switchers, so $CFD^*(1,t)/FD_D^*(1,t)$ identifies the LATE of switchers in group $G_t^*=1$ at $T=t$, and W_{TC} identifies the LATE of all switchers.

Finally, the Wald-TC estimand uses groups where the share of treated units is stable to infer the trends affecting $Y(0)$ and $Y(1)$, and apply those trends to the groups where the share of treated units changes. This strategy could fail, if groups whose average treatment changes between periods $t-1$ and t experience different trends on their outcomes than groups whose average treatment is stable. To assess whether this a serious concern, one can run placebo tests that are similar in spirit to those typically conducted in standard DID. Specifically, a testable implication of the conditional common trends assumption is that for every $d\in\{0,1\}$ and $t\geq 2$,

$$E(Y|D=d,G=g,T=t-1,G_{t-1}^*=0)-E(Y|D=d,G=g,T=t-2,G_{t-1}^*=0)$$

should not depend on g. The evolution of the mean outcome of treated (resp. untreated) units between $t-2$ and $t-1$ should be the same in all groups where the share of treated units is stable between $t-2$ and $t-1$. In particular, one can compare that evolution in groups where the share of treated units is also stable between $t-1$ and t, to that in groups where that share increases (resp. decreases). If the evolution of the mean outcome of treated and untreated units is the same between $t-2$ and $t-1$ in those three supergroups, this suggests that the identifying assumption underlying the Wald-TC estimand is plausible.

5 Extensions

5.1 Non-binary, ordered treatment

We now consider the case where the treatment takes a finite number of ordered values, $D \in \{0, 1, ..., \overline{d}\}$. To allow for this extension, we modify the stable treatment effect assumption.

Assumption 4O *(Stable treatment effect for ordered D) For all* $(d, g, t) \in \{1, ..., \overline{d}\} \times \{0, ..., \overline{g}\} \times \{1, ..., \overline{t}\}$, $E(Y(d)-Y(0)|G = g, T = t, D(t-1) = d) = E(Y(d)-Y(0)|G = g, T = t-1, D(t-1) = d)$.

We also need to modify the treatment effect parameters we consider. In lieu of Δ^{TR}, we consider $\text{ACR}^{TR} = E(Y(D) - Y(0))/E(D)$, the average causal response (ACR) on the treated. For any $(g, t) \in \{0, ..., \overline{g}\} \times \{0, ..., \overline{t}\}$, let $\text{ACR}^{TR}_{g,t} = E(Y_{g,t}(D) - Y_{g,t}(0))/E(D_{g,t})$ denote the ACR in group g and at period t. Let us also define the probability measure P^{TR} by $P^{TR}(A) = E(D1_A)/E(D)$ for any measurable set A. This probability measure generalizes the conditional probability $P(.|D = 1)$ for a binary treatment to non-binary treatments. We let E^{TR} and cov^{TR} denote the expectation and covariance operators associated to P^{TR}. Then note that

$$\text{ACR}^{TR} = E^{TR}\left[\text{ACR}^{TR}_{G,T}\right].$$

In lieu of Δ^S, we consider $\text{ACR}^S = E[\text{ACR}^S_{G,T}|S]$, where for all $(g, t) \in \{0, ..., \overline{g}\} \times \{1, ..., \overline{t}\}$,

$$\begin{aligned}\text{ACR}^S_{g,t} =& \sum_{d=1}^{\overline{d}} \frac{P(D_{g,t} \geq d) - P(D_{g,t-1} \geq d)}{E(D_{g,t}) - E(D_{g,t-1})} \Delta^S_{dgt},\\ \Delta^S_{dgt} =& E\left[Y_{g,t}(d) - Y_{g,t}(d-1)\right| \max(D(t), D(t-1)) \geq d > \min(D(t), D(t-1))\right].\end{aligned}$$

Δ^S_{dgt} is the average effect of going from $d - 1$ to d units of treatment, in period t, and among the units in group g whose treatment switches from strictly below to above d or from above to strictly below d between $T = t - 1$ and $T = t$. $\text{ACR}^S_{g,t}$ is a weighted average over d of the Δ^S_{dgt}, where each of the Δ^S_{dgt} is weighted proportionally to the size of the corresponding population of switchers. $\text{ACR}^S_{g,t}$ is similar to the ACR parameter considered in Angrist and Imbens (1995). ACR^S is a weighted average of the $\text{ACR}^S_{g,t}$ across groups and periods.

Finally, we generalize Assumptions 8 and 9 as follows.

Assumption 8O$_{\boldsymbol{k}}$ *(Random weights or homogeneous ACRs)* $cov^{TR}\left(v^{TR}_{k,G,T}, ACR^{TR}_{G,T}\right) = 0$.

Assumption 9O$_{\boldsymbol{k}}$ *(Random weights or homogeneous ACRs)* $cov\left(v^S_{k,G,T}, ACR^S_{G,T}|S\right) = 0$.

Theorem 3 below generalizes Theorem 1 and Corollary 2 to the case where the treatment takes a finite number of ordered values.

Theorem 3 *Suppose that Assumption 1 holds, $D \in \{0, ..., \overline{d}\}$, $k \in \{fe, fd\}$ and β_k is well-defined.*

1. *If Assumption 2 holds, then*
$$\beta_k = E^{TR}\left[v^{TR}_{k,G,T} ACR^{TR}_{G,T}\right]. \tag{7}$$
If Assumption $8O_k$ further holds, then $\beta_k = ACR^{TR}$.

2. *If Assumptions 2-3 and 4O hold, then*
$$\beta_k = E\left[v^{S}_{k,G,T} ACR^{S}_{G,T}|S\right]. \tag{8}$$
If Assumption $9O_k$ further holds, then $\beta_k = ACR^{S}$.

The first (resp. second) point of Theorem 3 shows that under Assumption 2 (resp. Assumption 2-3 and 4O), β_{fe} and β_{fd} identify weighted sums of the $(\text{ACR}^{TR}_{g,t})_{g,t}$ (resp. $(\text{ACR}^{S}_{g,t})_{g,t}$), where the weights are the same as those in Theorem 1. Because of this, and since the proof of Corollary 1 does not rely on the nature of the treatment, Corollary 1 directly applies to ordered treatments as well, by just replacing $V(v^{TR}_{G,T}|D = 1)$ by $V^{TR}(v^{TR}_{G,T})$. Finally, for every $k \in \{fe, fd\}$, under Assumption 8O_k (resp. 9O_k), β_k identifies ACR^{TR} (resp. ACR^{S}). Yet, as already explained in the binary treatment case, these assumptions may not be plausible.

Theorem 3 extends to a continuous treatment. In such instances, one can for instance show that under Assumptions 2-3 and an appropriate generalization of Assumption 4, β_{fe} and β_{fd} identify weighted sums of the same weighted averages of the derivative of potential outcomes with respect to treatment in group g and at time t as in Angrist et al. (2000).

Finally, Theorem 2 can easily be extended to the case where the treatment is not binary. One can show that W_{TC} identifies ACR^{S} under the same assumptions as that underlying Theorem 2, except that $E(D_{g,t}) = E(D_{g,t-1})$ should be replaced by $D_{g,t} \sim D_{g,t-1}$ in Assumption 10. We refer to de Chaisemartin and D'Haultfœuille (2018) for further details.

5.2 Including covariates

Often times, researchers also include a vector of covariates X as control variables in their regression. We show in this section that our results can be extended to this case. We start by redefining the two regressions we consider in this context.

Regression 1X *(Fixed-effects regression with covariates)*

Let β^X_{fe} and γ_{fe} denote the coefficients of $E(D|G,T)$ and X in an OLS regression of Y on a constant, $(1\{G = g\})_{1 \leq g \leq \overline{g}}$, $(1\{T = t\})_{1 \leq t \leq \overline{t}}$, $E(D|G,T)$, and X.

Regression 2X *(First-difference regression with covariates)*

Let β_{fd}^X *and* γ_{fd} *denote the coefficients of* $FD_D(G,T)$ *and* $FD_X(G,T)$ *in an OLS regression of* $FD_Y(G,T)$ *on a constant,* $(1\{T=t\})_{2\leq t\leq \bar{t}}$*,* $FD_D(G,T)$*, and* $FD_X(G,T)$*, conditional on* $T\geq 1$*.*

Then, we need to modify the common trends assumption as follows.

Assumption 2X$_{\boldsymbol{k}}$ *(Common trends for* β_k^X*)* $E(Y(0)-X'\gamma_k|G,T=t)-E(Y(0)-X'\gamma_k|G,T=t-1)$ *does not depend on* G *for all* $t\in\{1,...,\bar{t}\}$*.*

Assumptions 2X_{fe} and 2X_{fd} are implied by the linear and constant treatment effect models that are often invoked to justify the use of FE and FD regressions with covariates. For instance, the use of Regression 1X is often justified by the following model:

$$Y=\gamma_G+\lambda_T+\theta D+X'\gamma_{fe}+\varepsilon,\quad E(\varepsilon|G,T,D,X)=0. \tag{9}$$

Equation (9) implies Assumption 2X_{fe}, but it does not imply Assumption 2. This is why we consider the former common trends assumption instead of the latter in this subsection. Similarly, the linear and constant treatment effect model rationalizing the use of Regression 2X implies Assumption 2X_{fd} but not Assumption 2.

Assumption 2X_{fe} requires that once netted out from the partial linear correlation between the outcome and X, $Y(0)$ satisfies the common trends assumption. It may be more plausible than Assumption 2, for instance if there are group-specific trends affecting the outcome but if those group-specific trends can be captured by a linear model in X. A similar interpretation applies to Assumption 2X_{fd}.

Finally, we need to assume that the treatment has no effect on the covariates. Let $X(0)$ and $X(1)$ respectively denote the potential covariates of the same observation without and with the treatment.

Assumption 11 *(No treatment effect on the covariates)* $X(0)=X(1)=X$*.*

Assumption 11 will for instance hold if X is determined prior to the treatment.

Theorem 4 below generalizes Theorem 1 to the case where there are covariates in the regression.

Theorem 4 *Suppose that Assumption 1 holds,* D *is binary,* $k\in\{fe,fd\}$ *and* β_k *is well-defined.*

1. *If Assumptions* $2X_k$ *and 11 hold, then*

$$\beta_k^X=E\left[v_{k,G,T}^{TR}\Delta_{G,T}^{TR}|D=1\right].$$

If Assumption 8_k *further holds, then* $\beta_k^X=\Delta^{TR}$*.*

2. If Assumptions $2X_k$, 3, 4, and 11 hold, then

$$\beta_k^X = E\left[v_{k,G,T}^S \Delta_{G,T}^S | S\right].$$

If Assumption 9_k further holds, then $\beta_k^X = \Delta^S$.

Theorem 4 shows that under a modified version of the common trends assumption accounting for the covariates, β_k^X identifies the same weighted sum of the $(\Delta_{g,t}^{TR})_{g,t}$ as β_k in Theorem 1. If one further imposes Assumptions 3 and 4, β_k^X identifies the same weighted sum of the $(\Delta_{g,t}^S)_{g,t}$ as β_k in Theorem 1. Therefore, if the regression without covariates has some negative weights and one worries that treatment effect heterogeneity may be systematically related to the weights, the addition of covariates will not alleviate that concern. But adding covariates may still have some benefits. For instance, the common trends assumption underlying β_k^X may be more credible than that underlying β_k.

Finally, Theorem 2 can easily be extended to the case with covariates. Under versions of Assumptions 2', 3, and 10 written conditional on X, a conditional version of the Wald-TC estimand identifies Δ^S under the common support condition $\text{Supp}(X_{d,g,t}) = \text{Supp}(X)$. We refer to de Chaisemartin and D'Haultfœuille (2018) for further details.

5.3 2SLS regressions

Researchers have sometimes estimated 2SLS versions of regressions 1 and 2. Our main conclusions also apply to those regressions. Let β_{fe}^{2SLS} denote the coefficient of D in a 2SLS regression of Y on a constant, $(1\{G = g\})_{1 \leq g \leq \overline{g}}$, $(1\{T = t\})_{1 \leq t \leq \overline{t}}$, and D, using a variable Z constant within each group $\times$ period as the instrument for D. Z typically represents an incentive for treatment allocated at the group $\times$ period level. For instance, Duflo (2001) studies the effect of years of schooling on wages in Indonesia, using a primary school construction program as an instrument. Specifically, she estimates a 2SLS regression of wages on cohort and district of birth fixed effects and years of schooling, using the interaction of belonging to a cohort entering primary school after the program was completed and the number of schools constructed in one's district of birth as the instrument for years of schooling.

Following Imbens and Angrist (1994), for any $z \in \text{Supp}(Z)$ let $D(z)$ denote an observation's potential treatment if $Z = z$. β_{fe}^{2SLS} is the ratio of the coefficients of Z in the reduced-form regression of Y on group and period fixed effects and Z, and in the first-stage regression of D on group and period fixed effects and Z. As Z is constant within each group $\times$ period, those are regressions of Y and D on group and period fixed effects and $E(Z|G,T)$. Therefore, it follows from Theorem 1 that under a common trends assumption on $D(0)$, the first-stage coefficient of Z identifies a weighted sum of the average effects of the instrument on the treatment in each group and time period, with potentially many negative weights. Similarly, under a common trends assumption on $Y(D(0))$ instead of $Y(0)$, the reduced-form coefficient of Z identifies a weighted sum of the average effects of the instrument on the outcome, again with potentially many negative weights. For instance, in Duflo (2001), under a common trends assumption on $D(0)$, the number of years of schooling individuals would complete if zero new schools were constructed in their district, the first stage coefficient identifies a weighted sum of the effect of one new school on years of schoooling in every district, with many negative weights.[15]

Hence, it is only if the average effects of Z on Y and D are constant across groups and periods, or if a version of our random weights assumption adapted to those regressions holds, that the reduced-form and first-stage coefficients of Z respectively identify the average effect of Z on Y and D, thus implying that β_{fe}^{2SLS} identifies, under the conditions in Imbens and Angrist (1994), the LATE of D on Y among units that comply with the instrument.[16]

5.4 Dynamic treatment effects

Our main parameter of interest Δ^S is a weighted average of the period-t treatment effect among switchers switching from non treatment to treatment between $t-1$ and t, across values of t. Therefore, it is a measure of the instantaneous effect of the treatment. Though this is not the main focus of this paper, we briefly describe below how one can follow the same logic as in Section 4 to identify the dynamic effect of the treatment among switchers. For instance, one can identify the effect of the treatment among switchers one period after they have switched, provided that treatment monotonicity holds and the three following assumptions are satisfied:

1. For all $t \geq 2$, there exists $g \in \{0, ..., \overline{g}\}$ such that $E(D_{g,t}) = E(D_{g,t-2})$.
2. For all $(d, t) \in \{0, 1\} \times \{2, ..., \overline{t}\}$, $E(Y(d)|D(t-2) = d, T = t, G) - E(Y(d)|D(t-2) = d, T = t-2, G)$ does not depend on G.
3. For all $t \geq 2$, $E(D_{g,t-1}) \neq E(D_{g,t-2}) \Rightarrow E(D_{g,t}) = E(D_{g,t-1})$.

