American Economic Review 2020, 110(9): 2964–2996
https://doi.org/10.1257/aer.20181169

Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects†

By Clément de Chaisemartin and Xavier D'Haultfœuille*

Linear regressions with period and group fixed effects are widely used to estimate treatment effects. We show that they estimate weighted sums of the average treatment effects (*ATE*) *in each group and period, with weights that may be negative. Due to the negative weights, the linear regression coefficient may for instance be negative while all the ATEs are positive. We propose another estimator that solves this issue. In the two applications we revisit, it is significantly different from the linear regression estimator.* (*JEL* C21, C23, D72, J31, J51, L82)

A popular method to estimate the effect of a treatment on an outcome is to compare over time groups experiencing different evolutions of their exposure to treatment. In practice, this idea is implemented by estimating regressions that control for group and time fixed effects. Hereafter, we refer to those as two-way fixed effects (FE) regressions. We conducted a survey, and found that 19 percent of all empirical articles published by the *American Economic Review* (*AER*) between 2010 and 2012 have used a two-way FE regression to estimate the effect of a treatment on an outcome. When the treatment effect is constant across groups and over time, such regressions estimate that effect under the standard "common trends" assumption. However, it is often implausible that the treatment effect is constant. For instance, the minimum wage's effect on employment may vary across US counties, and may change over time. This paper examines the properties of two-way FE regressions when the constant effect assumption is violated.

We start by assuming that all observations in the same (g,t) cell have the same treatment and that the treatment is binary, as is for instance the case when the treatment is a county-level law. We consider the regression of $Y_{i,g,t}$, the outcome of unit i in group g at period t on group fixed effects, period fixed effects, and $D_{g,t}$, the treat-

*de Chaisemartin: University of California at Santa Barbara (email: clementdechaisemartin@ucsb.edu); D'Haultfœuille: CREST-ENSAE (email: xavier.dhaultfoeuille@ensae.fr). Thomas Lemieux was the coeditor for this article. We are very grateful to Olivier Deschênes, Guido Imbens, Peter Kuhn, Kyle Meng, Jesse Shapiro, Dick Startz, Doug Steigerwald, Clémence Tricaud, Gonzalo Vazquez-Bare, members of the UCSB econometrics research group, and seminar participants at Bergen, CIREQ Econometrics conference, CREST, Goteborg, Gothenburg, Groningen, ITAM, Pompeu Fabra, Stanford, SMU, Tinbergen Institute, UCL, UCLA, UC Davis, UCSB, USC, and Warwick for their helpful comments. Xavier D'Haultfœuille gratefully acknowledges financial support from the research grants Otelo (ANR-17-CE26-0015-041) and the Labex Ecodec: Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

†Go to https://doi.org/10.1257/aer.20181169 to visit the article page for additional materials and author disclosure statements.

ment in group g at period t. Let $\hat{\beta}_{fe}$ denote the coefficient of $D_{g,t}$, and let β_{fe} denote its expectation. Under the common trends assumption, we show that β_{fe} is equal to a weighted sum of the treatment effect in each treated (g,t) cell:

$$\beta_{fe} = E\left(\sum_{(g,t):D_{g,t}=1} W_{g,t}\Delta_{g,t}\right), \tag{1}$$

where $\Delta_{g,t}$ is the average treatment effect (ATE) in group g and period t and the weights $W_{g,t}$ sum to 1 but may be negative. Negative weights arise because $\hat{\beta}_{fe}$ is a weighted sum of several difference-in-differences (DID), which compare the evolution of the outcome between consecutive time periods across pairs of groups. However, the "control group" in some of those comparisons may be treated at both periods. Then, its treatment effect at the second period gets differenced out by the DID, hence the negative weights.

The negative weights are an issue when the ATEs are heterogeneous across groups or periods. Then, one could have that β_{fe} is negative while all the ATEs are positive. For instance, $1.5 \times 1 - 0.5 \times 4$, a weighted sum of 1 and 4, is strictly negative. Using the dataset of Gentzkow, Shapiro, and Sinkinson (2011), we find that 40 percent of the weights attached to β_{fe} are negative, so β_{fe} is not robust to heterogeneous effects.[1]

Researchers may want to know how serious that issue is in the application they consider. We show that conditional on all treatments, the absolute value of the expectation of $\hat{\beta}_{fe}$ divided by the standard deviation of the weights is equal to the minimal value of the standard deviation of the ATEs across the treated (g,t) cells under which the average treatment on the treated (ATT) may actually have the opposite sign than that coefficient. One can estimate that ratio to assess the robustness of the two-way FE coefficient. If that ratio is close to 0, that coefficient and the ATT can be of opposite signs even under a small and plausible amount of treatment effect heterogeneity. In that case, treatment effect heterogeneity would be a serious concern for the validity of that coefficient. On the contrary, if that ratio is very large, that coefficient and the ATT can only be of opposite signs under a very large and implausible amount of treatment effect heterogeneity.

Finally, we propose a new estimator, DID_{M}, that is valid even if the treatment effect is heterogeneous over time or across groups. It estimates the average treatment effect across all the (g,t) cells whose treatment changes from $t-1$ to t. It relies on common trends assumptions on both potential outcomes. Those conditions are partly testable, and we propose a test that amounts to looking at pretrends. This test differs from the standard event study pretrends test (see Autor 2003), which has been shown to be invalid when treatment effects are heterogeneous (see Abraham and Sun 2018). We show that our estimator is asymptotically normal. We compute it in the datasets of Gentzkow, Shapiro, and Sinkinson (2011) and Vella and Verbeek (1998), and in both cases we find that it is significantly different from $\hat{\beta}_{fe}$.[2] Our estimator can be used in applications where, for each pair of consecutive dates, there are

[1] Gentzkow, Shapiro, and Sinkinson (2011) does not estimate β_{fe}, but β_{fd}, the treatment coefficient in the first-difference regression defined below. Forty-six percent of the weights attached to β_{fd} are strictly negative.

[2] In both cases, our estimator is also significantly different from $\hat{\beta}_{fd}$.

groups whose treatment does not change. We estimate that this condition is satisfied for around 80 percent of the papers using two-way fixed effects regressions found in our survey of the *AER*.

Overall, our paper has implications for applied researchers estimating two-way fixed effects regressions. First, we recommend that they compute the weights attached to their regression and the ratio of $|\hat{\beta}_{fe}|$ divided by the standard deviation of the weights. To do so, they can use the *twowayfeweights* Stata package that is available from the SSC repository. If many weights are negative, and if the ratio is not very large, we recommend that they compute our new estimator, using the *fuzzydid* and *did_multiplegt* Stata packages, also available from the SSC repository (see de Chaisemartin, D'Haultfœuille, and Guyonvarch 2019, for explanations on how to use the former package).

We extend our results in several important directions. First, another commonly used regression is the first-difference regression of $Y_{g,t} - Y_{g,t-1}$, the change in the mean outcome in group g, on period fixed effects and on $D_{g,t} - D_{g,t-1}$, the change in the treatment. We let β_{fd} denote the expectation of the coefficient of $D_{g,t} - D_{g,t-1}$. We show that under common trends, β_{fd} also identifies a weighted sum of treatment effects, with potentially some negative weights. Second, in our online Appendix we show that our results extend to fuzzy designs, where the treatment varies within (g,t) cells, and to two-way fixed effects regressions with a nonbinary treatment and with covariates.

Our paper is related to the DID literature. Our main result generalizes Theorem 1 in de Chaisemartin and D'Haultfœuille (2018). When the data have two groups and two periods, the Wald-DID estimand considered therein is equal to β_{fe} and β_{fd}. Our results on β_{fe} and β_{fd} are thus extensions of that theorem to the case with multiple periods and groups.[3] Moreover, our DID_{M} estimator is related to the Wald-TC estimator with many groups and periods proposed in de Chaisemartin and D'Haultfœuille (2018), and to the multiperiod DID estimator proposed by Imai and Kim (2018). In Section III, we explain the differences between those three estimators.

More recently, Borusyak and Jaravel (2017), Abraham and Sun (2018), Athey and Imbens (2018), Callaway and Sant'Anna (2018), and Goodman-Bacon (2018) study the special case of staggered adoption designs, where the treatment of a group is weakly increasing over time. Those papers derive some important results specific to that design that we do not consider here. Still, some of the results in those papers are related to ours, and we describe precisely those connections later in the paper. The most important dimension on which our paper differs from those is that our results apply to any two-way fixed effects regressions, not only to those with staggered adoption. In our survey of the *AER* papers estimating two-way fixed effects regressions, less than 10 percent 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.

[3] 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'Haultfœuille (2015).

The paper is organized as follows. Section I introduces the setup. Section II presents our decomposition results. Section III introduces our alternative estimator. Section IV briefly describes some of the extensions covered in our online Appendix. Section V presents our survey of the articles published in the *AER*, and our two empirical applications. The data and codes are given in de Chaisemartin and D'Haultfœuille (2020b).

I. Setup

One considers observations that can be divided into G groups and T periods. For every $(g,t) \in \{1,\ldots,G\} \times \{1,\ldots,T\}$, let $N_{g,t}$ denote the number of observations in group g at period t, and let $N = \sum_{g,t} N_{g,t}$ be the total number of observations. The data may be an individual-level panel or repeated cross-section dataset where groups are, say, individuals' county of birth. The data could also be a cross section where cohort of birth plays the role of time. For instance, Duflo (2001) compares the schooling of different cohorts in Indonesia, some of which were exposed to a school construction program. It is also possible that for all (g,t), $N_{g,t} = 1$, e.g., a group is one individual or firm. All of the above are special cases of the data structure we consider.

One is interested in measuring the effect of a treatment on some outcome. Throughout the paper we assume that treatment is binary, but our results apply to any ordered treatment, as we show in online Appendix Section 3.2. Then, for every $(i,g,t) \in \{1,\ldots,N_{g,t}\} \times \{1,\ldots,G\} \times \{1,\ldots,T\}$, let $D_{i,g,t}$ and $(Y_{i,g,t}(0), Y_{i,g,t}(1))$ respectively denote the treatment status and the potential outcomes without and with treatment of observation i in group g at period t.

The outcome of observation i in group g and period t is $Y_{i,g,t} = Y_{i,g,t}(D_{i,g,t})$. For all (g,t), let

$$D_{g,t} = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} D_{i,g,t}, \qquad Y_{g,t}(0) = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} Y_{i,g,t}(0),$$

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

Here, $D_{g,t}$ denotes the average treatment in group g at period t, while $Y_{g,t}(0)$, $Y_{g,t}(1)$, and $Y_{g,t}$ respectively denote the average potential outcomes without and with treatment and the average observed outcome in group g at period t.

Throughout the paper, we maintain the following assumptions.

ASSUMPTION 1 (Balanced Panel of Groups): *For all* $(g,t) \in \{1,\ldots,G\} \times \{1,\ldots,T\}, N_{g,t} > 0$.

Assumption 1 requires 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 the denominators of some of the fractions below may then be equal to zero.

ASSUMPTION 2 (Sharp Design): *For all* $(g,t) \in \{1,\ldots,G\} \times \{1,\ldots,T\}$ *and* $i \in \{1,\ldots,N_{g,t}\}$, $D_{i,g,t} = D_{g,t}$.

Assumption 2 requires that units' treatments do not vary within each (g,t) cell, a situation we refer to as a sharp design. This is for instance satisfied when the treatment is a group-level variable, for instance a county or a state law. This is also mechanically satisfied when $N_{g,t} = 1$. In our survey in Section IIA, we find that almost 80 percent of the papers using two-way fixed effects regressions and published in the *AER* between 2010 and 2012 consider sharp designs. We focus on sharp designs because of their prevalence, but in online Appendix Section 2, we show that all the results in Sections II and III can be extended to fuzzy designs.

ASSUMPTION 3 (Independent Groups): *The vectors* $(Y_{g,t}(0), Y_{g,t}(1), D_{g,t})_{1\leq t\leq T}$ *are mutually independent.*

We consider $D_{g,t}$, $Y_{g,t}(0)$, $Y_{g,t}(1)$ as random variables. For instance, aggregate random shocks may affect the average potential outcomes of group g at period t. The treatment status of group g at period t may also be random. The expectations below are taken with respect to the distribution of those random variables. Assumption 3 allows for the possibility that the treatments and potential outcomes of a group may be correlated over time, but it requires that the potential outcomes and treatments of different groups be independent.

ASSUMPTION 4 (Strong Exogeneity): *For all* $(g,t) \in \{1,\ldots,G\} \times \{2,\ldots,T\}$, $E(Y_{g,t}(0) - Y_{g,t-1}(0)|D_{g,1},\ldots,D_{g,T}) = E(Y_{g,t}(0) - Y_{g,t-1}(0))$.

Assumption 4 requires that the shocks affecting a group's $Y_{g,t}(0)$ be mean independent of that group's treatment sequence. This rules out the possibility that a group gets treated because it experiences negative shocks, the so-called Ashenfelter's dip (see Ashenfelter 1978). Assumption 4 is related to the strong exogeneity condition in panel data models, which, as is well known, is necessary to obtain the consistency of the fixed effects estimator (see, e.g., Wooldridge 2002).

