Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey*

Clément de Chaisemartin† Xavier D'Haultfœuille‡

Abstract

Linear regressions with period and group fixed effects are widely used to estimate policies' effects: 26 of the 100 most cited papers published by the American Economic Review from 2015 to 2019 estimate such regressions. It has recently been show that those regressions may produce misleading estimates, if the policy's effect is heterogeneous between groups or over time, as is often the case. This survey reviews a fast-growing literature that documents this issue, and that proposes alternative estimators robust to heterogeneous effects.

Keywords: two-way fixed effects regressions, differences-in-differences, parallel trends, heterogeneous treatment effects, panel data, repeated-cross section data, policy evaluation.

JEL Codes: C21, C23

1 Introduction

A popular method to estimate the effect of a policy, or 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 regressing $Y_{g,t}$, the outcome in group g and at period t, on group fixed effects, period fixed effects, and $D_{g,t}$, the treatment of group g at period t. For instance, to measure the effect of the minimum wage on employment in the US, researchers have often regressed employment in county g and year t on county fixed effects, year fixed effects, and the minimum wage in county g and year t.

Such two-way fixed effects (TWFE) regressions are probably the most-commonly used technique in economics to measure the effect of a treatment on an outcome. de Chaisemartin and

*We are very grateful to Bruno Ferman, Jonathan Roth, and Kaspar Wüthrich for their helpful comments.

†Economics Department, Sciences Po, clement.dechaisemartin@sciencespo.fr

‡CREST-ENSAE, xavier.dhaultfoeuille@ensae.fr.

D'Haultfœuille (2021*a*) conducted a survey of the 20 papers with the most Google Scholar citations published by the American Economic Review in 2015, and of the similarly selected papers in 2016, 2017, 2018, and 2019. Of those 100 papers, 26 have estimated at least one TWFE regression to estimate the effect of a treatment on an outcome. TWFE regressions are also very commonly used in political science, sociology, and environmental sciences.

Researchers have long thought that TWFE estimators are equivalent to differences-in-differences (DID) estimators. With two groups and two periods, a DID estimator compares the outcome evolution from period 1 to 2 between a treatment group s that switches from untreated to treated, and a control group n that is untreated at both dates:

$$\text{DID} = Y_{s,2} - Y_{s,1} - (Y_{n,2} - Y_{n,1}) . \tag{1}$$

DID relies on a parallel trends assumption: in the absence of the treatment, both groups would have experienced the same outcome evolution. Specifically, for every $g \in \{s, n\}$ and $t \in \{1, 2\}$, let $Y_{g,t}(0)$ and $Y_{g,t}(1)$ denote the potential outcomes in group g at period t without and with the treatment, respectively.[1] Parallel trends requires that the expected evolution of the untreated outcome be the same in both groups:

$$E\left[Y_{s,2}(0) - Y_{s,1}(0)\right] = E\left[Y_{n,2}(0) - Y_{n,1}(0)\right] .$$

Under that assumption, DID is unbiased for the average treatment effect (ATE) in group s at period 2 (see, e.g., Abadie (2005)):

$$\begin{aligned}
E\left[\text{DID}\right] =& E\left[Y_{s,2} - Y_{s,1} - (Y_{n,2} - Y_{n,1})\right] \\
=& E\left[Y_{s,2}(1) - Y_{s,1}(0) - (Y_{n,2}(0) - Y_{n,1}(0))\right] \\
=& E\left[Y_{s,2}(1) - Y_{s,2}(0)\right] + E\left[Y_{s,2}(0) - Y_{s,1}(0)\right] - E\left[Y_{n,2}(0) - Y_{n,1}(0)\right] \\
=& E\left[Y_{s,2}(1) - Y_{s,2}(0)\right] ,
\end{aligned}$$

where the last equality follows from the parallel trends assumption. Parallel trends is partly testable, by comparing the outcome trends of groups s and n, before group s received the treatment. In practice, such pre-trends tests sometimes fail, but other times they indicate that the two groups were indeed on parallel paths before s got treated.[2]

Recent research has shown that unlike DID estimators, TWFE estimators are unbiased for an ATE if parallel trends holds, and if another assumption is satisfied: the treatment effect should

[1] Implicitly, this notation rules out dynamic treatment effects, and assumes that groups' potential outcomes only depend on their current treatment, not on their past treatments.

[2] Pre-trends tests come with caveats unveiled by a recent literature that is beyond the scope of this survey, see Kahn-Lang and Lang (2020), Bilinski and Hatfield (2018), and Roth (2019). Similarly, papers that have proposed relaxations of the parallel trends assumption (see, e.g., Rambachan and Roth, 2019; Freyaldenhoven et al., 2019) are also beyond the scope of this survey.

be constant, between groups and over time. Unlike parallel trends, this assumption is unlikely to hold, even approximately, in most of the applications where TWFE regressions have been used. For instance, the effect of the minimum wage on employment is likely to differ in counties with highly educated workers, and in counties with less educated workers.