Points 1 and 2 are similar to the stable groups and conditional common trends assumptions in Section 4. Point 3 requires that in all groups where some units switch treatment status between $t-2$ and $t-1$, the treatment rate remains stable between $t-1$ and t. If that assumption fails,

[15] New schools were constructed in every district, so this application falls into the heterogeneous adoption case.

[16] In the special case with two groups and two periods, a binary incentive for treatment, and where only group 1 in period 1 receives the incentive, de Chaisemartin (2010) and Hudson et al. (2015) show that in a 2SLS regression of Y on $1\{G = 1\}$, $1\{T = 1\}$ and D, using $Z = 1\{G = 1\}1\{T = 1\}$ as the instrument, the coefficient of D identifies a LATE under common trends assumptions on $Y(D(0))$ and $D(0)$. However, the discussion above shows that this result does not generalize to applications with multiple groups and periods, a non-binary instrument, or a fuzzy design where all groups receive heterogeneous levels of the incentive, as in Duflo (2001).

one cannot separate the period-t treatment effect of units that switched between $t-2$ and $t-1$ from the period-t treatment effect of units switching between $t-1$ and t. Though Point 3 is quite restrictive, it holds in the staggered adoption design. There, groups that switched from being untreated to treated between $t-2$ and $t-1$ remain treated between $t-1$ and t. Under treatment monotonicity and Points 1-3 above, one can show that the average effect of the treatment among switchers one period after they have switched is identified by a weighted average of Wald-TC estimands comparing the evolution of the outcome between t and $t-2$ in groups where $E(D_{g,t-1}) \neq E(D_{g,t-2})$ and $E(D_{g,t}) = E(D_{g,t-1})$, and in groups where $E(D_{g,t}) = E(D_{g,t-2})$.

6 Discussion on estimation and inference

In this section, we discuss briefly how weights and the alternative estimands presented above can be estimated, and how inference can be conducted.

6.1 Inference on the weights and functions of the weights

We start by assuming that the population of groups only bears the $\overline{g}+1$ groups we observe, and that for each group $\times$ period, we observe an i.i.d. sample of the population of that group $\times$ period of size $n_{g,t}$. Let $\widehat{E}(D|G,T)$ (resp. $\widehat{E}(D_{g,t})$) denote the empirical counterparts of $E(D|G,T)$ (resp. $E(D_{g,t})$). Then let $\widehat{\varepsilon}_{fe,g,t}$ denote the residual of the regression of $\widehat{E}(D|G,T)$ on group and time fixed effects. We can then estimate $w^{TR}_{fe,g,t}$ by

$$\widehat{w}^{TR}_{fe,g,t} = \frac{\widehat{P}(G=g,T=t)\widehat{E}(D_{g,t})\widehat{\varepsilon}_{fe,g,t}}{\sum_{g',t'} \widehat{P}(G=g',T=t')\widehat{E}(D_{g',t'})\widehat{\varepsilon}_{fe,g',t'}},$$

where $\widehat{P}(G=g,T=t) = n_{g,t}/\sum_{g',t'} n_{g',t'}$. The weights $w^{TR}_{fd,g,t}$ can be estimated similarly. The weights $w^{S}_{fd,g,t}$ and $w^{S}_{fe,g,t}$ can also be estimated similarly, with one caveat. Those weights depend on $s_{g,t} = \text{sgn}(E(D_{g,t}) - E(D_{g,t-1}))$, a discontinuous function of $(E(D_{g,t}), E(D_{g,t-1}))$. When the treatment is constant within each group $\times$ period, $s_{g,t}$ is known and does not need to be estimated. Otherwise, we consider the following estimator:

$$\widehat{s}_{g,t} = 1\{\widehat{E}(D_{g,t}) - \widehat{E}(D_{g,t-1}) > c_{n_{g,t}}\} - 1\{\widehat{E}(D_{g,t}) - \widehat{E}(D_{g,t-1}) < -c_{n_{g,t}}\}, \qquad (10)$$

for some $c_{n_{g,t}} > 0$. We introduce this threshold in order to obtain a consistent estimator of $s_{g,t}$, even in the case where $s_{g,t} = 0$. For $k \in \{fe, fd\}$, $\widehat{w}^{TR}_{k,g,t}$ is consistent and asymptotically normal as long as $n_{g,t} \to +\infty$, and $\widehat{w}^{S}_{k,g,t}$ is consistent and asymptotically normal as long as $n_{g,t} \to +\infty$, and $c_{n_{g,t}} \to 0$ and $\sqrt{n_{g,t}}c_{n_{g,t}} \to +\infty$ if $s_{g,t}$ has to be estimated. A similar result holds for estimators of functions of the weights such as $\underline{\sigma}^{TR}$ and $\underline{\sigma}^{S}$, because they are regular functions of $\widehat{\beta}_k$ and of the estimated weights. Inference can then be conducted using the bootstrap, by drawing samples from the samples of observations in each group $\times$ period.

A limitation of the modelling framework outlined in the previous paragraph is that when the treatment is constant within each group $\times$ period, $\widehat{E}(D|G,T) = E(D|G,T)$ and $\widehat{FD}_D(G,T) = FD_D(G,T)$, so we for instance have $\widehat{\varepsilon}_{fe,g,t} = \varepsilon_{fe,g,t}$. Then, if the probabilities $(P(G = g, T = t))_{g,t}$ are also known,[17] the population weights are known and do not need to be estimated. Accordingly, any function of the weights is known without statistical uncertainty. Then, an alternative way of introducing statistical uncertainty is to assume that the $\overline{g} + 1$ groups we observe are an i.i.d. sample from an infinite super-population of groups. Note that this modelling framework is in line with that underlying cluster-robust inference methods that are commonly used in DID analysis (see Bertrand et al., 2004). In that framework, we conjecture that it is possible to show that estimators of functions of the weights such as $\underline{\sigma}^{TR}$ and $\underline{\sigma}^{S}$ are still consistent and asymptotically normal, provided $\overline{g} + 1 \to +\infty$. Inference could then be conducted using the bootstrap. Specifically, one could first draw samples of groups from the $\overline{g} + 1$ groups in the sample, and then draw samples from the samples of observations in each group $\times$ period.[18]

6.2 Alternative estimand

Our alternative estimand can also be estimated by plug-in estimators, where G^*_t is replaced by $\widehat{G}^*_t = \widehat{s}_{G,t}$, with $\widehat{s}_{g,t}$ defined in (10). For instance, we estimate the weights Q and Q_d by

$$\widehat{Q} = \frac{\widehat{r}(G|\widehat{G}^*_{T+1}, T+1, T+1)}{\widehat{r}(G|\widehat{G}^*_{T+1}, T+1, T)}, \ \widehat{Q}_d = \frac{\widehat{r}_d(G|G^*_{T+1}, T+1, T+1)}{\widehat{r}_d(G|G^*_{T+1}, T+1, T)},$$

where

$$\begin{aligned}\widehat{r}(g|g^*, t, t') &= \widehat{P}(G = g|\widehat{G}^*_{t'} = g^*, T = t),\\ \widehat{r}_d(g|g^*, t, t') &= \widehat{P}(G = g|\widehat{G}^*_{t'} = g^*, T = t, D = d),\end{aligned}$$

and the estimated probabilities on the right-hand sides are just sample proportions. Following the same reasoning as in the proof of Theorem S6 in the supplement of de Chaisemartin and D'Haultfœuille (2018), we can show that the corresponding estimator of W_{TC} is consistent and asymptotically normal. Again, inference can be conducted using the bootstrap.

[17] This is often the case. For instance, in an application where groups are US counties, the proportion each county accounts for in the US population is known.

[18] This second step cannot be implemented when the data is aggregated at the group $\times$ period level.

7 Applicability, and applications

7.1 Applicability

We conducted a review of all papers published in the American Economic Review (AER) between 2010 and 2012 to assess the pervasiveness of two-way fixed effects regressions in economics. Over these three years, the AER published 337 papers. This excludes papers and proceedings, comments, replies, and presidential addresses. Out of these 337 papers, 33 or 9.8% of them estimate the FE or FD Regression, or other regressions resembling closely those regressions. When one withdraws from the denominator theory papers and lab experiments, the proportion of papers using these regressions raises to 19.1%.

Table 1: Papers using two-way fixed effects regressions published in the AER

	2010	2011	2012	Total
Papers using two-way fixed effects regressions	5	14	14	33
% of published papers	5.2%	12.2%	11.2%	9.8%
% of empirical papers, excluding lab experiments	12.8%	23.0%	19.2%	19.1%

Notes. This table reports the number of papers using two-way fixed effects regressions published in the AER from 2010 to 2012.

Table 2 shows descriptive statistics about the 33 2010-2012 AER papers estimating two-way fixed effects regressions. Panel A considers the estimation method used in these papers: 13 use the FE regression; six use the FD regression; six use regressions very similar to the FE or FD regression except that they have several treatment variables; three use the FE or FD 2SLS regression discussed in Subsection 5.3; five use other regressions that we deemed sufficiently close to the FE or FD regression to include them in our count.[19] Panel B shows whether the first point of the stable groups assumption is satisfied in those papers. For about a half of the papers, reading the paper was not enough to assess with certainty whether this assumption holds. We then tried to assess whether the assumption was presumably satisfied or presumably not satisfied based on the context. Overall, this assumption holds in 12 papers, it is presumably satisfied in 14, it is presumably not satisfied in five, and it is not satisfied or not applicable in two.

[19]For instance, two papers use regressions with three-way fixed-effects instead of two-way fixed effects.

Table 2: Descriptive statistics on two-way fixed effects papers

	# Papers
Panel A. Estimation method	
Fixed-effects OLS regression	13
First-difference OLS regression	6
Fixed-effects or first-difference OLS regression, with several treatment variables	6
Fixed-effects or first-difference 2LS regression	3
Other regression	5
Panel B. Point 1 of the stable groups assumption	
Satisfied	12
Presumably satisfied	14
Presumably not satisfied	5
Not satisfied or not applicable	2

Notes. This table reports the estimation method used in the 33 papers using two-way fixed effects regressions published in the AER between 2010 and 2012, and whether Point 1 of the stable groups assumption is satisfied.

In Appendix D, we review each of the 33 papers. We explain where two-way fixed effects regressions are used in the paper, and we detail our assessment of whether the stable groups assumption holds or not.

7.2 Applications

Enikolopov et al. (2011)

Enikolopov et al. (2011) study the effect of NTV, an independent TV channel introduced in 1996 in Russia, on the share of the electorate voting for opposition parties. NTV's coverage rate was heterogeneous across subregions: while a large fraction of the population received NTV in urbanized subregions, a smaller fraction received it in more rural subregions. The authors estimate the FE regression: they regress the share of votes for opposition parties in the 1995 and 1999 elections in Russian subregions on subregion fixed effects, an indicator for the 1999 election, and on the share of the population having access to NTV in each subregion at the time of the election. In 1995, the share of the population having access to NTV was equal to 0 in all subregions, while in 1999 it was strictly greater than 0 everywhere. Therefore, the authors' research design corresponds exactly to the heterogenous adoption design discussed in Section 3.

Enikolopov et al. (2011) find that $\widehat{\beta}_{fe} = 6.65$ (s.e.$= 1.40$). According to this regression, increasing the share of the population having access to NTV from 0 to 100% increases the share of votes for the opposition parties by 6.65 percentage points. Because $\bar{t} = 1$ and $T \perp\!\!\!\perp G$, $\widehat{\beta}_{fe} = \widehat{\beta}_{fd}$.

As no one was treated in 1995, treated observations and switchers in period 1 correspond to the same observations, so the weights attached to β_{fe} under the common trends assumption and under the common trends, treatment monotonicity, and stable treatment effect assumptions are also the same. We estimate the weights $(w^{TR}_{fe,g,1})_g$. 918 are strictly positive, while 1,020 are strictly negative. The negative weights sum to -2.26 (s.e. =0.06).[20] Finally, we find that $\widehat{\underline{\sigma}}^{TR}_{fe} = 0.91$ (95% level confidence interval=$[0.53, 1.28]$). Namely, β_{fe} and Δ^{TR} may be of opposite signs even if the standard deviation of the effect of NTV across subregions is below one percentage point.

Therefore, the causal interpretation of β_{fe} relies on Assumption 8_{fe}. This assumption is not warranted. First, the effect of NTV is unlikely to be constant across Russian subregions: that effect could for instance be higher in more rural areas, as fewer other sources of independent information may be available there. Moreover, a testable implication of the constant treatment effect assumption is rejected. We estimate $\widehat{\beta}_{fe}$ again, weighting the regression by subregions' population. We obtain $\widehat{\beta}_{fe} = 14.89$, more than twice its value in the unweighted regression, and the difference between the two coefficients is statistically significant (t-stat=2.46). Second, the weights $(w^{TR}_{fe,g,1})_g$ are not "randomly assigned" to subregions. For instance, the correlation between the weights and subregions' population is equal to 0.35 (t-stat=14.01), and the correlation between the weights and subregions' average wage is 0.05 (t-stat=2.22).

Finally, note that the share of people having access to NTV is strictly positive in every subregion in 1999, implying that there are no subregions where the treatment rate remains constant from 1995 to 1999. Hence, we cannot compute the Wald-TC estimator in this application.

Gentzkow et al. (2011)

Gentzkow et al. (2011) study the effect of newspapers on voters' turnout in US presidential elections between 1868 and 1928. They regress the first-difference of the turnout rate in county g between election years t and $t-1$ on state-year fixed effects and on the first difference of the number of newspapers available in that county. This regression corresponds to the first-difference regression, with state-year fixed effects as controls. As reproduced in Table 3 below, Gentzkow et al. (2011) find that $\widehat{\beta}_{fd} = 0.0026$ (s.e.$= 9 \times 10^{-4}$). According to this regression, one more newspaper increased voters' turnout by 0.26 percentage points. On the other hand, $\widehat{\beta}_{fe} = -0.0011$ (s.e.$= 0.0011$). $\widehat{\beta}_{fe}$ and $\widehat{\beta}_{fd}$ are significantly different (t-stat=2.86).