We now define the FE regression described in the introduction.[4]

REGRESSION 1 (Fixed Effects Regression): *Let* $\hat{\beta}_{fe}$ *denote the coefficient of* $D_{g,t}$ *in an OLS regression of* $Y_{i,g,t}$ *on group fixed effects, period fixed effects, and* $D_{g,t}$. *Let* $\beta_{fe} = E[\hat{\beta}_{fe}]$.[5]

For all g and t, let $N_{g,.} = \sum_{t=1}^{T} N_{g,t}$ and $N_{.,t} = \sum_{g=1}^{G} N_{g,t}$ respectively denote the total number of observations in group g and in period t. For any variable $X_{g,t}$ defined in each (g,t) cell, let $X_{g,.} = \sum_{t=1}^{T} (N_{g,t}/N_{g,.}) X_{g,t}$ denote the average value of $X_{g,t}$ in group g, let $X_{.,t} = \sum_{g=1}^{G} (N_{g,t}/N_{.,t}) X_{g,t}$ denote the average value of $X_{g,t}$

[4] Throughout the paper, we assume that $D_{g,t}$ in Regression 1 and $D_{g,t} - D_{g,t-1}$ in Regression 2 are not collinear with the other independent variables in those regressions, so $\hat{\beta}_{fe}$ and $\hat{\beta}_{fd}$ are well defined.

[5] As the independent variables in Regression 1 are constant within each (g,t) cell, Regression 1 is equivalent to a (g,t)-level regression of $Y_{g,t}$ on group and period fixed effects and $D_{g,t}$, weighted by $N_{g,t}$.

in period t, and let $X_{.,.} = \sum_{g,t}(N_{g,t}/N)X_{g,t}$ denote the average value of $X_{g,t}$. For instance, $D_{3,.}$ and $D_{.,2}$ respectively denote the average treatment in group 3 across time and in period 2 across groups, whereas $Y_{.,.}$ denotes the average value of the outcome across groups and time. Finally, for any variable $X_{g,t}$, we let $\mathbf{X}$ denote the vector $(X_{g,t})_{(g,t)\in\{1,\ldots,G\}\times\{1,\ldots,T\}}$ collecting the values of that variable in each (g,t) cell. For instance, $\mathbf{D}$ is the vector $(D_{g,t})_{(g,t)\in\{1,\ldots,G\}\times\{1,\ldots,T\}}$ collecting the treatments of all the (g,t) cells.

II. Two-Way Fixed Effects Regressions

A. *A Decomposition Result*

We study the FE regression under the following common trends assumption.

ASSUMPTION 5 (Common Trends): *For* $t \geq 2$, $E(Y_{g,t}(0) - Y_{g,t-1}(0))$ *does not vary across* g.

Assumption 5 requires that the expectation of the outcome without treatment follow the same evolution over time in every group. When t represents birth cohorts, Assumption 5 requires that the outcome difference between consecutive cohorts be the same across groups.

Let $N_1 = \sum_{i,g,t} D_{i,g,t}$ denote the number of treated units, let

$$\Delta^{TR} = \frac{1}{N_1} \sum_{(i,g,t):D_{g,t}=1} \left[Y_{i,g,t}(1) - Y_{i,g,t}(0)\right]$$

denote the average treatment effect across all treated units, and let $\delta^{TR} = E[\Delta^{TR}]$ denote the expectation of that parameter, hereafter referred to as the ATT. For any $(g,t) \in \{1,\ldots,G\} \times \{1,\ldots,T\}$, let

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

denote the ATE in cell (g,t). Note that δ^{TR} is equal to the expectation of a weighted average of the treated cells' $\Delta_{g,t}$,

$$\delta^{TR} = E\left[\sum_{g,t:D_{g,t}=1} \frac{N_{g,t}}{N_1} \Delta_{g,t}\right]. \tag{2}$$

Under the common trends assumption, we show that β_{fe} is also equal to the expectation of a weighted sum of the $\Delta_{g,t}$ terms, with potentially some negative weights.

Let $\varepsilon_{g,t}$ denote the residual of observations in cell (g,t) in the regression of $D_{g,t}$ on group and period fixed effects,[6]

$$D_{g,t} = \alpha + \gamma_g + \lambda_t + \varepsilon_{g,t}.$$

[6] $\varepsilon_{g,t}$ arises from a unit-level regression, where the dependent and independent variables only vary at the (g,t) level. Therefore, all the units in the same (g,t) cell have the same value of $\varepsilon_{g,t}$.

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

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

THEOREM 1: *Suppose that Assumptions 1–5 hold. Then*,[7]

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

This result implies that in general, $\beta_{fe} \neq \delta^{TR}$, so $\hat{\beta}_{fe}$ is a biased estimator of the ATT. To illustrate this, we consider a simple example of a staggered adoption design with two groups and three periods, and where the treatments are nonstochastic: group 1 is untreated at periods 1 and 2 and treated at period 3, while group 2 is untreated at period 1 and treated both at periods 2 and 3.[8] We also assume that $N_{g,t}/N_{g,t-1}$ does not vary across g: all groups experience the same growth of their number of observations from $t-1$ to t, a requirement that is for instance satisfied when the data is a balanced panel. Then, one can show that

$$\varepsilon_{g,t} = D_{g,t} - D_{g,.} - D_{.,t} + D_{.,.},$$

thus implying that

$$\varepsilon_{1,3} = 1 - 1/3 - 1 + 1/2 = 1/6,$$

$$\varepsilon_{2,2} = 1 - 2/3 - 1/2 + 1/2 = 1/3,$$

$$\varepsilon_{2,3} = 1 - 2/3 - 1 + 1/2 = -1/6.$$

The residual is negative in group 2 and period 3, because the regression predicts a treatment probability larger than one in that cell, a classic extrapolation problem with linear regressions. Then, under the common trends assumption, it follows from Theorem 1 and the fact that the treatments are nonstochastic that

$$\beta_{fe} = 1/2E\big[\Delta_{1,3}\big] + E\big[\Delta_{2,2}\big] - 1/2E\big[\Delta_{2,3}\big].$$

Here, β_{fe} is equal to a weighted sum of the ATEs in group 1 at period 3, group 2 at period 2, and group 2 at period 3, the three treated (g,t) cells. However, the weight assigned to each ATE differs from $1/3$, the proportion that each cell accounts for

[7] In the proof, we show the following, stronger result:

$$E\big[\hat{\beta}_{fe}|\mathbf{D}\big] = \sum_{(g,t):D_{g,t}=1}\frac{N_{g,t}}{N_1}w_{g,t}E\big[\Delta_{g,t}|\mathbf{D}\big].$$

[8] A similar example appears in Borusyak and Jaravel (2017).

in the population of treated observations. Therefore, β_{fe} is not equal to δ^{TR}. Perhaps more worryingly, not all the weights are positive: the weight assigned to the ATE in group 2, period 3 is strictly negative. Consequently, β_{fe} may be a very misleading measure of the treatment effect. Assume for instance that $E[\Delta_{1,3}] = E[\Delta_{2,2}] = 1$ and $E[\Delta_{2,3}] = 4$. At the period when they start receiving the treatment, both groups experience a modest positive ATE. But this effect builds over time and in period 3, one period after it has started receiving the treatment, group 2 now experiences a large ATE. Then,

$$\beta_{fe} = 1/2 \times 1 + 1 - 1/2 \times 4 = -1/2.$$

Therefore, β_{fe} is strictly negative, while $E[\Delta_{1,3}]$, $E[\Delta_{2,2}]$, and $E[\Delta_{2,3}]$ are all positive. More generally, the negative weights are an issue if the $E[\Delta_{g,t}]$ terms are heterogeneous, across groups or over time.[9] If $E[\Delta_{1,3}] = E[\Delta_{2,2}] = E[\Delta_{2,3}] = 1$, then $\beta_{fe} = 1 = \delta^{TR}$.

Here is some intuition as to why one weight is negative in this example. It follows from equation (A4) in the proof of Theorem 1 (see also Theorem 1 in Goodman-Bacon 2018) that in this simple example, $\beta_{fe} = (\text{DID}_1 + \text{DID}_2)/2$, with

$$\text{DID}_1 = E\big(Y_{2,2}\big) - E\big(Y_{2,1}\big) - \Big(E\big(Y_{1,2}\big) - E\big(Y_{1,1}\big)\Big),$$

$$\text{DID}_2 = E\big(Y_{1,3}\big) - E\big(Y_{1,2}\big) - \Big(E\big(Y_{2,3}\big) - E\big(Y_{2,2}\big)\Big).$$

The first DID compares the evolution of the mean outcome from period 1 to 2 in group 2 and in group 1. The second one compares the evolution of the mean outcome from period 2 to 3 in group 1 and in group 2. The control group in the second DID, group 2, is treated both in the pre- and in the post-period. Therefore, under the common trends assumption, it follows from Lemma 1 in Appendix A (a similar result appears in Lemma 1 of de Chaisemartin 2011 and in equation (13) of Goodman-Bacon 2018) that $\text{DID}_1 = E[\Delta_{2,2}]$, but

$$\text{DID}_2 = E\big[\Delta_{1,3}\big] - \Big(E\big[\Delta_{2,3}\big] - E\big[\Delta_{2,2}\big]\Big).$$

Note that, DID_2 is equal to the ATE in group 1, period 3, minus the change in group 2's ATE between periods 2 and 3. Intuitively, the mean outcome of groups 1 and 2 may follow different trends from period 2 to 3 either because group 1 becomes treated, or because group 2's ATE changes. The intuition that negative weights arise because $\hat{\beta}_{fe}$ uses treated observations as controls also appears in Borusyak and Jaravel (2017).

We now generalize the previous illustration by characterizing the (g,t) cells whose ATEs are weighted negatively by β_{fe}.

PROPOSITION 1: *Suppose that Assumption 1 holds and for all* $t \geq 2$, $N_{g,t}/N_{g,t-1}$ *does not vary across* g. *Then, for all* (g,t,t') *such that* $D_{g,t} = D_{g,t'}$

[9] On the other hand, β_{fe} does not rule out heterogeneous treatment effects within (g,t) cells, as it is identified by variations across (g,t) cells, and does not leverage any within-cell variation.

$= 1$, $D_{.,t} > D_{.,t'}$ *implies* $w_{g,t} < w_{g,t'}$. *Similarly, for all* (g, g', t) *such that* $D_{g,t} = D_{g',t} = 1, D_{g,.} > D_{g',.}$ *implies* $w_{g,t} < w_{g',t}$.

Proposition 1 shows that β_{fe} is more likely to assign a negative weight to periods where a large fraction of groups are treated, and to groups treated for many periods. Then, negative weights are a concern when treatment effects differ between periods with many versus few treated groups, or between groups treated for many versus few periods.

Proposition 1 has interesting implications in staggered adoption designs, a special case of sharp designs defined as follows.

ASSUMPTION 6 (Staggered Adoption Designs): *For all* g, $D_{g,t} \geq D_{g,t-1}$ *for all* $t \geq 2$.

Assumption 6 is satisfied in applications where groups adopt a treatment at heterogeneous dates (see, e.g., Athey and Stern 2002). In that design, Borusyak and Jaravel (2017) shows that β_{fe} is more likely to assign a negative weight to treatment effects at the last periods of the panel. This result is a special case of Proposition 1: in staggered adoption designs, $D_{.,t}$ is increasing in t, so Proposition 1 implies that $w_{g,t}$ is decreasing in t.[10] Proposition 1 also implies that in that design, groups that adopt the treatment earlier are more likely to receive some negative weights.

Finally, in staggered adoption designs, Athey and Imbens (2018) derives a decomposition of β_{fe} that resembles, but differs from, that in Theorem 1. They derive their decomposition under the assumption that the dates at which each group starts receiving the treatment are randomly assigned, while we derive ours under a common trends assumption.

B. *Robustness to Heterogeneous Treatment Effects*

Theorem 1 shows that in sharp designs with many groups and periods, $\hat{\beta}_{fe}$ may be a misleading measure of the treatment effect under the standard common trends assumption, if the treatment effect is heterogeneous across groups and time periods. In the corollary below, we propose two robustness measures that can be used to assess how serious that concern is.

Those robustness measures are defined conditional on $\mathbf{D}$, the vector stacking together the treatments of all the (g,t) cells. Specifically, for all $(g,t) \in \{1,\ldots,G\} \times \{1,\ldots,T\}$, let $\tilde{\Delta}_{g,t} = E(\Delta_{g,t}|\mathbf{D})$ denote the ATE in cell (g,t) conditional on $\mathbf{D}$,[11] let $\tilde{\Delta}^{TR} = E(\Delta^{TR}|\mathbf{D})$ denote the ATT conditional on $\mathbf{D}$, and let $\tilde{\beta}_{fe} = E(\hat{\beta}_{fe}|\mathbf{D})$. The first measure we consider is the minimal value of the standard deviation of the $\tilde{\Delta}_{g,t}$ terms under which one could have that $\tilde{\beta}_{fe}$ is of a different sign than $\tilde{\Delta}^{TR}$. Therefore, this summary measure applies to $\tilde{\beta}_{fe}$ and $\tilde{\Delta}^{TR}$,

[10] Borusyak and Jaravel (2017) assumes that the treatment effect of cell (g,t) only depends on the number of periods since group g has started receiving the treatment, whereas Proposition 1 does not rely on that assumption.