The realization that one of the most commonly used empirical method in social science relies on an often-implausible assumption has spurred a flurry of methodological papers diagnosing the seriousness of the issue, and proposing alternative estimators. This review aims to provide an overview of this recent literature, which has developed in such a quick and dynamic manner that some practitioners may have gotten lost in the whirlwind of new working papers. We start by giving an overview of the papers that have identified TWFE's regressions lack of robustness to heterogeneous treatment effects, and that have proposed diagnostic tools practitioners may use to assess the seriousness of this issue in their own application. We then give an overview of the papers that have proposed alternative estimators robust to heterogeneous treatment effects. When available, the Stata and R commands implementing the diagnostics tools and alternative estimators discussed in this review are referenced, and the basic syntax of the Stata command is provided. We refer the reader to the commands' help files for further details on their syntax.

2 TWFE regressions with heterogeneous treatment effects

2.1 TWFE regressions may not identify a convex combination of treatment effects

We consider observations that can be divided into G groups and T periods, respectively indexed by the placeholders g and t, which can refer to any group or time period. The data may be an individual-level panel, or repeated cross-sections where groups are, say, individuals' county of birth. The data could also be a cross-section where cohort of birth plays the role of time.

Let $\widehat{\beta}_{fe}$ denote the coefficient of $D_{g,t}$, the treatment in group g at period t, in an OLS regression of $Y_{i,g,t}$, the outcome of individual i in group g at period t, on group fixed effects, period fixed effects, and $D_{g,t}$:

$$Y_{i,g,t} = \alpha_g + \gamma_t + \widehat{\beta}_{fe} D_{g,t} + \epsilon_{g,t}.$$

de Chaisemartin and D'Haultfœuille (2020) show that under a parallel trends assumption on the potential outcome without treatment $Y_{g,t}(0)$,

$$E\left[\widehat{\beta}_{fe}\right] = E\left[\sum_{(g,t):D_{g,t}\neq 0} W_{g,t} TE_{g,t}\right]. \tag{2}$$

If the treatment is binary, $TE_{g,t} = Y_{g,t}(1) - Y_{g,t}(0)$, the ATE in group g at time t. If the treatment is discrete or continuous, $TE_{g,t} = (Y_{g,t}(D_{g,t}) - Y_{g,t}(0))/D_{g,t}$, the effect of moving the treatment

from 0 to $D_{g,t}$ scaled by $D_{g,t}$.[3] The $W_{g,t}$ are weights summing to 1, that are proportional to and of the same sign as

$$N_{g,t}(D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot}), \tag{3}$$

where $N_{g,t}$ is the number of observations in cell (g,t), $D_{g,\cdot}$ is the average treatment of group g across periods, $D_{\cdot,t}$ is the average treatment at period t across groups, and $D_{\cdot,\cdot}$ is the average treatment across groups and periods.[4]

Equations (2) and (3) have two important consequences. First, they imply that $\widehat{\beta}_{fe}$ may be biased for the average treatment effect across all treated (g,t) cells, the ATT. Indeed, Equation (3) shows that $W_{g,t}$ is not proportional to $N_{g,t}$, so $E\left[\widehat{\beta}_{fe}\right]$ does not assign to $TE_{g,t}$ a weight proportional to the population of cell (g,t). A special case where $\widehat{\beta}_{fe}$ is unbiased for the ATT under the parallel trends assumption alone is when i) the treatment is binary; ii) the design is staggered, meaning that groups can switch in but not out of the treatment; iii) there is no variation in treatment timing: all treated groups start receiving the treatment at the same date. Then, one can show that $D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot}$ is constant across the treated (g,t)s, so it follows from Equation (3) that $W_{g,t}$ is proportional to $N_{g,t}$, and Equation (2) then implies that $\widehat{\beta}_{fe}$ is unbiased for the ATT. However, conditions i)-iii) are seldom met in practice. $\widehat{\beta}_{fe}$ can also be unbiased for the ATT if one is ready to make more assumptions than just parallel trends. For instance, if one is also ready to assume that $D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot}$ is uncorrelated with $TE_{g,t}$, the treatment effects that are up- and down-weighted by $\widehat{\beta}_{fe}$ do not systematically differ, and one can then show that $\widehat{\beta}_{fe}$ is unbiased for the ATT (see Corollary 2 in de Chaisemartin and D'Haultfœuille, 2020).[5] Unfortunately, this no-correlation condition is often implausible. To see this, note that $D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot}$ is decreasing in $D_{g,\cdot}$, meaning that $\widehat{\beta}_{fe}$ downweights the treatment effect of groups with the highest average treatment. However, groups with the largest and lowest average treatment may have systematically different treatment effects. Similarly, $D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot}$ is decreasing in $D_{\cdot,t}$, and the treatment effects at time periods with the highest average treatment may also systematically differ from the treatment effects at time periods where the average treatment is lower.

Second, and perhaps more worryingly, Equation (3) implies that some of the weights $W_{g,t}$ may be negative. This means that in the minimum wage example, $\widehat{\beta}_{fe}$ could be estimating something

[3] de Chaisemartin and D'Haultfœuille (2020) derive Equation (2) assuming that groups' potential outcomes only depend on their current treatment, not on their past treatments. With dynamic effects, Equation (2) still holds if the treatment is binary and staggered, except that some of the $TE_{g,t}$s become effects of having been treated for more than one period.

