NBER WORKING PAPER SERIES

TWO-WAY FIXED EFFECTS ESTIMATORS WITH HETEROGENEOUS TREATMENT EFFECTS

Clément de Chaisemartin
Xavier D'Haultfoeuille

Working Paper 25904
http://www.nber.org/papers/w25904

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2019

Xavier D'Haultfoeuille 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). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects
Clément de Chaisemartin and Xavier D'Haultfoeuille
NBER Working Paper No. 25904
May 2019
JEL No. C21,C23

ABSTRACT

Linear regressions with period and group fixed effects are widely used to estimate treatment effects. We show that they identify 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 estimand may for instance be negative while all the ATEs are positive. In two articles that have used those regressions, half of the weights are negative. We propose another estimator that solves this issue. In one of the articles we revisit, it is of a different sign than the linear regression estimator.

Clément de Chaisemartin
Department of Economics
University of California at Santa Barbara
Santa Barbara, CA 93106
and NBER
clementdechaisemartin@ucsb.edu

Xavier D'Haultfoeuille
CREST
5 avenue Henry Le Chatelier
91764 Palaiseau cedex
FRANCE
xavier.dhaultfoeuille@ensae.fr

A data appendix is available at http://www.nber.org/data-appendix/w25904

1 Introduction

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 20% 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 identify that effect under the standard "common trends" assumption. However, it is often implausible that the treatment effect is constant. For instance, the effect of the minimum wage on employment may vary across US counties, and may change over time. The goal of this paper is to examine 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, as is for instance the case when the treatment is a county- or a state-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 value of the treatment in group g at period t. Let β_{fe} denote the expectation of the coefficient of $D_{g,t}$. Then, under the common trends assumption, we show that if the treatment is binary, β_{fe} identifies a weighted sum of the treatment effect in each group and at each period:

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

$\Delta_{g,t}$ is the average treatment effect (ATE) in group g and period t and the weights $W_{g,t}$s sum to one but may be negative. Negative weights arise because β_{fe} is a weighted average 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. We revisit Enikolopov et al. (2011) and Gentzkow et al. (2011), two articles that have estimated two-way FE regressions. In both cases, negative weights are extremely prevalent: around 50% of the weights attached to β_{fe} are negative, so β_{fe} is not robust to heterogeneous effects in those applications.

Researchers may want to know how serious that issue is in the application they consider. We show that the absolute value of β_{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 β_{fe}. One can estimate that ratio to assess the robustness of β_{fe}. If that ratio is close to 0, β_{fe} 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 β_{fe}. On the contrary, if that ratio is very large, β_{fe} 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 estimand that is valid even if the treatment effect is heterogeneous over time or across groups. This estimand identifies the average treatment effect across all the (g,t) cells whose treatment changes from $t-1$ to t. It relies on a variant of the standard common trends assumption. That condition is partly testable, and we propose a test that amounts to looking at pre-trends, as in a standard DID analysis. We propose a simple plug-in estimator of our estimand, and we show that it is asymptotically normal. In Gentzkow et al. (2011), our estimator is of a different sign than, and significantly different from, the estimator of β_{fe}.[1] Our estimator can be used in applications where, for each pair of consecutive dates, there are groups whose treatment does not change. We estimate that this condition is satisfied for around 80% 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 estimate the weights attached to their regression and the ratio of β_{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 that can also be downloaded from the SSC repository (for more explanations on how to use the former package, see de Chaisemartin et al., 2019).

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, we show that our results extend to fuzzy designs, where the treatment varies within (g,t) cells. Finally, in our Web Appendix, we show that our results also extend to two-way fixed effects regressions with a non-binary treatment and with covariates in the regression.

Our paper is related to the DID literature. Our main result generalizes Theorem 1 in de Chaisemartin and D'Haultfœuille (2018). When the data has two groups and two periods, the Wald-

[1]Note that the authors do not estimate β_{fe}, but β_{fd} defined below. Our estimator is of the same sign as theirs, but 66% larger and significantly different.

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.[2] Moreover, our Wald-TC estimand is related to the Wald-TC estimand with many groups and periods proposed in de Chaisemartin and D'Haultfœuille (2018), and to the multi-period DID estimand proposed by Imai and Kim (2018). In Section 3.3, we explain the differences between those three estimands.

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% 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.

The paper is organized as follows. Section 2 introduces the set-up. Section 3 presents our results in sharp designs where all observations in the same (g,t) cell have the same treatment. Section 4 extends those results to fuzzy designs. Section 5 presents some other extensions. Section 6 discusses inference. Section 7 presents our survey of the articles published in the AER, and our two empirical applications.

2 Set up

One considers observations that can be divided into $\overline{g}+1$ groups and $\overline{t}+1$ periods. For every $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, 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 for instance be an individual-level panel or repeated cross-section data set where groups are, say, individuals' county of birth. The data could also be a cross-section data set where cohort of birth plays the role of time. 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 Section 2.2 of the Web Appendix. Then, for every $(i,g,t) \in \{1,...,N_{g,t}\} \times \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, let

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

$D_{i,g,t}$ 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

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

$D_{g,t}$ 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.

We now define the FE regression described in the introduction.[3]

Regression 1 *(Fixed-effects regression)*

Let $\widehat{\beta}_{fe}$ 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 $\beta_{fe} = E\left(\widehat{\beta}_{fe}\right)$.[4]

Throughout the paper, we maintain the following assumption.

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

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.

Finally, for all g, let $N_{g,.} = \sum_{t=0}^{\overline{t}} N_{g,t}$ denote the total number of observations in group g. For all t, let $N_{.,t} = \sum_{g=0}^{\overline{g}} N_{g,t}$ denote the total number of observations in period t. For any variable $X_{g,t}$ defined in each (g,t) cell, let $X_{g,.} = \sum_{t=0}^{\overline{t}} (N_{g,t}/N_{g,.}) X_{g,t}$ denote the average value of $X_{g,t}$ in group g, let $X_{.,t} = \sum_{g=0}^{\overline{g}} (N_{g,t}/N_{.,t}) X_{g,t}$ denote the average value of $X_{g,t}$ 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.

[3] Throughout the paper, we assume that the regressors in Regression 1 and Regression 2 below are not perfectly collinear, so the coefficients of these regressions are well-defined.

[4] 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 fuzzy designs, using $D_{g,t}$ rather than the individual treatment ensures that β_{fe} is identified out of comparisons of groups experiencing different evolutions of their mean treatment over time, and not out of comparisons of observations with different treatments within the same (g,t) cell, as such comparisons may be plagued by selection bias.

3 Results in sharp designs

In this section, we focus on sharp designs with a non-stochastic treatment, as defined by the following assumption:

Assumption 2 *(Sharp designs with a non-stochastic treatment)*

1. *For all* $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, $D_{i,g,t} = D_{g,t}$ *for all* $i \in \{1,...,N_{g,t}\}$.
2. *For all* $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, $D_{g,t} = E(D_{g,t})$.

Point 1 of 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 $N_{g,t} = 1$ for all (g,t). In our literature review in Section 7.1, we find that almost 80% of the papers using two-way fixed effects regressions and published in the AER between 2010 and 2012 consider sharp designs. We first focus on this special case given its prevalence, before turning to fuzzy designs in the next section.

Point 2 of Assumption 2 requires that the treatment status of each (g,t) cell be non-stochastic. This assumption fails when the treatment is randomly assigned. However, two-way fixed effects regressions are more often used in observational than in experimental studies, as in the latter case one can use simpler regressions to estimate the treatment effect. In observational settings, whether the treatment should be considered as fixed or stochastic is less clear. Point 2 of Assumption 2 is in line with the modelling framework adopted in some articles studying DID, see e.g. Abadie (2005). And most importantly, we relax that assumption in Section 4 and in Section 3.1 of the Web Appendix. We just maintain it for now to ease the exposition.

3.1 A decomposition of β_{fe} as a weighted sum of ATEs under common trends

In this section, we study β_{fe} under the following common trends assumption.

Assumption 3 *(Common trends) For* $t \geq 1$, $E(Y_{g,t}(0)) - E(Y_{g,t-1}(0))$ *does not vary across* g.

The common trends assumption requires that the expectation of the outcome without the treatment follow the same evolution over time in every group. This assumption is sufficient for the DID estimand to identify the ATT in designs with two groups and two periods, and where only units in group 1 and period 1 get treated (see, e.g., Abadie, 2005).

For any $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, we denote the ATE in cell (g,t) as

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

Let Δ^{TR} denote the ATT. Δ^{TR} is a weighted average of the ATEs of the treated (g,t) cells:

$$\Delta^{TR} = \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \Delta_{g,t}, \tag{2}$$

where $N_1 = \sum_{i,g,t} D_{i,g,t}$ is the number of treated observations. We now show that under common trends, β_{fe} is also equal to a weighted sum of the $\Delta_{g,t}$s, 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:[5]

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

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 1-3 hold. Then,*

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

To illustrate this theorem, we consider a simple example of a staggered adoption design with two groups and three periods, and where group 0 is only treated at period 2, while group 1 is treated both at periods 1 and 2. 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

$$\begin{aligned}
\varepsilon_{0,2} &= 1 - 1/3 - 1 + 1/2 = 1/6, \\
\varepsilon_{1,1} &= 1 - 2/3 - 1/2 + 1/2 = 1/3, \\
\varepsilon_{1,2} &= 1 - 2/3 - 1 + 1/2 = -1/6.
\end{aligned}$$

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

$$\beta_{fe} = 1/2\Delta_{0,2} + \Delta_{1,1} - 1/2\Delta_{1,2}.$$

[5] $\varepsilon_{g,t}$ arises from a 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 $\varepsilon_{g,t}$.

β_{fe} is equal to a weighted sum of the ATEs in group 0 at period 2, group 1 at period 1, and group 1 at period 2, the three treated (g,t) cells.

However, the weight assigned to each ATE differs from the proportion that the corresponding cell accounts for in the population of treated observations. Therefore, β_{fe} is not equal to Δ^{TR}, the average treatment effect across all treated observations.

Perhaps more worryingly, not all the weights are positive: the weight assigned to the ATE in group 1 period 2 is strictly negative. Consequently, β_{fe} may be a very misleading measure of the treatment effect. Assume for instance that $\Delta_{0,2} = \Delta_{1,1} = 1$ and $\Delta_{1,2} = 4$. 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 2, one period after it has started receiving the treatment, group 1 now experiences a large ATE. Then,

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

β_{fe} is strictly negative, while $\Delta_{0,2}$, $\Delta_{1,1}$, and $\Delta_{1,2}$ are all positive. More generally, the negative weights are an issue if the $\Delta_{g,t}$s are heterogeneous, across groups or over time.[6] if $\Delta_{0,2} = \Delta_{1,1} = \Delta_{1,2} = 1$, then $\beta_{fe} = 1$, the average treatment effect across all treated observations.

Here is some intuition as to why one weight is negative in this example. It follows from Equation (8) in the proof of Theorem 1 (see also Theorem 1 in Goodman-Bacon, 2018) that in this simple example, $\beta_{fe} = (DID_1 + DID_2)/2$, with

$$\begin{aligned} DID_1 &= E(Y_{1,1}) - E(Y_{1,0}) - (E(Y_{0,1}) - E(Y_{0,0}))\,, \\ DID_2 &= E(Y_{0,2}) - E(Y_{0,1}) - (E(Y_{1,2}) - E(Y_{1,1}))\,. \end{aligned}$$

The first DID compares the evolution of the mean outcome from period 0 to 1 in group 1 and in group 0. The second one compares the evolution of the mean outcome from period 1 to 2 in group 0 and in group 1. The control group in the second DID, group 1, 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 $DID_1 = \Delta_{1,1}$, but

$$DID_2 = \Delta_{0,2} - (\Delta_{1,2} - \Delta_{1,1}).$$

DID_2 is equal to the average treatment effect in group 0 period 2, minus the change in the average treatment effect of group 1 between periods 2 and 1. Intuitively, the mean outcome of groups 0 and 1 may follow different trends from period 1 to 2 either because group 0 becomes treated, or because the treatment effect changes in group 1.

[6] 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.

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

Proposition 1 *Suppose that Assumptions 1-2 hold and for all $t \geq 1$ $N_{g,t}/N_{g,t-1}$ does not vary across g. Then, for all (g,t) such that $D_{g,t} = 1$, $w_{g,t}$ is decreasing in $D_{.,t}$ and $D_{g,.}$.*

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 4 *(Staggered adoption designs) For all g, $D_{g,t} \geq D_{g,t-1}$ for all $t \geq 1$.*

Assumption 4 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) show 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.[7] Proposition 1 also implies that in that design, groups that adopt the treatment earlier are more likely to receive some negative weights.

3.2 Assessing the robustness of β_{fe} to heterogeneous treatment effects

Theorem 1 shows that in sharp designs with many groups and periods, β_{fe} may be a misleading measure of the treatment effect under the standard common trends assumption, if the $\Delta_{g,t}$s are heterogeneous. In the corollary below, we propose two summary measures that can be used to assess how serious that concern is. For any variable $X_{g,t}$ defined in each (g,t) cell we let $\boldsymbol{X}$ denote the vector $(X_{g,t})_{(g,t)\in\{0,\ldots,\overline{g}\}\times\{0,\ldots,\overline{t}\}}$ collecting the values of that variable in each (g,t) cell. For instance, $\boldsymbol{\Delta}$ denotes the vector $(\Delta_{g,t})_{(g,t)\in\{0,\ldots,\overline{g}\}\times\{0,\ldots,\overline{t}\}}$ collecting the ATE in each of the (g,t) cells. Let

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

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

[7] Borusyak and Jaravel (2017) assume 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.

$\sigma(\boldsymbol{\Delta})$ is the standard deviation of the ATEs, and $\sigma(\boldsymbol{w})$ is the standard deviation of the $\boldsymbol{w}$-weights,[8] across the treated (g,t) cells. Let $n = \#\{(g,t) : D_{g,t} = 1\}$ denote the number of treated cells. For every $i \in \{1, ..., n\}$, let $w_{(i)}$ denote the ith largest of the weights of the treated cells: $w_{(1)} \geq w_{(2)} \geq ... \geq w_{(n)}$, and let $N_{(i)}$ and $\Delta_{(i)}$ be the number of observations and the ATE of the corresponding cell. Then, for any $k \in \{1, ..., n\}$, 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 1-3 hold.*

1. *If $\sigma(\boldsymbol{w}) > 0$, the minimal value of $\sigma(\boldsymbol{\Delta})$ compatible with β_{fe} and $\Delta^{TR} = 0$ is*

$$\underline{\sigma}_{fe} = \frac{|\beta_{fe}|}{\sigma(\boldsymbol{w})}.$$

2. *If $\beta_{fe} \neq 0$ and at least one of the $w_{g,t}$ weights is strictly negative, the minimal value of $\sigma(\boldsymbol{\Delta})$ compatible with β_{fe} and with $\Delta_{g,t}$ of a different sign than β_{fe} for all (g,t) is*

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

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

An estimator of $\underline{\sigma}_{fe}$ can be used to assess the robustness of β_{fe} to treatment effect heterogeneity across groups and periods. If $\underline{\sigma}_{fe}$ is close to 0, β_{fe} and Δ^{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 β_{fe}. On the contrary, if $\underline{\sigma}_{fe}$ is very large, β_{fe} and Δ^{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.