A large proportion of the weights attached to β_{fe} and β_{fd} under the common trends assump-

[20] To draw inference on the estimators we compute in this subsection, we use the bootstrap, clustered at the subregion level.

tion are negative. We estimate the weights $(w^{TR}_{fe,g,t})_{g,t}$: 6,363 are strictly positive, 4014 are strictly negative. The negative weights sum to -0.46 (s.e.=0.06), and $\widehat{\underline{\sigma}}^{TR}_{fe} = 4 \times 10^{-4}$ (95% level confidence interval=$[1 \times 10^{-5}, 0.00125]$). Namely, β_{fd} and ACR^{TR} may be of opposite signs if the standard deviation of the $(ACR^{TR}_{g,t})_{g,t}$ is equal to 0.0004. Similarly, 4,002 of the $(w^{TR}_{fd,g,t})$ weights are strictly positive and 6,376 are strictly negative. The negative weights sum to -1.28 (s.e.=0.14), and $\widehat{\underline{\sigma}}^{TR}_{fd} = 5 \times 10^{-4}$ (95% level confidence interval=$[1 \times 10^{-4}, 8 \times 10^{-4}]$). Therefore, under the common trends assumption the causal interpretation of β_{fe} and β_{fd} respectively relies on Assumption 8_{fe} and 8_{fd}. Those two assumptions are not warranted: as $\widehat{\beta}_{fe}$ and $\widehat{\beta}_{fd}$ significantly differ, at least one of them must be violated.

Under the common trends, treatment monotonicity, and stable treatment effect assumptions, 25% of the weights attached to β_{fe} are still negative, the negative weights sum to -0.22 (s.e.=0.02), and $\widehat{\underline{\sigma}}^{S}_{fe} = 4 \times 10^{-4}$ (95% level confidence interval=$[3 \times 10^{-5}, 0.00126]$). All the weights attached to β_{fd} are positive, and $\widehat{\underline{\sigma}}^{S}_{fe} = 0.0041$ (95% level confidence interval=$[0.001, 0.007]$). Under those three assumptions, β_{fd} is much more robust to heterogeneous treatment effects than β_{fe}: it identifies a convex combination of LATEs, and for it to not have the same sign as ACR^S, treatment effects should be ten times more heterogeneous than for β_{fe}.

However, for β_{fd} to identify a convex combination of LATEs, the stable treatment effect assumption has to hold, which may not be plausible in this context. This assumption requires that in counties with at least one newspaper in election year $t-1$, the effect of newspapers does not change between election years $t-1$ and t. However, newspapers' readership systematically decreases between consecutive elections. On average, across pairs of consecutive elections and counties, and restricting the sample to counties with at least one newspaper in the first of the two consecutive elections, the fraction of a county's population reading the newspapers divided by the county's number of newspapers decreases by 1.0 percentage point between two consecutive elections (t-stat=-6.70). As newspapers tend to be less widely read in election-year t than in election-year $t-1$, their effect may decrease between consecutive elections.

The stable groups assumption holds in this application: between each pair of consecutive elections, there are counties where the number of newspapers does not change. We can then estimate W_{TC}, the estimand proposed in Section 4. That estimand identifies the LATE of all switchers, and contrary to β_{fd} it does not rely on any restriction on treatment effect heterogeneity. Instead, it relies on Assumption 2', a conditional common trends assumption. As only 10% of county $\times$ election-year cells have 4 newspapers or more, in the estimation of $CFD^*(1,t)$ and $CFD^*(-1,t)$ we group the number of newspapers into 4 categories: 0, 1, 2, and more than 3.[21] Moreover, we include state-year fixed effects as controls in our estimation. We find that $\widehat{W}_{TC} = 0.0043$, with a standard error of 0.0015. $\widehat{W}_{TC}$ is 66% larger than $\widehat{\beta}_{fd}$, and the two estimators are significantly different at the 10% level (t-stat=1.69). $\widehat{W}_{TC}$ is also of a different sign than $\widehat{\beta}_{fe}$.

[21]Results are similar if we group the number of newspapers into 5 categories: 0, 1, 2, 3, and more than 4.

Finally, a placebo test suggests that the conditional common trends assumption is plausible in this application. We compute a placebo estimator, with the exact same formula as the actual Wald-TC estimator, up to two differences. First, instead of using the turnout in county g and election-year t as the outcome variable, we use the turnout in the same county in the previous election. Second, we restrict the sample to counties where the number of newspapers did not change between $t-2$ and $t-1$. Therefore, our placebo estimators compare the evolution of turnout from $t-2$ to $t-1$, between counties where the number of newspapers increased or decreased between $t-1$ and t and counties where that number remained stable, restricting the sample to counties where the number of newspapers remained stable from $t-2$ to $t-1$.

As shown in Table 3 below, our placebo estimator is close to, and not significantly different from, 0. This shows that counties where the number of newspapers increased or decreased between $t-1$ and t did not experience significantly different trends in turnout from $t-2$ to $t-1$ than counties where that number was stable. Our placebo estimator is estimated on a subset of the data: for each pair of time periods, we only keep counties where the number of newspapers did not change between $t-2$ and $t-1$ in the estimation of $CFD^*(1,t)$ and $CFD^*(-1,t)$. Still, almost 80% of the county $\times$ election-year observations are used in the computation of this placebo estimator. Moreover, when estimated on this subsample, the actual Wald-TC estimator is still large and significant, and it is very close to the Wald-TC estimator in the full sample.

Table 3: Estimates of the effect of one additional newspaper on turnout

	Estimate	Standard error	N
$\widehat{\beta}_{fd}$	0.0026	0.0009	15,627
$\widehat{\beta}_{fe}$	-0.0011	0.0011	16,872
$\widehat{W}_{TC}$	0.0043	0.0015	16,872
Placebo $\widehat{W}_{TC}$	-0.0009	0.0016	13,221
$\widehat{W}_{TC}$, on placebo subsample	0.0045	0.0019	13,221

Notes. This table reports estimates of the effect of one additional newspaper on turnout, as well as a placebo estimate of the conditional common trends assumption. All estimators are computed using the data used in Gentzkow et al. (2011), and with state-year fixed effects as controls. Standard errors are clustered at the district level. In the computation of the Wald-TC estimators, the number of newspapers is grouped into 4 categories: 0, 1, 2, and more than 3.

8 Conclusion

Almost 20% of empirical articles published in the AER between 2010 and 2012 use regressions with groups and period fixed effects to estimate treatment effects. While it is well-known that such regressions identify the treatment effect of interest if that effect is constant and if the standard common trends assumption is satisfied, those regressions have not yet been studied in a model allowing for treatment effect heterogeneity. In this paper, we start by showing that under the common trends assumption alone, two pervasive two-way fixed effects regressions identify weighted sums of the ATTs in each group and at each period. Many of the weights attached to those regressions may be negative: in two empirical applications, we find that more than 50% of the weights are negative. When many weights are negative, two-way fixed effects regressions are not robust to heterogeneous treatment effects across groups and periods: the coefficient of the treatment variable in those regressions may for instance be negative while the treatment effect is positive for every observation in the population.

Then, we consider two supplementary assumptions. The first one requires that in each group and for each pair of consecutive periods, the average treatment effect of units treated at period $t-1$ be stable from $t-1$ to t. The second one requires that in each group and for each pair of consecutive periods, the treatment follows a monotonic evolution from $t-1$ to t. Under the common trends assumption and those two supplementary assumptions, we show that our two-way fixed effects regressions identify weighted sums of the LATEs of switchers in each group and at each period, where switchers are units whose treatment changes between two consecutive time periods. Here again, some of the weights may be negative. However, in the special case with staggered adoption the weights are all positive.

Finally, we propose a new estimand. This estimand identifies the LATE of all switchers, and it does not rely on any treatment effect homogeneity condition. It can be used in applications where there are groups whose exposure to the treatment does not change between each pair of consecutive time periods. In one of the two applications we revisit, the corresponding estimator is very different from the two two-way fixed effects estimators we consider.

References

Aaronson, D., Agarwal, S. and French, E. (2012), 'The spending and debt response to minimum wage hikes', *American Economic Review* **102**(7), 3111–3139.

Abadie, A. (2005), 'Semiparametric difference-in-differences estimators', *Review of Economic Studies* **72**(1), 1–19.

Abraham, S. and Sun, L. (2018), 'Estimating dynamic treatment effects in event studies with heterogeneous treatment effects'.

Acemoglu, D., Cantoni, D., Johnson, S. and Robinson, J. A. (2011), 'The consequences of radical reform: The french revolution', *American Economic Review* **101**(7), 3286–3307.

Aizer, A. (2010), 'The gender wage gap and domestic violence', *American Economic Review* **100**(4), 1847–1859.

Algan, Y. and Cahuc, P. (2010), 'Inherited trust and growth', *American Economic Review* **100**(5), 2060–2092.

Anderson, S. T. and Sallee, J. M. (2011), 'Using loopholes to reveal the marginal cost of regulation: The case of fuel-economy standards', *American Economic Review* **101**(4), 1375–1409.

Angrist, J. D. (1998), 'Estimating the labor market impact of voluntary military service using social security data on military applicants', *Econometrica* **66**(2), 249–288.

Angrist, J. D., Graddy, K. and Imbens, G. W. (2000), 'The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish', *The Review of Economic Studies* **67**(3), 499–527.

Angrist, J. D. and Imbens, G. W. (1995), 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of the American Statistical Association* **90**, 431–442.

Athey, S. and Stern, S. (2002), 'The impact of information technology on emergency health care outcomes', *The RAND Journal of Economics* **33**(3), 399–432.

Bagwell, K. and Staiger, R. W. (2011), 'What do trade negotiators negotiate about? empirical evidence from the world trade organization', *American Economic Review* **101**(4), 1238–1273.

Bajari, P., Fruehwirth, J. C., Kim, K. I. and Timmins, C. (2012), 'A rational expectations approach to hedonic price regressions with time-varying unobserved product attributes: The price of pollution', *American Economic Review* **102**(5), 1898–1926.

Baum-Snow, N. and Lutz, B. F. (2011), 'School desegregation, school choice, and changes in residential location patterns by race', *American Economic Review* **101**(7), 3019–3046.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004), 'How much should we trust differences-in-differences estimates?', *The Quarterly Journal of Economics* **119**(1), 249–275.

Besley, T. and Mueller, H. (2012), 'Estimating the peace dividend: The impact of violence on house prices in northern ireland', *American Economic Review* **102**(2), 810–833.

Bloom, N., Sadun, R. and Van Reenen, J. (2012), 'Americans do it better: Us multinationals and the productivity miracle', *American Economic Review* **102**(1), 167–201.

Borusyak, K. and Jaravel, X. (2017), Revisiting event study designs. Working paper.

Brambilla, I., Lederman, D. and Porto, G. (2012), 'Exports, export destinations, and skills', *American Economic Review* **102**(7), 3406–3438.

Bustos, P. (2011), 'Trade liberalization, exports, and technology upgrading: Evidence on the impact of mercosur on argentinian firms', *American Economic Review* **101**(1), 304–340.

Callaway, B. and Sant'Anna, P. H. (2018), 'Difference-in-differences with multiple time periods and an application on the minimum wage and employment', *arXiv preprint arXiv:1803.09015* .

Chandra, A., Gruber, J. and McKnight, R. (2010), 'Patient cost-sharing and hospitalization offsets in the elderly', *American Economic Review* **100**(1), 193–213.

Chaney, T., Sraer, D. and Thesmar, D. (2012), 'The collateral channel: How real estate shocks affect corporate investment', *American Economic Review* **102**(6), 2381–2409.

Dafny, L., Duggan, M. and Ramanarayanan, S. (2012), 'Paying a premium on your premium? consolidation in the us health insurance industry', *American Economic Review* **102**(2), 1161–1185.

Dahl, G. B. and Lochner, L. (2012), 'The impact of family income on child achievement: Evidence from the earned income tax credit', *American Economic Review* **102**(5), 1927–1956.

de Chaisemartin, C. (2010), A note on instrumented difference in differences. Working paper.

de Chaisemartin, C. and D'Haultfoeuille, X. (2015), 'Fuzzy Differences-in-Differences', *ArXiv e-prints, eprint 1510.01757v2* .

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

de Chaisemartin, C., D'Haultfœuille, X. and Guyonvarch, Y. (2018), Fuzzy differences-in-differences with Stata. Working paper.

Dinkelman, T. (2011), 'The effects of rural electrification on employment: New evidence from south africa', *American Economic Review* **101**(7), 3078–3108.

Duflo, E. (2001), 'Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment', *American Economic Review* **91**(4), 795–813.

Duggan, M. and Morton, F. S. (2010), 'The effect of medicare part d on pharmaceutical prices and utilization', *American Economic Review* **100**(1), 590–607.

Duranton, G. and Turner, M. A. (2011), 'The fundamental law of road congestion: Evidence from us cities', *American Economic Review* **101**(6), 2616–2652.

Ellul, A., Pagano, M. and Panunzi, F. (2010), 'Inheritance law and investment in family firms', *American Economic Review* **100**(5), 2414–2450.

Enikolopov, R., Petrova, M. and Zhuravskaya, E. (2011), 'Media and political persuasion: Evidence from russia', *American Economic Review* **101**(7), 3253–3285.

Fang, H. and Gavazza, A. (2011), 'Dynamic inefficiencies in an employment-based health insurance system: Theory and evidence', *American Economic Review* **101**(7), 3047–77.

Faye, M. and Niehaus, P. (2012), 'Political aid cycles', *American Economic Review* **102**(7), 3516–3530.

Forman, C., Goldfarb, A. and Greenstein, S. (2012), 'The internet and local wages: A puzzle', *American Economic Review* **102**(1), 556–575.

Gentzkow, M., Shapiro, J. M. and Sinkinson, M. (2011), 'The effect of newspaper entry and exit on electoral politics', *American Economic Review* **101**(7), 2980–3018.

Goodman-Bacon, A. (2018), Difference-in-differences with variation in treatment timing, Technical report, Working Paper.

Hornbeck, R. (2012), 'The enduring impact of the american dust bowl: Short-and long-run adjustments to environmental catastrophe', *American Economic Review* **102**(4), 1477–1507.

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

Hudson, S., Hull, P. and Liebersohn, C. (2015), Interpreting instrumented difference-in-differences, Technical report, Working Paper (available upon request).

Imbens, G. W. and Angrist, J. D. (1994), ‘Identification and estimation of local average treatment effects’, *Econometrica* **62**(2), 467–475.

Imberman, S. A., Kugler, A. D. and Sacerdote, B. I. (2012), ‘Katrina’s children: Evidence on the structure of peer effects from hurricane evacuees’, *American Economic Review* **102**(5), 2048–2082.