[4] Equation (3) holds under the assumption that the number of observations follows the same evolution over time in every group.

[5] A special case of this "no-correlation" condition is if the treatment effect is constant, i.e. $TE_{g,t} = \delta$ for all (g,t). Then, it directly follows from Equation (2) that $E\left[\widehat{\beta}_{fe}\right] = \delta$. However, constant effect is most often an implausible assumption.

like 3 times the effect of the minimum wage on employment in Santa Clara county, minus 2 times the effect in Wayne county. Then, if raising the minimum wage by one dollar decreases employment by 5% in Santa Clara county and by 20% in Wayne county, one would have $E\left[\widehat{\beta}_{fe}\right] = 3 \times -0.05 - (2 \times -0.2) = 0.25$. $E\left[\widehat{\beta}_{fe}\right]$ would be positive, while the minimum wage's effect on employment is negative both in Santa Clara and in Wayne county. This example shows that $\widehat{\beta}_{fe}$ may not satisfy the "no-sign reversal property": $E\left[\widehat{\beta}_{fe}\right]$ could for instance be positive, even if the treatment effect is strictly negative in every (g,t). This phenomenon can only arise when some of the weights $W_{g,t}$ are negative: when all those weights are positive, $\widehat{\beta}_{fe}$ does satisfy the no-sign reversal property. Note that despite its intuitive appeal and its popularity among applied researchers, the no-sign reversal property is not grounded in statistical decision theory, unlike other commonly-used criteria to discriminate estimators such as the mean-squared error. Still, it is connected to the economic concept of Pareto efficiency. If an estimator satisfies "no-sign-reversal", the estimand attached to it can only be positive if the treatment is not Pareto-dominated by the absence of treatment, meaning that not everybody is hurt by the treatment. Conversely, the estimand can only be negative if the treatment does not Pareto-dominate the absence of treatment. On the other hand, if an estimator does not satisfy "no-sign-reversal", the estimand attached to it could for instance be positive, even if the treatment is Pareto-dominated by the absence of treatment.

Inasmuch as "no-sign-reversal" is a desirable property, it becomes interesting to understand when $\widehat{\beta}_{fe}$ may satisfy it. Equation (3) shows that with a binary treatment, the weights attached to $\widehat{\beta}_{fe}$ could all be positive. With a binary treatment, all the (g,t)s entering the summation in (2) must have $D_{g,t} = 1$, so for a weight $W_{g,t}$ to be strictly negative, one must have $1 + D_{.,.} < D_{g,.} + D_{.,t}$. This cannot happen if $D_{g,.} + D_{.,t} \leq 1$ for every (g,t). Accordingly, all the weights are likely to be positive when there is no group that is treated most of the time, and no time periods where most groups are treated. In staggered designs, this has led Jakiela (2021) to propose to drop the last periods of the data, those when $D_{.,t}$ is the highest, to mitigate or eliminate the negative weights. One could also drop the always-treated groups, if there are any.

On the other hand, Equation (3) shows that with a non-binary treatment, it becomes more likely that some of the weights $W_{g,t}$ are negative. Gentzkow et al. (2011) study the effect of the number of newspapers in county g and year t on turnout in presidential elections. Assume that in year t, county g has 1 newspaper ($D_{g,t} = 1$), which is below its average number of newspapers across years, equal, say, to 2 ($D_{g,.} = 2$). At the same time, the average number of newspapers across counties in year t is equal to 2 ($D_{.,t} = 2$), which is above the average number of newspapers across all counties and years, equal, say, to 1 ($D_{.,.} = 1$). Then, it follows from (3) that the weight assigned to the effect of newspapers in county g and year t is strictly negative. More generally, a necessary condition to have that all weights are positive is that in every period where the population's treatment is higher than its average across periods ($D_{.,t} \geq D_{.,.}$), the treatment of

each treated group must also be larger than its average across periods ($D_{g,t} \geq D_{g,\cdot}$ for all gs such that $D_{g,t} \neq 0$). In practice, it seems that with a non-binary treatment, $\widehat{\beta}_{fe}$ often has many large negative weights attached to it. de Chaisemartin and D'Haultfœuille (2020) review Gentzkow et al. (2011) and Enikolopov et al. (2011), two papers that have estimated TWFE regressions with a non-binary treatment. In both cases, more than half of the weights attached to $\widehat{\beta}_{fe}$ are negative, and the sum of the negative weights is large (-0.53 and -2.26, respectively).[6]

The `twowayfeweights` Stata (see de Chaisemartin, D'Haultfœuille and Deeb, 2019) and R (see Zhang and de Chaisemartin, 2021) commands compute the weights $W_{g,t}$ in (2). The basic syntax of the Stata command is:

```
twowayfeweights outcome groupid timeid treatment, type(feTR)
```