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

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

Assumption 5 *($\boldsymbol{w}$ uncorrelated with $\boldsymbol{\Delta}$)* $\sum_{(g,t):D_{g,t}=1} (N_{g,t}/N_1)(w_{g,t} - 1)(\Delta_{g,t} - \Delta^{TR}) = 0$.

Corollary 2 *If Assumptions 1-3 and 5 hold, then $\beta_{fe} = \Delta^{TR}$.*

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

Assumption 5 requires that the weights attached to β_{fe} be uncorrelated with the $\Delta_{g,t}$s. This is often implausible. For instance, periods with the highest proportion of treated units are also those with the lowest value of $w_{g,t}$, as shown in Proposition 1. But those periods may also be those with the largest treatment effect: more groups may choose to get treated at periods where the treatment effect is the highest. This would then induce a negative correlation between $\boldsymbol{w}$ and $\boldsymbol{\Delta}$. The plausibility of Assumption 5 can be assessed, by looking at whether $\boldsymbol{w}$ is correlated with a predictor of the treatment effect in each (g,t) cell. In the two applications we revisit in Section 7, this test is rejected.

3.3 An alternative estimand

In this section, we propose a new estimand that identifies a causal effect even if treatment effects are heterogeneous, across groups or over time. That causal effect is

$$\Delta^S = \sum_{(g,t):D_{g,t}\neq D_{g,t-1},t\geq 1} \frac{N_{g,t}}{N_S}\Delta_{g,t},$$

where $N_S = \sum_{(g,t):t\geq 1,D_{g,t}\neq D_{g,t-1}} N_{g,t}$. Δ^S is the average ATE of all switching cells. In staggered adoption designs, Δ^S is the average of the instantaneous 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 is identified by a weighted average of DID estimands. This result holds under Assumptions 6 and 7 below.

Assumption 6 *(Common trends for groups with the same treatment at $t-1$)*

1. *For all $t \geq 1$ and g such that $D_{g,t-1} = 0$, $E(Y_{g,t}(0)) - E(Y_{g,t-1}(0))$ does not vary across g.*
2. *For all $t \geq 1$ and g such that $D_{g,t-1} = 1$, $E(Y_{g,t}(1)) - E(Y_{g,t-1}(1))$ does not vary across g.*

Assumption 6 requires that between each pair of consecutive periods, the expectation of the outcome without treatment follow the same evolution over time in every group untreated at $t-1$, and that the expectation of the outcome with treatment follow the same evolution in every group treated at $t-1$. In staggered adoption designs, only Point 1 of Assumption 6 is necessary for Theorem 2 below to hold, and Point 1 of Assumption 6 is weaker than Assumption 3, the standard common trends assumption. Outside of staggered adoption designs, both Points 1 and 2 are necessary, so Assumptions 3 and 6 are non-nested. However, Points 1 and 2 are arguably a natural generalization of the identifying assumption in standard two-groups two-periods sharp DID designs (see Abadie, 2005): they require that groups sharing the same treatment at time $t-1$ would experience the same evolution of their mean outcome from $t-1$ to t if they did not change their treatment from $t-1$ to t. Importantly, Points 1 and 2 do not restrict treatment

effect heterogeneity for any group or period, as they only put restriction on one of the two potential outcomes.[9]

Assumption 7 *(Existence of "stable" groups) For all* $t \in \{1, ..., \overline{t}\}$*:*

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

2. *If there is at least one* $g \in \{0, ..., \overline{g}\}$ *such that* $D_{g,t-1} = 1, D_{g,t} = 0$*, then there exists at least one* $g' \neq g, g' \in \{0, ..., \overline{g}\}$ *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 group that switches from being untreated to treated, then there should be another group that remains untreated at both dates. The second point requires that between each pair of consecutive time periods, if there is a group that switches from being treated to untreated, then there should be another group that remains treated at both dates.

We can now define our estimand. For all $t \in \{1, ..., \overline{t}\}$ 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

$$DID_{+,t} = \sum_{g:D_{g,t}=1, D_{g,t-1}=0} \frac{N_{g,t}}{N_{1,0,t}} \left(E(Y_{g,t}) - E(Y_{g,t-1})\right) - \sum_{g:D_{g,t}=D_{g,t-1}=0} \frac{N_{g,t}}{N_{0,0,t}} \left(E(Y_{g,t}) - E(Y_{g,t-1})\right),$$

$$DID_{-,t} = \sum_{g:D_{g,t}=D_{g,t-1}=1} \frac{N_{g,t}}{N_{1,1,t}} \left(E(Y_{g,t}) - E(Y_{g,t-1})\right) - \sum_{g:D_{g,t}=0, D_{g,t-1}=1} \frac{N_{g,t}}{N_{0,1,t}} \left(E(Y_{g,t}) - E(Y_{g,t-1})\right).$$

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

$$W_{TC} = \sum_{t=1}^{\overline{t}} \left(\frac{N_{1,0,t}}{N_S} DID_{+,t} + \frac{N_{0,1,t}}{N_S} DID_{-,t} \right).$$

Theorem 2 *If Assumptions 1, 2, 6, and 7 hold, then* $W_{TC} = \Delta^S$*.*

Here is the intuition underlying our identification result. $DID_{+,t}$ compares the evolution of the mean outcome in two sets of groups: those switching from no treatment to treatment between

[9] Such restrictions would only arise if we imposed common trends on $Y(0)$ and $Y(1)$ for all groups and periods of time. In such a case, $\Delta_{g,t}$ would have to be additively separable in g and t, which still allows for heterogeneous treatment effects across groups and over time.

$t-1$ and t, and those remaining untreated. Because the groups in these two sets have the same treatment in $t-1$, under Assumption 6 $DID_{+,t}$ identifies the treatment effect in groups switching from no treatment to treatment. Similarly, $DID_{-,t}$ compares the evolution of the outcome in two sets of groups: those remaining treated between $t-1$ and t, and those switching from treatment to no treatment. Under Assumption 6, it identifies the treatment effect in groups switching from treatment to no treatment. Finally, W_{TC} is a weighted average of those DID estimands. Note that in staggered adoption designs, there are no groups whose treatment decreases over time, so W_{TC} is only a weighted average of the $DID_{+,t}$ estimands. The corresponding Wald-TC estimator is computed by the `fuzzydid` and `did_multipleGT` Stata packages.

W_{TC} is related to two other estimands. First, it is related to the Wald-TC estimand in point 2 of Theorem S1 in the Web Appendix of de Chaisemartin and D'Haultfœuille (2018), but the weighting of the $DID_{+,t}$ and $DID_{-,t}$ estimands differ. As a result, W_{TC} identifies Δ^S under weaker assumptions. W_{TC} is also related to the multi-period DID estimand proposed by Imai and Kim (2018). There are two differences between those estimands. First, the multi-period DID estimand is a weighted average of the $DID_{+,t}$ estimands, so it does not identify the treatment effect in the groups switching out of treatment. Second, Imai and Kim (2018) do not generalize their estimand to fuzzy designs, something we do in the next section, and to non-binary treatments, something we do in the Web Appendix (see Section 3.2 therein).

The Wald-TC uses groups whose treatment is stable to infer the trends that would have affected the groups whose treatment changes if their treatment had not changed. This strategy could fail, if groups whose treatment changes experience different trends than groups whose treatment is stable. To assess if this is a serious concern, we propose to use the following placebo estimand, 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 estimand is defined under a modified version of Assumption 7.

Assumption 8 *(Existence of "stable" groups for placebo tests) For all $t \in \{2,...,\overline{t}\}$:*

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

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

For all $t \in \{2,...,\overline{t}\}$ 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

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

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

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

Note that $DID_{+,t}^{pl}$ is not defined when there is no group such that $D_{g,t}=1, D_{g,t-1}=D_{g,t-2}=0$. In such instances, we let $DID_{+,t}^{pl}=0$. Similarly, let $DID_{-,t}^{pl}=0$ when there is no group such that $D_{g,t}=0, D_{g,t-1}=D_{g,t-2}=1$. Then, let

$$W_{TC}^{pl} = \sum_{t=2}^{\bar{t}} \left(\frac{N_{1,0,0,t}}{N_S^{pl}} DID_{+,t}^{pl} + \frac{N_{0,1,1,t}}{N_S^{pl}} DID_{-,t}^{pl} \right).$$

Theorem 3 *If Assumptions 1, 2, 6, and 8 hold, then* $W_{TC}^{pl}=0$.

$DID_{+,t}^{pl}$ 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 Assumption 6 holds, then $DID_{+,t}^{pl}=0$. Similarly, if Assumption 6 holds, $DID_{-,t}^{pl}=0$. Then, $W_{TC}^{pl}=0$ is a testable implication of Assumption 6, so finding $W_{TC}^{pl}\neq 0$ would imply that Assumption 6 is violated: groups that switch treatment experience different trends before that switch than the groups used to reconstruct their counterfactual trends when they switch.[10]

In staggered adoption designs, we have, for all $t\geq 2$,

$$D_{g,t} > D_{g,t-1} \Leftrightarrow D_{g,t} > D_{g,t-1} = D_{g,t-2},$$
$$D_{g,t} = D_{g,t-1} = 0 \Leftrightarrow D_{g,t} = D_{g,t-1} = D_{g,t-2} = 0.$$

Groups whose treatment increases from $t-1$ to t must have a stable treatment from $t-2$ to $t-1$, and groups untreated at t and $t-1$ must also be untreated at $t-2$. Then, Assumptions 7 and 8 are equivalent, and W_{TC}^{pl} closely mimicks W_{TC}: W_{TC}^{pl} is equal to W_{TC}, after replacing the mean outcome in group g and at period t by its lagged value. Outside of staggered adoption designs, some of the (g,t) cells that are used in the computation of W_{TC} are excluded from that of W_{TC}^{pl}, because their treatment changes from $t-2$ to $t-1$. Finally, W_{TC}^{pl} compares the trends of switching and stable groups one period before the switch. It is easy to define other placebo estimands comparing those trends, say, two or three periods before the switch. The corresponding placebo estimators are computed by the `fuzzydid` and `did_multipleGT` Stata package.

[10] See also Callaway and Sant'Anna (2018), who propose another placebo test in staggered adoption designs.

4 Results in fuzzy designs

In this section, the research design may be fuzzy: the treatment may vary within (g,t) cells. For instance, Enikolopov et al. (2011) study the effect of having access to an independent TV channel in Russia, and in each Russian region some people have access to that channel while other people do not. The treatment may also be stochastic.

As the numbers 1,...,$N_{g,t}$ assigned to the observations of a (g,t) cell play no role, one can always assume that those numbers are randomly assigned. Therefore, we assume hereafter that conditional on $\boldsymbol{D}$, the vector stacking together all the $D_{g,t}$s, all the variables indexed by i (e.g., $D_{i,g,t}$ or $Y_{i,g,t}(0)$) are identically distributed within each (g,t) cell.

4.1 Generalizing the decomposition of β_{fe} to fuzzy designs

With a stochastic treatment, we study β_{fe} under a modified common trends assumption.

Assumption 9 *(Common trends when the treatment is stochastic) For all* $t \geq 1$, $E(Y_{g,t}(0)|\boldsymbol{D}) - E(Y_{g,t-1}(0)|\boldsymbol{D})$ *almost surely does not vary across* g.

Assumption 9 requires that conditional on any realization of $\boldsymbol{D}$, the expectation of the outcome without the treatment follow the same evolution over time in every group. Assuming common trends conditional on the $D_{g,t}$s is necessary for identification due to the stochastic nature of the treatment, as we explain in further detail in Section 3.1 of the Web Appendix. Other articles that have studied DID with a stochastic treatment have also imposed conditional common trends assumptions similar to Assumption 9, see e.g. Assumption 1 in Conley and Taber (2011). Finally, note that under Assumption 2 (sharp designs with a non-stochastic treatment), Assumption 9 reduces to Assumption 3.

For any $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, let

$$\Delta^{TR}_{g,t}(\boldsymbol{D}) = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} E\left(Y_{i,g,t}(1) - Y_{i,g,t}(0) | D_{i,g,t} = 1, \boldsymbol{D}\right)$$

denote the average treatment effect on the treated (ATT) in cell (g,t), conditional on $\boldsymbol{D}$. Then

$$\Delta^{TR} = E\left(\sum_{g,t} \frac{N_{g,t} D_{g,t}}{N_1} \Delta^{TR}_{g,t}(\boldsymbol{D})\right),$$

which generalizes (2) when treatments are stochastic. Theorem 4 shows that β_{fe} is also equal to the expectation of a weighted sum of the $\Delta^{TR}_{g,t}(\boldsymbol{D})$s. Let

$$w^{TR}_{g,t} = \frac{\varepsilon_{g,t}}{\sum_{g,t} \frac{N_{g,t} D_{g,t}}{N_1} \varepsilon_{g,t}}.$$

Theorem 4 *Suppose that Assumptions 1 and 9 hold. Then,*

$$\beta_{fe} = E\left(\sum_{g,t} \frac{N_{g,t} D_{g,t}}{N_1} w_{g,t}^{TR} \Delta_{g,t}^{TR}(\boldsymbol{D})\right).$$

Theorem 4 shows that in fuzzy designs, β_{fe} is equal to the expectation of a weighted sum of the ATTs in each (g,t) cell. Again, some of the weights may be strictly negative. Note that under Assumption 2, Theorem 4 reduces to Theorem 1 in the previous section. One can also show that in sharp designs with a stochastic treatment, Theorem 1 still holds, conditional on any realization of $\boldsymbol{D}$. Overall, Theorem 4 generalizes Theorem 1 to situations where Assumption 3 fails, either because the design is not sharp, or because the treatment is stochastic. In staggered adoption designs, Athey and Imbens (2018) derive a decomposition of β_{fe} that resembles to, but differs from, that in Theorem 4. 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.

The weights have a simple expression in the following special case.

Assumption 10 *(Heterogenous adoption)* $\overline{t} = 1$ *and for all* $g \in \{0, ..., \overline{g}\}$, $D_{g,1} > D_{g,0} = 0$.

Assumption 10 is satisfied in applications with two time periods, and where all groups are fully untreated at $t = 0$ and partly treated at $t = 1$. This type of design often arises in practice, for instance when an innovation is heterogeneously adopted by various groups.