Mian, A. and Sufi, A. (2011), ‘House prices, home equity-based borrowing, and the us household leverage crisis’, *American Economic Review* **101**(5), 2132–2156.

Moser, P. and Voena, A. (2012), ‘Compulsory licensing: Evidence from the trading with the enemy act’, *American Economic Review* **102**(1), 396–427.

Simcoe, T. (2012), ‘Standard setting committees: Consensus governance for shared technology platforms’, *American Economic Review* **102**(1), 305–336.

Wang, S.-Y. (2011), ‘State misallocation and housing prices: Theory and evidence from china’, *American Economic Review* **101**(5), 2081–2107.

Zhang, X. M. and Zhu, F. (2011), ‘Group size and incentives to contribute: A natural experiment at chinese wikipedia’, *American Economic Review* **101**(4), 1601–1605.

A Proofs

A.1 Two useful lemmas

For any random variable R, and for all $(g,g',t,t') \in \{0,...,\overline{g}\}^2 \times \{0,...,\overline{t}\}^2$, let

$$DID_R(g,g',t,t') = E(R_{g,t}) - E(R_{g,t'}) - (E(R_{g',t}) - E(R_{g',t'})).$$

Our lemma relates the $DID_Y(g,g',t,t')$ estimands to the $\text{ACR}^{TR}_{g,t}$ and $\text{ACR}^{S}_{g,t}$ parameters.

Lemma 1 *Suppose that Assumption 1 holds and $D \in \{0,,...,\overline{d}\}$.*

1. *If Assumption 2 holds, for all $(g,g',t,t') \in \{0,...,\overline{g}\}^2 \times \{0,...,\overline{t}\}^2, g \neq g', t \neq t'$,*

$$DID_Y(g,g',t,t') = E(D_{g,t})ACR^{TR}_{g,t} - E(D_{g,t'})ACR^{TR}_{g,t'} - \left(E(D_{g',t})ACR^{TR}_{g',t} - E(D_{g',t'})ACR^{TR}_{g',t'}\right).$$

2. *If Assumptions 2, 3, and 4O hold, , for all $(g,g',t) \in \{0,...,\overline{g}\}^2 \times \{1,...,\overline{t}\}, g \neq g'$,*

$$DID_Y(g,g',t,t-1) = (E(D_{g,t}) - E(D_{g,t-1}))\, ACR^{S}_{g,t} - (E(D_{g',t}) - E(D_{g',t-1}))\, ACR^{S}_{g',t}.$$

In the special case where the treatment is binary, Lemma 1 can be rewritten as follows.

Lemma 2 *Suppose that Assumption 1 holds and D is binary.*

1. *If Assumption 2 holds, for all $(g,g',t,t') \in \{0,...,\overline{g}\}^2 \times \{0,...,\overline{t}\}^2, g \neq g', t \neq t'$,*

$$DID_Y(g,g',t,t') = E(D_{g,t})\Delta^{TR}_{g,t} - E(D_{g,t'})\Delta^{TR}_{g,t'} - \left(E(D_{g',t})\Delta^{TR}_{g',t} - E(D_{g',t'})\Delta^{TR}_{g',t'}\right).$$

2. *If Assumptions 2-4 hold, for all $(g,g',t) \in \{0,...,\overline{g}\}^2 \times \{1,...,\overline{t}\}, g \neq g'$,*

$$DID_Y(g,g',t,t-1) = (E(D_{g,t}) - E(D_{g,t-1}))\, \Delta^{S}_{g,t} - (E(D_{g',t}) - E(D_{g',t-1}))\, \Delta^{S}_{g',t}.$$

Proof of Lemma 1

1. We have

$$DID_Y(g,g',t,t') = E(Y_{g,t}) - E(Y_{g,t'}) - (E(Y_{g',t}) - E(Y_{g',t'})). \tag{11}$$

Moreover,

$$E(Y_{g,t}) = E(Y_{g,t}(0)) + E[Y_{g,t}(D) - Y_{g,t}(0)]. \tag{12}$$

The result follows by decomposing similarly the three other terms of $DID_Y(g,g',t,t')$, plugging these decompositions into (11), using Assumption 2, and finally using the definition of $\text{ACR}^{TR}_{g,t}$.

2. We prove the result when $E(D_{g,t}) \geq E(D_{g,t-1})$ and $E(D_{g',t}) \geq E(D_{g',t-1})$. The proof is similar in the three other cases. First,

$$\begin{aligned}E(Y_{g,t-1}) &= E(Y_{g,t-1}(0)) + \sum_{d=1}^{\bar{d}} P(D_{g,t-1}(t-1) = d)E(Y_{g,t-1}(d) - Y_{g,t-1}(0)|D(t-1) = d) \\ &= E(Y_{g,t-1}(0)) + \sum_{d=1}^{\bar{d}} P(D_{g,t}(t-1) = d)E(Y_{g,t}(d) - Y_{g,t}(0)|D(t-1) = d). \end{aligned} \tag{13}$$

where the second equality follows from Assumptions 3 and 4O. Similarly,

$$E(Y_{g,t}) = E(Y_{g,t}(0)) + \sum_{d=1}^{\bar{d}} P(D_{g,t}(t) = d)E(Y_{g,t}(d) - Y_{g,t}(0)|D(t) = d). \tag{14}$$

Combining Equations (13) and (14) yields

$$\begin{aligned}E(Y_{g,t}) - E(Y_{g,t-1}) =& E(Y_{g,t}(0)) - E(Y_{g,t-1}(0)) \\ &+ E\left[\sum_{d=1}^{\bar{d}} (Y_{g,t}(d) - Y_{g,t}(0)) \left(\mathbb{1}\{D_{g,t}(t) = d\} - \mathbb{1}\{D_{g,t}(t-1) = d\}\right)\right]. \end{aligned} \tag{15}$$

Now, remark that

$$\begin{aligned}&\sum_{d=1}^{\bar{d}} (Y_{g,t}(d) - Y_{g,t}(0)) \left(\mathbb{1}\{D_{g,t}(t) = d\} - \mathbb{1}\{D_{g,t}(t-1) = d\}\right) \\ =& \sum_{d=1}^{\bar{d}} (Y_{g,t}(d) - Y_{g,t}(d-1)) \left(\mathbb{1}\{D_{g,t}(t) \geq d\} - \mathbb{1}\{D_{g,t}(t-1) \geq d\}\right) \\ =& \sum_{d=1}^{\bar{d}} (Y_{g,t}(d) - Y_{g,t}(d-1)) \mathbb{1}\{D_{g,t}(t) \geq d > D_{g,t}(t-1)\}. \end{aligned} \tag{16}$$

The first equality follows by summation by parts. The second uses the fact that under Assumption 3, $E(D_{g,t}) \geq E(D_{g,t-1})$ implies that $D_{g,t}(t) \geq D_{g,t}(t-1)$. Then,

$$\begin{aligned}&E(Y_{g,t}) - E(Y_{g,t-1}) - (E(Y_{g,t}(0)) - E(Y_{g,t-1}(0))) \\ =& \sum_{d=1}^{\bar{d}} P(D_{g,t}(t) \geq d > D_{g,t}(t-1)) E\left(Y_{g,t}(d) - Y_{g,t}(d-1)|D_{g,t}(t) \geq d > D_{g,t}(t-1)\right) \\ =& \sum_{d=1}^{\bar{d}} \left[P(D_{g,t} \geq d) - P(D_{g,t-1} \geq d)\right] E\left(Y_{g,t}(d) - Y_{g,t}(d-1)|D_{g,t}(t) \geq d > D_{g,t}(t-1)\right) \\ =& (E(D_{g,t}) - E(D_{g,t-1}))\text{ACR}^S_{g,t}. \end{aligned} \tag{17}$$

The first equality follows from plugging Equation (16) into Equation (15). The second uses the fact that under Assumption 3, $E(D_{g,t}) \geq E(D_{g,t-1})$ implies that $D_{g,t}(t) \geq D_{g,t}(t-1)$.

One can follow the same steps to show that

$$E(Y_{g',t}) - E(Y_{g',t-1}) - (E(Y_{g',t}(0)) - E(Y_{g',t-1}(0))) = (E(D_{g',t}) - E(D_{g',t-1}))\text{ACR}^S_{g',t}. \quad (18)$$

Finally, the result follows by taking the difference between Equations (17) and (18), and using Assumption 2.

Proof of Lemma 2

The first statement of the lemma follows from the first statement of Lemma 1, once noted that when D is binary, $ACR^{TR}_{g,t} = \Delta^{TR}_{g,t}$ for all $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$. The second statement follows from the second statement of Lemma 1, once noted that when D is binary, Assumptions 4 and 4O are the same, and $ACR^S_{g,t} = \Delta^S_{g,t}$ for all $(g,t) \in \{0,...,\overline{g}\} \times \{1,...,\overline{t}\}$.

A.2 Proof of Theorem 1

Proof of the first statement.

Proof for the fixed-effect regression.

It follows from the Frisch-Waugh theorem and the definition of $\varepsilon_{fe,G,T}$ that

$$\beta_{fe} = \frac{\text{cov}(\varepsilon_{fe,G,T}, Y)}{\text{cov}(\varepsilon_{fe,G,T}, E(D|G,T))}. \quad (19)$$

Note that the denominator must be different from 0, otherwise β_{fe} would not be well defined. As a result,

$$E(\varepsilon_{fe,G,T}E(D|G,T)) = \text{cov}(\varepsilon_{fe,G,T}, E(D|G,T)) \neq 0. \quad (20)$$

Now, by definition of $\varepsilon_{fe,G,T}$ again,

$$E[\varepsilon_{fe,G,T}|G] = 0,\ E[\varepsilon_{fe,G,T}|T] = 0. \quad (21)$$

Then,

$$\begin{aligned}
\text{cov}(\varepsilon_{fe,G,T}, Y) &= E[\varepsilon_{fe,G,T}E(Y|G,T)] \\
&= E\left[\varepsilon_{fe,G,T}DID_Y(G,0,T,0)\right] \\
&= E\left[\varepsilon_{fe,G,T}\left(E(D|G,T)\Delta^{TR}_{G,T} - E(D_{G,0})\Delta^{TR}_{G,0} - E(D_{0,T})\Delta^{TR}_{0,T} + E(D_{0,0})\Delta^{TR}_{0,0}\right)\right] \\
&= E\left[\varepsilon_{fe,G,T}E(D|G,T)\Delta^{TR}_{G,T}\right] \\
&= E\left[\varepsilon_{fe,G,T}D\Delta^{TR}_{G,T}\right]. \quad (22)
\end{aligned}$$

The first equality follows from the law of iterated expectations. (21) implies that

$$E\left[\varepsilon_{fe,G,T}(-E(Y_{G,0})-E(Y_{0,T})+E(Y_{0,0}))\right]=0,$$

hence the second equality. The third equality follows from the first point of Lemma 2. (21) implies that

$$E\left[\varepsilon_{fe,G,T}(-E(D_{G,0})\Delta^{TR}_{G,0}-E(D_{0,T})\Delta^{TR}_{0,T}+E(D_{0,0})\Delta^{TR}_{0,0})\right]=0,$$

hence the fourth equality. The fifth equality follows from the law of iterated expectations. Combining (19), (20), and (22), we obtain

$$\begin{aligned}\beta_{fe}&=\frac{E\left[\varepsilon_{fe,G,T}D\Delta^{TR}_{G,T}\right]}{E(\varepsilon_{fe,G,T}E(D|G,T))}\\&=\frac{E\left[v^{TR}_{fe,G,T}D\Delta^{TR}_{G,T}\right]}{E(D)}\\&=E\left[v^{TR}_{fe,G,T}\Delta^{TR}_{G,T}|D=1\right],\end{aligned}$$

where the second equality follows from the definition of $v^{TR}_{fe,G,T}$.

Proof for the first-difference regression.

With a slight abuse of notation, let, $E(Y_{G,T})$ and $E(D_{G,T})$ denote respectively $E(Y_{g,t})$ and $E(D_{g,t})$ evaluated at $(g,t)=(G,T)$.

It follows from Frisch-Waugh theorem and the definition of $\varepsilon_{fd,G,T}$ that

$$\beta_{fd}=\frac{\text{cov}(\varepsilon_{fd,G,T},FD_Y(G,T)|T\geq 1)}{\text{cov}(\varepsilon_{fd,G,T},FD_D(G,T)|T\geq 1)}. \quad (23)$$

Note that the denominator must be different from 0, otherwise β_{fd} would not be well defined. Now, by definition of $\varepsilon_{fd,G,T}$ again,

$$E[\varepsilon_{fd,G,T}|T]=0. \quad (24)$$

Then,

$$\begin{aligned}
&\text{cov}(\varepsilon_{fd,G,T}, FD_Y(G,T)|T\geq 1)\\
=&E[\varepsilon_{fd,G,T}DID_Y(G,0,T,T-1)|T\geq 1]\\
=&E\left[\varepsilon_{fd,G,T}\left(E(D_{G,T})\Delta^{TR}_{G,T}-E(D_{G,T-1})\Delta^{TR}_{G,T-1}-E(D_{0,T})\Delta^{TR}_{0,T}+E(D_{0,T-1})\Delta^{TR}_{0,T-1}\right)|T\geq 1\right]\\
=&E\left[\varepsilon_{fd,G,T}\left(E(D_{G,T})\Delta^{TR}_{G,T}-E(D_{G,T-1})\Delta^{TR}_{G,T-1}\right)|T\geq 1\right]\\
=&\sum_{g=0}^{\overline{g}}\sum_{t=1}^{\overline{t}}E(\varepsilon_{fd,G,T}1\{G=g\}1\{T=t\}|T\geq 1)\left(E(D_{g,t})\Delta^{TR}_{g,t}-E(D_{g,t-1})\Delta^{TR}_{g,t-1}\right)\\
=&\sum_{g=0}^{\overline{g}}\sum_{t=0}^{\overline{t}}E\left[\varepsilon_{fd,G,T}1\{G=g\}\left(1\{T=t\}-1\{T=t+1\}\right)|T\geq 1\right]E(D_{g,t})\Delta^{TR}_{g,t}\\
=&\sum_{g=0}^{\overline{g}}\sum_{t=0}^{\overline{t}}\left(\varepsilon_{fd,g,t}\frac{P(G=g,T=t)}{P(T\geq 1)}-\varepsilon_{fd,g,t+1}\frac{P(G=g,T=t+1)}{P(T\geq 1)}\right)E(D_{g,t})\Delta^{TR}_{g,t}\\
=&\frac{1}{P(T\geq 1)}\sum_{g=0}^{\overline{g}}\sum_{t=0}^{\overline{t}}P(G=g,T=t)\widetilde{v}^{TR}_{fd,g,t}E(D_{g,t})\Delta^{TR}_{g,t}\\
=&\frac{1}{P(T\geq 1)}E\left[\widetilde{v}^{TR}_{fd,G,T}D\Delta^{TR}_{G,T}\right].
\end{aligned}$$

Eq. (24) implies that

$$E[\varepsilon_{fd,G,T}(E(Y_{0,T})-E(Y_{0,T-1}))|T\geq 1]=0,$$

hence the first equality. The second equality follows from the first point of Lemma 2. Eq. (24) implies again that

$$E[\varepsilon_{fd,G,T}(E(D_{0,T})\Delta^{TR}_{0,T}-E(D_{0,T-1})\Delta^{TR}_{0,T-1})|T\geq 1]=0,$$

hence the third equality. The fifth equality follows from a summation by part, the sixth holds because $\varepsilon_{fd,g,0}=0$, the seventh follows from the definition of $\tilde{v}^{TR}_{fd,g,t}$, and the eighth from the law of iterated expectations.