The decomposition in (2) is the main result in de Chaisemartin and D'Haultfœuille (2020). Related results have appeared earlier in Theorems S1 and S2 of the Supplementary Material of de Chaisemartin and D'Haultfœuille (2015). Borusyak and Jaravel (2017) consider the case with a binary and staggered treatment. In their Lemma 1 and Proposition 1, they assume that the treatment effect varies with the duration elapsed since one has started receiving the treatment but does not vary across groups and over time. Then, they show that $\widehat{\beta}_{fe}$ estimates a weighted sum of effects, that may assign negative weights to long-run treatment effects. Their Appendix C also contains a result related to that in Equation (2).[7]

2.2 The origin of the problem: forbidden comparisons

2.2.1 Forbidden comparisons when the treatment is binary and the design is staggered

Goodman-Bacon (2021) shows that when the treatment is binary and the design is staggered, meaning that groups can switch in but not out of treatment, we have

$$\widehat{\beta}_{fe} = \sum_{g \neq g', t<t'} v_{g,g',t,t'} \text{DID}_{g,g',t,t'}, \tag{4}$$

where $DID_{g,g',t,t'}$ is a DID comparing the outcome evolution of two groups g and g' from a pre period t to a post period t', and where $v_{g,g',t,t'}$ are non-negative weights summing to one, with $v_{g,g',t,t'} > 0$ if and only if g switches treatment between t and t' while g' does not.[8] Some of

[6] Gentzkow et al. (2011) do not estimate the TWFE regression defined above, but a first-difference version of that regression. Many weights attached to their regression are negative, and the sum of negative weights is even larger than for the TWFE regression.

[7] Prior to that, Chernozhukov et al. (2013) had shown that one-way FE regressions may be biased for the average treatment effect, though unlike TWFE regressions they always estimate a convex combination of effects.

[8] Goodman-Bacon (2021) actually decomposes $\widehat{\beta}_{fe}$ as a weighted average of DIDs between cohorts of groups becoming treated at the same date, and between periods of time where their treatment remains constant. One can then further decompose his decomposition, as we do here.

the $DID_{g,g',t,t'}$s in Equation (4) compare a group switching treatment from t to t' to a group untreated at both dates, while other $DID_{g,g',t,t'}$s compare a switching group to a group treated at both dates. The negative weights in (2) originate from this second type of DIDs.

To see that, let us consider a simple example, first introduced by Borusyak and Jaravel (2017),[9] with two groups and three periods. Group e, the early-treated group, is untreated at period 1 and treated at periods 2 and 3. Group ℓ, the late-treated group, is untreated at periods 1 and 2 and treated at period 3. In this simple example, Equation (4) reduces to

$$\widehat{\beta}_{fe} = (\text{DID}_{e,\ell,1,2} + \text{DID}_{\ell,e,2,3})/2, \tag{5}$$

with

$$\begin{aligned}\text{DID}_{e,\ell,1,2} &= Y_{e,2} - Y_{e,1} - (Y_{\ell,2} - Y_{\ell,1}),\\ \text{DID}_{\ell,e,2,3} &= Y_{\ell,3} - Y_{\ell,2} - (Y_{e,3} - Y_{e,2}).\end{aligned}$$

$\text{DID}_{e,\ell,1,2}$ compares the period-1-to-2 outcome evolution of group e, that switches from untreated to treated from period 1 to 2, to the outcome evolution of group ℓ that is untreated at both periods. Thus, $\text{DID}_{e,\ell,1,2}$ is similar to the DID estimator in Equation (1), and under parallel trends it is unbiased for the treatment effect in group e at period 2:

$$E\left[\text{DID}_{e,\ell,1,2}\right] = E\left[TE_{e,2}\right]. \tag{6}$$

$\text{DID}_{\ell,e,2,3}$, on the other hand, compares the period-2-to-3 outcome evolution of group ℓ, that switches from untreated to treated from period 2 to 3, to the outcome evolution of group e that is treated at both dates. At both periods, e's outcome is its treated potential outcome, which is equal to the sum of its untreated outcome and its treatment effect. Accordingly,

$$Y_{e,3} - Y_{e,2} = Y_{e,3}(0) + TE_{e,3} - (Y_{e,2}(0) + TE_{e,2}).$$

On the other hand, group ℓ is only treated at period 3, so

$$Y_{\ell,3} - Y_{\ell,2} = Y_{\ell,3}(0) + TE_{\ell,3} - Y_{\ell,2}(0).$$

Taking the expectation of the difference between the two previous equations,

$$E\left[\text{DID}_{\ell,e,2,3}\right] = E\left[TE_{\ell,3} - TE_{e,3} + TE_{e,2}\right], \tag{7}$$

where $E\left[Y_{e,3}(0) - Y_{e,2}(0)\right]$ and $E\left[Y_{\ell,3}(0) - Y_{\ell,2}(0)\right]$ cancel out under the parallel trends assumption. Finally, it follows from Equations (5), (6), and (7) that

$$E\left[\widehat{\beta}_{fe}\right] = E\left[1/2TE_{\ell,3} + TE_{e,2} - 1/2TE_{e,3}\right]. \tag{8}$$

[9] Borusyak and Jaravel (2017) have also coined the "forbidden comparisons" expression that we borrow here.