Proposition 2 *If Assumptions 1, 9, and 10 hold and* $N_{g,1}/N_{g,0}$ *does not vary across* g, *then*[11]

$$w_{g,1}^{TR} = \frac{D_{g,1} - D_{.,1}}{\sum_{g=0}^{\overline{g}} \frac{N_{g,1}}{N_1}(D_{g,1} - D_{.,1})^2}.$$

Proposition 2 shows that in the heterogeneous adoption design, β_{fe} assigns negative weights to the period-one ATT of groups with a mean treatment lower than the mean treatment in the full population. The reason why negative weights arise is intuitive. With two periods, the FE regression is equivalent to a regression of the first difference of the outcome on the period-one treatment in each group. This regression compares the evolution of the outcome in more- and less-treated groups. Doing so, it subtracts the treatment effect of the less-treated groups, hence the negative weights. Negative weights are a concern if the ATTs of the less- and more-treated groups systematically differ. This could be the case if treatment is determined by a Roy selection model. Then, the groups with the highest proportion of treated units could also be those where the ATT is the highest. On the other hand, if the proportion of treated units is randomly assigned to each group, negative weights are not a concern.[12]

[11] Under Assumption 10, $D_{g,0} = 0$, so $w_{g,0}^{TR}$ does not enter in the decomposition in Theorem 4.

[12] Corollary 2 extends directly to fuzzy settings. Corollary 1 also extends to such cases, though the sensitivity measure we propose is now a random variable, as it is a function of $\boldsymbol{D}$. We discuss this issue in more details in Section 4.1 of the Web Appendix.

4.2 Generalizing the Wald-TC estimand to fuzzy designs

In this section, we show that in fuzzy designs, a causal effect similar to the causal effect Δ^S defined above is identified by an estimand similar to the Wald-TC estimand defined above. In other words, we generalize Theorem 2 to fuzzy designs.

We start by introducing some new notation. For all $(i,g,t) \in \{1,...,N_{g,t}\} \times \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$ and for every $t' \in \{0,...,\overline{t}\}$, let $D_{i,g,t}(t')$ denote the period-t' treatment of unit i in group g and period t. When the data are an individual-level panel, those variables are observed: $D_{i,g,t}(t') = D_{i,g,t'}$, $D_{i,g,t}(t')$ is just the treatment of unit i in group g at time t'. In sharp designs, those variables are also observed. As the treatment is assigned at the (g,t)-level, the period-t' treatment of unit i in group g and period t is the same as the treatment status of units in group g at period t': $D_{i,g,t}(t') = D_{g,t'}$. On the other hand, in fuzzy designs, and when the data are repeated cross-sections or a cross-section where cohort of birth plays the role of time, only $D_{i,g,t}(t) = D_{i,g,t}$ is observed. In the latter case, the $D_{i,g,t}(t')$s are potential treatments: $D_{i,g,t}(t')$ is the treatment status of individual i in group g and cohort t if she had been born in cohort t'.

As we assume that the random variables attached to each observation are identically distributed within each (g,t) cell, $\big((D_{1,g,t}(t'))_{t' \in \{0,...,\overline{t}\}}, Y_{1,g,t}(0), Y_{1,g,t}(1)\big)$, the treatments and potential outcomes of the first observation in cell (g,t) have the same distribution as the treatments and potential outcomes of all observations in the same cell. We make the following assumptions.

Assumption 11 *(Treatment monotonicity) For all* $(g,t) \in \{1,...,\overline{g}\} \times \{1,...,\overline{t}\}$,

1. $P(D_{1,g,t}(t-1) = 1) = P(D_{1,g,t-1}(t-1) = 1)$.
2. *Either* $P(D_{1,g,t}(t) \geq D_{1,g,t}(t-1)) = 1$, *or* $P(D_{1,g,t}(t) \leq D_{1,g,t}(t-1)) = 1$.

First, Point 1 requires that in each group, the probability of getting treated at time $t-1$ is the same among units observed at periods t and $t-1$. If the data is a balanced panel, the units observed at different periods are the same, so this requirement is automatically satisfied: $P(D_{1,g,t}(t-1) = 1) = P(D_{1,g,t-1} = 1)$. With cohort-of-birth data, this requirement could fail: for instance, the treatment probability of units born in year $t-1$ may differ from the treatment probability of the cohort born in t if it had been born in $t-1$. This requirement is plausible when consecutive cohorts have similar propensities for treatment. Then, Point 2 requires that in each group, either the treatments of all units observed at period t weakly increase from period $t-1$ to t, or their treatments weakly decrease. This is similar in spirit to the monotonicity assumption in Imbens and Angrist (1994), with time as the instrument. With panel data, one can assess from the data whether Point 2 holds or not.[13] With repeated-cross-sections or cohort-of-birth data,

[13] This test may reveal that Point 2 fails, but our results still hold if the treatments satisfy the threshold crossing Equation (3.2) in de Chaisemartin and D'Haultfœuille (2018), which is weaker than Point 2 of Assumption 11.

Point 2 is untestable. Note that Assumption 11 is automatically satisfied under Assumption 2 (sharp designs with a non-stochastic treatment).

Assumption 12 *(Conditional common trends) For all $t \geq 1$,*

1. $E\left(Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0\right) - E\left(Y_{1,g,t-1}(0)|D_{1,g,t-1}(t-1)=0\right)$ *does not vary across g.*
2. $E\left(Y_{1,g,t}(1)|D_{1,g,t}(t-1)=1\right) - E\left(Y_{1,g,t-1}(1)|D_{1,g,t-1}(t-1)=1\right)$ *does not vary across g.*

Point 1 requires that for all $t \geq 1$, the evolution from $t-1$ to t of the mean $Y(0)$ among units untreated at $t-1$ be the same in every group. Similarly, Point 2 requires that for all $t \geq 1$, the evolution from $t-1$ to t of the mean $Y(1)$ of units treated at $t-1$ be the same in every group. Note that under Assumption 2, Assumption 12 reduces to Assumption 6.

Assumption 13 *(Existence of "stable" groups) For all $t \in \{1, ..., \overline{t}\}$:*

1. *If there is at least one $g \in \{0, ..., \overline{g}\}$ such that $0 < E(D_{g,t-1}) \neq E(D_{g,t})$, then there is at least one $g' \neq g, g' \in \{0, ..., \overline{g}\}$ such that $0 < E(D_{g',t-1}) = E(D_{g',t})$.*
2. *If there is at least one $g \in \{0, ..., \overline{g}\}$ such that $1 > E(D_{g,t-1}) \neq E(D_{g,t})$, then there is at least one $g' \neq g, g' \in \{0, ..., \overline{g}\}$ such that $1 > E(D_{g',t-1}) = E(D_{g',t})$.*

Point 1 requires that between each pair of consecutive time periods, if there is a group with some treated units at $t-1$ whose mean treatment changes from $t-1$ to t, then there should be another group with some treated units at $t-1$ whose mean treatment remains stable from $t-1$ to t. Similarly, Point 2 requires that if there is a group with some untreated units at $t-1$ whose mean treatment changes from $t-1$ to t, then there should be another group with some untreated units at $t-1$ whose mean treatment remains stable from $t-1$ to t. Again, under Assumption 2, Assumption 13 reduces to Assumption 7.

In fuzzy designs, the causal effect identified by our estimand is

$$\tilde{\Delta}^S = \sum_{(g,t),t\geq 1} \frac{N_{g,t}(P(D_{1,g,t}(t) \neq D_{1,g,t}(t-1))}{P_S} E(Y_{1,g,t}(1) - Y_{1,g,t}(0)|D_{1,g,t}(t) \neq D_{1,g,t}(t-1)),$$

where $P_S = \sum_{(g,t),t\geq 1} N_{g,t}(P(D_{1,g,t}(t) \neq D_{1,g,t}(t-1)))$. $\tilde{\Delta}^S$ is a weighted average of the LATE of switchers in each (g,t) cell, where each (g,t) cell receives a weight proportional to its expected number of switchers. It is also the probability limit of switchers' LATE,

$$\frac{1}{\#\{(i,g,t) : t \geq 1, D_{i,g,t}(t) \neq D_{i,g,t}(t-1)\}} \sum_{(i,g,t):D_{i,g,t}(t)\neq D_{i,g,t}(t-1),t\geq 1} (Y_{i,g,t}(1) - Y_{i,g,t}(0)),$$

if all the (g,t) cells grow large and units are independent within cells.

We can now define our Wald-TC estimand in fuzzy designs. First, for any $t \geq 1$, let

$$\begin{aligned}
N_{+,t} &= \sum_{g:E(D_{g,t})>E(D_{g,t-1})} N_{g,t}, \\
N_{-,t} &= \sum_{g:E(D_{g,t})<E(D_{g,t-1})} N_{g,t}, \\
N_{=,t} &= \sum_{g:E(D_{g,t})=E(D_{g,t-1})} N_{g,t}
\end{aligned}$$

denote the number of observations at period t in groups whose mean treatment respectively increases, decreases, and remains stable between periods $t-1$ and t. Let also

$$N_{1,=,t} = \sum_{g:E(D_{g,t})=E(D_{g,t-1})} N_{g,t} E(D_{g,t})$$

denote the expected number of treated observations at period t in groups whose mean treatment remains stable between periods $t-1$ and t. Then, let

$$\begin{aligned}
\delta_{0,t} &= \sum_{g:E(D_{g,t})=E(D_{g,t-1})} \frac{N_{g,t}(1-E(D_{g,t}))}{N_{=,t}-N_{1,=,t}} \left(E(Y_{1,g,t}|D_{1,g,t}=0) - E(Y_{1,g,t-1}|D_{1,g,t-1}=0)\right), \\
\delta_{1,t} &= \sum_{g:E(D_{g,t})=E(D_{g,t-1})} \frac{N_{g,t}E(D_{g,t})}{N_{1,=,t}} \left(E(Y_{1,g,t}|D_{1,g,t}=1) - E(Y_{1,g,t-1}|D_{1,g,t-1}=1)\right)
\end{aligned}$$

respectively denote the change of the mean outcome of untreated and treated observations from $t-1$ to t, in groups whose mean treatment remains stable between $t-1$ and t. Then, let

$$\begin{aligned}
p_{+,t-1} &= \sum_{g:E(D_{g,t})>E(D_{g,t-1})} \frac{N_{g,t}E(D_{g,t-1})}{N_{+,t}}, \\
p_{-,t-1} &= \sum_{g:E(D_{g,t})<E(D_{g,t-1})} \frac{N_{g,t}E(D_{g,t-1})}{N_{-,t}}.
\end{aligned}$$

$p_{+,t-1}$ (resp. $p_{-,t-1}$) is the weighted average of the treatment probabilities at period $t-1$ of groups whose mean treatment increases (resp. decreases) between periods $t-1$ and t, where groups are weighted by their number of observations at period t.

Our estimand is a weighted average of the following estimands:

$$\begin{aligned}
W_{TC,+,t} &= \frac{\sum_{g:E(D_{g,t})>E(D_{g,t-1})} \frac{N_{g,t}}{N_{+,t}} \left(E(Y_{g,t}) - (E(Y_{g,t-1}) + (1-p_{+,t-1})\delta_{0,t} + p_{+,t-1}\delta_{1,t})\right)}{\sum_{g:E(D_{g,t})>E(D_{g,t-1})} \frac{N_{g,t}}{N_{+,t}} \left(E(D_{g,t}) - E(D_{g,t-1})\right)}, \\
W_{TC,-,t} &= \frac{\sum_{g:E(D_{g,t})<E(D_{g,t-1})} \frac{N_{g,t}}{N_{-,t}} \left(E(Y_{g,t}) - (E(Y_{g,t-1}) + (1-p_{-,t-1})\delta_{0,t} + p_{-,t-1}\delta_{1,t})\right)}{\sum_{g:E(D_{g,t})<E(D_{g,t-1})} \frac{N_{g,t}}{N_{-,t}} \left(E(D_{g,t-1}) - E(D_{g,t})\right)}.
\end{aligned}$$

Under Assumption 2, $W_{TC,+,t} = DID_{+,t}$. Indeed, one then has $p_{+,t-1} = 0$, as all groups whose treatment increases between $t-1$ and t are fully untreated at period $t-1$. Moreover,

$$\delta_{0,t} = \sum_{g:D_{g,t}=D_{g,t-1}=0} \frac{N_{g,t}}{N_{0,0,t}} \left(E(Y_{g,t}) - E(Y_{g,t-1})\right),$$

as the groups whose mean treatment is stable from $t-1$ to t are fully untreated or fully treated at both dates, and the latter groups do not have untreated units. Likewise, $W_{TC,-,t} = DID_{-,t}$. Finally, let

$$\begin{aligned}
w_{+,t} =& \frac{\sum_{g:E(D_{g,t})>E(D_{g,t-1})} N_{g,t} \left(E(D_{g,t}) - E(D_{g,t-1})\right)}{\sum_{t=1}^{\bar{t}} \sum_{g:E(D_{g,t})\neq E(D_{g,t-1})} N_{g,t} \left|E(D_{g,t}) - E(D_{g,t-1})\right|}, \\
w_{-,t} =& \frac{\sum_{g:E(D_{g,t})<E(D_{g,t-1})} N_{g,t} \left(E(D_{g,t-1}) - E(D_{g,t})\right)}{\sum_{t=1}^{\bar{t}} \sum_{g:E(D_{g,t})\neq E(D_{g,t-1})} N_{g,t} \left|E(D_{g,t}) - E(D_{g,t-1})\right|}, \\
W_{TC} =& \sum_{t=1}^{\bar{t}} \left(w_{+,t} W_{TC,+,t} + w_{-,t} W_{TC,-,t}\right). \qquad (4)
\end{aligned}$$

Under Assumption 2, the Wald-TC estimand above is equal to that defined in Section 3.

Theorem 5 *Suppose that Assumption 1 and Assumptions 11-13 hold, then* $W_{TC} = \tilde{\Delta}^S$.

Theorem 5 generalizes Theorem 2 to fuzzy designs. Here is the reasoning underlying that result. First, in all the groups where the mean treatment is stable between $t-1$ and t,

$$\begin{aligned}
&E(Y_{1,g,t}|D_{1,g,t} = 0) - E(Y_{1,g,t-1}|D_{1,g,t-1} = 0) \\
=&E(Y_{1,g,t}(0)|D_{1,g,t}(t) = 0) - E(Y_{1,g,t-1}(0)|D_{1,g,t-1}(t-1) = 0) \\
=&E(Y_{1,g,t}(0)|D_{1,g,t}(t-1) = 0) - E(Y_{1,g,t-1}(0)|D_{1,g,t-1}(t-1) = 0).
\end{aligned}$$

Indeed, treatment monotonicity implies that in those groups, $D_{1,g,t}(t) = D_{1,g,t}(t-1)$. The previous display and Assumption 12 then imply that $\delta_{0,t}$ identifies the evolution of $Y_{1,g,t}$ that units untreated at $t-1$ would have experienced from $t-1$ to t if they had remained untreated at t. Similarly, $\delta_{1,t}$ identifies the evolution of $Y_{1,g,t}$ that units treated at $t-1$ would have experienced if they had remained treated. Then, in groups where the proportion of treated units increases,

$$E(Y_{g,t-1}) + (1 - p_{+,t-1})\delta_{0,t} + p_{+,t-1}\delta_{1,t}$$

identifies the mean outcome one would have observed at period t if the proportion of treated units had not increased. Consequently, the numerator of $W_{TC,+,t}$ compares the period-t outcome in those groups to what this outcome would have been if some switchers had not become treated. The denominator of $W_{TC,+,t}$ identifies the proportion of switchers, so $W_{TC,+,t}$ identifies their LATE. Similarly, $W_{TC,-,t}$ identifies the LATE of switchers in groups where the proportion of treated units diminishes from $t-1$ to t. Eventually, W_{TC} averages those LATEs across periods. Finally, Assumption 12 can be tested by computing a placebo estimand generalizing W_{TC}^{pl} to fuzzy designs, if for any $t \geq 2$ there are groups whose mean treatment is stable from $t-2$ to t.