A similar reasoning yields

$$\text{cov}(\varepsilon_{fd,G,T}, FD_D(G,T)|T\geq 1)=\frac{1}{P(T\geq 1)}E\left[\widetilde{v}^{TR}_{fd,G,T}D\right].$$

The end of the proof follows exactly as for $k=fe$.

Proof of the second statement.

We first prove Equation (4) and two other equalities that we use in the proof. For all $(g,t)\in\{0,...,\overline{g}\}\times\{1,...,\overline{t}\}$, we have

$$\begin{aligned}
|E(D_{g,t})-E(D_{g,t-1})|=&|P(D(t)=1|G=g,T=t)-P(D(t-1)=1|G=g,T=t-1)|\\
=&|P(D(t)=1|G=g,T=t)-P(D(t-1)=1|G=g,T=t)|\\
=&P(S|G=g,T=t),
\end{aligned}$$

where the equalities follow from Assumption 3. Then,

$$\begin{aligned}E[f(G,T)1_S] &= E[f(G,T)P(S|G,T)]\\ &= \sum_{g=0}^{\bar{g}}\sum_{t=1}^{\bar{t}} P(G=g,T=t)\left|E(D_{g,t})-E(D_{g,t-1})\right| f(g,t).\end{aligned} \tag{25}$$

The first equality follows from the law of iterated expectations, and the second follows from the fact that $P(S|G=g,T=0)=0$ and from the previous display. Similarly,

$$P(S)=\sum_{g=0}^{\bar{g}}\sum_{t=1}^{\bar{t}} P(G=g,T=t)\left|E(D_{g,t})-E(D_{g,t-1})\right|. \tag{26}$$

(4) follows by dividing (25) by (26).

Proof for the fixed-effect regression.

We have

$$\begin{aligned}
&\operatorname{cov}(\varepsilon_{fe,G,T},Y)\\
=&E[\varepsilon_{fe,G,T}(E(Y|G,T)-E(Y|G=0,T))]\\
=&E\left(\sum_{t=0}^{\bar{t}} E(\varepsilon_{fe,G,T}1\{T=t\}|G)\left[E(Y|G,T=t)-E(Y|G=0,T=t)\right]\right)\\
=&E\left(\sum_{t=1}^{\bar{t}}\left[\sum_{t'\geq t} E(\varepsilon_{fe,G,T}1\{T=t'\}|G)\right] DID_Y(G,0,t,t-1)\right)\\
=&E\left(\sum_{t=1}^{\bar{t}} E(\varepsilon_{fe,G,T}1\{T\geq t\}|G)\left[FD_D(G,t)\Delta^S_{G,t}-FD_D(0,t)\Delta^S_{0,t}\right]\right)\\
=&E\left(\sum_{t=1}^{\bar{t}} E(\varepsilon_{fe,G,T}1\{T\geq t\}|G)FD_D(G,t)\Delta^S_{G,t}\right)-\sum_{t=1}^{\bar{t}} E(\varepsilon_{fe,G,T}1\{T\geq t\})FD_D(0,t)\Delta^S_{0,t}\\
=&\sum_{g=0}^{\bar{g}}\sum_{t=1}^{\bar{t}} P(G=g,T=t)|FD_D(g,t)|\frac{s_{g,t}E[\varepsilon_{fe,G,T}1\{G=g,T\geq t\}]}{P(G=g,T=t)}\Delta^S_{g,t} &(27)\\
=&E\left[\widetilde{v}^S_{fe,G,T}\Delta^S_{G,T}1_S\right]. &(28)
\end{aligned}$$

The first equality follows from the law of iterated expectation and (21). The third equality follows from summation by part and (21). The fourth equality follows from the second point of Lemma 2. (27) follows from (21) and some algebra. (28) follows from the definition of $\widetilde{v}^S_{fe,g,t}$ and from (25).

Similarly,

$$
\begin{aligned}
& \operatorname{cov}(\varepsilon_{fe,G,T}, E(D|G,T)) \\
=& \sum_{g=0}^{\overline{g}} \sum_{t=1}^{\overline{t}} P(G=g,T=t)|FD_D(g,t)| \frac{s_{g,t} E[\varepsilon_{fe,G,T} 1\{G=g, T \geq t\}]}{P(G=g,T=t)} \qquad (29) \\
=& E\left[\widetilde{v}^S_{fe,G,T} 1_S\right]. \qquad (30)
\end{aligned}
$$

Combining (19), (28), and (30) yields

$$\beta_{fe} = \frac{E\left[\widetilde{v}^S_{fe,G,T} \Delta^S_{G,T} 1_S\right]}{E\left[\widetilde{v}^S_{fe,G,T} 1_S\right]}.$$

Finally, the result follows from the definition of $v^S_{fe,G,T}$, after dividing the numerator and the denominator in the right-hand side of the previous display by $P(S)$.

Proof for the first-difference regression.

First,

$$
\begin{aligned}
\operatorname{cov}(\varepsilon_{fd,G,T}, FD_Y(G,T)|T \geq 1) =& E[\varepsilon_{fd,G,T} DID_Y(G,0,T,T-1)|T \geq 1] \\
=& E[\varepsilon_{fd,G,T} \left(FD_D(G,T)\Delta^S_{G,T} - FD_D(0,T)\Delta^S_{0,T}\right)|T \geq 1] \\
=& E[\varepsilon_{fd,G,T} FD_D(G,T)\Delta^S_{G,T}|T \geq 1] \\
=& E[\varepsilon_{fd,G,T} s_{G,T}|FD_D(G,T)|\Delta^S_{G,T}|T \geq 1] \\
=& \frac{E[\varepsilon_{fd,G,T} s_{G,T} \Delta^S_{G,T} 1_S]}{P(T \geq 1)}.
\end{aligned}
$$

The first equality follows from (24). The second equality follows from the second point of Lemma 2. The third equality follows from (24). The fifth equality follows from (25).

Similarly,

$$\operatorname{cov}(\varepsilon_{fd,G,T}, FD_D(G,T)|T \geq 1) = \frac{E[\varepsilon_{fd,G,T} s_{G,T} 1_S]}{P(T \geq 1)}.$$

The result follows combining (23) and the two previous displays.

A.3 Proof of Proposition 1

1. We have

$$P(v^{TR}_{fe,G,T} \geq 0|D=1) = \sum_{g=0}^{\overline{g}} \sum_{t=0}^{\overline{t}} 1\{v^{TR}_{fe,g,t} \geq 0\} P(G=g,T=t|D=1).$$

Therefore, $P(v^{TR}_{fe,G,T} \geq 0|D=1) = 1$ is equivalent to having that $v^{TR}_{fe,g,t} P(D=1|G=g,T=t) \geq 0$ for all $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$. A few lines of algebra show that $E(\varepsilon_{fe,G,T} E(D|G,T)) =$

$V(\varepsilon_{fe,G,T})$, which is strictly positive if β_{fe} is well-defined. Then, to study the sign of $v^{TR}_{fe,g,t}P(D=1|G=g,T=t)$, it is sufficient to study the sign of $\varepsilon_{fe,g,t}1\{E(D_{g,t})>0\}$. Under Assumption 6, $G \perp\!\!\!\perp T$. If $G \perp\!\!\!\perp T$, one can show that

$$\varepsilon_{fe,g,t} = E(D_{g,t}) - E(D_{g,.}) - E(D_{.,t}) + E(D). \tag{31}$$

Then, Assumptions 5 and 6 imply that for all $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$,

$$\begin{aligned}&\varepsilon_{fe,g,t}1\{E(D_{g,t})>0\}\\ =&1\{t\geq a_g\}\left(1-\frac{\overline{t}+1-a_g}{\overline{t}+1}-\frac{1}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}\leq t\}+\frac{1}{\overline{t}+1}\sum_{t'=0}^{\overline{t}}\frac{1}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}\leq t'\}\right).\end{aligned}$$

Thus, $\varepsilon_{fe,g,t}1\{E(D_{g,t})>0\}=0$ for all (g,t) such that $t<a_g$. For all (g,t) such that $t \geq a_g$, $t \mapsto \varepsilon_{fe,g,t}1\{E(D_{g,t})>0\}$ is decreasing in t. Moreover, $g \mapsto \varepsilon_{fe,g,\overline{t}}$ is minimized at $g=g(e)$. Therefore, $\varepsilon_{fe,g,t}1\{E(D_{g,t})>0\}$ is strictly negative for some (g,t) if and only if

$$\frac{a_{g(e)}}{\overline{t}+1}-\frac{1}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}\leq \overline{t}\}+\frac{1}{\overline{t}+1}\sum_{t'=0}^{\overline{t}}\frac{1}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}\leq t'\}<0.$$

2. Here as well, $P(v^{TR}_{fd,G,T}\geq 0|D=1)=1$ is equivalent to having that $v^{TR}_{fd,g,t}P(D=1|G=g,T=t)\geq 0$ for all $(g,t)\in\{0,...,\overline{g}\}\times\{0,...,\overline{t}\}$. Moreover, under the assumptions of the proposition, to study the sign of $v^{TR}_{fd,g,t}P(D=1|G=g,T=t)$, it is sufficient to study the sign of $u_{g,t}=\tilde{v}^{TR}_{fd,g,t}1\{E(D_{g,t})>0\}$. For all $(g,t)\in\{0,...,\overline{g}\}\times\{0,...,\overline{t}\}$,

$$\begin{aligned}&u_{g,t}\\ =&1\{t\geq a_g\}(\varepsilon_{fd,g,t}-\varepsilon_{fd,g,t+1})\\ =&1\{t\geq a_g\}\left[1\{t\geq 1\}(E(D_{g,t})-E(D_{g,t-1})-E(D_{.,t})+E(D_{.,t-1}))\right.\\ &\left.-1\{t\leq \overline{t}-1\}(E(D_{g,t+1})-E(D_{g,t})-E(D_{.,t+1})+E(D_{.,t}))\right]\\ =&1\{t\geq a_g\}\left[1\{t\geq 1\}1\{t=a_g\}-\frac{1\{t\geq 1\}}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}=t\}+\frac{1\{t\leq\overline{t}-1\}}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}=t+1\}\right].\end{aligned}$$

The first equality follows from the definition of $\tilde{v}^{TR}_{fd,g,t}$, and from Assumptions 6 and 5. The second equality follows from the fact that for all $(g,t)\in\{0,...,\overline{g}\}\times\{0,...,\overline{t}+1\}$,

$$\varepsilon_{fd,g,t}=1\{1\leq t\leq\overline{t}\}(E(D_{g,t})-E(D_{g,t-1})-E(D_{.,t})+E(D_{.,t-1})).$$

The third equality follows from Assumptions 5 and 6.

If $a_g\geq\overline{t}$ for every $g\in\{0,...,\overline{g}\}$, it follows from the previous display that $v_{g,t}\geq 0$ for every $(g,t)\in\{0,...,\overline{g}\}\times\{0,...,\overline{t}\}$. Conversely, assume that at least one group adopts before $\overline{t}$. Then,

$a_{g(e)} = t_0 < \bar{t}$. Now, we reason by contradiction. Assume that $u_{g(e),t_0+1}, ..., u_{g(e),\bar{t}} \geq 0$. $u_{g(e),\bar{t}} \geq 0$ implies $\frac{1}{\bar{g}+1}\sum_{g'=0}^{\bar{g}} 1\{a_{g'} = \bar{t}\} = 0$. Then, $u_{g(e),\bar{t}-1} \geq 0$ implies $\frac{1}{\bar{g}+1}\sum_{g'=0}^{\bar{g}} 1\{a_{g'} = \bar{t}-1\} = 0$. And so on and so forth. Finally, $v_{g(e),t_0+1} \geq 0$ implies $\frac{1}{\bar{g}+1}\sum_{g'=0}^{\bar{g}} 1\{a_{g'} = t_0+1\} = 0$. Therefore, all groups must have $a_g = a_{g(e)}$. But then, $E(D_{g,t}) - E(D_{g,t-1}) = 1\{t = a_{g(e)}\}$ if $a_{g(e)} \geq 1$, and $E(D_{g,t}) - E(D_{g,t-1}) = 0$ otherwise. This contradicts the fact that β_{fd} is well-defined. Therefore, at least one of the $(u_{g(e),t_0+1}, ..., u_{g(e),\bar{t}})$ must be strictly negative.