[11] $\tilde{\Delta}_{g,t}$ may differ from $E(\Delta_{g,t})$. To see this, let us consider a simple example where $T = 2$. Then, under Assumption 3, one has $\tilde{\Delta}_{g,t} = E(\Delta_{g,t}|D_{g,1}, D_{g,2})$. One may for instance have $E(\Delta_{g,1}|D_{g,1} = 0, D_{g,2} = 0) < E(\Delta_{g,1}|D_{g,1} = 1, D_{g,2} = 1)$, if a group is more likely to be treated if her treatment effect is initially high.

rather than β_{fe} and δ^{TR}, the unconditional expectations of $\hat{\beta}_{fe}$ and Δ^{TR} on which we have focused so far. However, one can show that when G, the number of groups, goes to infinity, $\tilde{\beta}_{fe} - \beta_{fe}$ and $\tilde{\Delta}^{TR} - \delta^{TR}$ both converge to 0. So if the number of groups is large, $\tilde{\beta}_{fe}$ and $\tilde{\Delta}^{TR}$ should not differ much from β_{fe} and δ^{TR}, and our robustness measure "almost" applies to β_{fe} and δ^{TR}.

Let

$$\sigma(\tilde{\mathbf{\Delta}}) = \left(\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \left(\tilde{\Delta}_{g,t} - \tilde{\Delta}^{TR} \right)^2 \right)^{1/2},$$

$$\sigma(\mathbf{w}) = \left(\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \left(w_{g,t} - 1 \right)^2 \right)^{1/2},$$

where $\sigma(\tilde{\mathbf{\Delta}})$ is the standard deviation of the conditional ATEs, and $\sigma(\mathbf{w})$ is the standard deviation of the $\mathbf{w}$-weights,[12] across the treated (g,t) cells. Let $n = \#\{(g,t): D_{g,t} = 1\}$ denote the number of treated cells. For every $i \in \{1, \ldots, n\}$, let $w_{(i)}$ denote the ith largest of the weights of the treated cells: $w_{(1)} \geq w_{(2)} \geq \cdots \geq w_{(n)}$, and let $N_{(i)}$ and $\tilde{\Delta}_{(i)}$ be the number of observations and the conditional ATE of the corresponding cell. Then, for any $k \in \{1, \ldots, n\}$, let $P_k = \sum_{i \geq k} N_{(i)}/N_1$, $S_k = \sum_{i \geq k} (N_{(i)}/N_1) w_{(i)}$, and $T_k = \sum_{i \geq k} (N_{(i)}/N_1) w_{(i)}^2$.

COROLLARY 1: *Suppose that Assumptions 1–5 hold.*

(*i*) *If* $\sigma(\mathbf{w}) > 0$, *the minimal value of* $\sigma(\tilde{\mathbf{\Delta}})$ *compatible with* $\tilde{\beta}_{fe}$ *and* $\tilde{\Delta}^{TR} = 0$ *is*

$$\underline{\sigma}_{fe} = \frac{|\tilde{\beta}_{fe}|}{\sigma(\mathbf{w})}.$$

(*ii*) *If* $\tilde{\beta}_{fe} \neq 0$ *and at least one of the* $w_{g,t}$ *weights is strictly negative, the minimal value of* $\sigma(\tilde{\mathbf{\Delta}})$ *compatible with* $\tilde{\beta}_{fe}$ *and with* $\tilde{\Delta}_{g,t}$ *of a different sign than* $\tilde{\beta}_{fe}$ *for all* (g,t) *is*

$$\underline{\underline{\sigma}}_{fe} = \frac{|\tilde{\beta}_{fe}|}{\left[T_s + S_s^2/(1 - P_s) \right]^{1/2}},$$

where $s = \min\{i \in \{1, \ldots, n\}: w_{(i)} < -S_{(i)}/(1 - P_{(i)})\}$.

Note that $\underline{\sigma}_{fe}$ and $\underline{\underline{\sigma}}_{fe}$ can be estimated simply by replacing $\tilde{\beta}_{fe}$ by $\hat{\beta}_{fe}$. An estimator of $\underline{\sigma}_{fe}$ can be used to assess the robustness of $\hat{\beta}_{fe}$ to treatment effect heterogeneity across groups and periods. If $\underline{\sigma}_{fe}$ is close to 0, $\tilde{\beta}_{fe}$ and $\tilde{\Delta}^{TR}$ can be of opposite signs even under a small and plausible amount of treatment effect heterogeneity. In that case, treatment effect heterogeneity would be a serious concern for the validity of $\hat{\beta}_{fe}$. On the contrary, if $\underline{\sigma}_{fe}$ is very large, $\tilde{\beta}_{fe}$ and $\tilde{\Delta}^{TR}$ can only be of opposite signs under a very large and implausible amount of treatment effect heterogeneity. Then, treatment effect heterogeneity is less of a concern.

[12] One can show that $\sum_{(g,t):D_{g,t}=1} (N_{g,t}/N_1) w_{g,t} = 1$.

Similarly, if $\underline{\underline{\sigma}}_{fe}$ is close to 0, one may have, say, $\tilde{\beta}_{fe} > 0$, while $\tilde{\Delta}_{g,t} \leq 0$ for all (g,t), even if the dispersion of the $\tilde{\Delta}_{g,t}$ terms is relatively small. Notice that $\underline{\underline{\sigma}}_{fe}$ is only defined if at least one of the weights is strictly negative: if all the weights are positive, then one cannot have that $\tilde{\beta}_{fe}$ is of a different sign than all the $\tilde{\Delta}_{g,t}$ terms.

When some of the weights $w_{g,t}$ are negative, $\hat{\beta}_{fe}$ may still be robust to heterogeneous treatment effects across groups and periods, provided the assumption below is satisfied.

ASSUMPTION 7 (**w** Uncorrelated with $\tilde{\mathbf{\Delta}}$): $E[\sum_{(g,t):D_{g,t}=1}(N_{g,t}/N_1)(w_{g,t}-1) \times (\tilde{\Delta}_{g,t} - \tilde{\Delta}^{TR})] = 0$.

COROLLARY 2: *If Assumptions 1–5 and 7 hold, then* $\beta_{fe} = \delta^{TR}$.

Assumption 7 requires that the weights attached to the fixed effects estimator be uncorrelated with the conditional ATEs in the treated (g,t) cells. This is often implausible. For instance, groups treated the most are also those with the lowest value of $w_{g,t}$, as shown in Proposition 1. But those groups could also be those with the largest treatment effect. This would then induce a negative correlation between **w** and $\tilde{\mathbf{\Delta}}$. The plausibility of Assumption 7 can be assessed, by looking at whether **w** is correlated with a predictor of the treatment effect in each (g,t) cell. In the two applications we revisit in Section V, this test is rejected.

C. *Extension to the First-Difference Regression*

Instead of Regression 1, many articles have estimated the first-difference regression defined below.

REGRESSION 2 (First-Difference Regression): *Let* $\hat{\beta}_{fd}$ *denote the coefficient of* $D_{g,t} - D_{g,t-1}$ *in an OLS regression of* $Y_{g,t} - Y_{g,t-1}$ *on period fixed effects and* $D_{g,t} - D_{g,t-1}$, *among observations for which* $t \geq 2$. *Let* $\beta_{fd} = E[\hat{\beta}_{fd}]$.

When $T = 2$ and $N_{g,2}/N_{g,1}$ does not vary across g, meaning that all groups experience the same growth of their number of units from period 1 to 2, one can show that $\hat{\beta}_{fe} = \hat{\beta}_{fd}$. But, $\hat{\beta}_{fe}$ differs from $\hat{\beta}_{fd}$ if $T > 2$ or $N_{g,2}/N_{g,1}$ varies across g.

We start by showing that a result similar to Theorem 1 also applies to $\hat{\beta}_{fd}$. For any $(g,t) \in \{1,\ldots,G\} \times \{2,\ldots,T\}$, let $\varepsilon_{fd,g,t}$ denote the residual of observations in group g and at period t in the regression of $D_{g,t} - D_{g,t-1}$ on period fixed effects, among observations for which $t \geq 2$. For any $g \in \{1,\ldots,G\}$, let $\varepsilon_{fd,g,1} = \varepsilon_{fd,g,T+1} = 0$. One can show that if the regressors in Regression 2 are not perfectly collinear,

$$\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1}\left(\varepsilon_{fd,g,t} - \frac{N_{g,t+1}}{N_{g,t}}\varepsilon_{fd,g,t+1}\right) \neq 0.$$

Then we define

$$w_{fd,g,t} = \frac{\varepsilon_{fd,g,t} - \frac{N_{g,t+1}}{N_{g,t}}\varepsilon_{fd,g,t+1}}{\sum_{(g,t):D_{g,t}=1}\frac{N_{g,t}}{N_1}\left(\varepsilon_{fd,g,t} - \frac{N_{g,t+1}}{N_{g,t}}\varepsilon_{fd,g,t+1}\right)}.$$

THEOREM 2: *Suppose that Assumptions 1–5 hold. Then*,

$$\beta_{fd} = E\left[\sum_{(g,t):D_{g,t}=1}\frac{N_{g,t}}{N_1}w_{fd,g,t}\Delta_{g,t}\right].$$

Theorem 2 shows that under Assumption 5, β_{fd} is equal to a weighted sum of the ATEs in each treated (g,t) cell with potentially some strictly negative weights, just as β_{fe}. We now characterize the (g,t) cells whose ATEs are weighted negatively by β_{fd}. To do so, we focus on staggered adoption designs, as outside of this case it is more difficult to characterize those cells. Our characterization relies on the fact that for every $t \in \{2,\ldots,T\}$, $\varepsilon_{fd,g,t} = D_{g,t} - D_{g,t-1} - (D_{.,t} - D_{.,t-1})$. Here, $\varepsilon_{fd,g,t}$ is the difference between the change of the treatment in group g between $t-1$ and t, and the average change of the treatment across all groups.

PROPOSITION 2: *Suppose that Assumptions 1–2 and 6 hold and for all* g, $N_{g,t}$ *does not depend on* t. *Then, for all* (g,t) *such that* $D_{g,t} = 1$, $w_{fd,g,t} < 0$ *if and only if* $D_{g,t-1} = 1$ *and* $D_{.,t} - D_{.,t-1} > D_{.,t+1} - D_{.,t}$ (*with the convention that* $D_{.,T+1} = D_{.,T}$).

Proposition 2 shows that for all $t \in \{2,\ldots,T-1\}$ such that the increase in the proportion of treated units is larger from $t-1$ to t than from t to $t+1$, the period-t ATE of groups already treated in $t-1$ receives a negative weight. Moreover, if, at period T, at least one group becomes treated, the ATE of groups already treated in $T-1$ also receives a negative weight. Therefore, the treatment effect arising at the date when a group starts receiving the treatment does not receive a negative weight, only long-run treatment effects do. Then, negative weights are a concern when instantaneous and long-run treatment effects may differ. Proposition 2 also shows that the prevalence of negative weights depends on how the number of groups that start receiving the treatment at date t evolves with t. Assume for instance that this number decreases with t: many groups start receiving the treatment at date 1, a bit less start at date 2, etc., a case hereafter referred to as the "more early adopters" case. Then, if $N_{g,t}$ is constant across (g,t), $D_{.,t} - D_{.,t-1}$ is decreasing in t, and all the long-run treatment effects receive negative weights, except maybe those of period T if $D_{.,T} = D_{.,T-1}$. Conversely, assume that the number of groups that start receiving the treatment at date t increases with t: few groups start receiving the treatment at date 1, a bit more start at date 2, etc., a case hereafter referred to as the "more late adopters" case. Then, if $N_{g,t}$ is constant across (g,t), $D_{.,t} - D_{.,t-1}$ is increasing in t, and only the period-T long-run treatment effects receive negative weights. Overall, negative weights are much more prevalent in the "more early adopters" than in the "more late adopters" case.

We now come back to general sharp designs where the treatment may not follow a staggered adoption. Let $\tilde{\beta}_{fd} = E(\hat{\beta}_{fd}|\mathbf{D})$ denote the expectation of $\hat{\beta}_{fd}$ conditional on the vector of treatment assignments $\mathbf{D}$. Just as for $\tilde{\beta}_{fe}$, one can show that the minimal value of $\sigma(\tilde{\mathbf{\Delta}})$ compatible with $\tilde{\beta}_{fd}$ and $\tilde{\Delta}^{TR} = 0$ is $\underline{\sigma}_{fd} = |\tilde{\beta}_{fd}|/\sigma(\mathbf{w}_{fd})$, where

$$\sigma(\mathbf{w}_{fd}) = \left(\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1}(w_{fd,g,t} - 1)^2\right)^{1/2}$$

is the standard deviation of the $\mathbf{w}_{fd}$-weights. One can also show that $\underline{\underline{\sigma}}_{fd}$, the minimal value of $\sigma(\tilde{\mathbf{\Delta}})$ compatible with $\tilde{\beta}_{fd}$ and $\tilde{\Delta}_{g,t}$ of a different sign than $\tilde{\beta}_{fd}$ for all (g,t), has the same expression as $\underline{\sigma}_{fe}$, except that one needs to replace the weights $w_{g,t}$ by the weights $w_{fd,g,t}$ in its definition. Estimators of $\underline{\sigma}_{fe}$ and $\underline{\sigma}_{fd}$ (or $\underline{\underline{\sigma}}_{fe}$ and $\underline{\underline{\sigma}}_{fd}$) can then be used to determine which of $\hat{\beta}_{fe}$ or $\hat{\beta}_{fd}$ is more robust to heterogeneous treatment effects.