5 Extensions

5.1 A decomposition of β_{fd} as a weighted sum of ATEs under common trends

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

Regression 2 *(First-difference regression)*

Let $\widehat{\beta}_{fd}$ 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 $t \geq 1$. Let $\beta_{fd}=E\left(\widehat{\beta}_{fd}\right)$.

When $\bar{t}=1$ and $N_{g,1}/N_{g,0}$ does not vary across g, meaning that all groups experience the same growth of their number of units from period 0 to 1, one can show that $\beta_{fe}=\beta_{fd}$. β_{fe} differs from β_{fd} if $\bar{t}>1$ or $N_{g,1}/N_{g,0}$ varies across g.

We start by showing that a result similar to Theorem 1 also applies to β_{fd}. For any $(g,t) \in \{0,...,\overline{g}\} \times \{1,...,\bar{t}\}$, 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 a constant and period fixed effects, among observations for which $t \geq 1$. For any $g \in \{0,...,\overline{g}\}$, let $\varepsilon_{fd,g,0}=0$, $\varepsilon_{fd,g,\bar{t}+1}=0$. 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 6 *Suppose that Assumptions 1-3 hold. Then,*

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

Theorem 6 shows that under Assumption 3, 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 $t \in \{1,...,\bar{t}\}$, $\varepsilon_{fd,g,t}=D_{g,t}-D_{g,t-1}-(D_{.,t}-D_{.,t-1})$. $\varepsilon_{fd,g,t}$ 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 3 *Suppose that Assumptions 1-2 and 4 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 $D_{.,\bar{t}+1}=D_{.,\bar{t}}$).*

Proposition 3 shows that for all $t \in \{1, ..., \bar{t}-1\}$ 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 the proportion of treated units increases from $\bar{t}-1$ to $\bar{t}$, the period-$\bar{t}$ ATE of groups already treated in $\bar{t}-1$ 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 3 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 $\bar{t}$ if $D_{.,\bar{t}} = D_{.,\bar{t}-1}$. 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-$\bar{t}$ 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.

Just as for β_{fe}, one can show that the minimal value of $\sigma(\mathbf{\Delta})$ compatible with β_{fd} and $\Delta^{TR}=0$ is $\underline{\sigma}_{fd} = |\beta_{fd}|/\sigma(\boldsymbol{w_{fd}})$, where

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

is the standard deviation of the weights attached to β_{fd}. One can also show that $\underline{\underline{\sigma}}_{fd}$, the minimal value of $\sigma(\mathbf{\Delta})$ compatible with β_{fd} and $\Delta_{g,t}$ of a different sign than β_{fd} for all (g,t), has the same expression as $\underline{\underline{\sigma}}_{fe}$, 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 β_{fe} or β_{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 14 *($\boldsymbol{w_{fd}}$ uncorrelated with $\mathbf{\Delta}$)* $\sum_{(g,t):D_{g,t}=1}(N_{g,t}/N_1)(w_{fd,g,t}-1)(\Delta_{g,t}-\Delta^{TR})=0$.

Note that under the common trends assumption, one can jointly test Assumption 14 and Assumption 5, the assumption that the weights attached to β_{fe} are uncorrelated with the $\Delta_{g,t}$s: if $\widehat{\beta}_{fe}$ and $\widehat{\beta}_{fd}$ are significantly different, at least one of these two assumptions must fail. In the second application we revisit in Section 7, $\widehat{\beta}_{fe}$ and $\widehat{\beta}_{fd}$ are significantly different.

5.2 Dynamic treatment effects

With panel data, one may want to allow for the possibility that a unit's contemporaneous outcome also depends on her past values of the treatment. To do so, one just needs to replace, in all of the above, $Y_{i,g,t}(0)$ (resp. $Y_{i,g,t}(1)$)) by $Y_{i,g,t}(0, D_{i,t-1,g}, ..., D_{i,0,g})$ (resp. $Y_{i,t,g}(1, D_{i,t-1,g}, ..., D_{i,0,g})$), the counterfactual outcome of unit i in group g at period t if she is untreated (resp. treated) at that period, and her past treatments are equal to their actual values. Then, β_{fe} and β_{fd} still identify uninterpretable parameters: unsurprisingly, allowing for dynamic treatment effects does not make those estimands more robust to heterogeneous treatment effects. We do not use the $Y_{i,g,t}(0, D_{i,t-1,g}, ..., D_{i,0,g})$ and $Y_{i,t,g}(1, D_{i,t-1,g}, ..., D_{i,0,g})$ notation throughout the paper because it does not apply to cross-section data sets where cohort of birth plays the role of time.

On the other hand, allowing for dynamic treatment effects may alter the plausibility of the common trends assumption underlying our Wald-TC estimand. To simplify the discussion, let us focus on sharp designs and let us assume that the treatment is not stochastic. Then, Assumption 6 can be rewritten as:

Assumption 15 *(Common trends for groups with the same treatment at $t-1$)*

1. *For all $t \geq 1$ and g such that $D_{g,t-1} = 0$, $E(Y_{g,t}(0, 0, D_{g,t-2}, ..., D_{g,0}))-E(Y_{g,t-1}(0, D_{g,t-2}, ..., D_{g,0}))$ does not vary across g.*

2. *For all $t \geq 1$ and g such that $D_{g,t-1} = 1$, $E(Y_{g,t}(1, 1, D_{g,t-2}, ..., D_{g,0}))-E(Y_{g,t-1}(1, D_{g,t-2}, ..., D_{g,0}))$ does not vary across g.*

Assumption 15 may not be plausible if past treatments can affect the outcome at time t. For instance, two groups g and g' that were both untreated at $t-1$ may have had different treatments at $t-2$, and then their outcome could follow different trends from $t-1$ to t.

Note that this issue is absent in staggered adoption designs. Then, only Point 1 of Assumption 15 is necessary for Theorem 2 to hold. Moreover, $D_{g,t-1} = 0$ implies $D_{g,t-2} = ... = D_{g,0} = 0$. Therefore, Point 1 of Assumption 15 requires that $E(Y_{g,t}(0, 0, 0, ..., 0)) - E(Y_{g,t-1}(0, 0, ..., 0))$ be independent of g for all untreated groups at $t-1$. This requirement does not impose any restriction on dynamic treatment effects. Therefore, our Wald-TC estimand is fully robust to dynamic treatment effects in that design.

Outside of that special case, one can identify Δ^S under assumptions that restrict dynamic treatment effects less than Assumption 15. For instance, we consider the following assumption.

Assumption 16 *(Common trends for groups with the same treatments at $t-1$ and $t-2$)*

1. *For all $t \geq 2$, $(d, d') \in \{0, 1\}^2$, and g, $Y_{g,t}(d, d', D_{g,t-2}, ..., D_{g,0})$ does not depend on $(D_{g,t-2}, D_{g,t-3}, ..., D_{g,0})$ and is denoted $Y_{g,t}(d, d')$.*

2. *For all* $t \geq 2$ *and* $(d, d') \in \{0,1\}^2$*, for all* g *such that* $D_{t-1,g} = d, D_{t-2,g} = d'$*,* $E(Y_{g,t}(d,d)) - E(Y_{g,t-1}(d,d'))$ *does not vary across* g*.*

3. *For* $t = 1$*, Assumption 15 holds.*

Under Point 1 of Assumption 16, $D_{g,t-2}$ or earlier treatments cannot have an effect on the period-t outcome, but $D_{g,t-1}$ may have an effect. Point 2 requires that the mean outcome would have followed the same evolution from period $t-1$ to t in all the groups sharing the same treatments at periods $t-2$ and $t-1$, if those groups had kept the same treatment in period t as in period $t-1$. Under Assumption 16, Δ^S is identified by an estimand that differs from, but is similar in spirit to, our Wald-TC estimand. That estimand is a weighted average of four type of DID estimands: DID estimands comparing groups with $D_{g,t} = 1, D_{g,t-1} = 0, D_{g,t-2} = 0$ to groups with $D_{g,t} = 0, D_{g,t-1} = 0, D_{g,t-2} = 0$, DID estimands comparing groups with $D_{g,t} = 1, D_{g,t-1} = 0, D_{g,t-2} = 1$ to groups with $D_{g,t} = 0, D_{g,t-1} = 0, D_{g,t-2} = 1$, etc. Finally, note that one can identify Δ^S under assumptions that restrict dynamic treatment effects even less than Assumption 16, for instance by assuming that $D_{g,t-1}$ and $D_{g,t-2}$ may have an effect on the period-t outcome, and that groups with the same treatments at $t-1$, $t-2$, and $t-3$ experience the same trends.

Finally, one may also be interested in estimating dynamic treatment effects. Let us focus hereafter on staggered adoption designs. Abraham and Sun (2018) show that the event-study regression with groups and periods fixed effects and lags of the treatment is not robust to heterogeneous treatment effects, just as the FE regression we study. On the other hand, the Wald-TC estimand identifies the average instantaneous effect arising at the time when a group starts receiving the treatment, across all groups that become treated at some point. It is easy to show that the average dynamic effect arising, say, one period after those groups have started receiving the treatment, is also identified, by an estimand closely related to the Wald-TC and under a modified version of Assumption 15. That estimand is a weighted average of DID estimands comparing the evolution of the outcome from $t-1$ to $t+1$, between groups becoming treated at period t, and groups that are still untreated at period $t+1$. That estimand differs from other estimands of dynamic treatment effects that have been proposed in staggered adoption designs. The estimand in Callaway and Sant'Anna (2018) uses groups that never become treated as their control group, that in Abraham and Sun (2018) uses the groups that become treated at the last period, while ours uses all groups not yet treated at period $t+1$. Our control group is larger, so the corresponding estimator may be more precise than those other estimators. The estimators of dynamic treatment effects we propose in staggered adoption designs are computed by the `fuzzydid` and `did_multipleGT` Stata packages. Outside of staggered adoption designs, estimating dynamic effects is still feasible but more challenging, and deserves more than an informal discussion. This question is therefore left for future work.

5.3 Other extensions

In Section 2 of the Web Appendix, we consider some other important extensions. We sketch the four most important here. First, in sharp designs, we show that under common trends and the supplementary assumption that the treatment effect is stable over time, β_{fe} and β_{fd} identify weighted sums of the ATEs of switching cells. All the weights attached to β_{fd} are positive. Some of the weights attached to β_{fe} may still be negative, but in the special case with staggered adoption and $N_{g,t}/N_{g,t-1}$ independent of g, all the weights are positive. Second, our decompositions of β_{fe} and β_{fd} as weighted sums of the treatment effect in each (g,t) cell extend to two-way fixed effects regressions with some controls, under a modified version of the common trends assumption accounting for the controls. Importantly, the weights in the decompositions remain the same. Third, our decompositions of β_{fe} and β_{fd} also extend to two-way fixed effects regressions with a non-binary treatment. Here again, the weights remain almost the same. Finally, the Wald-TC estimand can be extended to have some controls, or a non-binary treatment.

6 Estimation and inference

In this section, we assume that Assumption 2 (sharp designs with a non-stochastic treatment) holds. We discuss estimation and inference outside of that case in Section 4 of the Web Appendix.

First, we consider estimation of, and inference on, $\underline{\sigma}_{fe}$, $\underline{\underline{\sigma}}_{fe}$, $\underline{\sigma}_{fd}$, and $\underline{\underline{\sigma}}_{fd}$. Under Assumption 2, the weights do not have to be estimated. Then, $\underline{\sigma}_{fe}$ is simply estimated by:

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

$\underline{\underline{\sigma}}_{fe}$, $\underline{\sigma}_{fd}$, and $\underline{\underline{\sigma}}_{fd}$ are estimated similarly. To draw inference on $\underline{\sigma}_{fe}$, let se_{fe} denote the standard error of $\widehat{\beta}_{fe}$ clustered at the group level. We then consider confidence intervals of the form

$$\text{CI}_{1-\alpha}(\underline{\sigma}_{fe}) = \left[\widehat{\underline{\sigma}}_{fe} - z_{1-\alpha/2}\frac{\text{se}_{fe}}{\sigma(\boldsymbol{w})}, \widehat{\underline{\sigma}}_{fe} + z_{1-\alpha/2}\frac{\text{se}_{fe}}{\sigma(\boldsymbol{w})}\right],$$

where z_β denotes the quantile of order β of a standard normal variable.

Second, we consider estimation of, and inference on, Δ^S. Let

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

Then the Wald-TC estimator is given by

$$\widehat{W}_{TC} = \sum_{t=1}^{\bar{t}} \left(\frac{N_{1,0,t}}{N_S} \widehat{DID}_{+,t} + \frac{N_{0,1,t}}{N_S} \widehat{DID}_{-,t} \right).$$

Here, we take the convention that $\widehat{DID}_{+,t} = 0$ if $\min(N_{1,0,t}, N_{0,0,t}) = 0$, and similarly for $\widehat{DID}_{-,t} = 0$. If Assumption 7 holds, let

$$Z_{g,\bar{g}} = \frac{\bar{g}+1}{N_S} \sum_{t=1}^{\bar{t}} N_{g,t} \left\{ -\frac{N_{1,0,t}}{N_{0,0,t}} I_{1,g,t} - I_{2,g,t} + I_{3,g,t} + \frac{N_{0,1,t}}{N_{1,1,t}} I_{4,g,t} \right\} (Y_{g,t} - Y_{g,t-1}),$$

where $I_{k,g,t} = 1\{D_{g,t-1} + 2D_{g,t} = k-1\}$, with the convention that $0/0 = 0$. Let us also define $\overline{Z}_{\bar{g}} = \sum_{g=0}^{\bar{g}} Z_{g,\bar{g}}/(\bar{g}+1)$ and $\widehat{\sigma}^2 = \sum_{g=0}^{\bar{g}} (Z_{g,\bar{g}} - \overline{Z}_{\bar{g}})^2/(\bar{g}+1)$. We then consider confidence intervals of the form

$$\mathrm{CI}_{1-\alpha}(\Delta^S) = \left[\widehat{W}_{TC} - z_{1-\alpha/2} \frac{\widehat{\sigma}}{\sqrt{\bar{g}}}, \widehat{W}_{TC} + z_{1-\alpha/2} \frac{\widehat{\sigma}}{\sqrt{\bar{g}}} \right].$$

We now establish the asymptotic properties of $\widehat{\underline{\sigma}}_{fe}$, $\widehat{W}_{TC}$ and the corresponding confidence intervals. We consider an asymptotic framework where the number of groups $\bar{g}$ tends to infinity. Let $\widetilde{Y}_{g,\bar{g}} = \sum_{t=0}^{\bar{t}} N_{g,t} \varepsilon_{g,t} Y_{g,t}$.[14] We make the following assumptions.