3. It follows from (30) and from $\text{cov}(\varepsilon_{fe,G,T}, E(D|G,T)) = V(\varepsilon_{fe,G,T})$ that under the assumptions of the proposition, $E(\tilde{v}^S_{fe,G,T}|S) > 0$, so $P(v^S_{fe,G,T} \geq 0|D=1) = 1$ is equivalent to having that $\tilde{v}^S_{fe,g,t}P(S|G=g,T=t) \geq 0$ for all $(g,t) \in \{0,...,\bar{g}\} \times \{0,...,\bar{t}\}$. Because $P(S|G=g,T=t) \neq 0$ only if $t = a_g$, the sign of $\tilde{v}^S_{fe,g,t}P(S|G=g,T=t)$ is not zero only if $t = a_g$. Moreover, since $G \perp\!\!\!\perp T$ and $s_{g,a_g} = 1$, this sign is then equal to that of $E[\varepsilon_{fe,g,T}|T \geq a_g]$. Then, using (31), we have, for all $g \in \{0,...,\bar{g}\}$ such that $a_g \leq \bar{t}$,

$$
\begin{aligned}
(\bar{t}+1-a_g)E[\varepsilon_{fe,g,T}|T \geq a_g] =& \sum_{t=a_g}^{\bar{t}} \left[E(D_{g,t}) - E(D_{g,.}) - E(D_{.,t}) + E(D)\right] \\
=&(\bar{t}+1-a_g)(1+E(D)) - \frac{(\bar{t}+1-a_g)^2}{\bar{t}+1} - \sum_{t=a_g}^{\bar{t}} E(D_{.,t}) \\
=&(\bar{t}+1-a_g)\left(E(D) + \frac{a_g}{\bar{t}+1}\right) - \sum_{t=a_g}^{\bar{t}} E(D_{.,t}) \\
=&a_g\left[1 - E(D) - \frac{a_g}{\bar{t}+1}\right] + \sum_{t=0}^{a_g-1} E(D_{.,t}) \\
=&a_g\left[1 - \frac{a_g}{\bar{t}+1}\right] + \sum_{t=0}^{\bar{t}} E(D_{.,t})\left(1\{t < a_g\} - \frac{a_g}{\bar{t}+1}\right).
\end{aligned}
$$

This last expression is minimized when $E(D_{.,t}) = 0$ when $t < a_g$, $E(D_{.,t}) = 1$ otherwise. In such a case,

$$
(\bar{t}+1-a_g)E[\varepsilon_{fe,g,T}|T \geq a_g] = a_g\left[1 - \frac{a_g}{\bar{t}+1}\right] + \sum_{t=a_g}^{\bar{t}} \frac{-a_g}{\bar{t}+1} = 0.
$$

4. Here as well, $P(v^S_{fd,G,T} \geq 0|S) = 1$ is equivalent to having that $v^S_{fd,g,t}P(S|G=g,T=t) \geq 0$ for all $(g,t) \in \{0,...,\bar{g}\} \times \{1,...,\bar{t}\}$. Moreover, to study the sign of $v^S_{fd,g,t}P(S|G=g,T=t)$, it is sufficient to study the sign of $s_{g,t}\varepsilon_{fd,g,t}P(S|G=g,T=t)$. For all $(g,t) \in \{0,...,\bar{g}\} \times \{1,...,\bar{t}\}$,

$$
\begin{aligned}
&s_{g,t}\varepsilon_{fd,g,t} \\
=&s_{g,t}\left[E(D_{g,t}) - E(D_{g,t-1}) - E(D_{.,t}) + E(D_{.,t-1})\right] \\
=&1\{t = a_g\}\left[1 - E(D_{.,t}) + E(D_{.,t-1})\right] \geq 0.
\end{aligned}
$$

The second equality follows from Assumption 5. This proves the result.

A.4 Proof of Proposition 2

First, notice that under Assumption 7, $S = \{D = 1\}$ and $s_{g,1} = 1$ for all g. Therefore, for all $(k, g) \in \{fe, fd\} \times \{0, ..., \overline{g}\}$,

$$v^{TR}_{k,g,1} = v^{S}_{k,g,1} = \frac{\varepsilon_{k,g,1} E(D)}{E(D\varepsilon_{k,G,T})}.$$

Moreover,

$$P(v^{S}_{k,G,T} < 0|S) = P(v^{TR}_{k,G,T} < 0|D = 1, T = 1) = P(v^{TR}_{k,G,T} < 0|D = 1).$$

Let us reason by contradiction and suppose that $P(v^{TR}_{k,G,T} < 0|D = 1, T = 1) = 0$. Let us also suppose that $E(D\varepsilon_{k,G,T}) > 0$ (the proof is symmetric if instead $E(D\varepsilon_{k,G,T}) < 0$). Then:

$$\begin{aligned}
0 &= P(v^{TR}_{k,G,T} < 0|D = 1, T = 1) \\
&= P(\varepsilon_{k,G,1} < 0|D = 1, T = 1) \\
&= \frac{E\left[D1\{\varepsilon_{k,G,1} < 0\}|T = 1\right]}{E(D_{.,1})} \\
&= \frac{E\left[E(D|G, T = 1)1\{\varepsilon_{k,G,1} < 0\}|T = 1\right]}{E(D_{.,1})} \\
&\geq \frac{\left[\min_g E(D_{g,1})\right] P(\varepsilon_{k,G,1} < 0|T = 1)}{E(D_{.,1})}.
\end{aligned}$$

The second equality follows from the fact that if $E(D\varepsilon_{k,G,T}) > 0$ and conditional on $T = 1$, $v^{TR}_{k,G,T} < 0$ is equivalent to having $\varepsilon_{k,G,1} < 0$. The fourth equality follows from the law of iterated expectations. The fifth equality follows from the fact that $E(D|G, T = 1) \geq \min_g E(D_{g,1})$.

The previous display implies that $P(\varepsilon_{k,G,1} < 0|T = 1) = 0$. By definition of $\varepsilon_{k,g,t}$, $E(\varepsilon_{k,G,1}|T = 1) = 0$. Hence, $P(\varepsilon_{k,G,1} = 0|T = 1) = 1$. But then

$$E(D\varepsilon_{k,G,T}) = E(D\varepsilon_{k,G,1}1\{T = 1\}) = 0.$$

This implies that β_{fe} is not well defined, a contradiction.

A.5 Proof of Corollary 1

We prove only the first statement, as the second statement can be proven by following the exact same steps. If the assumptions of the corollary hold and if $\Delta^{TR} = 0$, then

$$\begin{cases} \beta_k &= E\left[v^{TR}_{k,G,T}\Delta^{TR}_{G,T}|D = 1\right], \\ 0 &= E\left[\Delta^{TR}_{G,T}|D = 1\right]. \end{cases}$$

These two conditions and the Cauchy-Schwarz inequality imply

$$|\beta_k| = \left|\text{cov}\left(v^{TR}_{k,G,T}, \Delta^{TR}_{G,T}\middle| D=1\right)\right| \leq V\left(v^{TR}_{k,G,T}|D=1\right)^{1/2} V\left(\Delta^{TR}_{G,T}|D=1\right)^{1/2}.$$

Hence, $\sigma^{TR} \geq \underline{\sigma}^{TR}_k$.

Now, we prove that we can rationalize this lower bound. Let us define

$$\Delta^{TR}_{G,T} = \frac{\beta_k\left(v^{TR}_{k,G,T}-1\right)}{V\left(v^{TR}_{k,G,T}|D=1\right)}.$$

Then

$$\Delta^{TR} = E\left[\Delta^{TR}_{G,T}|D=1\right] = \frac{\beta_k}{V\left(v^{TR}_{k,G,T}|D=1\right)} E\left[v^{TR}_{k,G,T}-1|D=1\right] = 0.$$

Similarly,

$$\begin{aligned}
E[v^{TR}_{k,G,T}\Delta^{TR}_{G,T}|D=1] &= \frac{\beta_k}{V\left(v^{TR}_{k,G,T}|D=1\right)} E\left[{v^{TR}_{k,G,T}}^2 - v^{TR}_{k,G,T}|D=1\right] \\
&= \frac{\beta_k}{V\left(v^{TR}_{k,G,T}|D=1\right)} V\left(v^{TR}_{k,G,T}|D=1\right) \\
&= \beta_k.
\end{aligned}$$

This proves the result.

A.6 Proof of Theorem 2

For all $(g^*, t) \in \{-1, 0, 1\} \times \{1, ..., \bar{t}\}$, let $\mathcal{G}_{g^*,t} = \{g : \text{sgn}\left(FD_D(g,t)\right) = g^*\}$. First, note that for all $t \geq 1$,

$$E(Y|G^*_t = 1, T = t) = \sum_{g \in \mathcal{G}_{1,t}} r(g|1,t,t) E(Y_{g,t}).$$

Similarly,

$$\begin{aligned}
E(QY|G^*_t = 1, T = t-1) &= \sum_{g \in \mathcal{G}_{1,t}} r(g|1,t,t-1) \frac{r(g|1,t,t)}{r(g|1,t,t-1)} E(Y_{g,t-1}) \\
&= \sum_{g \in \mathcal{G}_{1,t}} r(g|1,t,t) E(Y_{g,t-1}).
\end{aligned}$$

Therefore,

$$E(Y|G^*_t = 1, T = t) - E(QY|G^*_t = 1, T = t-1) = \sum_{g \in \mathcal{G}_{1,t}} r(g|1,t,t)[E(Y_{g,t}) - E(Y_{g,t-1})]. \quad (32)$$

Moreover,

$$
\begin{aligned}
&E(Y_{g,t}) - E(Y_{g,t-1}) \\
=&E(Y_{g,t}(1) - Y_{g,t}(0)|S)P(S_{g,t}) + E(Y_{g,t}(1)|D(t-1) = 1)P(D(t-1) = 1|G = g, T = t) \\
-&E(Y_{g,t-1}(1)|D(t-1) = 1)P(D(t-1) = 1|G = g, T = t-1) \\
+&E(Y_{g,t}(0)|D(t-1) = 0)P(D(t-1) = 0|G = g, T = t) \\
-&E(Y_{g,t-1}(0)|D(t-1) = 0)P(D(t-1) = 0|G = g, T = t-1) \\
=&\Delta^S_{g,t}P(S_{g,t}) + E(Y_{g,t}(1) - Y_{g,t-1}(1)|D(t-1) = 1)E(D_{g,t-1}) \\
&+ E(Y_{g,t}(0) - Y_{g,t-1}(0)|D(t-1) = 0)(1 - E(D_{g,t-1})). \qquad (33)
\end{aligned}
$$

where the first equality follows from Assumption 3.3 and the second from Assumption 3.1 and 3.2.

Now, by Assumption 10, $\delta_{d,t}$ is well-defined for all $d \in \text{Supp}(D|G^*_t = 1, T = t-1)$. Moreover, by a similar reasoning as that used to obtain (32),

$$
\begin{aligned}
E(Q\delta_{D,t}|G^*_t = 1, T = t-1) &= E(Q(D\delta_{1,t} + (1-D)\delta_{0,t}|G^*_t = 1, T = t-1) \\
&= \sum_{g\in\mathcal{G}_{1,t}} r(g|1,t,t)[E(D_{g,t-1})\delta_{1,t} + (1 - E(D_{g,t-1}))\delta_{0,t}]. \qquad (34)
\end{aligned}
$$

Next, we have, for any $g \in \mathcal{G}_{1,t}$,

$$
\begin{aligned}
\delta_{d,t} &= E(Y|D = d, G^*_t = 0, T = t) - E(Q_dY|D = d, G^*_t = 0, T = t-1) \\
&= \sum_{g'\in\mathcal{G}_{0,t}} r_d(g'|0,t,t)E(Y_{g',t}(d) - Y_{g',t-1}(d)|D(t-1) = d) \\
&= \sum_{g'\in\mathcal{G}_{0,t}} r_d(g'|0,t,t)E(Y_{g,t}(d) - Y_{g,t-1}(d)|D(t-1) = d) \\
&= E(Y_{g,t}(d) - Y_{g,t-1}(d)|D(t-1) = d).
\end{aligned}
$$

The first equality follows from a similar reasoning as that used to obtain (32). The third follows from Assumption 2'.

Therefore, in view of (32), (33) and (34), we obtain

$$
\begin{aligned}
&E(Y|G^*_t = 1, T = t) - E(QY|G^*_t = 1, T = t-1) \\
=&\sum_{g\in\mathcal{G}_{1,t}} r(g|1,t,t)\Delta^S_{g,t}P(S_{g,t}) + E(Q\delta_{D,t}|G^*_t = 1, T = t-1).
\end{aligned}
$$

Rearranging the previous display and using the definition of $CFD^*(1,t)$ yields

$$CFD^*(1,t) = \sum_{g\in\mathcal{G}_{1,t}} r(g|1,t,t)P(S_{g,t})\Delta^S_{g,t}.$$

Then, because $P(G_t^*=1,T=t)r(g|1,t,t)=P(G=g,T=t)$ for all $g\in\mathcal{G}_{1,t}$,

$$\omega_{1,t}\frac{CFD^*(1,t)}{FD_D^*(1,t)}=\frac{\sum_{g\in\mathcal{G}_{1,t}}P(G=g,T=t)P(S_{g,t})\Delta^S_{g,t}}{\sum_{t'=1}^{\overline{t}}\sum_{g'\in\{-1,1\}}g'FD_D^*(g',t')P(G_t^*=g',T=t')}.$$

Reasoning as above, we obtain

$$\begin{aligned}CFD^*(-1,t)&=-\sum_{g\in\mathcal{G}_{-1,t}}r(g|-1,t,t)P(S_{g,t})\Delta^S_{g,t},\\ FD_D^*(1,t)&=\sum_{g\in\mathcal{G}_{1,t}}r(g|1,t,t)P(S_{g,t}),\\ FD_D^*(-1,t)&=-\sum_{g\in\mathcal{G}_{-1,t}}r(g|-1,t,t)P(S_{g,t}).\end{aligned}$$

Combining all these and the fact that $P(S_{g,t})=0$ for all $g\in\mathcal{G}_{0,t}$ yields

$$\sum_{g\in\{-1,1\}}\omega_{g,t}\frac{CFD^*(g,t)}{FD_D^*(g,t)}=\frac{\sum_{g=0}^{\overline{g}}P(G=g,T=t)P(S_{g,t})\Delta^S_{g,t}}{\sum_{t'=1}^{\overline{t}}\sum_{g'=0}^{\overline{g}}P(G=g',T=t)P(S_{g',t})}.$$

The result follows by summing over $t\in\{1,...,\overline{t}\}$ and by definition of Δ^S.

A.7 Proof of Theorem 3

The reasoning is exactly the same as in Theorem 1, except that we rely on Lemma 1 instead of Lemma 2.

A.8 Proof of Theorem 4

Proof of the first statement for β_{fe}.

First, remark that β_{fe}^X is the coefficient of $E(D|G,T)$ in the regression of $Y-X'\gamma_{fe}$ on group and time fixed effects and $E(D|G,T)$. Therefore, by the Frisch-Waugh theorem,

$$\beta_{fe}^X=\frac{\text{cov}(\varepsilon_{fe,G,T},Y-X'\gamma_{fe})}{\text{cov}(\varepsilon_{fe,G,T},E(D|G,T))}.$$

Then, reasoning as in the proof of Theorem 1, we obtain

$$\text{cov}(\varepsilon_{fe,G,T},Y-X'\gamma_{fe})=E\left[\varepsilon_{fe,G,T}DID_{Y-X'\gamma_{fe}}(G,0,T,0)\right].$$

Now, under Assumptions 2X$_{fe}$ and 11, we can follow the same steps as those used to establish the first point of Lemma 2 to show that

$$DID_{Y-X'\gamma_{fe}}(g,0,t,0)=E(D_{g,t})\Delta^{TR}_{g,t}-E(D_{g,0})\Delta^{TR}_{g,0}-\left(E(D_{0,t})\Delta^{TR}_{0,t}-E(D_{0,0})\Delta^{TR}_{0,0}\right).$$

Then, the proof follows exactly as that of the first statement of Theorem 1 for β_{fe}.