Finally, and similarly to the result shown in Corollary 2 for β_{fe}, β_{fd} is equal to δ^{TR} under common trends and the following assumption.

ASSUMPTION 8 ($\mathbf{w}_{fd}$ Uncorrelated with $\tilde{\Delta}$): $E[\sum_{(g,t):D_{g,t}=1}(N_{g,t}/N_1)(w_{fd,g,t} - 1) \times (\Delta_{g,t} - \Delta^{TR})] = 0$.

Note that under the common trends assumption, one can jointly test Assumption 8 and Assumption 7, the assumption that the weights attached to β_{fe} are uncorrelated with the $\Delta_{g,t}$ terms: if $\hat{\beta}_{fe}$ and $\hat{\beta}_{fd}$ are significantly different, at least one of these two assumptions must fail. In the two applications we revisit in Section V, $\hat{\beta}_{fe}$ and $\hat{\beta}_{fd}$ are significantly different.

III. An Alternative Estimator

In this section, we show that it is possible to estimate a well-defined causal effect even if treatment effects are heterogeneous across groups or over time. Let

$$\delta^S = E\left[\frac{1}{N_S}\sum_{(i,g,t):t\geq 2, D_{g,t}\neq D_{g,t-1}} \left[Y_{i,g,t}(1) - Y_{i,g,t}(0)\right]\right],$$

with $N_S = \sum_{(g,t):t\geq 2, D_{g,t}\neq D_{g,t-1}} N_{g,t}$. The term δ^S is the ATE of all switching cells. In staggered adoption designs, δ^S is the average of the treatment effect at the time when a group starts receiving the treatment, across all groups that become treated at some point.

We now show that δ^S can be unbiasedly estimated by a weighted average of DID estimators. This result holds under the following supplementary assumptions.

ASSUMPTION 9 (Strong Exogeneity for $Y(1)$): *For all* $(g,t) \in \{1,\ldots,G\} \times \{2,\ldots,T\}$, $E(Y_{g,t}(1) - Y_{g,t-1}(1)|D_{g,1},\ldots,D_{g,T}) = E(Y_{g,t}(1) - Y_{g,t-1}(1))$.

Assumption 9 is the equivalent of Assumption 4, for the potential outcome with treatment. It requires that the shocks affecting a group's $Y_{g,t}(1)$ be mean independent of that group's treatment sequence.

ASSUMPTION 10 (Common Trends for $Y(1)$): *For* $t \geq 2$, $E(Y_{g,t}(1) - Y_{g,t-1}(1))$ *does not vary across* g.

Again, Assumption 10 is the equivalent of Assumption 5, for the potential outcome with treatment. It requires that between each pair of consecutive periods, the expectation of the outcome with treatment follow the same evolution over time in every group. Assumptions 9 and 10 ensure that one can reconstruct the potential outcome that groups leaving the treatment between $t-1$ and t would have experienced if they had remained treated. In staggered adoption designs, Assumptions 9 and 10 are not necessary for identification, because no group leaves the treatment. Together, Assumptions 5 and 10 imply that the ATE follows the same evolution over time in every group: $E(\Delta_{g,t}) = \eta_t + \theta_g$.[13] This still allows for heterogeneous treatment effects across groups and over time.[14]

ASSUMPTION 11 (Existence of "Stable" Groups): *For all* $t \in \{2, \ldots, T\}$*:*

(*i*) *If there is at least one* $g \in \{1, \ldots, G\}$ *such that* $D_{g,t-1} = 0$, $D_{g,t} = 1$, *then there exists at least one* $g' \neq g, g' \in \{1, \ldots, G\}$ *such that* $D_{g',t-1} = D_{g',t} = 0$.

(*ii*) *If there is at least one* $g \in \{1, \ldots, G\}$ *such that* $D_{g,t-1} = 1, D_{g,t} = 0$, *then there exists at least one* $g' \neq g, g' \in \{1, \ldots, G\}$ *such that* $D_{g',t-1} = D_{g',t} = 1$.

The first point of the stable groups assumption requires that between each pair of consecutive time periods, if there is a "joiner" (i.e., a group switching from being untreated to treated), then there should be another group that is untreated at both dates. The second point requires that between each pair of consecutive time periods, if there is a "leaver" (i.e., a group switching from being treated to untreated), then there should be another group that is treated at both dates.

Notice that under Assumption 11, groups' treatments are not independent, so Assumption 3 cannot hold. Accordingly, we replace Assumption 3 by Assumption 12. Assumption 12 requires that conditional on its own treatments, a group's outcomes be mean independent of the other groups' treatments. It is weaker than Assumption 3. Assumption 11 is necessary to show that our estimator is unbiased, but it is not necessary to show that it is consistent. Accordingly, in Section 5 of the online Appendix, we show that our estimator is consistent under Assumption 3. For every $g \in \{1, \ldots, G\}$, let $\mathbf{D}_g = (D_{1,g}, \ldots, D_{T,g})$.

ASSUMPTION 12 (Mean Independence between a Group's Outcome and Other Groups Treatments): *For all* g *and* t, $E(Y_{g,t}(0)|\mathbf{D}) = E(Y_{g,t}(0)|\mathbf{D}_g)$ *and* $E(Y_{g,t}(1)|\mathbf{D}) = E(Y_{g,t}(1)|\mathbf{D}_g)$.

[13] It should be possible to weaken Assumptions 9–10, in particular to account for dynamic effects where $\Delta_{g,t}$ may depend on $(D_{g,1}, \ldots, D_{g,t-1})$. This introduces complications that are beyond the scope of this paper, but that we address in de Chaisemartin and D'Haultfœuille (2020a).

[14] Imposing Assumptions 9 and 10 does not change the decompositions obtained in Theorems 1 and 2; $Y_{g,t}(1)$ is observed for all the treated (g,t) cells entering these decompositions, so those assumptions do not bring identifying information for those cells.

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

$$N_{d,d',t} = \sum_{g:D_{g,t}=d,D_{g,t-1}=d'} N_{g,t} \tag{3}$$

denote the number of observations with treatment d' at period $t-1$ and d at period t. Let

$$\text{DID}_{+,t} = \sum_{g:D_{g,t}=1,D_{g,t-1}=0} \frac{N_{g,t}}{N_{1,0,t}}\big(Y_{g,t} - Y_{g,t-1}\big) - \sum_{g:D_{g,t}=D_{g,t-1}=0} \frac{N_{g,t}}{N_{0,0,t}}\big(Y_{g,t} - Y_{g,t-1}\big),$$

$$\text{DID}_{-,t} = \sum_{g:D_{g,t}=D_{g,t-1}=1} \frac{N_{g,t}}{N_{1,1,t}}\big(Y_{g,t} - Y_{g,t-1}\big) - \sum_{g:D_{g,t}=0,D_{g,t-1}=1} \frac{N_{g,t}}{N_{0,1,t}}\big(Y_{g,t} - Y_{g,t-1}\big).$$

Note that $\text{DID}_{+,t}$ is not defined when there is no group such that $D_{g,t}=1$, $D_{g,t-1}=0$, or no group such that $D_{g,t}=0, D_{g,t-1}=0$. In such instances, we let $\text{DID}_{+,t}=0$. Similarly, let $\text{DID}_{-,t}=0$ when there is no group such that $D_{g,t}=1, D_{g,t-1}=1$ or no group such that $D_{g,t}=0, D_{g,t-1}=1$. Finally, let

$$\text{DID}_{\text{M}} = \sum_{t=2}^{T}\left(\frac{N_{1,0,t}}{N_S}\,\text{DID}_{+,t} + \frac{N_{0,1,t}}{N_S}\,\text{DID}_{-,t}\right).$$

THEOREM 3: *If Assumptions 1, 2, 4, 5, and 9–12 hold*, $E[\text{DID}_{\text{M}}] = \delta^S$.

In online Appendix Section 5, we also show that when G goes to infinity, DID_{M} is a consistent and asymptotically normal estimator of δ^S. The DID_{M} estimator is computed by the *fuzzydid* and *did_multiplegt* Stata packages.

Here is the intuition underlying Theorem 3. The estimator $\text{DID}_{+,t}$ compares the evolution of the mean outcome between $t-1$ and t in two sets of groups: the joiners, and those remaining untreated. Under Assumptions 4 and 5, $\text{DID}_{+,t}$ estimates the joiners' treatment effect. Similarly, $\text{DID}_{-,t}$ compares the evolution of the outcome between $t-1$ and t in two sets of groups: those remaining treated, and the leavers. Under Assumptions 9 and 10, it estimates the leavers' treatment effect. Finally, DID_{M} is a weighted average of those DID estimators. Note that in staggered designs, there are no groups whose treatment decreases over time, so DID_{M} is only a weighted average of the $\text{DID}_{+,t}$ estimators. Note also that one can separately estimate the joiners' and the leavers' treatment effect, by computing separately weighted averages of the $\text{DID}_{+,t}$ and $\text{DID}_{-,t}$ estimators. The former estimator only relies on Assumptions 4 and 5, while the latter only relies on Assumptions 9 and 10.

Note that, DID_{M} is related to two other estimators. First, it is related to the Wald-TC estimator in point 2 of Theorem S1 in the online Appendix of de Chaisemartin and D'Haultfœuille (2018), but the weighting of $\text{DID}_{+,t}$ and $\text{DID}_{-,t}$ therein differs. As a result, DID_{M} estimates Δ^S under weaker assumptions. Second, DID_{M} is related to

the multiperiod DID estimator in Imai and Kim (2018). However, the multiperiod DID estimator is a weighted average of the $DID_{+,t}$, so it does not estimate the leavers' treatment effect, and applies to a smaller population. Besides, Imai and Kim (2018) do not establish the properties of their estimator. Finally, they do not generalize it to nonbinary treatments, something we do in online Appendix Section 4.

There may be a bias-variance trade-off between DID_M and the two-way fixed effects regression estimators. For instance, assume that Regression 1 is correctly specified:

$$\begin{aligned} Y_{g,t}(0) &= \alpha_g + \lambda_t + \varepsilon_{g,t}, \\ Y_{g,t}(1) - Y_{g,t}(0) &= \delta, \\ E(\varepsilon_{g,t}|\mathbf{D}) &= 0. \end{aligned}$$

Then, if the errors $\varepsilon_{g,t}$ are homoskedastic and uncorrelated, it follows from the Gauss-Markov theorem that $\hat{\beta}_{fe}$ is the linear estimator of δ, the constant treatment effect parameter, with the lowest variance. As DID_M is also an unbiased linear estimator of δ, the variance of $\hat{\beta}_{fe}$ must be lower than that of DID_M. With heteroskedastic or correlated errors, one can construct examples where the variance of $\hat{\beta}_{fe}$ is higher than that of DID_M, but this still suggests that DID_M may often have a larger variance than that of $\hat{\beta}_{fe}$, as we find in our applications in Section V.

Note that, DID_M uses groups whose treatment is stable to infer the trends that would have affected switchers if their treatment had not changed. This strategy could fail, if switchers experience different trends than groups whose treatment is stable. To assess if this is a serious concern, we propose to use the following placebo estimator, that essentially compares the outcome's evolution from $t-2$ to $t-1$, in groups that switch and do not switch treatment between $t-1$ and t. This placebo estimator is defined under a modified version of Assumption 11.

ASSUMPTION 13 (Existence of "Stable" Groups for the Placebo Test): *For all* $t \in \{3,\ldots,T\}$*:*

(*i*) *If there is at least one* $g \in \{1,\ldots,G\}$ *such that* $D_{g,t-2} = D_{g,t-1} = 0$ *and* $D_{g,t} = 1$*, then there exists at least one* $g' \neq g, g' \in \{1,\ldots,G\}$ *such that* $D_{g',t-2} = D_{g',t-1} = D_{g',t} = 0$.

(*ii*) *If there is at least one* $g \in \{1,\ldots,G\}$ *such that* $D_{g,t-2} = D_{g,t-1} = 1$, $D_{g,t} = 0$, *then there exists at least one* $g' \neq g$, $g' \in \{1,\ldots,G\}$ *such that* $D_{g',t-2} = D_{g',t-1} = D_{g',t} = 1$.