Assumption 17 *The vectors $(Y_{g,0}, ..., Y_{g,\bar{t}})_{g \geq 0}$ are mutually independent.*

Assumption 18 *We have $\sup_{\bar{g}:0 \leq g \leq \bar{g}} \sum_{t=0}^{\bar{t}} N_{g,t} \varepsilon_{g,t}^2 < +\infty$ and there exists $\delta > 0$ such that $\sup_{\bar{g}:0 \leq g \leq \bar{g}} E(\widetilde{Y}_{g,\bar{g}}^{2+\delta}) < +\infty$. Moreover, as $\bar{g}$ tends to infinity, $\sum_{g=0}^{\bar{g}} \sum_{t=0}^{\bar{t}} N_{g,t} \varepsilon_{g,t} E[Y_{g,t}]/(\bar{g}+1)$, $\sum_{g=0}^{\bar{g}} \sum_{t=0}^{\bar{t}} N_{g,t} \varepsilon_{g,t}^2/(\bar{g}+1)$, $\sum_{g=0}^{\bar{g}} V(\widetilde{Y}_{g,\bar{g}})/(\bar{g}+1)$ and $\sigma(\boldsymbol{w})$ converge respectively to $B \neq 0$, $J > 0$, $H > 0$ and $\sigma_\infty(\boldsymbol{w}) > 0$.*

Assumption 19 *There exists $\delta > 0$ such that $\sup_{\bar{g}:0 \leq g \leq \bar{g}} E\left(|Z_{g,\bar{g}}|^{2+\delta}\right) < +\infty$. Moreover, as $\bar{g}$ tends to infinity, $\sum_{g=0}^{\bar{g}} V(Z_{g,\bar{g}})/(\bar{g}+1)$ converges to $\sigma^2 > 0$.*

Assumption 17 requires the outcome to be independent across groups, but it allows for serial correlation within groups (see, e.g. Bertrand et al., 2004). Note that this assumption does not require that the outcome be identically distributed across groups. This is important, to allow the $\Delta_{g,t}$ to vary across g. Assumptions 18 and 19 respectively impose the (uniform) existence of moments of order $2+\delta$ of $\widetilde{Y}_{g,\bar{g}}$ and $Z_{g,\bar{g}}$, and that some non-random averages converge as $\bar{g}$ tends to infinity. These assumptions ensure that we can apply a law of large number and a central limit theorem to $(\widetilde{Y}_{g,\bar{g}})_{g \geq 0}$ and $(Z_{g,\bar{g}})_{g \geq 0}$, which are independent but not identically distributed.

[14] $\widetilde{Y}_{g,\bar{g}}$ depends on $\bar{g}$ because $\varepsilon_{g,t}$ depends on $\bar{g}$, though we have let this dependency implicit before.

Theorem 7 *Suppose that Assumptions 1-3 and 17-18 hold. Then, as $\overline{g}$ tends to infinity,*

$$\sqrt{\overline{g}}\left(\widehat{\underline{\sigma}}_{fe} - \underline{\sigma}_{fe}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{H}{J^2\sigma_\infty^2(\boldsymbol{w})}\right).$$

Moreover, $\limsup_{\overline{g}\to\infty} \Pr\left(\underline{\sigma}_{fe} \in CI_{1-\alpha}(\underline{\sigma}_{fe})\right) \geq 1-\alpha$.

Theorem 8 *Suppose that Assumptions 1, 2, 6-7, 17 and 19 hold. Then, as $\overline{g}$ tends to infinity,*

$$\sqrt{\overline{g}}\left(\widehat{W}_{TC} - \Delta^S\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Moreover, $\limsup_{\overline{g}\to\infty} \Pr\left(\Delta^S \in CI_{1-\alpha}(\Delta^S)\right) \geq 1-\alpha$.

Theorems 7 and 8 show that $\widehat{\underline{\sigma}}_{fe}$ and $\widehat{W}_{TC}$ are asymptotically normal estimators of $\underline{\sigma}_{fe}$ and Δ^S when the number of groups tends to infinity, provided the outcome is independent across groups. As is usually the case for estimators constructed using independent but not identically distributed random variables (see e.g. Liu and Singh, 1995), the asymptotic variances of $\widehat{\underline{\sigma}}_{fe}$ and $\widehat{W}_{TC}$ can only be conservatively estimated. As a result, the confidence intervals we propose are asymptotically conservative.

7 Applicability, and applications

7.1 Applicability

We conducted a review of all papers published in the American Economic Review (AER) between 2010 and 2012 to assess the 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% of them estimate the FE or FD Regression, or other regressions resembling closely those regressions. When one withdraws from the denominator theory papers and lab experiments, the proportion of papers using these regressions raises to 19.1%.

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

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

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

Table 2 shows descriptive statistics about the 33 2010-2012 AER papers estimating two-way fixed effects regressions. Panel A shows that 13 use the FE regression; six use the FD regression; six use regressions the FE or FD regression with several treatment variables; three use the FE or FD 2SLS regression discussed in Section 2.4 of the Web Appendix; five use other regressions that we deemed sufficiently close to the FE or FD regression to include them in our count.[15] Panel B shows that more than three fourths of those papers consider sharp designs, while less than one fourth consider fuzzy designs. Finally, Panel C assesses whether there are groups whose exposure to the treatment remains stable between each pair of consecutive time periods in those applications, implying that the Wald-TC estimator can be computed. For about a half of the papers, reading the paper was not enough to assess this with certainty. We then assessed whether they presumably have stable groups or not. Overall, 12 papers have stable groups, 14 presumably have stable groups, five presumably do not have stable groups, and two do not have stable groups.

In Section 5 of the Web Appendix, 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 or not.

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

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

	# Papers
Panel A. Estimation method	
Fixed-effects OLS regression	13
First-difference OLS regression	6
Fixed-effects or first-difference OLS regression, with several treatment variables	6
Fixed-effects or first-difference 2LS regression	3
Other regression	5
Panel B. Research design	
Sharp design	26
Fuzzy design	7
Panel C. Are there stable groups?	
Yes	12
Presumably yes	14
Presumably no	5
No	2

Notes. 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.

7.2 Applications

Enikolopov et al. (2011)

Enikolopov et al. (2011) study the effect of NTV, an independent TV channel introduced in 1996 in Russia, on the share of the electorate voting for opposition parties. NTV's coverage rate was heterogeneous across subregions: while a large fraction of the population received NTV in urbanized subregions, a smaller fraction received it in more rural subregions. The authors estimate the FE regression: they regress the share of votes for opposition parties in the 1995 and 1999 elections in Russian subregions on subregion fixed effects, an indicator for the 1999 election, and on the share of the population having access to NTV in each subregion at the time of the election. In 1995, the share of the population having access to NTV was equal to 0 in all subregions, while in 1999 it was strictly greater than 0 everywhere. Therefore, the authors' research design corresponds exactly to the heterogenous adoption design discussed in Section 4. Enikolopov et al. (2011) find that $\widehat{\beta}_{fe} = 6.65$ (s.e.= 1.40). According to this regression, increasing the share of the population having access to NTV from 0 to 100% increases the share

of votes for the opposition parties by 6.65 percentage points. Because there are only two time periods in the data and the regression is not weighted by subregions' populations, $\widehat{\beta}_{fe} = \widehat{\beta}_{fd}$.

We use the `twowayfeweights` Stata package, downloadable with its help file from the SSC repository, to estimate the weights attached to $\widehat{\beta}_{fe}$. In 1995, all the weights are equal to zero because NTV does not exist yet. In 1999, 918 weights (47.4%) are strictly positive, while 1,020 (52.6%) are strictly negative. The negative weights sum to -2.26. $\widehat{\underline{\sigma}}_{fe} = 0.91$ (95% level confidence interval=$[0.49, 1.28]$):[16] β_{fe} and Δ^{TR} may be of opposite signs if the standard deviation of the effect of NTV across subregions is above 0.91 percentage point. $\widehat{\underline{\sigma}}_{=fe} = 1.23$ (95% level confidence interval=$[0.67, 1.72]$): β_{fe} may be of a different sign than all the $\Delta^{TR}_{g,1999}$s if the standard deviation of the effect of NTV across subregions is above 1.23 percentage point, which does not seem to be an implausible amount of treatment effect heterogeneity.

Therefore, β_{fe} can only receive a causal interpretation if the effect of NTV is constant across subregions, or if the weights attached to it are uncorrelated with the intensity of that effect in each subregion (Assumption 5). These assumptions are not warranted. First, we estimate $\widehat{\beta}_{fe}$ again, weighting the regression by subregions' populations. We obtain $\widehat{\beta}_{fe} = 14.89$, more than twice its value in the unweighted regression, and the difference between the coefficients is significant (t-stat=2.46). Therefore, we can reject the null that the treatment effect is constant: if the treatment effect was constant across subregions, the weighting would not matter so both the unweighted and the weighted regressions would estimate the same parameter. Second, the weights attached to $\widehat{\beta}_{fe}$ are correlated with variables that are likely to be themselves associated with the intensity of the effect in each subregion. For instance, the correlation between the weights and subregions' populations is equal to 0.35 (t-stat=14.01). The effect of NTV may be higher in less populated subregions, as those regions are more rural and fewer other sources of information may be available there. This would lead to a violation of Assumption 5.

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

Gentzkow et al. (2011)

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

[16] The confidence intervals of $\underline{\sigma}_{fe}$ and $\underline{\sigma}_{=fe}$ are computed using the bootstrap, clustered at the subregion level.

one more newspaper increased voters' turnout by 0.26 percentage points. On the other hand, $\widehat{\beta}_{fe} = -0.0011$ (s.e.= 0.0011). $\widehat{\beta}_{fe}$ and $\widehat{\beta}_{fd}$ are significantly different (t-stat=2.86).

We estimate the weights attached to $\widehat{\beta}_{fe}$. 6,180 are strictly positive, 4,198 are strictly negative. The negative weights sum to -0.47. $\underline{\widehat{\sigma}}_{fe} = 4 \times 10^{-4}$ (95% level confidence interval=$[2 \times 10^{-5}, 0.0012]$),[17] meaning that β_{fe} and the ATT may be of opposite signs if the standard deviation of the ATTs across groups and time periods is equal to 0.0004.[18] $\underline{\widehat{\sigma}}_{fe} = 7 \times 10^{-4}$ (95% level confidence interval=$[3 \times 10^{-5}, 0.0026]$), meaning that β_{fe} may be of a different sign than all the ATTs if the standard deviation of the ATTs across groups and time periods is equal to 0.0007. We also estimate the weights attached to $\widehat{\beta}_{fd}$. 4,790 are strictly positive, and 5,588 are strictly negative. The negative weights sum to -1.30. $\underline{\widehat{\sigma}}_{fd} = 5 \times 10^{-4}$ (95% level confidence interval=$[1 \times 10^{-4}, 8 \times 10^{-4}]$), and $\underline{\widehat{\sigma}}_{fd} = 7 \times 10^{-4}$ (95% level confidence interval=$[1 \times 10^{-4}, 0.0012]$).

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×election-year cell (Assumptions 5 and 14, respectively). This is not warranted. First, as $\widehat{\beta}_{fe}$ and $\widehat{\beta}_{fd}$ significantly differ, Assumptions 5 and 14 cannot jointly hold. Moreover, the weights attached to $\widehat{\beta}_{fe}$ and $\widehat{\beta}_{fd}$ 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 $\widehat{\beta}_{fd}$ and t, the year variable, is equal to -0.07 (t-stat=-3.33). The effect of newspapers may be different in the last than in the first years of the panel. For instance, new means of communication, like the radio, appear in the end of the period under consideration, and may diminish the effect of newspapers. This would lead to a violation of Assumption 14.

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 Wald-TC estimand, that accounts for the fact that the number of newspapers is not binary (see section 2.2 of our Web Appendix, where we define this modified estimand). We include state-year fixed effects as controls in our estimation. We find that $\widehat{W_{TC}} = 0.0043$, with a standard error of 0.0015. $\widehat{W_{TC}}$ is 66% larger than $\widehat{\beta}_{fd}$, and the two estimators are significantly different at the 10% level (t-stat=1.69). $\widehat{W_{TC}}$ is also of a different sign than $\widehat{\beta}_{fe}$.

Our Wald-TC estimand only relies on a common trends assumption. To assess its plausibility, we estimate W_{TC}^{pl}, the placebo estimand introduced in Section 3.3.[19] As shown in Table 3 below, our placebo estimator is small and not significantly different from 0, meaning that counties

[17] The confidence intervals of $\underline{\sigma}_{fe}$, $\underline{\sigma}_{fe}$, $\underline{\sigma}_{fd}$, and $\underline{\sigma}_{fd}$ are derived using a bootstrap clustered at the county level.

[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 Section 2.2 of our Web Appendix, rather than the ATT.

[19] Again, we need to slightly modify W_{TC}^{pl} to account for the fact that the number of newspapers is not binary.

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% of the county $\times$ election-year observations are used in the computation of the placebo estimator. Moreover, when reestimated on this subsample, the Wald-TC estimator is very close to the Wald-TC estimator in the full sample.

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

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

Notes. This table reports estimates of the effect of one additional newspaper on turnout, as well as a placebo estimate of the common trends assumption underlying W_{TC}. Estimators are computed using the data of Gentzkow et al. (2011), with state-year fixed effects as controls. Standard errors are clustered by county. To compute the Wald-TC estimators, the number of newspapers is grouped into 4 categories: 0, 1, 2, and more than 3.

8 Conclusion

Almost 20% of empirical articles published in the AER between 2010 and 2012 use regressions with groups and period fixed effects to estimate treatment effects. In this paper, we show that under a common trends assumption, those regressions identify weighted sums of the treatment effect in each group and at each period. The weights may be negative: in two empirical applications, we find that more than 50% 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 coefficient of the treatment variable in those regressions is negative while the treatment effect is positive in every group and time period. We therefore propose a new estimand to address this problem. This estimand identifies the effect of the treatment 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` Stata package, and can be used in applications where there are groups whose exposure to the treatment does not change between each pair of consecutive time periods. In one of the two applications we revisit, the corresponding estimator is significantly and economically different from the two-way fixed effects estimators.

References

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

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

Athey, S. and Imbens, G. W. (2018), Design-based analysis in difference-in-differences settings with staggered adoption, Technical report, National Bureau of Economic Research.

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

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

Billingsley, P. (1986), *Probability and measure*, 2nd edn, John Wiley & Sons.

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

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

Conley, T. G. and Taber, C. R. (2011), 'Inference with "difference in differences" with a small number of policy changes', *The Review of Economics and Statistics* **93**(1), 113–125.

de Chaisemartin, C. (2011), Fuzzy differences in differences. Working Paper 2011-10, Center for Research in Economics and Statistics.

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

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

de Chaisemartin, C., D'Haultfœuille, X. and Guyonvarch, Y. (2019), 'Fuzzy differences-in-differences with Stata', *Stata Journal* **Forthcoming**.