In this simple example, Equation (2) reduces to (8). The right-hand side of Equation (8) is a weighted sum of three ATEs where one ATE receives a negative weight. As the previous derivation shows, this negative weight comes from the fact $\widehat{\beta}_{fe}$ leverages $\text{DID}_{\ell,e,2,3}$, a DID comparing a group switching from untreated to treated to an always treated group.

If one is ready to assume that the treatment effect does not change over time, $TE_{e,3} = TE_{e,2}$, and (7) simplifies to

$$E\left[\text{DID}_{\ell,e,2,3}\right] = E\left[TE_{\ell,3}\right]. \tag{9}$$

Then, the negative weight in (7) disappears, and $\widehat{\beta}_{fe}$ estimates a weighted average of treatment effects. This extends beyond this simple example: Theorem S2 of the Web Appendix of de Chaisemartin and D'Haultfœuille (2020) and Equation (16) of Goodman-Bacon (2021) show that in staggered adoption designs with a binary treatment, $\widehat{\beta}_{fe}$ estimates a convex combination of effects, if the treatment effect does not change over time but may still vary across groups. This conclusion, however, no longer holds if the treatment is not binary or the design is not staggered. Moreover, assuming constant treatment effects over time is often implausible as this rules out both dynamic treatment effects and calendar time effects.

The decomposition in Equation (4) is key to understand why $\widehat{\beta}_{fe}$ may not identify a convex combination of treatment effects. On the other hand, it cannot be used to assess if $\widehat{\beta}_{fe}$ does indeed estimate a convex combination of effects in a given application. Consider an example similar to that above, but with a third group n that remains untreated from period 1 to 3. In this second example, the decomposition in (4) now indicates that $\widehat{\beta}_{fe}$ assigns a weight equal to 1/6 to DIDs comparing switchers to always treated. On the other hand, all the weights in (2) are positive in this second example. This phenomenon can also arise in real data sets. In the data of Stevenson and Wolfers (2006) used by Goodman-Bacon (2021) in his empirical application, if one restricts the sample to states that are not always treated and to the first ten years of the panel, all the weights in (2) are positive, but the sum of the weights in (4) on DIDs comparing switchers to always treated is equal to 0.06. Beyond these examples, one can show that having DIDs comparing switchers to always treated in (4) is necessary but not sufficient to have negative weights in (2). Similarly, the sum of the weights on DIDs comparing switchers to always treated in (4) is always larger than the absolute value of the sum of the negative weights in (2). The reason why Equation (4) "overestimates" the negative weights in (2) is that as soon as there are three distinct treatment dates, there is not a unique way of decomposing $\widehat{\beta}_{fe}$ as a weighted average of DIDs, and there exists other decompositions than Equation (4) putting less weight on DIDs using always treated as controls.[10]

[10]To see that, let $t_0 < t_1 < t_2$ be three dates, let e be an early-treated group becoming treated at t_1, let ℓ be a late-treated group becoming treated at t_2, and let n be a group untreated yet at t_2. Let $\underline{v} =$

The `bacondecomp` Stata (see Goodman-Bacon et al., 2019) and R (see Flack and Edward, 2020) commands compute the $\text{DID}_{g,g',t,t'}$s entering in (4), the weights assigned to them, as well as the sum of the weights on $\text{DID}_{g,g',t,t'}$s using an always treated as control. The basic syntax of the `bacondecomp` Stata command is:

```
bacondecomp outcome treatment, ddetail
```

2.2.2 More forbidden comparisons when the design is not staggered or treatment is not binary

When the treatment does not follow a staggered design, or when it is not binary, $\widehat{\beta}_{fe}$ may leverage another type of comparison: it may compare the outcome evolution of a group m whose treatment increases more to the outcome evolution of a group ℓ whose treatment increases less. Such comparisons are also not robust to heterogeneous effects, as shown by de Chaisemartin and D'Haultfœuille (2018). For instance, from period 1 to 2 group m may be going from 0 to 2 units of treatment while group ℓ goes from 0 to 1 unit. To simplify, assume that in both groups, potential outcomes are linear in the number of treatment units:

$$\begin{aligned} Y_{m,t}(d) =& Y_{m,t}(0) + \delta_m d \\ Y_{\ell,t}(d) =& Y_{m,t}(0) + \delta_\ell d, \end{aligned}$$

with strictly positive slopes in both groups but a slope thrice as large in group ℓ: $\delta_\ell = 3\delta_m > 0$. Then, under parallel trends,

$$\begin{aligned} & E\left[Y_{m,2} - Y_{m,1} - (Y_{\ell,2} - Y_{\ell,1})\right] \\ =& E\left[Y_{m,2}(2) - Y_{m,1}(0) - (Y_{\ell,2}(1) - Y_{\ell,1}(0))\right] \\ =& E\left[Y_{m,2}(0) + 2\delta_m - Y_{m,1}(0) - (Y_{\ell,2}(0) + \delta_\ell - Y_{\ell,1}(0))\right] \\ =& E\left[Y_{m,2}(0) - Y_{m,1}(0)\right] - E\left[Y_{\ell,2}(0) - Y_{\ell,1}(0)\right] + 2\delta_m - \delta_\ell \\ =& -\delta_m < 0. \end{aligned}$$