For all $t \in \{2,\ldots,T\}$ and for all $(d,d',d'') \in \{0,1\}^3$, let

$$N_{d,d',d'',t} = \sum_{g:D_{g,t}=d,D_{g,t-1}=d',D_{g,t-2}=d''} N_{g,t}$$

denote the number of observations with treatment status d'' at period $t-2$, d' at period $t-1$, and d at period t. Let

$$N_S^{\text{pl}} = \sum_{(g,t):t\geq 3, D_{g,t}\neq D_{g,t-1}=D_{g,t-2}} N_{g,t},$$

$$\text{DID}_{+,t}^{\text{pl}} = \sum_{g:D_{g,t}=1, D_{g,t-1}=D_{g,t-2}=0} \frac{N_{g,t}}{N_{1,0,0,t}}\left(Y_{g,t-1}-Y_{g,t-2}\right) - \sum_{g:D_{g,t}=D_{g,t-1}=D_{g,t-2}=0} \frac{N_{g,t}}{N_{0,0,0,t}}\left(Y_{g,t-1}-Y_{g,t-2}\right),$$

$$\text{DID}_{-,t}^{\text{pl}} = \sum_{g:D_{g,t}=D_{g,t-1}=D_{g,t-2}=1} \frac{N_{g,t}}{N_{1,1,1,t}}\left(Y_{g,t-1}-Y_{g,t-2}\right) - \sum_{g:D_{g,t}=0, D_{g,t-1}=D_{g,t-2}=1} \frac{N_{g,t}}{N_{0,1,1,t}}\left(Y_{g,t-1}-Y_{g,t-2}\right).$$

When there is no group such that $D_{g,t}=1, D_{g,t-1}=D_{g,t-2}=0$ or no group such that $D_{g,t}=D_{g,t-1}=D_{g,t-2}=0$, we let $\text{DID}_{+,t}^{\text{pl}}=0$, and we adopt the same convention for $\text{DID}_{-,t}^{\text{pl}}=0$. Let

$$\text{DID}_{\text{M}}^{\text{pl}} = \sum_{t=3}^{T}\left(\frac{N_{1,0,0,t}}{N_S^{\text{pl}}}\,\text{DID}_{+,t}^{\text{pl}} + \frac{N_{0,1,1,t}}{N_S^{\text{pl}}}\,\text{DID}_{-,t}^{\text{pl}}\right).$$

THEOREM 4: *If Assumptions 1, 2, 4, 5, 9, 10, 12, and 13 hold, then* $E\left[DID_M^{pl}\right]=0$.

The $\text{DID}_{+,t}^{\text{pl}}$ estimator compares the evolution of the mean outcome from $t-2$ to $t-1$ in two sets of groups: those untreated at $t-2$ and $t-1$ but treated at t, and those untreated at $t-2$, $t-1$, and t. If Assumptions 4 and 5 hold, then $E[\text{DID}_{+,t}^{\text{pl}}]=0$. Similarly, if Assumptions 9 and 10 hold, $E[\text{DID}_{-,t}^{\text{pl}}]=0$. Then, $E[\text{DID}_{\text{M}}^{\text{pl}}]=0$ is a testable implication of Assumptions 4, 5, 9, and 10, so finding $\text{DID}_{\text{M}}^{\text{pl}}$ significantly different from 0 would imply that those assumptions are violated: groups that switch treatment experience different trends before that switch than the groups used to reconstruct their counterfactual trends when they switch.[15] Note that $\text{DID}_{\text{M}}^{\text{pl}}$ compares the trends of switching and stable groups one period before the switch. One can define other placebo estimators comparing those trends, say, two or three periods before the switch. The $\text{DID}_{\text{M}}^{\text{pl}}$ estimator and all those other placebo estimators are computed by the *did_multiplegt* Stata package.

[15] See also Callaway and Sant'Anna (2018), which proposes another placebo test in staggered adoption designs.

IV. Extensions

In this section, we briefly review some of the extensions in our online Appendix. First, we show that the decomposition of β_{fe} in Theorem 1 can be extended to fuzzy designs where the treatment varies within (g,t) cells and to applications with a nonbinary treatment.[16] In fuzzy designs or with a nonbinary treatment, the weights in Theorem 1 remain essentially unchanged.

We also consider two-way fixed effects regressions with covariates. Specifically, we study the coefficient of $D_{g,t}$ in a regression of $Y_{i,g,t}$ on group and period fixed effects, $D_{g,t}$, and a vector of covariates $X_{g,t}$. We show that a result very similar to Theorem 1 applies to that coefficient, up to two differences. First, including covariates allows for different trends across groups, provided those differential trends are fully accounted for by a linear model in $X_{g,t} - X_{g,t-1}$, the change in a group's covariates. Specifically, instead of Assumptions 4 and 5, one needs to assume that

$$E\big(Y_{g,t}(0)\big|\mathbf{D}_g,\mathbf{X}_g\big) - E\big(Y_{g,t-1}(0)\big|\mathbf{D}_g,\mathbf{X}_g\big) = \big(X_{g,t} - X_{g,t-1}\big)'\gamma + \lambda_t,$$

for some vector γ and constant λ_t, and where $\mathbf{X}_g = (X_{g,1},\ldots,X_{g,T})$. Importantly, when the covariates are group-specific linear trends, the equation above is equivalent to

$$E\big(Y_{g,t}(0)\big|\mathbf{D}_g,\mathbf{X}_g\big) - E\big(Y_{g,t-1}(0)\big|\mathbf{D}_g,\mathbf{X}_g\big) = \gamma_g + \lambda_t,$$

meaning that from $t-1$ to t, the evolution of $Y(0)$ in group g should deviate from its group-specific linear trend γ_g by an amount λ_t common to all groups. Second, the residual $\varepsilon_{g,t}$ in the weights in Theorem 1 has to be replaced by $\varepsilon^X_{g,t}$, the residual of observations in cell (g,t) in the regression of $D_{g,t}$ on group and period fixed effects and $X_{g,t}$. Some of the corresponding weights may still be negative, as in Theorem 1. Overall, two-way fixed effects regressions with covariates may rely on a more plausible common trends assumptions than those without covariates, but they still require that the treatment effect be homogeneous, across time and between groups.

Third, we show that under the common trends assumption and the assumption that the ATE of a (g,t) cell does not change over time, β_{fe} and β_{fd} identify weighted sums of the ATEs of the (g,t) cells whose treatment changes between $t-1$ and t. In sharp designs, the weights attached to β_{fd} are all positive, while for β_{fe}, the same only holds in staggered adoption designs.

Fourth, we show that our DID_M estimator can easily be extended to nonbinary, discrete treatments. Then, we define it as a weighted average of DID terms comparing the evolution of the outcome in groups whose treatment went from d to d' between $t-1$ and t and in groups with a treatment of d at both dates, across all possible values of d, d', and t.

Finally, our *twowayfeweights*, *fuzzydid*, and *did_multiplegt* Stata packages can handle all of those extensions.

[16] The decomposition of β_{fd} in Theorem 2 can also be extended to all of those cases.

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
Percent of published papers	5.2	12.2	11.2	9.8
Percent of empirical papers, excluding lab experiments	12.8	23.0	19.2	19.1

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

Table 2—Descriptive Statistics on Two-Way Fixed Effects Papers

	Number of 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. Research design	
Sharp design	26
Fuzzy design	7
Panel C. Are there stable groups?	
Yes	12
Presumably yes	14
Presumably no	5
No	2

Note: This table reports the estimation method and the research design used in the 33 papers using two-way fixed effects regressions published in the *AER* from 2010 to 2012, and whether those papers have stable groups.

V. Applicability, and Applications

A. *Applicability*

We conducted a review of all papers published in the *American Economic Review* (*AER*) between 2010 and 2012 to assess the importance of two-way fixed effects regressions in economics. Over these three years, the *AER* published 337 papers. Out of these 337 papers, 33 or 9.8 percent 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 percent.

Table 2 shows descriptive statistics about the 33 2010–2012 *AER* papers estimating two-way fixed effects regressions. Panel A shows that 13 use the FE regression; 6 use the FD regression; 6 use the FE or FD regression with several treatment variables; 3 use the FE or FD 2SLS regression discussed in online Appendix Section 3.4; 5 use other regressions that we deemed sufficiently close to the FE or FD regression to include them in our count.[17] Panel B shows that more than three-fourths of those

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

papers consider sharp designs, while less than one-fourth consider fuzzy designs. Finally, panel C assesses whether, in those applications, there are groups whose exposure to the treatment remains stable between each pair of consecutive time periods, the condition that has to be met to be able to compute the DID_M estimator. For about one-half of the papers, reading the paper was not enough to assess this with certainty. We then assessed whether they presumably have stable groups. Overall, 12 papers have stable groups, 14 presumably have stable groups, 5 presumably do not have stable groups, and 2 do not have stable groups.

In online Appendix Section 6, 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 design is a sharp or a fuzzy design, and of whether the stable groups assumption holds.

B. *Application to Gentzkow, Shapiro, and Sinkinson (2011)*

Gentzkow, Shapiro, and Sinkinson (2011) studies 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 $t-1$ and t on state-year fixed effects and on the first difference of the number of newspapers available in that county. This corresponds to Regression 2, with state-year fixed effects as controls. As reproduced in Table 3, Gentzkow, Shapiro, and Sinkinson (2011) finds that $\hat{\beta}_{fd} = 0.0026$ (standard error $= 9 \times 10^{-4}$). According to this regression, one more newspaper increased voters' turnout by 0.26 percentage points. On the other hand, $\hat{\beta}_{fe} = -0.0011$ (standard error $= 0.0011$). Here, $\hat{\beta}_{fe}$ and $\hat{\beta}_{fd}$ are significantly different (t-statistic $= 2.86$).

We use the *twowayfeweights* Stata package, downloadable with its help file from the SSC repository, to estimate the weights attached to $\hat{\beta}_{fe}$. We find that 60 percent are strictly positive, 40 percent are strictly negative. The negative weights sum to -0.53. We find $\underline{\hat{\sigma}}_{fe} = 3 \times 10^{-4}$, meaning that β_{fe} and the ATT may be of opposite signs if the standard deviation of the ATEs across all the treated (g,t) cells is equal to 0.0003.[18] Further, $\underline{\hat{\sigma}}_{fe} = 7 \times 10^{-4}$, meaning that β_{fe} may be of a different sign than the ATEs of all the treated (g,t) cells if the standard deviation of those ATEs is equal to 0.0007. We also estimate the weights attached to $\hat{\beta}_{fd}$. Here, 54 percent are strictly positive, and 46 percent are strictly negative. The negative weights sum to -1.43. We find $\underline{\hat{\sigma}}_{fd} = 4 \times 10^{-4}$, and $\underline{\hat{\sigma}}_{fd} = 6 \times 10^{-4}$.

Therefore, β_{fe} and β_{fd} can only receive a causal interpretation if the weights attached to them are uncorrelated with the intensity of the treatment effect in each county $\times$ election-year cell (Assumptions 7 and 8, respectively). This is not warranted. First, as $\hat{\beta}_{fe}$ and $\hat{\beta}_{fd}$ significantly differ, Assumptions 7 and 8 cannot jointly hold. Moreover, the weights attached to $\hat{\beta}_{fe}$ and $\hat{\beta}_{fd}$ are correlated with variables that are likely to be themselves associated with the intensity of the treatment effect in each cell. For instance, the correlation between the weights attached to $\hat{\beta}_{fd}$ and t, the year variable, is equal to -0.06 (t-statistic $= -3.28$). The effect of newspapers may be different in the last than in the first years of the panel. For instance, new

[18] The number of newspapers is not binary, so strictly speaking, in this application the parameter of interest is the average causal response parameter introduced in online Appendix Section 3.2, rather than the ATT.

Table 3—Estimates of the Effect of One Additional Newspaper on Turnout

	Estimate	Standard error	Observations
$\hat{\beta}_{fd}$	0.0026	0.0009	15,627
$\hat{\beta}_{fe}$	−0.0011	0.0011	16,872
DID_M	0.0043	0.0014	16,872
$\text{DID}_\text{M}^\text{pl}$	−0.0009	0.0016	13,221
DID_M, 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 common trends assumption underlying DID_M. Estimators are computed using the data of Gentzkow, Shapiro, and Sinkinson (2011), with state-year fixed effects as controls. Standard errors are clustered by county. To compute the DID_M estimators, the number of newspapers is grouped into 4 categories: 0, 1, 2, and more than 3.

means of communication, like the radio, appear in the end of the period under consideration, and may diminish the effect of newspapers.[19] This would lead to a violation of Assumption 8.

The stable groups assumption holds: between each pair of consecutive elections, there are counties where the number of newspapers does not change. We use the *fuzzydid* Stata package, downloadable with its help file from the SSC repository, to estimate a modified version of our DID_M estimator, that accounts for the fact that the number of newspapers is not binary (see online Appendix Section 3.2, where we define this modified estimator). We include state-year fixed effects as controls in our estimation. We find that $\text{DID}_\text{M} = 0.0043$, with a standard error of 0.0014. Therefore, DID_M is 66 percent larger than $\hat{\beta}_{fd}$, and the two estimators are significantly different at the 10 percent level (t-statistic $= 1.77$); DID_M is also of a different sign than $\hat{\beta}_{fe}$.