Proof of the first statement for β_{fd}.

First, remark that β^X_{fd} is the coefficient of $FD_D(G,T)$ in the regression of $FD_{Y-X'\gamma_{fd}}(G,T)$ on time fixed effects and $FD_D(G,T)$. Therefore, by the Frisch-Waugh theorem,

$$\beta^X_{fd} = \frac{\text{cov}(\varepsilon_{fd,G,T}, FD_{Y-X'\gamma_{fd}}(G,T)|T \geq 1)}{\text{cov}(\varepsilon_{fd,G,T}, FD_D(G,T)|T \geq 1)}.$$

Then, reasoning as in the proof of Theorem 1, we obtain

$$\text{cov}(\varepsilon_{fd,G,T}, FD_{Y-X'\gamma_{fd}}(G,T)|T \geq 1) = E\left[\varepsilon_{fd,G,T} DID_{Y-X'\gamma_{fd}}(G,0,T,T-1)|T \geq 1\right].$$

Now, as above with $k = fe$,

$$DID_{Y-X'\gamma_{fd}}(g,0,t,t-1) = E(D_{g,t})\Delta^{TR}_{g,t} - E(D_{g,t-1})\Delta^{TR}_{g,t-1} - \left(E(D_{0,t})\Delta^{TR}_{0,t} - E(D_{0,t-1})\Delta^{TR}_{0,t-1}\right).$$

Then, the proof follows exactly as that of the first statement of Theorem 1 for β_{fd}.

Proof of the second statement.

We only sketch the proof of the result for β_{fe}. As above, the reasoning is similar to that in the proof of the second statement of Theorem 1, but replacing Y by $Y - X'\gamma_{fe}$. In particular, under Assumptions 2X$_{fe}$, 4, and 11, one can follow the same steps as those used to establish the second point of Lemma 2 to show that

$$DID_{Y-X'\gamma_{fe}}(g,0,t,t-1) = \left(E(D_{g,t}) - E(D_{g,t-1})\right)\Delta^S_{g,t} - \left(E(D_{0,t}) - E(D_{0,t-1})\right)\Delta^S_{0,t}.$$

A.9 Proof of Theorem 5

Reasoning as in the proof of Theorem 2,

$$\begin{aligned} &DID^*_Y(1,t) \\ =& \sum_{g \in \mathcal{G}_{1,t}} r(g|1,t,t)\left[E(Y_{g,t}) - E(Y_{g,t-1}) - (E(Y|G^*_t = 0, T = t) - E(QY|G^*_t = 0, T = t-1))\right], \end{aligned}$$

and

$$E(Y|G^*_t = 0, T = t) - E(QY|G^*_t = 0, T = t-1) = \sum_{g' \in \mathcal{G}_{0,t}} r(g'|0,t,t)\left(E(Y_{g',t}) - E(Y_{g',t-1})\right).$$

Thus,

$$DID^*_Y(1,t) = \sum_{(g,g') \in \mathcal{G}_{1,t} \times \mathcal{G}_{0,t}} r(g|1,t,t) r(g'|0,t,t) DID_Y(g,g',t,t-1), \tag{35}$$

where $DID(g, g', t, t-1)$ is defined as above in Lemma 1. By definition of $\mathcal{G}_{0,t}$, we have that under Assumption 3, for all $g' \in \mathcal{G}_{0,t}$, $P(S_{g',t}) = 0$. Then, by Lemma 2, we have, for all $(g, g') \in \mathcal{G}_{1,t} \times \mathcal{G}_{0,t}$,

$$DID_Y(g, g', t, t-1) = P(S_{g,t})\Delta^S_{g,t}.$$

Combining this equation with (35), we obtain

$$DID^*_Y(1, t) = \sum_{g \in \mathcal{G}_{1,t}} r(g|1, t, t)P(S_{g,t})\Delta^S_{g,t}.$$

A similar reasoning yields

$$DID^*_Y(-1, t) = -\sum_{g \in \mathcal{G}_{-1,t}} r(g|-1, t, t)P(S_{g,t})\Delta^S_{g,t}.$$

The rest of the proof is the same as that of Theorem 2.

For Online Publication

B Further discussion

B.1 Can common trends hold with heterogeneous treatment effects?

Throughout the paper, we assume that groups experience common trends, but that the effect of the treatment may be heterogeneous between groups and / or over time. We now discuss two examples where this may happen. We then argue that the mechanisms behind these examples are fairly general. Thus treatment effects are often likely to be heterogeneous, even when common trends are plausible.

First, assume one wants to learn the effect of the minimum wage on the employment levels of some US counties. For simplicity, let us assume that the minimum wage can only take two values, a low and a high value. Also, let us assume that there are only two periods, the 90s and the 2000s. Between these two periods, the amount of competition from China for the US industry increased substantially. Thus, for the common trends assumption to hold for counties A and B, the effect of that increase in competition should be the same in those two counties, in the counterfactual state of the world where A and B have a low minimum wage at both dates. For that to be true, the economy of those two counties should be pretty similar. For instance, if A has a very service-oriented economy, while B has a very industry-oriented economy, it is unlikely that their employment levels will react similarly to Chinese competition.

Now, if the economies of A and B are similar, they should also have similar effects of the minimum wage on employment, thus implying that the treatment effect is homogenous between groups. On the other hand, the treatment effect may vary over time. For instance, the drop in the employment levels of A and B due to Chinese competition will probably be higher if their minimum wage is high than if their minimum wage is low. This is equivalent to saying that the effect of the minimum wage on employment diminishes from the first to the second period: due to Chinese competition in the second period, the minimum wage may have a more negative effect on employment then.

Second, assume one wants to learn the effect of a job training program implemented in some US counties on participants' wages. Let us suppose that individuals self-select into the training according to a Roy model:

$$D = 1\{Y(1) - Y(0) > c_{G,T}\}, \tag{36}$$

where $c_{g,t}$ represents the cost of the training for individuals in county g and period t. Here, the common trends condition requires that average wages without the training follow the same evolution in all counties. As above, for this to hold counties used in the analysis should have similar economies, so let us assume that those counties are actually identical copies of each other: at each period, their distribution of wages without and with the training is the same. Therefore, $(g,t) \mapsto E(Y(1) - Y(0)|G = g, T = t)$ is constant. However, $c_{g,t}$ may vary across counties and over time: some counties may subsidize the training more than others, and some counties may change their subsidies over time. Then, $(g,t) \mapsto \Delta^{TR}_{g,t} = E(Y(1) - Y(0)|Y(1) - Y(0) > c_{g,t})$ will not be constant, despite the fact that all counties in the sample have similar economies and experience similar trends on their wages. Similarly, $(g,t) \mapsto \Delta^{S}_{g,t}$ will also not be constant.

Overall, when the treatment is assigned at the group $\times$ period level as in the minimum wage example, the economic restrictions underlying the common trends assumption may also imply homogeneous treatment effect between groups. However, those restrictions usually do not imply that the treatment effect is constant over time. Moreover, when the treatment is assigned at the individual level, as in the job training example, the economic restrictions underlying the common trends assumption neither imply homogeneous treatment effects between groups, nor homogeneous treatment effects over time.

B.2 Assumptions 2' and 4 in a Roy model

We consider the following dynamic model on potential outcomes:

$$Y_t(d) = a_{d,G} + b_{d,t} + e_{d,t},$$

where we suppose that $e_{d,t}$ follows a random walk. Namely, $e_{d,t} = e_{d,t-1} + \eta_{d,t}$, where the $(\eta_{0,t}, \eta_{1,t})_{t\geq 0}$ are assumed i.i.d. conditional on G and $E(\eta_{d,t}|G) = 0$. Note that we leave the dependence between $\eta_{0,t}$ and $\eta_{1,t}$ unrestricted. This model corresponds for instance to the popular "Restricted Income Profile" (RIP) model if $Y_t(d)$ is the logarithm of wages.

Then suppose that selection into treatment corresponds to the following Roy model:

$$D(t) = 1\{Y_t(1) - Y_t(0) > c_{G,t}\}.$$

In this framework, Assumption 2' holds. To see this, consider $d = 1$ (the proof is similar for $d = 0$) and remark that

$$\begin{aligned}
&E(Y(1)|D(t-1) = 1, T = t, G = g) - E(Y(1)|D(t-1) = 1, T = t-1, G = g)\\
=&b_{1,t} - b_{1,t-1} + E(\eta_{1,t}|e_{1,t-1} - e_{0,t-1} > c_{g,t-1} - a_{1,g} + a_{0,g} - b_{1,t-1} + b_{0,t-1}, G = g)\\
=&b_{1,t} - b_{1,t-1} + E(\eta_{1,t}|G = g)\\
=&b_{1,t} - b_{1,t-1}.
\end{aligned}$$

The second line follows from the fact that $e_{1,t-1} - e_{0,t-1}$ does not depend on the $(\eta_{d,t'})_{t'<t}$. The third line follows from $E(\eta_{d,t}|G) = 0$.

On the other hand, Assumption 4 is not satisfied in general in this model. By what precedes,

$$\begin{aligned}&E(Y(1)-Y(0)|D(t-1)=1,T=t,G=g)-E(Y(1)-Y(0)|D(t-1)=1,T=t-1,G=g)\\=&b_{1,t}-b_{1,t-1}-(b_{0,t}-b_{0,t-1}),\end{aligned}$$

which is not zero in general.

C Another alternative estimand

In this section, we show that under Assumptions 2, 3, 4, and the first point of Assumption 10, Δ^S is identified. For any random variable R and for all $(d,g,t)\in\{0,1\}\times\{-1,1\}\times\{1,...,\overline{t}\}$, let

$$\begin{aligned}DID^*_R(g,t)=&E(R|G^*_t=g,T=t)-E(QR|G^*_t=g,T=t-1)\\&-\left(E(R|G^*_t=0,T=t)-E(QR|G^*_t=0,T=t-1)\right).\end{aligned}$$

$DID^*_R(g,t)$ compares the evolution of the mean of R in the supergroups $G^*_t=g$ and $G^*_t=0$ between $T=t-1$ and $T=t$, after reweighting observations in period $t-1$ by Q to ensure that groups' distribution is the same in periods $t-1$ and t. Our estimand is defined by

$$W_{DID}=\sum_{t=1}^{\overline{t}}\sum_{g\in\{-1,1\}}\omega_{g,t}\frac{DID^*_Y(g,t)}{DID^*_D(g,t)}.$$

W_{DID} is a weighted average of the Wald-DID estimands studied in de Chaisemartin and D'Haultfœuille (2018). Theorem 2 below shows that W_{DID} identifies Δ^S even if the second point of Assumption 10 fails. On the other hand, W_{DID} relies on the stable treatment effect assumption, which restricts treatment effect heterogeneity over time, contrary to W_{TC}.

Theorem 5 *Suppose that Assumption 1 holds and D is binary. If Assumptions 2, 3, 4, and the first point of Assumption 10 hold, then $W_{DID}=\Delta^S$.*

D Detailed literature review

We now review the 33 papers that use two-way fixed effects or closely related regressions that we found in our literature review. For each paper, we use the following presentation:
Authors (year), Title. *Where the two-way fixed effects estimator is used in the paper.* Description of the two-way fixed effects estimator used in the paper, and how it relates to Regression 1 or 2. Assessment of whether the stable groups assumption holds in this paper.

1. **Chandra et al. (2010), Patient Cost-Sharing and Hospitalization Offsets in the Elderly.** *First line of Tables 2 and 3.*

In the regressions in the first line of Tables 2 and 3, the outcomes (e.g. a measure of utilization for plan p in month t) are regressed on plan fixed effects, month fixed effects, and an indicator of whether plan p had increased copayments in month t (see regression equation at the bottom of page 198). This regression corresponds to Regression 1. The period analyzed runs from January 2000 to September 2003. The stable groups assumption is satisfied until January 2002, when the HMO plans also become treated.

2. **Duggan and Morton (2010), The Effect of Medicare Part D on Pharmaceutical Prices and Utilization.** *Tables 2 and 3.*
   In regression Equation (1), the dependent variable is the change in the price of drug j between 2003 and 2006, the explanatory variables are the Medicare market share for drug j in 2003, and some control variables. This regression corresponds to Regression 2, with some control variables. The stable groups assumption is presumably not satisfied: it seems unlikely that there are drugs whose Medicare market share in 2003 is equal to 0.

3. **Aizer (2010), The Gender Wage Gap and Domestic Violence.** *Table 2.*
   In regression Equation (2), the dependent variable is the log of female assaults among females of race r in county c and year t, and the explanatory variables are race, year, county, race × year, race × county, and county × year fixed effects, as well as the gender wage gap in county c, year t, and race r, and some control variables. This regression is a "three-way fixed effects" version of Regression 1, with some control variables. The stable groups assumption is presumably satisfied: it seems likely that between each pair of consecutive years, there are counties where the gender wage gap does not change.

4. **Algan and Cahuc (2010), Inherited Trust and Growth.** *Figure 4.*
   Figure 4 presents a regression of changes in income per capita from 1935 to 2000 on changes in inherited trust over the same period and a constant. This regression corresponds to Regression 2. The stable groups assumption is satisfied: there are countries where inherited trust does not change from 1935 to 2000.

5. **Ellul et al. (2010), Inheritance Law and Investment in Family Firms.** *Table 7.*
   In the regressions presented in Table 7, the dependent variable is the capital expenditure of firm j in year t, and the explanatory variables are firm fixed effects, an indicator for whether year t is a succession period for firm j, some controls, and three treatment variables: the interaction of the succession indicator with the level of investor protection in the country where firm j is located, the interaction of the succession indicator with the level of inheritance laws permissiveness in the country where firm j is located, and the interaction of the succession indicator with the level of inheritance laws permissiveness and the level of investor protection in the country where firm j is located. This regression is similar to Regression 1 with controls, except that it has three treatment variables. The stable

groups assumption is presumably not satisfied: for instance, it seems unlikely that there are countries with no investor protection at all.

6. **Bustos (2011), Trade Liberalization, Exports, and Technology Upgrading: Evidence on the Impact of MERCOSUR on Argentinean Firms.** *Tables 3 to 12.*
   In regression Equation (11), the dependent variable is the change in exporting status of firm i in sector j between 1992 and 1996, and the explanatory variables are the change in trade tariffs in Brasil for products in sector j over the same period, and some control variables. This regression corresponds to Regression 2, with some controls. The stable groups assumption is presumably satisfied: it seems likely that there are sectors where trade tariffs in Brasil did not change between 1992 and 1996.