$\min(v_{\ell,e,t_1,t_2}, v_{e,n,t_0,t_2}) > 0$. One has

$$\text{DID}_{\ell,e,t_1,t_2} = \text{DID}_{\ell,n,t_0,t_2} - \text{DID}_{e,n,t_0,t_2} + \text{DID}_{e,\ell,t_0,t_1}. \tag{10}$$

Then, it follows from Equation (10) that

$$\begin{aligned} & v_{\ell,e,t_1,t_2}\text{DID}_{\ell,e,t_1,t_2} + v_{e,n,t_0,t_2}\text{DID}_{e,n,t_0,t_2} \\ =& (v_{\ell,e,t_1,t_2} - \underline{v})\text{DID}_{\ell,e,t_1,t_2} + \underline{v}\text{DID}_{\ell,n,t_0,t_2} + \underline{v}\text{DID}_{e,\ell,t_0,t_1} + (v_{e,n,t_0,t_2} - \underline{v})\text{DID}_{e,n,t_0,t_2}. \end{aligned} \tag{11}$$

Plugging Equation (11) into Equation (4) will yield a different decomposition of $\widehat{\beta}_{fe}$ as a weighted average of DIDs. But the weight on DIDs using always-treated as controls is equal to v_{ℓ,e,t_1,t_2} in the left-hand-side of Equation (11), and to $(v_{\ell,e,t_1,t_2} - \underline{v})$ in its right-hand side. Accordingly, this new decomposition puts strictly less weight than Equation (4) on DIDs using always-treated as controls.

This DID's expectation is strictly negative, despite the fact that both groups have a strictly positive treatment effect. Intuitively, group ℓ's treatment increases half as much as group m's, but its treatment effect is three times larger, so its outcome ends up increasing more.

Self-selection into treatment would rather suggest that group m, who receives more treatment units, benefits more from the treatment. If $\delta_m = 2\delta_\ell > 0$,

$$E\left[Y_{m,2} - Y_{m,1} - (Y_{\ell,2} - Y_{\ell,1})\right] = 3\delta_\ell.$$

In that case, the DID comparing m to ℓ would estimate an effect that is larger than the maximum of the two effects, and twice as large as the average causal response, which is equal to $3\delta_\ell/2$.

2.3 Decomposition results for other TWFE regression coefficients

2.3.1 Extensions of the decomposition in Equation (2)

A decomposition similar to (2) can be obtained for TWFE regressions with control variables, and for first-difference regressions where the outcome's first difference is regressed on the treatment's first difference and period fixed effects. de Chaisemartin and D'Haultfœuille (2020) also derive decompositions similar to (2), for $\widehat{\beta}_{fe}$ and for the first-difference coefficient, under common trends and under the assumption that the treatment effect does not change over time. The weights in all those decompositions are also computed by the `twowayfeweights` Stata and R commands.

2.3.2 Dynamic TWFE regressions

In staggered designs with a binary treatment, Sun and Abraham (2021) study dynamic TWFE regressions, also called event-study regressions:

$$Y_{g,t} = \gamma_g + \lambda_t + \sum_{\ell=-K, \ell\neq -1}^{L} \widehat{\beta}_\ell 1\{F_g = t - \ell\} + \varepsilon_{g,t}, \tag{12}$$

where F_g is the first period at which group g is treated. In words, the outcome is regressed on group and period fixed effects, and relative-time indicators $1\{F_g = t - \ell\}$ equal to 1 if group g started receiving the treatment ℓ periods ago. For $\ell \geq 0$, $\widehat{\beta}_\ell$ is supposed to estimate the cumulative effect of $\ell + 1$ treatment periods. For $\ell \leq -2$, $\widehat{\beta}_\ell$ is supposed to be a placebo coefficient testing the parallel trends assumption, by comparing the outcome trends of groups that will and will not start receiving the treatment in $|\ell|$ periods.

Actually, Sun and Abraham (2021) show that under parallel trends, for $\ell \geq 0$,

$$E\left[\widehat{\beta}_\ell\right] = E\left[\sum_g w_{g,\ell} TE_g(\ell) + \sum_{\ell' \neq \ell} \sum_g w_{g,\ell'} TE_g(\ell')\right], \tag{13}$$