Our DID_M estimator only relies on a common trends assumption. To assess its plausibility, we compute $\text{DID}_\text{M}^\text{pl}$, the placebo estimator introduced in Section III.[20] As shown in Table 3, our placebo estimator is small and not significantly different from 0, meaning 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 consecutive time periods $t-1$ and t, we only keep counties where the number of newspapers did not change between $t-2$ and $t-1$. Still, almost 80 percent of the county $\times$ election-year observations are used in the computation of the placebo estimator. Moreover, when reestimated on this subsample, the DID_M estimator is very close to the DID_M estimator in the full sample.

C. *The Effect of Union Membership on Wages*

A number of articles have estimated the effect of union membership on wages using panel data and controlling for workers' fixed effects. For instance, Jakubson

[19] In fact, Gentzkow, Shapiro, and Sinkinson (2011) analyzes the 1868 to 1928 period separately from later periods, because the growth of the radio may have changed newspapers' effects.

[20] Again, we need to slightly modify $\text{DID}_\text{M}^\text{pl}$ to account for the fact that the number of newspapers is not binary.

TABLE 4—ESTIMATES OF THE UNION PREMIUM

	Estimate	Standard error	Observations
$\hat{\beta}_{fe}$	0.107	0.030	4,360
$\hat{\beta}_{fd}$	0.060	0.032	3,815
DID_M	0.041	0.034	3,815
DID_M^{pl}	0.094	0.038	3,101
$DID_M^{pl,2}$	−0.041	0.030	2,458
$DID_M^{pl,3}$	−0.004	0.033	1,881

Notes: This table reports estimates of the effect of the union premium, as well as placebo estimators of the common trends assumption. Estimators are computed using the data of Vella and Verbeek (1998). Standard errors are clustered at the worker level.

(1991) has found a 8.3 percent union membership premium using that strategy, in a sample of American males from the PSID followed from 1976 to 1980. Vella and Verbeek (1998) estimates a similar regression and find similar results, in a sample of young American males from the NLSY followed from 1980 to 1987.[21]

We use the data in Vella and Verbeek (1998) to compute various estimators of the union wage premium. As union status is often measured with error (see, e.g., Freeman 1984; Card 1996), we discard changes in union status happening twice in three consecutive years. Specifically, for individuals with $D_{i,t-1} = 0, D_{i,t} = 1$, and $D_{i,t+1} = 0$, we replace $D_{i,t}$ by 0. Similarly, for individuals with $D_{i,t-1} = 1, D_{i,t} = 0$, and $D_{i,t+1} = 1$, we replace $D_{i,t}$ by 1. Doing so, we discard half of the union status changes in the initial data.[22]

We start by estimating a two-way fixed effects regression of wages on union membership with worker and year fixed effects. Table 4 shows that $\hat{\beta}_{fe} = 0.107$ (standard error $= 0.030$), a result close to that of the worker fixed effects regressions in Jakubson (1991) and Vella and Verbeek (1998).

Then, we estimate the weights attached to $\hat{\beta}_{fe}$. Here, 820 are strictly positive, 196 are strictly negative, but the negative weights only sum to −0.01. Still, $\underline{\hat{\sigma}}_{fe} = 0.097$, meaning that β_{fe} and the ATT may be of opposite signs if the standard deviation of the treatment effect across the unionized worker × year observations is equal to 0.097, a substantial but still possible amount of heterogeneity. The weights are negatively correlated with workers' years of schooling (correlation $= -0.12$, t-statistic $= -1.88$). The union premium may be lower for more educated workers (see Freeman and Medoff 1984), as they may be less substitutable than less educated ones. Then, $\hat{\beta}_{fe}$ may overestimate δ^{TR}, the average union premium across all unionized worker × year observations. We also find that $\hat{\beta}_{fd} = 0.060$ (standard error $= 0.032$) and that $\hat{\beta}_{fe}$ and $\hat{\beta}_{fd}$ significantly differ (t-statistic $= 1.91$),[23] thus casting further doubt on Assumptions 7 and 8.

[21] The fixed effects regression is not the main specification in Vella and Verbeek (1998). The authors favor instead a dynamic selection model.

[22] Keeping the original data does not change much the results presented below, except that the placebo estimator $DID_M^{pl,2}$ becomes significant.

[23] The standard error of $\hat{\beta}_{fe} - \hat{\beta}_{fd}$ is computed with a worker-level clustered bootstrap.

The stable groups assumption holds: between each pair of consecutive years, there are workers whose union membership status does not change. We therefore compute our DID_M estimator. Table 4 shows that it is equal to 0.041 (standard error $=$ 0.034). In this case DID_M is significantly different from $\hat{\beta}_{fe}$ (t-statistic $=$ 2.60) and $\hat{\beta}_{fd}$ (t-statistic $=$ 2.36).[24] As discussed in Section III, we can also estimate separately the union premium for workers joining and leaving a union, something that was previously done by Freeman (1984). The joiners' effect estimate is equal to 0.059 (standard error $=$ 0.053), the leavers' effect is equal to 0.021 (standard error $=$ 0.044), and the two estimates do not significantly differ (t-statistic $=$ 0.55).

The estimator DID_M relies on a common trends assumption. To assess its plausibility, we compute DID_M^{pl}, the placebo estimator introduced in Section III; DID_M^{pl} compares the wage growth of workers changing and not changing their union status one period before that change. We also compute $DID_M^{pl,2}$ and $DID_M^{pl,3}$, two other placebo estimators performing the same comparison two and three periods before the change. As shown in Table 4, DID_M^{pl} is large, positive, and significant (t-statistic $=$ 2.49). On the other hand $DID_M^{pl,2}$ and $DID_M^{pl,3}$ are smaller and insignificant. Workers that become unionized start experiencing a differential positive pretrend one year before becoming unionized. This differential pretrend mostly comes from union joiners: for them, the placebo estimator is equal to 0.119 (standard error $=$ 0.051), while for union leavers the placebo is smaller (0.061) and insignificant (standard error $=$ 0.057). Therefore, the placebos suggest that even the already small and insignificant DID_M estimator may overestimate the union premium, due to a positive pretrend. In fact, the estimate of leavers' effect, for which there is no evidence of a pretrend, is very close to 0. Overall, our results indicate that there may not be a significant union wage premium.

VI. Conclusion

Almost 20 percent of empirical articles published in the *AER* between 2010 and 2012 use regressions with groups and period fixed effects to estimate treatment effects. In this paper, we show that under a common trends assumption, those regressions estimate weighted sums of the treatment effect in each group and period. The weights may be negative: in one application, we find that more than 40 percent of the weights are negative. The negative weights are an issue when the treatment effect is heterogeneous, between groups or over time. Then, one could have that the treatment's coefficient in those regressions is negative while the treatment effect is positive in every group and time period. We therefore propose a new estimator to address this problem. This estimator estimates the treatment effect in the groups that switch treatment, at the time when they switch. It does not rely on any treatment effect homogeneity condition. It is computed by the *fuzzydid* and *did_multiplegt* Stata packages. In the two applications we revisit, this estimator is significantly and economically different from the two-way fixed effects estimators.

[24] The standard errors of $\hat{\beta}_{fe} - DID_M$ and $\hat{\beta}_{fd} - DID_M$ are computed with a worker-level clustered bootstrap.

Appendix A. Proofs

One Useful Lemma

Our results rely on the following lemma.

LEMMA 1: *If Assumptions 1–5 hold, for all* $(g,g',t,t') \in \{1,\dots,G\}^2 \times \{1,\dots,T\}^2$,

$$E\big(Y_{g,t}|\mathbf{D}\big) - E\big(Y_{g,t'}|\mathbf{D}\big) - \Big(E\big(Y_{g',t}|\mathbf{D}\big) - E\big(Y_{g',t'}|\mathbf{D}\big)\Big)$$
$$= D_{g,t}E\big(\Delta_{g,t}|\mathbf{D}\big) - D_{g,t'}E\big(\Delta_{g,t'}|\mathbf{D}\big) - \Big(D_{g',t}E\big(\Delta_{g',t}|\mathbf{D}\big) - D_{g',t'}E\big(\Delta_{g',t'}|\mathbf{D}\big)\Big).$$

PROOF OF LEMMA 1:

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

$$\begin{aligned} E\big(Y_{g,t}|\mathbf{D}\big) &= E\left(\frac{1}{N_{g,t}}\sum_{i=1}^{N_{g,t}} Y_{i,g,t}|\mathbf{D}\right) \\ &= E\left(\frac{1}{N_{g,t}}\sum_{i=1}^{N_{g,t}} \big(Y_{i,g,t}(0) + D_{i,g,t}\big(Y_{i,g,t}(1) - Y_{i,g,t}(0)\big)\big)|\mathbf{D}\right) \\ &= E\big(Y_{g,t}(0)|\mathbf{D}\big) + D_{g,t}E\big(\Delta_{g,t}|\mathbf{D}\big) \\ &= E\big(Y_{g,t}(0)|\mathbf{D}_g\big) + D_{g,t}E\big(\Delta_{g,t}|\mathbf{D}\big), \end{aligned}$$

where the third equality follows from Assumption 2, and the fourth from Assumption 3. Therefore,

$$\begin{aligned} &E\big(Y_{g,t}|\mathbf{D}\big) - E\big(Y_{g,t'}|\mathbf{D}\big) - \Big(E\big(Y_{g',t}|\mathbf{D}\big) - E\big(Y_{g',t'}|\mathbf{D}\big)\Big) \\ &= E\big(Y_{g,t}(0) - Y_{g,t'}(0)|\mathbf{D}_g\big) - E\big(Y_{g',t}(0) - Y_{g',t'}(0)|\mathbf{D}_{g'}\big) \\ &\quad + D_{g,t}E\big(\Delta_{g,t}|\mathbf{D}\big) - D_{g,t'}E\big(\Delta_{g,t'}|\mathbf{D}\big) - \Big(D_{g',t}E\big(\Delta_{g',t}|\mathbf{D}\big) - D_{g',t'}E\big(\Delta_{g',t'}|\mathbf{D}\big)\Big) \\ &= E\big(Y_{g,t}(0) - Y_{g,t'}(0)\big) - E\big(Y_{g',t}(0) - Y_{g',t'}(0)\big) \\ &\quad + D_{g,t}E\big(\Delta_{g,t}|\mathbf{D}\big) - D_{g,t'}E\big(\Delta_{g,t'}|\mathbf{D}\big) - \Big(D_{g',t}E\big(\Delta_{g',t}|\mathbf{D}\big) - D_{g',t'}E\big(\Delta_{g',t'}|\mathbf{D}\big)\Big) \\ &= D_{g,t}E\big(\Delta_{g,t}|\mathbf{D}\big) - D_{g,t'}E\big(\Delta_{g,t'}|\mathbf{D}\big) - \Big(D_{g',t}E\big(\Delta_{g',t}|\mathbf{D}\big) - D_{g',t'}E\big(\Delta_{g',t'}|\mathbf{D}\big)\Big), \end{aligned}$$

where the second equality follows from Assumption 4, and the third from Assumption 5. ∎

PROOF OF THEOREM 1:

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

$$E\left(\hat{\beta}_{fe}\middle|\mathbf{D}\right) = \frac{\sum_{g,t} N_{g,t}\varepsilon_{g,t}E\left(Y_{g,t}\middle|\mathbf{D}\right)}{\sum_{g,t} N_{g,t}\varepsilon_{g,t}D_{g,t}}. \tag{A1}$$

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

$$\sum_{t=1}^{T} N_{g,t}\varepsilon_{g,t} = 0 \quad \text{for all } g \in \{1,\ldots,G\}, \tag{A2}$$

$$\sum_{g=1}^{G} N_{g,t}\varepsilon_{g,t} = 0 \quad \text{for all } t \in \{1,\ldots,T\}. \tag{A3}$$

Then,

$$\begin{aligned}
&\sum_{g,t} N_{g,t}\varepsilon_{g,t}E\left(Y_{g,t}\middle|\mathbf{D}\right)\\
&= \sum_{g,t} N_{g,t}\varepsilon_{g,t}\left(E\left(Y_{g,t}\middle|\mathbf{D}\right) - E\left(Y_{g,1}\middle|\mathbf{D}\right) - E\left(Y_{1,t}\middle|\mathbf{D}\right) + E\left(Y_{1,1}\middle|\mathbf{D}\right)\right) \qquad \text{(A4)}\\
&= \sum_{g,t} N_{g,t}\varepsilon_{g,t}\Big(D_{g,t}E\left(\Delta_{g,t}\middle|\mathbf{D}\right) - D_{g,1}E\left(\Delta_{g,1}\middle|\mathbf{D}\right)\\
&\qquad\qquad - D_{1,t}E\left(\Delta_{1,t}\middle|\mathbf{D}\right) + D_{1,1}E\left(\Delta_{1,1}\middle|\mathbf{D}\right)\Big)\\
&= \sum_{g,t} N_{g,t}\varepsilon_{g,t}D_{g,t}E\left(\Delta_{g,t}\middle|\mathbf{D}\right)\\
&= \sum_{(g,t):D_{g,t}=1} N_{g,t}\varepsilon_{g,t}E\left(\Delta_{g,t}\middle|\mathbf{D}\right). \qquad \text{(A5)}
\end{aligned}$$