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

Frank, M. and Wolfe, P. (1956), 'An algorithm for quadratic programming', *Naval research logistics quarterly* **3**(1-2), 95–110.

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

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

Gut, A. (1992), 'The weak law of large numbers for arrays', *Statistics & probability letters* **14**(1), 49–52.

Imai, K. and Kim, I. S. (2018), 'On the use of two-way fixed effects regression models for causal inference with panel data'.

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

Liu, R. Y. and Singh, K. (1995), 'Using iid bootstrap inference for general non-iid models', *Journal of statistical planning and inference* **43**(1-2), 67–75.

van der Vaart, A. (2000), *Asymptotics Statistics*, Cambridge University Press.

A Proofs

Two useful lemmas

In the sharp designs considered in Section 3, our results rely on the following lemma.

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

$$E(Y_{g,t}) - E(Y_{g,t'}) - (E(Y_{g',t}) - E(Y_{g',t'})) = D_{g,t}\Delta_{g,t} - D_{g,t'}\Delta_{g,t'} - (D_{g',t}\Delta_{g',t} - D_{g',t'}\Delta_{g',t'}).$$

In the fuzzy designs considered in Section 4, our results rely on the following lemma.

Lemma 2 *If Assumptions 1 and 9 hold, for all* $(g, g', t, t') \in \{0, ..., \overline{g}\}^2 \times \{0, ..., \overline{t}\}^2$,

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

Proof of Lemma 1

For all $(g, t) \in \{0, ..., \overline{g}\} \times \{0, ..., \overline{t}\}$,

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

where the second equality follows from Assumption 2. Therefore,

$$\begin{aligned}&E(Y_{g,t}) - E(Y_{g,t'}) - (E(Y_{g',t}) - E(Y_{g',t'}))\\ =&E\left(Y_{g,t}(0)\right) - E\left(Y_{g,t'}(0)\right) - (E\left(Y_{g',t}(0)\right) - E\left(Y_{g',t'}(0)\right))\\ +&D_{g,t}\Delta_{g,t} - D_{g,t'}\Delta_{g,t'} - (D_{g',t}\Delta_{g',t} - D_{g',t'}\Delta_{g',t'})\\ =&D_{g,t}\Delta_{g,t} - D_{g,t'}\Delta_{g,t'} - (D_{g',t}\Delta_{g',t} - D_{g',t'}\Delta_{g',t'}),\end{aligned}$$

where the second equality follows from Assumption 3.

Proof of Lemma 2

For all $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$,

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

where the second and third equalities follow from the fact that the treatment of observations in the same cell are identically distributed. The end of the proof is very similar to that of Lemma 1, except that one needs to invoke Assumption 9 instead of Assumption 3.

Proof of Theorem 1

It follows from the Frisch-Waugh theorem, the definition of $\varepsilon_{g,t}$, and Assumption 2 that

$$\beta_{fe} = \frac{\sum_{g,t} N_{g,t}\varepsilon_{g,t}E(Y_{g,t})}{\sum_{g,t} N_{g,t}\varepsilon_{g,t}D_{g,t}}. \tag{5}$$

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

$$\sum_{t=0}^{\overline{t}} N_{g,t}\varepsilon_{g,t} = 0 \text{ for all } g \in \{0,...,\overline{g}\}, \tag{6}$$

$$\sum_{g=0}^{\overline{g}} N_{g,t}\varepsilon_{g,t} = 0 \text{ for all } t \in \{0,...,\overline{t}\}. \tag{7}$$

Then,

$$\begin{aligned}
&\sum_{g,t} N_{g,t}\varepsilon_{g,t}E(Y_{g,t}) \\
=& \sum_{g,t} N_{g,t}\varepsilon_{g,t}\left(E(Y_{g,t}) - E(Y_{g,0}) - E(Y_{0,t}) + E(Y_{0,0})\right) \qquad (8) \\
=& \sum_{g,t} N_{g,t}\varepsilon_{g,t}\left(D_{g,t}\Delta_{g,t} - D_{g,0}\Delta_{g,0} - D_{0,t}\Delta_{0,t} + D_{0,0}\Delta_{0,0}\right) \\
=& \sum_{g,t} N_{g,t}\varepsilon_{g,t}D_{g,t}\Delta_{g,t} \\
=& \sum_{(g,t):D_{g,t}=1} N_{g,t}\varepsilon_{g,t}\Delta_{g,t}. \qquad (9)
\end{aligned}$$

The first and third equalities follow from Equations (6) and (7). The second equality follows from Lemma 1. The fourth equality follows from Assumption 2. Finally, Assumption 2 implies that

$$\sum_{g,t} N_{g,t}\varepsilon_{g,t}D_{g,t} = \sum_{(g,t):D_{g,t}=1} N_{g,t}\varepsilon_{g,t}. \tag{10}$$

Combining (5), (9), and (10) yields the result.

Proof of Proposition 1

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

$$D_{\cdot,\cdot} = \sum_{g=0}^{\bar{g}} \frac{N_{g,\cdot}}{N} D_{g,\cdot} = \sum_{t=0}^{\bar{t}} \frac{N_{\cdot,t}}{N} D_{\cdot,t},$$

it is easy to see that $1 - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot}$ is decreasing in $D_{g,\cdot}$ and $D_{\cdot,t}$ for all (g,t).

Proof of Corollary 1

Proof of the first point

We start by proving the first point. If the assumptions of the corollary hold and $\Delta^{TR} = 0$, then

$$\begin{cases} \beta_{fe} &= \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta_{g,t}, \\ 0 &= \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \Delta_{g,t}. \end{cases}$$

These two conditions and the Cauchy-Schwarz inequality imply

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

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

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

$$\Delta^{TR}_{g,t} = \frac{\beta_{fe}\,(w_{g,t} - 1)}{\sigma^2(\boldsymbol{w})}.$$

Then,

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

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

Similarly,

$$\begin{aligned}\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} \frac{\beta_{fe}\left(w_{g,t}-1\right)}{\sigma^2(\boldsymbol{w})} &= \frac{\beta_{fe}}{\sigma^2(\boldsymbol{w})} \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t}\left(w_{g,t}-1\right) \\ &= \frac{\beta_{fe}}{\sigma^2(\boldsymbol{w})} \sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \left(w_{g,t}-1\right)^2 \\ &= \beta_{fe},\end{aligned}$$

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 $\beta_{fe} > 0$. We seek to solve:

$$\min_{\Delta_{(1)},\ldots,\Delta_{(n)}} \sum_{i=1}^{n} \frac{N(i)}{N_1} \left(\Delta_{(i)} - \Delta^{TR}\right)^2 \quad \text{s.t. } \beta_{fe} = \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} w_{(i)} \Delta_{(i)},$$
$$\Delta_{(i)} \leq 0 \text{ for all } i \in \{1,\ldots,n\}.$$

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 $\sum_{i=1}^{n} \frac{N_{(i)}}{N_1} w_{(i)} \Delta_{(i)} \leq 0 < \beta_{fe}$. On the other hand, if $w_{(n)} < 0$, this set is non-empty since it includes $(0,\ldots,0,\beta_{fe}/(P_{(n)} w_{(n)}))$.

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

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

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

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

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

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

$$\Delta_{(i)} = \min(\Delta^{TR} + \lambda w_{(i)}, 0). \tag{11}$$

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

As a result, $\Delta^{TR} + \lambda w_{(i)}$ is decreasing in i, and because $x \mapsto \min(x, 0)$ is increasing, $\Delta_{(i)}$ is also decreasing in i. Then $\Delta_{(n)} < 0$: otherwise one would have $\Delta_{(i)} = 0$ for all i which would imply $\beta_{fe} = 0$, a contradiction. Let $s = \min\{i \in \{1, ..., n\} : \Delta_{(i)} < 0\}$. Using again (11), we get

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

Therefore,

$$\Delta^{TR} = \frac{\lambda S_s}{1 - P_s}. \tag{12}$$

Hence, plugging Δ in (11), we obtain that for all $i \geq s$,

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

Finally, using again (11), we obtain

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

Thus,

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

Then, using what precedes,

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

The result follows, once noted that Equations (11) and (12) imply that $s = \min\{i \in \{1, ..., n\} : w_{(i)} < -S_{(i)}/(1 - P_{(i)})\}$.

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

$$\underline{\sigma}_{fe} = \min_{\Delta'_{(1)} \leq 0, \ldots, \Delta'_{(n)} \leq 0} \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} {\Delta'_{(i)}}^2 - \left(\sum_{i=1}^{n} \frac{N_{(i)}}{N_1} \Delta'_{(i)} \right)^2 \text{ s.t. } \sum_{i=1}^{n} \frac{N_{(i)}}{N_1} w_{(i)} \Delta'_{(i)} = \beta'_{fe}.$$

This is the same program as before, with β'_{fe} instead of β_{fe}. Therefore, by the same reasoning as before, we obtain

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

Proof of Corollary 2

We have

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

Under the assumptions of the corollary, Theorem 1 holds, hence the first equality. The second equality follows from Assumption 5. By the definition of $w_{g,t}$, $\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{g,t} = 1$, hence the third equality.

Proof of Theorem 2

Let t be greater than 1, and let us assume for now that there exists at least one g_1 such that $D_{g_1,t-1} = 0$ and $D_{g_1,t} = 1$. Then Assumption 7 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

$$\begin{aligned} E[Y_{g,t} - Y_{g,t-1}] =& E[Y_{g,t}(1) - Y_{g,t}(0)] + E[Y_{g,t}(0) - Y_{g,t-1}(0)] \\ =& \Delta_{g,t} + E[Y_{g,t}(0) - Y_{g,t-1}(0)]. \end{aligned} \tag{13}$$

Under Assumption 6, for all $t \geq 1$ there exists a real number $\psi_{0,t}$ such that for all g such that $D_{g,t-1} = 0$, $E(Y_{g,t}(0) - Y_{g,t-1}(0)) = \psi_{0,t}$. Then,

$$\begin{aligned}
N_{1,0,t} DID_{+,t} = & \sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t}\Delta_{g,t} + \sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t} E[Y_{g,t}(0) - Y_{g,t-1}(0)] \\
& - \frac{N_{1,0,t}}{N_{0,0,t}} \sum_{g:D_{g,t}=D_{g,t-1}=0} N_{g,t} E[Y_{g,t}(0) - Y_{g,t-1}(0)] \\
= & \sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t}\Delta_{g,t} + \psi_{0,t} \sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t} \\
& - \psi_{0,t} \frac{N_{1,0,t}}{N_{0,0,t}} \sum_{g:D_{g,t}=D_{g,t-1}=0} N_{g,t} \\
= & \sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t}\Delta_{g,t}, \qquad (14)
\end{aligned}$$

where the first equality follows by (13), the second by Assumption 6 and the third by the fact that

$$\sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t} = \frac{N_{1,0,t}}{N_{0,0,t}} \sum_{g:D_{g,t}=D_{g,t-1}=0} N_{g,t} = N_{1,0,t}.$$

If there exists no g such that $D_{g,t-1} = 0$ and $D_{g,t} = 1$, (14) still holds, as we have let $DID_{+,t} = 0$ in this case.

A similar reasoning yields

$$N_{0,1,t} DID_{-,t} = \sum_{g:D_{g,t}=0,D_{g,t-1}=1} N_{g,t}\Delta_{g,t}.$$

As a result,

$$\begin{aligned}
W_{TC} = & \frac{1}{N_S} \sum_{t=1}^{\bar{t}} \left(\sum_{g:D_{g,t}=1,D_{g,t-1}=0} N_{g,t}\Delta_{g,t} + \sum_{g:D_{g,t}=0,D_{g,t-1}=1} N_{g,t}\Delta_{g,t} \right) \\
= & \Delta^S.
\end{aligned}$$

Proof of Theorem 3

Let t be greater than 2, and let us assume for now that 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 8 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,

$$\begin{aligned} N_{1,0,0,t} DID^{pl}_{+,t} = & \sum_{g:D_{g,t}=1,D_{g,t-1}=D_{g,t-2}=0} N_{g,t} E[Y_{g,t-1}(0) - Y_{g,t-2}(0)] \\ & - \frac{N_{1,0,0,t}}{N_{0,0,0,t}} \sum_{g:D_{g,t}=D_{g,t-1}=D_{g,t-2}=0} N_{g,t} E[Y_{g,t-1}(0) - Y_{g,t-2}(0)] \\ = & \psi_{0,t-1} \sum_{g:D_{g,t}=1,D_{g,t-1}=D_{g,t-2}=0} N_{g,t} - \psi_{0,t-1} \frac{N_{1,0,0,t}}{N_{0,0,0,t}} \sum_{g:D_{g,t}=D_{g,t-1}=D_{g,t-2}=0} N_{g,t} \\ = & 0. \end{aligned} \tag{15}$$

The second equality follows by Assumption 6, where $\psi_{0,t-1}$ is defined in the proof of Theorem 2. The third equality follows from the fact that

$$\sum_{g:D_{g,t}=1,D_{g,t-1}=D_{g,t-2}=0} N_{g,t} = \frac{N_{1,0,0,t}}{N_{0,0,0,t}} \sum_{g:D_{g,t}=D_{g,t-1}=D_{g,t-2}=0} N_{g,t} = N_{1,0,0,t}.$$

If there exists no g such that $D_{g,t-2} = D_{g,t-1} = 0$ and $D_{g,t} = 1$, (15) still holds, as we have let $DID^{pl}_{+,t} = 0$ in this case.

A similar reasoning yields $N_{0,1,1,t} DID_{-,t} = 0$. As a result, $W^{pl}_{TC} = 0$.

Proof of Theorem 4

The proof of that result is very similar to the proof of Theorem 1.

$$\begin{aligned} \beta_{fe} &= E\left(\frac{\sum_{g,t} N_{g,t} \varepsilon_{g,t} Y_{g,t}}{\sum_{g,t} N_{g,t} \varepsilon_{g,t} D_{g,t}} \right) \\ &= E\left(\frac{\sum_{g,t} N_{g,t} \varepsilon_{g,t} E(Y_{g,t}|\boldsymbol{D})}{\sum_{g,t} N_{g,t} \varepsilon_{g,t} D_{g,t}} \right) \\ &= E\left(\frac{\sum_{g,t} N_{g,t} \varepsilon_{g,t} \left(E(Y_{g,t}|\boldsymbol{D}) - E(Y_{g,0}|\boldsymbol{D}) - E(Y_{0,t}|\boldsymbol{D}) + E(Y_{0,0}|\boldsymbol{D}) \right)}{\sum_{g,t} N_{g,t} \varepsilon_{g,t} D_{g,t}} \right) \\ &= E\left(\frac{\sum_{g,t} N_{g,t} \varepsilon_{g,t} \left(D_{g,t} \Delta^{TR}_{g,t}(\boldsymbol{D}) - D_{g,0} \Delta^{TR}_{g,0}(\boldsymbol{D}) - D_{0,t} \Delta^{TR}_{0,t}(\boldsymbol{D}) + D_{0,0} \Delta^{TR}_{0,0}(\boldsymbol{D}) \right)}{\sum_{g,t} N_{g,t} \varepsilon_{g,t} D_{g,t}} \right) \\ &= E\left(\frac{\sum_{g,t} N_{g,t} \varepsilon_{g,t} D_{g,t} \Delta^{TR}_{g,t}(\boldsymbol{D})}{\sum_{g,t} N_{g,t} \varepsilon_{g,t} D_{g,t}} \right). \end{aligned}$$

The second equality follows from the law of iterated expectations. The third and fifth equalities follow from Equations (6) and (7). The fourth equality follows from Lemma 2.