where $TE_g(\ell)$ is the cumulative effect of $\ell+1$ treatment periods in group g, and $w_{g,\ell}$ and $w_{g,\ell'}$ are weights such that $\sum_g w_{g,\ell} = 1$ and $\sum_g w_{g,\ell'} = 0$ for every ℓ'.[11] The first summation in the right-hand side of Equation (13) is a weighted sum across groups of the cumulative effect of $\ell+1$ treatment periods, with weights summing to 1 but that may be negative. This first summation resembles that in the decomposition of the "static" TWFE coefficient in (2), and it implies that $\widehat{\beta}_\ell$ may be biased if the cumulative effect of $\ell+1$ treatment periods varies across groups. The second summation is a weighted sum, across $\ell' \neq \ell$ and groups, of the cumulative effect of $\ell'+1$ treatment periods in group g, with weights summing to 0. This second summation was not present in the decomposition of the static TWFE coefficient. Importantly, its presence implies that $\widehat{\beta}_\ell$, which is supposed to estimate the cumulative effect of $\ell+1$ treatment periods, may in fact be contaminated by the effects of $\ell'+1$ treatment periods. As $\sum_g w_{g,\ell'} = 0$ for every ℓ', this second summation disappears if $TE_g(\ell')$ does not vary across groups, but it is often implausible that the treatment effect does not vary across groups.

For $\ell \leq -2$, and without assuming parallel trends, Sun and Abraham (2021) show that $\widehat{\beta}_\ell$ estimates the sum of two terms. As intended, the first term measures deviations from parallel trends between groups that will and will not start receiving the treatment in $|\ell|$ periods. But the second term is similar to the second summation in the right-hand side of Equation (13): a weighted sum, across $\ell' \geq 0$ and groups, of the cumulative effect of $\ell'+1$ treatment periods in group g, with weights summing to zero. Due to the presence of this second term, the expectation of $\widehat{\beta}_\ell$ may differ from zero even if parallel trends holds, and it may be equal to zero even if parallel trends fails. Thus, an important consequence of the results in Sun and Abraham (2021) is that in the presence of heterogeneous treatment effects, $\widehat{\beta}_\ell$ cannot be used to test for parallel trends.

The `eventstudyweights` Stata command (see Sun, 2020) computes the weights attached to event-study regressions. Its basic syntax is:

```
eventstudyweights {rel_time_list}, absorb(i.groupid i.timeid)
cohort(first_treatment) rel_time(ry),
```

where `rel_time_list` is the list of relative-time indicators $1\{F_g = t - \ell\}$ included in (13), `first_treatment` is a variable equal to the period when group g got treated for the first time, and `ry` is a variable equal to `timeid` minus `first_treatment`, the number of periods elapsed since group g started receiving the treatment.

[11] Equation (13) follows from Proposition 3 in Sun and Abraham (2021), assuming no binning and that the treatment does not have an effect after $L+1$ periods of exposure. A slight difference is that the decomposition in Sun and Abraham (2021) gathers groups that started receiving the treatment at the same period into cohorts. Their decomposition can then be further decomposed, as we do here.

2.3.3 TWFE regressions with more than one treatment

Another case of interest is TWFE regressions with several treatments. For instance, to estimate separately the effect of medical and recreational marijuana laws on consumption, one may regress marijuana consumption in state g and year t on state and year fixed effects, on whether state g has a recreational marijuana law in year t, and on whether state g has a medical law in year t. de Chaisemartin and D'Haultfœuille (2021*b*) show that the coefficient on a given treatment identifies a weighted sum of that treatment's effect across (g,t)s, with weights summing to 1 but that may be negative, plus weighted sums of the effects of the other treatments in the regression, with weights summing to 0. In the example above, the coefficient on recreational laws may be contaminated by the effect of medical laws. The weights attached to TWFE regressions with several treatments are also computed by the `twowayfeweights` Stata and R commands.

3 Alternative estimators robust to heterogeneous treatment effects

In this section, we review several recently-proposed alternatives to TWFE regressions. We restrict our attention to estimators relying on parallel trends assumptions, like TWFE regressions, but that do not restrict treatment effect heterogeneity between groups and over time, unlike TWFE regressions. This excludes papers that have assumed randomized treatment timing (see, e.g., Athey and Imbens, 2022; Roth and Sant'Anna, 2021) or sequential treatment randomization (see, e.g., Bojinov et al., 2021), rather than parallel trends. Intuitively, all the estimators below carefully choose valid control groups, to avoid making the "forbidden comparisons" that render TWFE estimators non-robust to heterogeneous treatment effects.

3.1 Estimators ruling out dynamic effects

We first consider estimators that rule out dynamic effects, i.e. that assume that a group's current outcome only depends on its current treatment, not on its past treatments. Then, with a binary treatment de Chaisemartin and D'Haultfœuille (2020) propose to use the DID_M estimator, a weighted average, across t, of two types of DIDs:

1. a DID comparing the $t-1$ to t outcome evolution of groups going from untreated to treated from $t-1$ to t, the "switchers in", and of groups untreated at both dates.

2. a DID comparing the $t-1$ to t outcome evolution of groups treated at both dates, and of groups going from treated to untreated from $t-1$ to t, the "switchers out".