The first and third equalities follow from equations (A2) and (A3). The second equality follows from Lemma 1. The fourth equality follows from Assumption 2. Finally, Assumption 2 implies that

$$\sum_{g,t} N_{g,t}\varepsilon_{g,t}D_{g,t} = \sum_{(g,t):D_{g,t}=1} N_{g,t}\varepsilon_{g,t}. \tag{A6}$$

Combining (A1), (A5), (A6) yields

$$E\left(\hat{\beta}_{fe}\middle|\mathbf{D}\right) = \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}E\left(\Delta_{g,t}\middle|\mathbf{D}\right). \tag{A7}$$

Then, the result follows from the law of iterated expectations. ∎

PROOF OF PROPOSITION 1:

If for all $t \geq 2$, $N_{g,t}/N_{g,t-1}$ does not depend on t, then it follows from the first order conditions attached to Regression 1 and a few lines of algebra that $\varepsilon_{g,t} = D_{g,t} - D_{g,.} - D_{.,t} + D_{.,.}$. Therefore, $w_{g,t}$ is proportional to $D_{g,t} - D_{g,.} - D_{.,t} + D_{.,.}$. Then, for all (g,t,t') such that $D_{g,t} = D_{g,t'}$

$= 1$, $D_{.,t} > D_{.,t'}$ implies $w_{g,t} < w_{g,t'}$. Similarly, for all (g, g', t) such that $D_{g,t} = D_{g',t} = 1, D_{g,.} > D_{g',.}$ implies $w_{g,t} < w_{g',t}$. ∎

PROOF OF COROLLARY 1:

Proof of the First Point.—If the assumptions of the corollary hold and $\tilde{\Delta}^{TR} = 0$, then

$$\begin{cases} \tilde{\beta}_{fe} = \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} \tilde{\Delta}_{g,t}, \\ 0 = \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \tilde{\Delta}_{g,t}, \end{cases}$$

where the first equality follows from (A7). These two conditions and the Cauchy-Schwarz inequality imply

$$|\tilde{\beta}_{fe}| = \left| \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \big(w_{g,t} - 1\big)\big(\tilde{\Delta}_{g,t} - \tilde{\Delta}^{TR}\big) \right| \leq \sigma(\mathbf{w})\sigma\big(\tilde{\boldsymbol{\Delta}}\big).$$

Hence, $\sigma(\tilde{\boldsymbol{\Delta}}) \geq \underline{\sigma}_{fe}$.

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

$$\tilde{\Delta}^{TR}_{g,t} = \frac{\tilde{\beta}_{fe}\big(w_{g,t} - 1\big)}{\sigma^2(\mathbf{w})}.$$

Then,

$$\tilde{\Delta}^{TR} = \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \frac{\tilde{\beta}_{fe}\big(w_{g,t} - 1\big)}{\sigma^2(\mathbf{w})} = \frac{\tilde{\beta}_{fe}}{\sigma^2(\mathbf{w})} \left(\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} - 1 \right) = 0,$$

as it follows from the definition of $w_{g,t}$ that $\sum_{(g,t):D_{g,t}=1} (N_{g,t}/N_1) w_{g,t} = 1$.

Similarly,

$$\begin{aligned} \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} \frac{\tilde{\beta}_{fe}\big(w_{g,t} - 1\big)}{\sigma^2(\mathbf{w})} &= \frac{\tilde{\beta}_{fe}}{\sigma^2(\mathbf{w})} \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\big(w_{g,t} - 1\big) \\ &= \frac{\tilde{\beta}_{fe}}{\sigma^2(\mathbf{w})} \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \big(w_{g,t} - 1\big)^2 \\ &= \tilde{\beta}_{fe}, \end{aligned}$$

where the second equality follows again from the fact that $\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} = 1$.

Proof of the Second Point.—We first suppose that $\tilde{\beta}_{fe} > 0$. We seek to solve

$$\min_{\tilde{\Delta}_{(1)},\ldots,\tilde{\Delta}_{(n)}} \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \big(\tilde{\Delta}_{(i)} - \tilde{\Delta}^{TR}\big)^2,$$

subject to

$$\tilde{\beta}_{fe} = \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} w_{(i)} \tilde{\Delta}_{(i)}, \qquad \tilde{\Delta}_{(i)} \leq 0 \quad \text{for all } i \in \{1, \ldots, n\}.$$

This is a quadratic programming problem, with a matrix that is symmetric positive but not definite. Hence, by Frank and Wolfe (1956) and the fact that the linear term in the quadratic problem is 0, the solution exists if and only if the set of constraints is not empty. If $w_{(n)} \geq 0$, the set of constraints is empty because $\sum_{i=1}^{n} (N_{(i)}/N_1) w_{(i)} \tilde{\Delta}_{(i)} \leq 0 < \tilde{\beta}_{fe}$. On the other hand, if $w_{(n)} < 0$, this set is non-empty since it includes $(0, \ldots, 0, \tilde{\beta}_{fe}/(P_{(n)} w_{(n)}))$.

We now derive the corresponding bound. For that purpose, remark that

$$\sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \left(\tilde{\Delta}_{(i)} - \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \tilde{\Delta}_{(i)} \right)^2 = \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \tilde{\Delta}_{(i)}^2 - \left(\sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \tilde{\Delta}_{(i)} \right)^2.$$

The Karush-Kuhn-Tucker necessary conditions for optimality are that for all i:

$$\begin{aligned} \tilde{\Delta}_{(i)} &= \tilde{\Delta}^{TR} + \lambda w_{(i)} - \gamma_{(i)}, \\ \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} w_{(i)} \tilde{\Delta}_{(i)} &= \tilde{\beta}_{fe}, \\ \gamma_{(i)} &\geq 0, \\ \gamma_{(i)} \tilde{\Delta}_{(i)} &= 0, \end{aligned}$$

where $\tilde{\Delta}^{TR} = \sum_{i=1}^{n} (N_{(i)}/N_1) \tilde{\Delta}_{(i)}$, 2λ is the Lagrange multiplier of the constraint $\sum_{i=1}^{n} (N_{(i)}/N_1) w_{(i)} \tilde{\Delta}_{(i)} = \tilde{\beta}_{fe}$ and $2 (N_{(i)}/N_1) \gamma_{(i)}$ is the Lagrange multiplier of the constraint $\tilde{\Delta}_{(i)} \leq 0$.

These constraints imply that $\tilde{\Delta}_{(i)} = 0$ if and only if $\tilde{\Delta}^{TR} + \lambda w_{(i)} \geq 0$. Therefore, if $\tilde{\Delta}^{TR} + \lambda w_{(i)} < 0$, $\tilde{\Delta}_{(i)} \neq 0$ so $\gamma_{(i)} = 0$, and $\tilde{\Delta}_{(i)} = \tilde{\Delta}^{TR} + \lambda w_{(i)}$. Therefore,

$$\tilde{\Delta}_{(i)} = \min\left(\tilde{\Delta}^{TR} + \lambda w_{(i)}, 0 \right). \tag{A8}$$

This equation implies that $\tilde{\Delta}_{(i)} \leq \tilde{\Delta}^{TR} + \lambda w_{(i)}$, which in turn implies that $\tilde{\Delta}^{TR} \leq \tilde{\Delta}^{TR} + \lambda$, so $\lambda \geq 0$.

As a result, $\tilde{\Delta}^{TR} + \lambda w_{(i)}$ is decreasing in i, and because $x \mapsto \min(x, 0)$ is increasing, $\tilde{\Delta}_{(i)}$ is also decreasing in i. Then $\tilde{\Delta}_{(n)} < 0$: otherwise one would have $\tilde{\Delta}_{(i)} = 0$ for all i which would imply $\tilde{\beta}_{fe} = 0$, a contradiction. Let $s = \min\{i \in \{1, \ldots, n\} : \tilde{\Delta}_{(i)} < 0\}$. Using again (A8), we get

$$\tilde{\Delta}^{TR} = \sum_{i \geq s} \frac{N_{(i)}}{N_1} \tilde{\Delta}_{(i)} = P_s \tilde{\Delta}^{TR} + \lambda S_s.$$

Therefore,

$$\tilde{\Delta}^{TR} = \frac{\lambda S_s}{1 - P_s}. \tag{A9}$$

Hence, plugging $\tilde{\Delta}^{TR}$ in (A8), we obtain that for all $i \geq s$,

$$\tilde{\Delta}_{(i)} = \lambda\left\{\frac{S_s}{1 - P_s} + w_{(i)}\right\}.$$

Finally, using again (A8), we obtain

$$\tilde{\beta}_{fe} = \sum_{i \geq s} \frac{N_{(i)}}{N_1} w_{(i)} \tilde{\Delta}_{(i)} = \lambda\left\{\frac{S_s^2}{1 - P_s} + T_s\right\}.$$

Thus,

$$\lambda = \frac{\tilde{\beta}_{fe}}{T_s + S_s^2/(1 - P_s)}.$$

Then, using what precedes,

$$\begin{aligned}
\underline{\sigma}_{fe}^2 &= \sum_{i \geq s} \frac{N_{(i)}}{N_1} \left(\lambda w_{(i)}\right)^2 + \sum_{i < s} \frac{N_{(i)}}{N_1} \left(\tilde{\Delta}^{TR}\right)^2 \\
&= \lambda^2 T_s + (1 - P_s)\left(\frac{\lambda S_s}{1 - P_s}\right)^2 \\
&= \lambda^2\left[T_s + \frac{S_s^2}{1 - P_s}\right] \\
&= \frac{\tilde{\beta}_{fe}^2}{T_s + S_s^2/(1 - P_s)}.
\end{aligned}$$

The result follows, once noted that equations (A8) and (A9) imply that $s = \min\{i \in \{1, \ldots, n\} : w_{(i)} < -S_{(i)}/(1 - P_{(i)})\}$.

Finally, consider the case $\tilde{\beta}_{fe} < 0$. By letting $\tilde{\Delta}'_{(i)} = -\tilde{\Delta}_{(i)}$ and $\tilde{\beta}'_{fe} = -\tilde{\beta}_{fe}$, we have

$$\underline{\sigma}_{fe} = \min_{\tilde{\Delta}'_{(1)} \leq 0, \ldots, \tilde{\Delta}'_{(n)} \leq 0} \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \tilde{\Delta}'^{\,2}_{(i)} - \left(\sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \tilde{\Delta}'_{(i)}\right)^2$$

subject to

$$\sum_{i=1}^{n} \frac{N_{(i)}}{N_1} w_{(i)} \tilde{\Delta}'_{(i)} = \tilde{\beta}'_{fe}.$$

This is the same program as before, with $\tilde{\beta}'_{fe}$ instead of $\tilde{\beta}_{fe}$. Therefore, by the same reasoning as before, we obtain

$$\underline{\sigma}_{fe}^2 = \frac{\left(\tilde{\beta}'_{fe}\right)^2}{T_s + S_s^2/(1 - P_s)} = \frac{\tilde{\beta}_{fe}^2}{T_s + S_s^2/(1 - P_s)}. \blacksquare$$

PROOF OF COROLLARY 2:

We have

$$\begin{aligned}\beta_{fe} &= E\left(\sum_{(g,t):D_{g,t}=1}\frac{N_{g,t}}{N_1}w_{g,t}\tilde{\Delta}_{g,t}\right)\\ &= E\left(\left(\sum_{(g,t):D_{g,t}=1}\frac{N_{g,t}}{N_1}w_{g,t}\right)\tilde{\Delta}^{TR}\right)\\ &= E\left(\tilde{\Delta}^{TR}\right)\\ &= \delta^{TR}.\end{aligned}$$

The first equality follows from the law of iterated expectations and (A7). The second equality follows from Assumption 7. By the definition of $w_{g,t}$, $\sum_{(g,t):D_{g,t}=1}(N_{g,t}/N_1)\,w_{g,t} = 1$, hence the third equality. The fourth equality follows from the law of iterated expectations. ∎

PROOF OF THEOREM 2:

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

$$E\left(\hat{\beta}_{fd}\middle|\mathbf{D}\right) = \frac{\sum_{(g,t):t\geq 2}N_{g,t}\varepsilon_{fd,g,t}\left(E\left(Y_{g,t}\middle|\mathbf{D}\right) - E\left(Y_{g,t-1}\middle|\mathbf{D}\right)\right)}{\sum_{(g,t):t\geq 2}N_{g,t}\varepsilon_{fd,g,t}\left(D_{g,t} - D_{g,t-1}\right)}. \tag{A10}$$

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

$$\sum_{g=1}^{G}N_{g,t}\varepsilon_{fd,g,t} = 0 \quad \text{for all } t \in \{2,\ldots,T\}. \tag{A11}$$

Then,

$$\begin{aligned}&\sum_{(g,t):t\geq 2}N_{g,t}\varepsilon_{fd,g,t}\left(E\left(Y_{g,t}\middle|\mathbf{D}\right) - E\left(Y_{g,t-1}\middle|\mathbf{D}\right)\right) \qquad (A12)\\ &= \sum_{(g,t):t\geq 2}N_{g,t}\varepsilon_{fd,g,t}\Big(E\left(Y_{g,t}\middle|\mathbf{D}\right) - E\left(Y_{g,t-1}\middle|\mathbf{D}\right)\\ &\qquad\qquad - E\left(Y_{1,t}\middle|\mathbf{D}\right) - E\left(Y_{1,t-1}\middle|\mathbf{D}\right)\Big)\\ &= \sum_{(g,t):t\geq 2}N_{g,t}\varepsilon_{fd,g,t}\left(D_{g,t}\tilde{\Delta}_{g,t} - D_{g,t-1}\tilde{\Delta}_{g,t-1} - D_{1,t}\tilde{\Delta}_{1,t} + D_{1,t-1}\tilde{\Delta}_{1,t-1}\right)\\ &= \sum_{(g,t):t\geq 2}N_{g,t}\varepsilon_{fd,g,t}\left(D_{g,t}\tilde{\Delta}_{g,t} - D_{g,t-1}\tilde{\Delta}_{g,t-1}\right)\\ &= \sum_{g,t}\left(N_{g,t}\varepsilon_{fd,g,t} - N_{g,t+1}\varepsilon_{fd,g,t+1}\right)D_{g,t}\tilde{\Delta}_{g,t}\\ &= \sum_{(g,t):D_{g,t}=1}N_{g,t}\left(\varepsilon_{fd,g,t} - \frac{N_{g,t+1}}{N_{g,t}}\varepsilon_{fd,g,t+1}\right)\tilde{\Delta}_{g,t}.\end{aligned}$$

The first and third equalities follow from (A11). The second equality follows from Lemma 1. The fourth equality follows from a summation by part, and from the fact $\varepsilon_{fd,g,1} = \varepsilon_{fd,g,T+1} = 0$. The fifth equality follows from Assumption 2.