7. **Anderson and Sallee (2011), Using Loopholes to Reveal the Marginal Cost of Regulation: The Case of Fuel-Economy Standards.** *Table 5 Column 2.*
   In the regression in Table 5 Column (2), the dependent variable is an indicator for whether a car sold is a flexible fuel vehicle, and the explanatory variables are state and month fixed effects, the percent ethanol availability in each month $\times$ state, and some controls. This regression corresponds to Regression 1. The stable groups assumption is presumably satisfied: it seems likely that between each pair of consecutive months, there are states where the percent ethanol availability does not change.

8. **Bagwell and Staiger (2011), What Do Trade Negotiators Negotiate About? Empirical Evidence from the World Trade Organization.** *Table 3, OLS columns.*
   In regression equations (15a) and (15b), the dependent variable is the ad valorem tariff level bound by country c on product g, while the explanatory variables are country and product fixed effects, and two treatment variables which vary at the country $\times$ product level. These regressions are similar to Regression 1, except that they have two treatment variables. The stable groups assumption is not applicable here, as none of the two sets of fixed effects included in the regression correspond to an ordered variable.

9. **Zhang and Zhu (2011), Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia.** *Tables 3 and 4, Columns 4-6.*
   In the regression in, say, Table 3 Column (4), the dependent variable is the total number of contributions to Wikipedia by individual i at period t, regressed on individual fixed effects, an indicator for whether period t is after the Wikipedia block, the interaction of this indicator and a measure of social participation by individual i, and some controls. This regression corresponds to Regression 1 with some controls. The stable groups assumption is satisfied: there are individuals with a social participation measure equal to 0.

10. **Hotz and Xiao (2011), The Impact of Regulations on the Supply and Quality**

of Care in Child Care Markets. *Table 7, Columns 4 and 5.*

In Regression Equation (1), the dependent variable is the outcome for market m in state s and year t, and the explanatory variables are state and year fixed effects, various measures of regulations in state s in year t, and some controls. This regression corresponds to Regression 1 with several treatment variables and with some controls. The stable groups assumption is presumably satisfied: between each pair of consecutive years, it is likely that there are states whose regulations do not change.

11. **Mian and Sufi (2011), House Prices, Home Equity-Based Borrowing, and the US Household Leverage Crisis.** *Tables 2 and 3.*

    In Regression Equation (1), the dependent variable is the change in homeowner leverage from 2002 to 2006 for individual i living in zip code z in MSA m, instrumented by MSA-level housing supply elasticity. This regression is the 2SLS version of Regression 2, with some controls. The stable groups assumption is presumably not satisfied: it is unlikely that some MSAs have an housing supply elasticity equal to 0.

12. **Wang (2011), State Misallocation and Housing Prices: Theory and Evidence from China.** *Table 5, Panel A.*

    In regression Equation (15), the dependent variable is the quantity of housing services in household i's residence in year t, while the explanatory variables are an indicator for period t being after the reform, a measure of mismatch in household i, the interaction of the measure of mismatch and the time indicator, and some controls. This regression is similar to Regression 1 with some controls, except that it has a measure of mismatch in household i instead of household fixed effects. The stable groups assumption is presumably satisfied: it is likely that some households have a mismatch equal to 0.

13. **Duranton and Turner (2011), The Fundamental Law of Road Congestion: Evidence from US Cities.** *Table 5.*

    In the regressions presented in, say, the first column of Table 5, the dependent variable is the change in vehicle kilometers traveled in MSA s between decades t and t-1, and the explanatory variables are the change in kilometers of roads in MSA s between decades t and t-1, and decade effects. This regression corresponds to Regression 2. The stable groups assumption is presumably satisfied: it is likely that between each pair of consecutive decades, there are some MSAs where the kilometers of roads do not change.

14. **Acemoglu et al. (2011), The Consequences of Radical Reform: The French Revolution.** *Table 3.*

    In regression Equation (1), the dependent variable is urbanization in polity j at time t, while the explanatory variables are time and polity fixed effects, and the number of years of French presence in polity j interacted with the time effects. This regression corresponds to Regression 1. The stable groups assumption is satisfied as there are several polities that did not experience any year of French presence.

15. **Baum-Snow and Lutz (2011), School Desegregation, School Choice, and Changes in Residential Location Patterns by Race.** *Tables 2 to 6.*
In regression Equation (1), the dependent variable is, say, whites public school enrolment in MSA j in year t, while the explanatory variables are MSA and region $\times$ time fixed effects, and an indicator for whether MSA j is desegregated. This regression corresponds to Regression 1 with controls. The stable groups assumption is satisfied: between each pair of consecutive years, there are MSAs whose desegregation status does not change.

16. **Dinkelman (2011), The Effects of Rural Electrification on Employment: New Evidence from South Africa.** *Tables 4 and 5 Columns 5-8, Table 8 Columns 3-4, Table 9 Column 2, and Table 10 Columns 2, 4, and 6.*
In regression Equation (3), the dependant variable is, say, the first difference of the female employment rate for community j between periods 0 and 1, and the explanatory variables are district fixed effects, the change of electrification status of community j between periods 0 and 1, and some statistical controls. The land gradient in community j is used as an instrument for the change in electrification. This regression corresponds to the 2SLS version of Regression 2 with some controls. The stable groups assumption is presumably satisfied: it is likely that there are communities whose land gradient is 0.

17. **Enikolopov et al. (2011), Media and Political Persuasion: Evidence from Russia.** *Table 3.*
In regression Equation (5), the dependent variable is the share of votes for party j in election-year t and subregion s, and the explanatory variables are subregion and election fixed effects, and the share of people having access to NTV in subregion s in election-year t. This regression corresponds to Regression 1. The stable groups assumption is not satisfied: the share of people having access to NTV strictly increases in all regions between 1995 and 1999, the two elections used in the analysis.

18. **Fang and Gavazza (2011), Dynamic Inefficiencies in an Employment-Based Health Insurance System: Theory and Evidence.** *Tables 2, 3, 5, and 6, Column 3.*
In regression Equation (7), the dependent variable is the health expenditures of individual j working in industry i in period t and region r, and the explanatory variables are individual effects, region specific time effects, and the job tenure of individual j. The death rate of establishments in industry i in period t and region r is used as an instrument for the job tenure of individual j. This regression is the 2SLS version of Regression 2 with controls. The stable groups assumption is presumably satisfied: between each pair of consecutive

years, it is likely that there are some industry $\times$ region pairs where the death rate of establishments does not change.

19. **Gentzkow et al. (2011), The Effect of Newspaper Entry and Exit on Electoral Politics.** *Tables 2 and 3.*
    In regression Equation (2), the dependent variable is the change in voter turnout in county c between elections year t and t-1, and the explanatory variables are state $\times$ year effects, and the change in the number of newspapers in county c between t and t-1. This regression corresponds to Regression 2 with controls. The stable groups assumption is satisfied: between each pair of consecutive years, there are some counties where the number of newspapers does not change.

20. **Bloom et al. (2012), Americans Do IT Better: US Multinationals and the Productivity Miracle.** *Table 2, Columns 6-8.*
    In the regression in, say, Column 6 of Table 2, the dependent variable is the log of output per worker in firm i in period t, while the explanatory variables are firms and time fixed effects, the log of the amount of IT capital per employee $\ln(C/L)$, the interaction of $\ln(C/L)$ and an indicator for whether the firm is owned by a US multinational, the interaction of $\ln(C/L)$ and an indicator for whether the firm is owned by a non-US multinational, and some controls. This regression is similar to Regression 1 with some controls, except that it has three treatment variables. The stable groups assumption is presumably satisfied: between each pair of consecutive years, it is likely that there are some firms where the amount of IT capital per employee $\ln(C/L)$ does not change.

21. **Simcoe (2012), Standard Setting Committees: Consensus Governance for Shared Technology Platforms.** *Table 4, Columns 1-3.*
    In regression Equation (5), the dependent variable is a measure of time to consensus for project i submitted to committee j, while the explanatory variables are an indicator for projects submitted to the standards track, a measure of distributional conflict, the interaction of the standards track and distributional conflict, and some controls variables. This regression is similar to Regression 1 with some controls, except that it has a measure of distributional conflict instead of committee fixed effects. The stable groups assumption is presumably not satisfied: it is unlikely that there is any committee where the measure of distributional conflict is equal to 0.

22. **Moser and Voena (2012), Compulsory Licensing: Evidence from the Trading with the Enemy Act.** *Table 2.*
    In the regression equation in the beginning of Section III, the dependent variable is the number of patents by US inventors in patent class c at period t, and the explanatory variables are patent class and time fixed effects, the interaction of period t being after

the trading with the enemy act and the number of licensed patents in class c, and some control variables. This regression corresponds to Regression 1 with some controls. The stable groups assumption is satisfied: there are patent classes where no patent was licensed.

23. **Forman et al. (2012), The Internet and Local Wages: A Puzzle.** *Tables 2 and 4.* In regression Equation (1), the dependent variable is the difference between log wages in 2000 and 1995 in county i, and the explanatory variables are Internet investment by businesses in county i in 2000, and control variables. This regression corresponds to Regression 2 with some controls. The stable groups assumption is satisfied: there are counties with no Internet investment in 2000.

24. **Besley and Mueller (2012), Estimating the Peace Dividend: The Impact of Violence on House Prices in Northern Ireland.** *Table 1, Columns 3 and 5-7.* In regression Equation (1), the dependent variable is the price of houses in region r at time t, while the explanatory variables are region and time fixed effects, and the number of people killed because of the civil war in region r at time t-1. This regression corresponds to Regression 1. The stable groups assumption is presumably satisfied: between each pair of consecutive years, it is likely that there are some regions where the number of people killed because of the civil war does not change.

25. **Dafny et al. (2012), Paying a Premium on Your Premium? Consolidation in the US Health Insurance Industry.** *Table 3.* In regression Equation (3), the dependent variable is the the concentration of the hospital industry in market m and year t, and explanatory variables are time fixed effects, market fixed effects, and the change in concentration in market m induced by a merger interacted with an indicator for t being after the merger. This regression corresponds to Regression 1. The stable groups assumption is satisfied: there are many markets where the merger did not change concentration.

26. **Hornbeck (2012), The Enduring Impact of the American Dust Bowl: Short- and Long-Run Adjustments to Environmental Catastrophe.** *Table 2.* In regression Equation (1), the dependent variable is, say, the change in log land value in county c between period t and 1930, and the explanatory variables are state $\times$ year fixed effects, the share of county c in high erosion regions, the share of county c in medium erosion regions, and some control variables. This regression is similar to Regression 1 with controls, except that it has two treatment variables. The stable groups assumption is satisfied: many counties have 0% of their land situated in medium or high erosion regions.

27. **Bajari et al. (2012), A Rational Expectations Approach to Hedonic Price Regressions with Time-Varying Unobserved Product Attributes: The Price of**

Pollution. *Table 5.*

In, say, the first regression equation in the bottom of page 1915, the dependent variable is the change in the price of house j between sales 2 and 3, and the explanatory variables are the change in various pollutants in the area around house j between sales 2 and 3, and some controls. This regression is similar to Regression 2 with controls, except that it has several treatment variables. The stable groups assumption is presumably satisfied: it is likely that for each pair of consecutive sales, there are houses where the level of each pollutant does not change.

28. **Dahl and Lochner (2012), The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit.** *Table 3.*

    In regression Equation (4), the dependent variable is the change in test scores for child i between years a and a-1, while the explanatory variables are the change in the EITC income of her family and some controls, and the change in the expected EITC income of her family based on her family income in year a-1 is used to instrument for the actual change of her family's EITC income. This regression is a 2SLS version of Regression 2 with controls, except that it does not have years fixed effects. The stable groups assumption is presumably satisfied: it is likely that for each pair of consecutive years, there are children whose family's expected EITC income does not change.

29. **Imberman et al. (2012), Katrina's Children: Evidence on the Structure of Peer Effects from Hurricane Evacuees.** *Tables 3-6.*

    In regression Equation (1), the dependent variable is the test score of student i in school j in grade g and year t, and the explanatory variables are grade, school, year, and grade $\times$ year fixed effects, the fraction of Katrina evacuee students received by school j in grade g and year t, and some controls. This regression is a three-way fixed effects version of Regression 1. The stable groups assumption is satisfied: there are schools that did not receive any Katrina evacuee.

30. **Chaney et al. (2012), The Collateral Channel: How Real Estate Shocks Affect Corporate Investment.** *Table 5.*

    In regression Equation (1), the dependent variable is the value of investment in firm i and year t divided by the lagged book value of properties, plants, and equipments (PPE), and the explanatory variables are firm and time fixed effects and the market value of firm i in year t divided by its lagged PPE, and some controls. This regression corresponds to Regression 1, with some controls. The stable groups assumption is presumably satisfied: it is likely that between each pair of consecutive years, there are firms whose market value divided by their lagged PPE does not change.

31. **Aaronson et al. (2012), The Spending and Debt Response to Minimum Wage**

Hikes. *Tables 1, 2, and 5.*
In regression Equation (1), the outcome variable is, say, income of household i at period t, and the explanatory variables are household and time fixed effects, and the minimum wage in the state where household i lives in period t. This regression corresponds to Regression 1. The stable groups assumption is satisfied: between each pair of consecutive periods, there are states where the minimum wage does not change.

32. **Brambilla et al. (2012), Exports, Export Destinations, and Skills.** *Table 5.*
In the regression in, say, the first column of Table 2, the dependent variable is a measure of skills in the labor force employed by firm i in industry j at period t, and the explanatory variables are firm and industry $\times$ period fixed effects, the ratio of exports to sales in firm i at period t, and some controls. This regression corresponds to Regression 1, with some controls. The stable groups assumption is presumably satisfied: it is likely that between each pair of consecutive periods, there are firms whose ratio of exports to sales does not change.

33. **Faye and Niehaus (2012), Political Aid Cycles.** *Table 3, Columns 4 and 5, and Tables 4 and 5.*
In regression Equation (2), the dependent variable is the amount of donations received by receiver r from donor d in year t, and the explanatory variables are donor $\times$ receiver fixed effects, an indicator for whether there is an election in country r in year t, a measure of alignment between the ruling political parties in countries r and d, and the interaction of the election indicator and the measure of alignment. This regression corresponds to Regression 1. The stable groups assumption is presumably not satisfied: it is unlikely that there are donor-receiver pairs that are perfectly unaligned.