Proof of Proposition 2

Assuming that $N_{g,1}/N_{g,0}$ does not vary across g ensures that there exists a strictly positive real number ϕ such that $N_{g,1}/N_{g,0} = \phi$. Then,

$$\begin{aligned}
\varepsilon_{g,1} =& D_{g,1} - D_{g,.} - D_{.,1} + D_{.,.} \\
=& D_{g,1} - \left(\frac{N_{g,0}}{N_{g,.}} D_{g,0} + \frac{N_{g,1}}{N_{g,.}} D_{g,1}\right) - D_{.,1} + \left(\frac{N_{.,0}}{N} D_{.,0} + \frac{N_{.,1}}{N} D_{.,1}\right) \\
=& D_{g,1} - \left(\frac{1}{1+\phi} D_{g,0} + \frac{\phi}{1+\phi} D_{g,1}\right) - D_{.,1} + \left(\frac{1}{1+\phi} D_{.,0} + \frac{\phi}{1+\phi} D_{.,1}\right) \\
=& \frac{1}{1+\phi} \left(D_{g,1} - D_{g,0} - D_{.,1} + D_{.,0}\right), \qquad (16)
\end{aligned}$$

where the first and third equalities follow from the fact $N_{g,1}/N_{g,0}$ does not vary across g.

Then, the definition of $w_{g,1}^{TR}$, Equation (16) and Assumption 10 imply that

$$\begin{aligned}
w_{g,1}^{TR} =& \frac{(D_{g,1} - D_{.,1})}{\sum_{g=0}^{\overline{g}} \frac{N_{g,1}}{N_1} (D_{g,1} - D_{.,1}) D_{g,1}} \\
=& \frac{(D_{g,1} - D_{.,1})}{\sum_{g=0}^{\overline{g}} \frac{N_{g,1}}{N_1} (D_{g,1} - D_{.,1})^2}.
\end{aligned}$$

Proof of Theorem 5

In what follows, we consider a t greater than 1.

We first show that for all (g,t) such that $E(D_{g,t}) > E(D_{g,t-1})$, $P(D_{1,g,t}(t) \geq D_{1,g,t}(t-1)) = 1$. We prove the result by contradiction. If $P(D_{1,g,t}(t) \geq D_{1,g,t}(t-1)) \neq 1$ then Point 2 of Assumption 11 implies that $P(D_{1,g,t}(t) \leq D_{1,g,t}(t-1)) = 1$, which implies that $E(D_{1,g,t}(t)) \leq E(D_{1,g,t}(t-1))$. Point 1 of Assumption 11 then implies that $E(D_{1,g,t}(t)) \leq E(D_{1,g,t-1}(t-1))$, which is equivalent to having $E(D_{g,t}) \leq E(D_{g,t-1})$, a contradiction. Similarly, one can show that for all (g,t) such that $E(D_{g,t}) = E(D_{g,t-1})$, $P(D_{1,g,t}(t) = D_{1,g,t}(t-1)) = 1$.

Then, for all (g,t) such that $E(D_{g,t}) > E(D_{g,t-1})$,

$$\begin{aligned}
E(D_{g,t}) - E(D_{g,t-1}) =& P(D_{1,g,t}(t) = 1) - P(D_{1,g,t-1}(t-1) = 1) \\
=& P(D_{1,g,t}(t) = 1) - P(D_{1,g,t}(t-1) = 1) \\
=& P(D_{1,g,t}(t) = 1, D_{1,g,t}(t-1) = 0). \qquad (17)
\end{aligned}$$

The second equality follows from Point 1 of Assumption 11. The third follows from the fact that as shown above, $E(D_{g,t}) > E(D_{g,t-1})$ implies that $P(D_{1,g,t}(t) \geq D_{1,g,t}(t-1)) = 1$.

Then, under Assumption 12, for all $t \geq 1$ and $d \in \{0,1\}$, there exist real numbers $\alpha_{d,t}$ such that

$$E(Y_{1,g,t}(d)|D_{1,g,t}(t-1)=d) - E(Y_{1,g,t-1}(d)|D_{1,g,t-1}(t-1)=d) = \alpha_{d,t} \tag{18}$$

Now, for all (g,t) such that $E(D_{g,t}) > E(D_{g,t-1})$,

$$\begin{aligned}
&E(Y_{g,t}) - E(Y_{g,t-1}) \\
=&E(Y_{1,g,t}) - E(Y_{1,g,t-1}) \\
=&E(Y_{1,g,t}(1)|D_{1,g,t}(t-1)=1)P(D_{1,g,t}(t-1)=1) + E(Y_{1,g,t}(0)|D_{1,g,t}(t)=0)P(D_{1,g,t}(t)=0) \\
+&E(Y_{1,g,t}(1)|D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1)P(D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1) \\
-&E(Y_{1,g,t-1}(1)|D_{1,g,t-1}(t-1)=1)P(D_{1,g,t-1}(t-1)=1) - E(Y_{1,g,t-1}(0)|D_{1,g,t-1}(t-1)=0)P(D_{1,g,t-1}(t-1)=0) \\
=&E(Y_{1,g,t}(1)|D_{1,g,t}(t-1)=1)P(D_{1,g,t}(t-1)=1) + E(Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0)P(D_{1,g,t}(t-1)=0) \\
-&E(Y_{1,g,t-1}(1)|D_{1,g,t-1}(t-1)=1)P(D_{1,g,t-1}(t-1)=1) - E(Y_{1,g,t-1}(0)|D_{1,g,t-1}(t-1)=0)P(D_{1,g,t-1}(t-1)=0) \\
+&E(Y_{1,g,t}(1) - Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1)P(D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1) \\
=&E(Y_{1,g,t}(1)|D_{1,g,t}(t-1)=1)P(D_{1,g,t-1}(t-1)=1) + E(Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0)P(D_{1,g,t-1}(t-1)=0) \\
-&E(Y_{1,g,t-1}(1)|D_{1,g,t-1}(t-1)=1)P(D_{1,g,t-1}(t-1)=1) - E(Y_{1,g,t-1}(0)|D_{1,g,t-1}(t-1)=0)P(D_{1,g,t-1}(t-1)=0) \\
+&E(Y_{1,g,t}(1) - Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1)P(D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1) \\
=&\alpha_{1,t}E(D_{g,t-1}) + \alpha_{0,t}(1 - E(D_{g,t-1})) \\
+&E(Y_{1,g,t}(1) - Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1)P(D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1).
\end{aligned} \tag{19}$$

The first equality follows from the fact that observations are identically distributed within each (g,t) cell. The second equality follows from the law of iterated expectations, and from the fact that as shown above, $E(D_{g,t}) > E(D_{g,t-1})$ implies that $P(D_{1,g,t}(t) \geq D_{1,g,t}(t-1)) = 1$. The third equality follows after adding and subtracting $E(Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1)P(D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1)$, and using again the fact that $P(D_{1,g,t}(t) \geq D_{1,g,t}(t-1)) = 1$. The fourth equality follows from Point 1 of Assumption 11. The fifth equality follows from (18). Then,

$$\begin{aligned}
&\delta_{0,t} \\
=&\sum_{g:E(D_{g,t})=E(D_{g,t-1})} \frac{N_{g,t}(1-E(D_{g,t}))}{N_{=,t} - N_{1,=,t}} \left(E(Y_{1,g,t}(0)|D_{1,g,t}(t)=0) - E(Y_{1,g,t-1}(0)|D_{1,g,t-1}(t-1)=0)\right) \\
=&\sum_{g:E(D_{g,t})=E(D_{g,t-1})} \frac{N_{g,t}(1-E(D_{g,t}))}{N_{=,t} - N_{1,=,t}} \left(E(Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0) - E(Y_{1,g,t-1}(0)|D_{1,g,t-1}(t-1)=0)\right) \\
=&\alpha_{0,t}.
\end{aligned} \tag{20}$$

The second equality follows from the fact that $E(D_{g,t}) = E(D_{g,t-1})$ implies that $P(D_{1,g,t}(t) = D_{1,g,t}(t-1)) = 1$. The third equality follows from (18) and the definitions of $N_{=,t}$ and $N_{1,=,t}$. Similarly,

$$\delta_{1,t} = \alpha_{1,t}. \tag{21}$$

Therefore,

$$\begin{aligned}
&\sum_{g:E(D_{g,t})>E(D_{g,t-1})} \frac{N_{g,t}}{N_{+,t}} \left(E(Y_{g,t}) - (E(Y_{g,t-1}) + (1-p_{+,t-1})\delta_{0,t} + p_{+,t-1}\delta_{1,t})\right) \\
&= \sum_{g:E(D_{g,t})>E(D_{g,t-1})} \frac{N_{g,t}}{N_{+,t}} \left(\alpha_{1,t}E(D_{g,t-1}) + \alpha_{0,t}(1-E(D_{g,t-1}))\right) - p_{+,t-1}\alpha_{1,t} - (1-p_{+,t-1})\alpha_{0,t} \\
&+ \sum_{g:E(D_{g,t})>E(D_{g,t-1})} \frac{N_{g,t}}{N_{+,t}} E(Y_{1,g,t}(1) - Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1)P(D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1) \\
&= \sum_{g:E(D_{g,t})>E(D_{g,t-1})} \frac{N_{g,t}}{N_{+,t}} E(Y_{1,g,t}(1) - Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1)P(D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1). \qquad (22)
\end{aligned}$$

The first equality follows from (19), (20), and (21). The second equality follows from the definition of $p_{+,t-1}$.

Then, it follows from (22), (17) and the definition of $W_{TC,+,t}$ that

$$\begin{aligned}
&W_{TC,+,t} \\
&= \sum_{g:E(D_{g,t})>E(D_{g,t-1})} \frac{N_{g,t}P(D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1)}{P_{S,+,t}} E(Y_{1,g,t}(1) - Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1), \qquad (23)
\end{aligned}$$

where $P_{S,+,t} = \sum_{g:E(D_{g,t})>E(D_{g,t-1})} N_{g,t}P(D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1)$.

Similarly, one can show that

$$\begin{aligned}
&W_{TC,-,t} \\
&= \sum_{g:E(D_{g,t})<E(D_{g,t-1})} \frac{N_{g,t}P(D_{1,g,t}(t-1)=1, D_{1,g,t}(t)=0)}{P_{S,-,t}} E(Y_{1,g,t}(1) - Y_{1,g,t}(0)|D_{1,g,t}(t-1)=0, D_{1,g,t}(t)=1), \qquad (24)
\end{aligned}$$

where $P_{S,-,t} = \sum_{g:E(D_{g,t})<E(D_{g,t-1})} N_{g,t}P(D_{1,g,t}(t-1)=1, D_{1,g,t}(t)=0)$.

Finally, the result follows after summing (23) and (24) over $t \geq 1$, once noted that

$$\sum_{t=1}^{\bar{t}} \sum_{g:E(D_{g,t})\neq E(D_{g,t-1})} N_{g,t}|E(D_{g,t}) - E(D_{g,t-1})| = P_S.$$

Proof of Theorem 6

It follows from the Frisch-Waugh theorem, the definition of $\varepsilon_{fd,g,t}$, and Assumption 2 that

$$\beta_{fd} = \frac{\sum_{(g,t):t\geq 1} N_{g,t}\varepsilon_{fd,g,t}\left(E(Y_{g,t}) - E(Y_{g,t-1})\right)}{\sum_{(g,t):t\geq 1} N_{g,t}\varepsilon_{fd,g,t}\left(D_{g,t} - D_{g,t-1}\right)}. \qquad (25)$$

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

$$\sum_{g=0}^{\bar{g}} N_{g,t}\varepsilon_{fd,g,t} = 0 \text{ for all } t \in \{1, ..., \bar{t}\}. \qquad (26)$$

Then,

$$
\begin{aligned}
&\sum_{(g,t):t\geq 1} N_{g,t}\varepsilon_{fd,g,t}\left(E(Y_{g,t})-E(Y_{g,t-1})\right)\\
&=\sum_{(g,t):t\geq 1} N_{g,t}\varepsilon_{fd,g,t}\left(E(Y_{g,t})-E(Y_{g,t-1})-E(Y_{0,t})+E(Y_{0,t-1})\right)\\
&=\sum_{(g,t):t\geq 1} N_{g,t}\varepsilon_{fd,g,t}\left(D_{g,t}\Delta_{g,t}-D_{g,t-1}\Delta_{g,t-1}-D_{0,t}\Delta_{0,t}+D_{0,t-1}\Delta_{0,t-1}\right)\\
&=\sum_{(g,t):t\geq 1} N_{g,t}\varepsilon_{fd,g,t}\left(D_{g,t}\Delta_{g,t}-D_{g,t-1}\Delta_{g,t-1}\right)\\
&=\sum_{g,t}\left(N_{g,t}\varepsilon_{fd,g,t}-N_{g,t+1}\varepsilon_{fd,g,t+1}\right)D_{g,t}\Delta_{g,t}\\
&=\sum_{(g,t):D_{g,t}=1} N_{g,t}\left(\varepsilon_{fd,g,t}-\frac{N_{g,t+1}}{N_{g,t}}\varepsilon_{fd,g,t+1}\right)\Delta_{g,t}. \qquad (27)
\end{aligned}
$$

The first and third equalities follow from (26). The second equality follows from Lemma 1. The fourth equality follows from a summation by part, and from the fact $\varepsilon_{fd,g,0}=\varepsilon_{fd,g,\bar{t}+1}=0$. The fifth equality follows from Assumption 2.

A similar reasoning yields

$$
\sum_{(g,t):t\geq 1} N_{g,t}\varepsilon_{fd,g,t}\left(D_{g,t}-D_{g,t-1}\right)=\sum_{(g,t):D_{g,t}=1} N_{g,t}\left(\varepsilon_{fd,g,t}-\frac{N_{g,t+1}}{N_{g,t}}\varepsilon_{fd,g,t+1}\right). \qquad (28)
$$

Combining (25), (27), and (28) yields the result.

Proof of Proposition 3

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 4 and if $N_{g,t}$ does not vary across t, one has that for all (g,t) such that $D_{g,t}=1, 1\leq t\leq \bar{t}-1$, $w_{fd,g,t}$ is proportional to $1-D_{g,t-1}-(2D_{.,t}-D_{.,t-1}-D_{.,t+1})$. $D_{.,t}-D_{.,t-1}\leq 1$, and under Assumption 4 $D_{.,t}-D_{.,t+1}\leq 0$, 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 $D_{g,t}=1, 1\leq t\leq \bar{t}-1$, $w_{fd,g,t}$ is strictly negative if and only if $D_{g,t-1}=1$ and $2D_{.,t}-D_{.,t-1}-D_{.,t+1}>0$.