A similar reasoning yields

$$\text{(A13)} \qquad \sum_{(g,t):t\geq 2} N_{g,t}\varepsilon_{fd,g,t}\big(D_{g,t} - D_{g,t-1}\big) = \sum_{(g,t):D_{g,t}=1} N_{g,t}\left(\varepsilon_{fd,g,t} - \frac{N_{g,t+1}}{N_{g,t}}\varepsilon_{fd,g,t+1}\right).$$

Combining (A10), (A12), (A13), and the law of iterated expectations yields the result. ∎

PROOF OF PROPOSITION 2:

It follows from the first order conditions attached to Regression 2 and a few lines of algebra that $\varepsilon_{fd,g,t} = D_{g,t} - D_{g,t-1} - D_{.,t} + D_{.,t-1}$. Therefore, under Assumption 6 and if $N_{g,t}$ does not vary across t, one has that for all (g,t) such that $D_{g,t} = 1$, $1 \leq t \leq T-1$, $w_{fd,g,t}$ is proportional to $1 - D_{g,t-1} - (2D_{.,t} - D_{.,t-1} - D_{.,t+1})$. Now, $D_{.,t} - D_{.,t-1} \leq 1$, and under Assumption 6 $D_{.,t} - D_{.,t+1} \leq 0$, so $1 - D_{g,t-1} - (2D_{.,t} - D_{.,t-1} - D_{.,t+1})$ can only be strictly negative if $D_{g,t-1} = 1$. Then, for all (g,t) such that $D_{g,t} = 1$, $1 \leq t \leq T-1$, $w_{fd,g,t}$ is strictly negative if and only if $D_{g,t-1} = 1$ and $2D_{.,t} - D_{.,t-1} - D_{.,t+1} > 0$.

Similarly, when $t = T$, under the same assumptions as above, one has that for all g such that $D_{g,T} = 1$, $w_{fd,g,T}$ is proportional to $1 - D_{g,T-1} - (D_{.,T} - D_{.,T-1})$. Now, $D_{.,T} - D_{.,T-1} \leq 1$, so $1 - D_{g,T-1} - (D_{.,T} - D_{.,T-1})$ can only be strictly negative if $D_{g,T-1} = 1$. Then, $w_{fd,g,T}$ is strictly negative if and only if $D_{g,T-1} = 1$ and $D_{.,T} - D_{.,T-1} > 0$.

Finally, when $t = 1$, one has that for all g such that $D_{g,1} = 1$, $D_{g,2} = 1$ under Assumption 6, so $w_{fd,g,1}$ is proportional to $D_{.,2} - D_{.,1}$, which is greater than 0 under Assumption 6. ∎

PROOF OF THEOREM 3:

First, by definition of DID_M,

$$\text{(A14)} \qquad E\big(\text{DID}_\text{M}\big) = \sum_{t=2}^{T} E\left(\left(\frac{N_{1,0,t}}{N_S}E\big(\text{DID}_{+,t}\big|\mathbf{D}\big) + \frac{N_{0,1,t}}{N_S}E\big(\text{DID}_{-,t}\big|\mathbf{D}\big)\right)\right).$$

Let t be greater than 2, and let us focus for now on the case where there is at least one g_1 such that $D_{g_1,t-1} = 0$ and $D_{g_1,t} = 1$. Then Assumption 11 ensures that there is a least another group g_2 such that $D_{g_2,t-1} = D_{g_2,t} = 0$. For every g such that $D_{g,t-1} = 0$ and $D_{g,t} = 1$, we have

$$\text{(A15)} \qquad E\big(Y_{g,t} - Y_{g,t-1}\big|\mathbf{D}\big) = E\big(\Delta_{g,t}\big|\mathbf{D}\big) + E\big(Y_{g,t}(0) - Y_{g,t-1}(0)\big|\mathbf{D}\big).$$

Under Assumptions 12, 4, and 5, for all $t \geq 2$, there exists a real number $\psi_{0,t}$ such that for all g,

$$\text{(A16)} \qquad \begin{aligned} E\big(Y_{g,t}(0) - Y_{g,t-1}(0)\big|\mathbf{D}\big) &= E\big(Y_{g,t}(0) - Y_{g,t-1}(0)\big|\mathbf{D}_g\big) \\ &= E\big(Y_{g,t}(0) - Y_{g,t-1}(0)\big) = \psi_{0,t}. \end{aligned}$$

Then,

$$
\begin{aligned}
(\text{A17}) \qquad & N_{1,0,t} E\big(\text{DID}_{+,t} \big| \mathbf{D}\big) \\
&= \sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t} E\big(\Delta_{g,t} \big| \mathbf{D}\big) \\
&\quad + \sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t} E\big(Y_{g,t}(0) - Y_{g,t-1}(0) \big| \mathbf{D}\big) \\
&\quad - \frac{N_{1,0,t}}{N_{0,0,t}} \sum_{g:D_{g,t}=D_{g,t-1}=0} N_{g,t} E\big(Y_{g,t}(0) - Y_{g,t-1}(0) \big| \mathbf{D}\big) \\
&= \sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t} E\big(\Delta_{g,t} \big| \mathbf{D}\big) \\
&\quad + \psi_{0,t} \left(\sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t} - \frac{N_{1,0,t}}{N_{0,0,t}} \sum_{g:D_{g,t}=D_{g,t-1}=0} N_{g,t} \right) \\
&= \sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t} E\big(\Delta_{g,t} \big| \mathbf{D}\big).
\end{aligned}
$$

The first equality follows by (A15), the second by (A16), and the third after some algebra. If there is no g such that $D_{g,t-1} = 0$ and $D_{g,t} = 1$, (A17) still holds, as $\text{DID}_{+,t} = 0$ in this case.

A similar reasoning yields

$$
(\text{A18}) \qquad N_{0,1,t} E\big(\text{DID}_{-,t} \big| \mathbf{D}\big) = \sum_{g:D_{g,t}=0,D_{g,t-1}=1} N_{g,t} E\big(\Delta_{g,t} \big| \mathbf{D}\big).
$$

Plugging (A17) and (A18) into (A14) yields

$$
\begin{aligned}
E\big(\text{DID}_{\text{M}}\big) &= \sum_{t=2}^{T} E\left(E\left(\frac{1}{N_S} \left(\sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t} \Delta_{g,t} + \sum_{g:D_{g,t}=0,D_{g,t-1}=1} N_{g,t} \Delta_{g,t} \right) \middle| \mathbf{D} \right) \right) \\
&= \delta^S. \blacksquare
\end{aligned}
$$

PROOF OF THEOREM 4:

First, as with DID_{M}, we have

$$
(\text{A19}) \quad E\big(\text{DID}^{\text{pl}}_{\text{M}}\big) = \sum_{t=3}^{T} E\left(\left(\frac{N_{1,0,0,t}}{N^{\text{pl}}_S} E\big(\text{DID}^{\text{pl}}_{+,t} \big| \mathbf{D}\big) + \frac{N_{0,1,1,t}}{N^{\text{pl}}_S} E\big(\text{DID}^{\text{pl}}_{-,t} \big| \mathbf{D}\big) \right) \right).
$$

Let t be greater than 3, and let us for now focus on the case where there exists at least one g_1 such that $D_{g_1,t-2} = D_{g_1,t-1} = 0$ and $D_{g_1,t} = 1$. Then Assumption 13 ensures that there is a least another group g_2 such that $D_{g_2,t-2} = D_{g_2,t-1} = D_{g_2,t} = 0$. Then,

$$
\begin{aligned}
\text{(A20)} \quad & N_{1,0,0,t} E\left(\text{DID}^{\text{pl}}_{+,t} \middle| \mathbf{D}\right) \\
&= \sum_{g:D_{g,t}=1,D_{g,t-1}=D_{g,t-2}=0} N_{g,t} E\left(Y_{g,t-1}(0) - Y_{g,t-2}(0) \middle| \mathbf{D}\right) \\
&\quad - \frac{N_{1,0,0,t}}{N_{0,0,0,t}} \sum_{g:D_{g,t}=D_{g,t-1}=D_{g,t-2}=0} N_{g,t} E\left(Y_{g,t-1}(0) - Y_{g,t-2}(0) \middle| \mathbf{D}\right) \\
&= \psi_{0,t-1} \left(\sum_{g:D_{g,t}=1,D_{g,t-1}=D_{g,t-2}=0} N_{g,t} - \frac{N_{1,0,0,t}}{N_{0,0,0,t}} \sum_{g:D_{g,t}=D_{g,t-1}=D_{g,t-2}=0} N_{g,t} \right) \\
&= 0.
\end{aligned}
$$

The second equality follows by (A16), and the third follows after some algebra. If there exists no g such that $D_{g,t-2} = D_{g,t-1} = 0$ and $D_{g,t} = 1$, (A20) still holds, as $\text{DID}^{\text{pl}}_{+,t} = 0$ in this case.

A similar reasoning yields

$$
\text{(A21)} \qquad N_{0,1,1,t} E\left(\text{DID}^{\text{pl}}_{-,t} \middle| \mathbf{D}\right) = 0.
$$

The result follows after plugging (A20) and (A21) into (A19). ∎

REFERENCES

Abraham, Sarah, and Liyang Sun. 2018. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." Unpublished.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics* 60 (1): 47–57.

Athey, Susan, and Guido W. Imbens. 2018. "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption." NBER Working Paper 24963.

Athey, Susan, and Scott Stern. 2002. "The Impact of Information Technology on Emergency Health Care Outcomes." *RAND Journal of Economics* 33 (3): 399–432.

Autor, David H. 2003. "Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing." *Journal of Labor Economics* 21 (1): 1–42.

Borusyak, Kirill, and Xavier Jaravel. 2017. "Revisiting Event Study Designs." Unpublished.

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2018. "Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment." arXiv e-print 1803.09015.

Card, David. 1996. "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis." *Econometrica* 64 (4): 957–79.

de Chaisemartin, Clément. 2011. "Fuzzy Differences in Differences." Center for Research in Economics and Statistics Working Paper 2011-10.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2015. "Fuzzy Differences-in-Differences." arXiv e-print 1510.01757v2.

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

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020a. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." arXiv:2007.04267

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020b. "Replication Data for: Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E118363V1.

de Chaisemartin, Clément, Xavier D'Haultfœuille, and Yannick Guyonvarch. 2019. "Fuzzy Differences-in-Differences with Stata." *Stata Journal* 19 (2): 435–58.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–813.

Frank, Marguerite, and Philip Wolfe. 1956. "An Algorithm for Quadratic Programming." *Naval Research Logistics Quarterly* 3 (1–2): 95–110.

Freeman, Richard B. 1984. "Longitudinal Analyses of the Effects of Trade Unions." *Journal of Labor Economics* 2 (1): 1–26.

Freeman, Richard B., and James L. Medoff. 1984. "What Do Unions Do?" *ILR Review* 38 (2): 244–63.

Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson. 2011. "The Effect of Newspaper Entry and Exit on Electoral Politics." *American Economic Review* 101 (7): 2980–3018.

Goodman-Bacon, Andrew. 2018. "Difference-in-Differences with Variation in Treatment Timing." Unpublished.

Imai, Kosuke, and In Song Kim. 2018. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." Unpublished.

Jakubson, George. 1991. "Estimation and Testing of the Union Wage Effect Using Panel Data." *Review of Economic Studies* 58 (5): 971–91.

Vella, Francis, and Marno Verbeek. 1998. "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men." *Journal of Applied Econometrics* 13 (2): 163–83.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.