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

Finally, when $t=0$, one has that for all g such that $D_{g,0}=1$, $w_{fd,g,0}$ is proportional to $D_{.,1}-D_{.,0}$, which is greater than 0 under Assumption 4.

Proof of Theorem 7

Remark, as in the proof of Theorem 1, that

$$\begin{aligned}\widehat{\beta}_{fe} - \beta_{fe} =& \frac{\sum_{g,t} N_{g,t}\varepsilon_{g,t}(Y_{g,t} - E(Y_{g,t}))}{\sum_{g,t} N_{g,t}\varepsilon_{g,t}D_{g,t}} \\ =& \frac{\frac{1}{\overline{g}+1}\sum_{g=0}^{\overline{g}}(\widetilde{Y}_{g,\overline{g}} - E(\widetilde{Y}_{g,\overline{g}}))}{\frac{1}{\overline{g}+1}\sum_{g,t} N_{g,t}\varepsilon_{g,t}^2}. \end{aligned} \tag{29}$$

By Assumption 18, the denominator tends to J. Now, by the triangular inequality, convexity of $x \mapsto x^{2+\delta}$ on $\mathbb{R}^+$ and Jensen's inequality,

$$|\widetilde{Y}_{g,\overline{g}} - E(\widetilde{Y}_{g,\overline{g}})|^{2+\delta} \leq 2^{1+\delta}\left(|\widetilde{Y}_{g,\overline{g}}|^{2+\delta} + E(|\widetilde{Y}_{g,\overline{g}}|^{2+\delta})\right). \tag{30}$$

Thus,

$$\sum_{g=0}^{\overline{g}} E[|\widetilde{Y}_{g,\overline{g}} - E(\widetilde{Y}_{g,\overline{g}})|^{2+\delta}]/(\overline{g}+1)^{2+\delta} \leq \frac{2^{2+\delta}}{(\overline{g}+1)^{1+\delta}} \sup_{\overline{g}:0\leq g\leq \overline{g}} E(|\widetilde{Y}_{g,\overline{g}}|^{2+\delta}).$$

Therefore, letting $\widetilde{V}_{\overline{g}} = (1/(\overline{g}+1))\sum_{g=0}^{\overline{g}} V(\widetilde{Y}_{g,\overline{g}})$,

$$\frac{\sum_{g=0}^{\overline{g}} E[|\widetilde{Y}_{g,\overline{g}} - E(\widetilde{Y}_{g,\overline{g}})|^{2+\delta}]/(\overline{g}+1)^{2+\delta}}{V\left(\sum_{g=0}^{\overline{g}} \widetilde{Y}_{g,\overline{g}}/(\overline{g}+1)\right)^{1+\delta/2}} \leq \frac{2^{2+\delta}}{(\overline{g}+1)^{\delta/2}} \frac{\sup_{\overline{g}:0\leq g\leq \overline{g}} E(|\widetilde{Y}_{g,\overline{g}}|^{2+\delta})}{\widetilde{V}_{\overline{g}}^{1+\delta/2}}.$$

By Assumption 18, the last term on the right-hand side converges to a finite value. Thus, the right-hand side tends to zero as $\overline{g}$ tends to infinity, and by the Lyapunov central limit theorem for triangular arrays (see,e.g. Billingsley, 1986, Theorem 27.3),

$$\frac{\sum_{g=0}^{\overline{g}}(\widetilde{Y}_{g,\overline{g}} - E(\widetilde{Y}_{g,\overline{g}})/\sqrt{\overline{g}+1})}{\widetilde{V}_{\overline{g}}^{1/2}} \xrightarrow{d} \mathcal{N}(0,1).$$

By Assumption 18, $\widetilde{V}_{\overline{g}}$ tends to H. Then, using (29), we obtain

$$\sqrt{\overline{g}+1}\left(\widehat{\beta}_{fe} - \beta_{fe}\right) \xrightarrow{d} \mathcal{N}(0, H/J^2). \tag{31}$$

Moreover, by Assumption 18 and in view of (5), $\beta_{fe} \to B/J \neq 0$. Then, by the uniform delta method (see, e.g. van der Vaart, 2000, Theorem 3.8), (31) still holds if we replace $\widehat{\beta}_{fe} - \beta_{fe}$ by $|\widehat{\beta}_{fe}| - |\beta_{fe}|$. The first result follows by definition of $\widehat{\underline{\sigma}}_{fe}$ and $\underline{\sigma}_{fe}$ and the convergence of $\sigma(\boldsymbol{w})$ to $\sigma_\infty(\boldsymbol{w}) > 0$.

To prove the result on the confidence interval, first remark that $(\overline{g}+1)\text{se}_{fe}^2 = H_{\overline{g}}/J_{\overline{g}}^2$, with

$$\begin{aligned} H_{\overline{g}} &= \frac{1}{\overline{g}+1}\sum_{g=0}^{\overline{g}}\left[\sum_{t=0}^{\bar{t}} N_{g,t}(Y_{g,t} - \widehat{\beta}_{fe}\varepsilon_{g,t})\varepsilon_{g,t}\right]^2, \\ J_{\overline{g}} &= \frac{1}{\overline{g}+1}\sum_{g,t} N_{g,t}\varepsilon_{g,t}^2. \end{aligned}$$

Moreover, letting $u_{g,\overline{g}} = \sum_{t=0}^{\overline{t}} N_{g,t}\varepsilon_{g,t}^2$, we have

$$\begin{aligned}
H_{\overline{g}} =& \frac{1}{\overline{g}+1}\sum_{g=0}^{\overline{g}}\left[\widetilde{Y}_{g,\overline{g}} - \widehat{\beta}_{fe}u_{g,\overline{g}}\right]^2 \\
=& \frac{1}{\overline{g}+1}\sum_{g=0}^{\overline{g}}(\widetilde{Y}_{g,\overline{g}} - E(\widetilde{Y}_{g,\overline{g}}))^2 + \frac{2}{\overline{g}+1}\sum_{g=0}^{\overline{g}}(\widetilde{Y}_{g,\overline{g}} - E(\widetilde{Y}_{g,\overline{g}}))(E(\widetilde{Y}_{g,\overline{g}}) - \widehat{\beta}_{fe}u_{g,\overline{g}}) \\
&+ \frac{1}{\overline{g}+1}\sum_{g=0}^{\overline{g}}(E(\widetilde{Y}_{g,\overline{g}}) - \widehat{\beta}_{fe}u_{g,\overline{g}})^2 \\
=& T_{1\overline{g}} + T_{2\overline{g}} + T_{3\overline{g}}. \qquad (32)
\end{aligned}$$

This implies that

$$\left|\frac{\widehat{\underline{\sigma}}_{fe} - \underline{\sigma}_{fe}}{\mathrm{se}_{fe}/\sigma(\boldsymbol{w})}\right| \leq \sqrt{\overline{g}+1}\left|\frac{\widehat{\underline{\sigma}}_{fe} - \underline{\sigma}_{fe}}{\frac{\sqrt{T_{1\overline{g}}+T_{2\overline{g}}}}{J_{\overline{g}}\sigma(\boldsymbol{w})}}\right|. \qquad (33)$$

By Assumption 18, we have

$$\sup_{\overline{g}:0\leq g\leq\overline{g}} E\left[\left(\widetilde{Y}_{g,\overline{g}}E(\widetilde{Y}_{g,\overline{g}})\right)^2\right] \leq \left(\sup_{\overline{g}:0\leq g\leq\overline{g}}|E(\widetilde{Y}_{g,\overline{g}})|\right)^2 \sup_{\overline{g}:0\leq g\leq\overline{g}} E\left[\widetilde{Y}_{g,\overline{g}}^2\right] < +\infty$$

Then, by the weak law of large numbers for triangular arrays (see, e.g. Gut, 1992),

$$\frac{2}{\overline{g}+1}\sum_{g=0}^{\overline{g}}(\widetilde{Y}_{g,\overline{g}} - E(\widetilde{Y}_{g,\overline{g}}))E(\widetilde{Y}_{g,\overline{g}}) \xrightarrow{\mathbb{P}} 0.$$

By the same law of large numbers, and since $\sup_{\overline{g}:0\leq g\leq\overline{g}}|u_{g,\overline{g}}| < +\infty$, by Assumption 18, $2\sum_{g=0}^{\overline{g}}(\widetilde{Y}_{g,\overline{g}} - E(\widetilde{Y}_{g,\overline{g}}))u_{g,\overline{g}}/(\overline{g}+1) \xrightarrow{\mathbb{P}} 0$ Then, by definition of $T_{2\overline{g}}$ and because $\widehat{\beta}_{fe} = O_p(1)$, $T_{2\overline{g}} \xrightarrow{\mathbb{P}} 0$.

Next, using (30) and Assumption 18,

$$\sup_{\overline{g}:0\leq g\leq\overline{g}} E\left|\widetilde{Y}_{g,\overline{g}} - E(\widetilde{Y}_{g,\overline{g}})\right|^{2+\delta} < +\infty.$$

Again, this implies by the weak law of large numbers for triangular arrays that $T_{1\overline{g}} - \widetilde{V}_{\overline{g}} \xrightarrow{\mathbb{P}} 0$. Moreover, by Assumption 18, $\widetilde{V}_{\overline{g}} \to H$. Next, $J_{\overline{g}}$ is equal to $\sum_{g,t} N_{g,t}D_{g,t}\varepsilon_{g,t}/(\overline{g}+1)$ and thus converges to J. Combining all the previous results, we obtain

$$\frac{\sqrt{T_{1g}+T_{2g}}}{J_g\sigma(\boldsymbol{w})} \xrightarrow{\mathbb{P}} \frac{H^{1/2}}{J\sigma_\infty(\boldsymbol{w})}.$$

Thus, by the first result of the theorem and Slutsky's theorem,

$$\lim_{\overline{g}\to\infty}\Pr\left(\left|\frac{\widehat{\underline{\sigma}}_{fe} - \underline{\sigma}_{fe}}{\frac{\sqrt{T_{1g}+T_{2g}}}{J_g\sigma(\boldsymbol{w})}}\right| \leq z_{1-\alpha/2}\right) = 1-\alpha.$$

The result follows by (33).

Proof of Theorem 8

Some algebra and the definition of $\widehat{W}_{TC}$ show that

$$\widehat{W}_{TC} - \tilde{\Delta}^S = \frac{1}{\overline{g}+1} \sum_{g=0}^{\overline{g}} (Z_{g,\overline{g}} - E(Z_{g,\overline{g}})). \tag{34}$$

Using the same reasoning as in Theorem 7 and Assumption 19, we obtain

$$\frac{\sum_{g=0}^{\overline{g}} (Z_{g,\overline{g}} - E(Z_{g,\overline{g}}))/\sqrt{\overline{g}+1}}{V_{\overline{g}}^{1/2}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $V_{\overline{g}} = (1/(\overline{g}+1)) \sum_{g=0}^{\overline{g}} V(Z_{g,\overline{g}})$. The first result then follows by (34) and Assumption 19.

To prove the second part, let us first define

$$\sigma_{\overline{g}}^2 = \frac{1}{\overline{g}+1} \sum_{g=0}^{\overline{g}} E(Z_{g,\overline{g}}^2) - E(\overline{Z}_{\overline{g}})^2.$$

By the same reasoning as in Theorem 7, $E\left[|Z_{g,\overline{g}}^2 - E(Z_{g,\overline{g}}^2)|^{1+\delta/2}\right] \leq 2^{1+\delta/2} E\left[|Z_{g,\overline{g}}|^{2+\delta}\right]$. Hence, by Assumption 19 and the weak law of large numbers for triangular arrays,

$$\frac{1}{\overline{g}+1} \sum_{g=0}^{\overline{g}} (Z_{g,\overline{g}}^2 - E(Z_{g,\overline{g}}^2)) \xrightarrow{\mathbb{P}} 0. \tag{35}$$

Reasoning similarly on $Z_{g,\overline{g}}$ instead of $Z_{g,\overline{g}}^2$, we also get that $\overline{Z}_{\overline{g}} - E(\overline{Z}_{\overline{g}}) \to 0$ in probability. Furthermore, by Assumption 19 again, $|E(\overline{Z}_{\overline{g}})| \leq \sup_g E(|Z_{g,\overline{g}}|) < +\infty$. This and $\overline{Z}_{\overline{g}} - E(\overline{Z}_{\overline{g}}) \to 0$ imply that $\overline{Z}_{\overline{g}} + E(\overline{Z}_{\overline{g}}) = O_p(1)$. Hence, $\overline{Z}_{\overline{g}}^2 - E(\overline{Z}_{\overline{g}})^2 = o_p(1)$. Together with (35), this implies

$$\widehat{\sigma}^2 - \sigma_{\overline{g}}^2 \xrightarrow{\mathbb{P}} 0. \tag{36}$$

Moreover, by convexity,

$$\left(\frac{1}{\overline{g}+1} \sum_{g=0}^{\overline{g}} E(Z_{g,\overline{g}}) \right)^2 \leq \frac{1}{\overline{g}+1} \sum_{g=0}^{\overline{g}} E(Z_{g,\overline{g}})^2,$$

and thus $\sigma_{\overline{g}}^2 \geq V_{\overline{g}}$. Hence, $\widehat{\sigma}^2 \geq V_{\overline{g}} \left(1 + (\widehat{\sigma}^2 - \sigma_{\overline{g}}^2)/\sigma_{\overline{g}}^2\right)$, which implies that

$$\left| \frac{\widehat{W}_{TC} - \Delta^S}{\widehat{\sigma}} \right| \leq \left| \frac{\sum_{g=0}^{\overline{g}} (Z_{g,\overline{g}} - E(Z_{g,\overline{g}}))/\sqrt{\overline{g}+1}}{\sqrt{V_{\overline{g}} \left(1 + \frac{\widehat{\sigma}^2 - \sigma_{\overline{g}}^2}{\sigma_{\overline{g}}^2}\right)}} \right|. \tag{37}$$

Because $V_{\overline{g}} \to \sigma^2 > 0$, $\sigma^2_{\overline{g}} \geq V_{\overline{g}}$ implies that $\liminf \sigma^2_{\overline{g}} > 0$. Combined with (36), we obtain $(\widehat{\sigma}^2 - \sigma^2_{\overline{g}})/\sigma^2_{\overline{g}} = o_p(1)$. Hence, by Sluski's lemma, the right-hand side of (37) tends to a standard normal variable. As a result,

$$\limsup_{\overline{g}} \Pr\left(\Delta^S \in \mathrm{CI}_{1-\alpha}\right) \geq \limsup_{\overline{g}} \Pr\left(\left|\frac{\widehat{W}_{TC} - \Delta^S}{\sqrt{V_{\overline{g}}\left(1 + \frac{\widehat{\sigma}^2 - \sigma^2_{\overline{g}}}{\sigma^2_{\overline{g}}}\right)}}\right| \leq z_{1-\alpha/2}\right) = 1 - \alpha.$$