The first DID compares the outcome evolution of groups switching from untreated to treated, and groups untreated at both dates. It is therefore similar to the DID estimator in Equation

(1), and it is unbiased for the treatment effect of the switching-in groups at period t, under a parallel trends assumption on the untreated outcome $Y_{g,t}(0)$. The second DID compares the outcome evolution of groups treated at both dates, and groups switching from treated to untreated. It is again similar to the DID estimator in Equation (1), switching "treatment" and "non-treatment". Then, one can show that this second DID is unbiased for the treatment effect of the switching-out groups at period t, under a parallel trends assumption on the treated outcome $Y_{g,t}(1)$. Accordingly, DID_M relies on parallel trends assumptions on both $Y_{g,t}(0)$ and $Y_{g,t}(1)$, which implies parallel trends on the treatment effect $Y_{g,t}(1)-Y_{g,t}(0)$. de Chaisemartin and D'Haultfœuille (2020) propose placebo estimators that one can use to test those two assumptions. The placebos compare the outcome trends of switchers and non-switchers, before the switchers switch.

The DID_M estimator can easily be extended to non-binary treatments taking a finite number of values. Then, it is a weighted average, across d and t, of DIDs comparing the $t-1$ to t outcome evolution of groups whose treatment goes from d to some other value from $t-1$ to t, and of groups with a treatment equal to d at both dates, normalized by the intensity of the treatment change experienced by the switchers. For instance, in Gentzkow et al. (2011), a county going from 2 to 4 newspapers is compared to a county with 2 newspapers at both dates.

The DID_M estimator is computed by the `did_multiplegt` Stata (see de Chaisemartin, D'Haultfœuille and Guyonvarch, 2019) and R (see Zhang and de Chaisemartin, 2020) commands. The basic syntax of the Stata command is:

```
did_multiplegt outcome groupid timeid treatment
```

The multi-period DID estimator in Imai and Kim (2021) is related to the DID_M estimator. It can be used with a binary treatment, to estimate the switchers-in's treatment effect.

3.2 Estimators allowing for dynamic effects...

3.2.1 ... When the treatment is binary and the design is staggered.

With dynamic effects, group g's outcome at time t is allowed to depend on her past treatments. Then, researchers have proposed to replace the parallel trends assumption on the potential outcome without treatment $Y_{g,t}(0)$ by a parallel trends assumption on the outcome without having ever been treated $Y_{g,t}(\mathbf{0}_t)$, where $\mathbf{0}_t$ is a vector of t zeros (see Callaway and Sant'Anna, 2021; Sun and Abraham, 2021): for all $g \neq g'$,

$$E\left[Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})\right] = E\left[Y_{g',t}(\mathbf{0}_t) - Y_{g',t-1}(\mathbf{0}_{t-1})\right]. \tag{14}$$

We now review the estimators that have been proposed by Callaway and Sant'Anna (2021), Sun and Abraham (2021), and Borusyak et al. (2021) for binary and staggered treatments, under the parallel trends assumption in Equation (14).

The estimators proposed by Callaway and Sant'Anna (2021)

In a staggered adoption design, groups can be aggregated into cohorts that start receiving the treatment at the same period. Callaway and Sant'Anna (2021) define their parameters of interest as $TE_{c,c+\ell}$, the average treatment effect at period $c+\ell$ of the cohort that started receiving the treatment at period c, for every $c \geq 2$ and $\ell \geq 0$ such that $\ell + c \leq T$. To estimate, say, $TE_{c,c}$, the treatment effect at time c of the cohort that started receiving the treatment at c, they propose to compare c's outcome evolution from $c-1$ to c to the same outcome evolution among all never-treated groups n, assuming for now that such groups exist. This DID estimator is unbiased:

$$
\begin{aligned}
&E\left[Y_{c,c} - Y_{c,c-1} - (Y_{n,c} - Y_{n,c-1})\right] \\
=&E\left[Y_{c,c}(\mathbf{0}_{c-1}, 1) - Y_{c,c-1}(\mathbf{0}_{c-1}) - (Y_{n,c}(\mathbf{0}_c) - Y_{n,c-1}(\mathbf{0}_{c-1}))\right] \\
=&E\left[Y_{c,c}(\mathbf{0}_{c-1}, 1) - Y_{c,c}(\mathbf{0}_c)\right] + E\left[Y_{c,c}(\mathbf{0}_c) - Y_{c,c-1}(\mathbf{0}_{c-1}) - (Y_{n,c}(\mathbf{0}_c) - Y_{n,c-1}(\mathbf{0}_{c-1}))\right] \\
=&E\left[Y_{c,c}(\mathbf{0}_{c-1}, 1) - Y_{c,c}(\mathbf{0}_c)\right],
\end{aligned}
$$

where the last equality follows from Equation (14). More generally, to estimate c's effect at period $c+\ell$, Callaway and Sant'Anna (2021) propose to compare its $c-1$ to $c+\ell$ outcome evolution to that of the never treated groups.

Callaway and Sant'Anna (2021) also propose estimators of more aggregated treatment effects, such as the average effect of having been treated for $\ell+1$ periods across all groups reaching that treatment horizon, or the average treatment effect across all treated units that do not belong to a group treated throughout the panel. When there is no group that is always treated, this latter parameter is in fact the ATT.

Callaway and Sant'Anna (2021) also propose estimators similar to those above, but that use the not-yet-treated instead of the never-treated as controls. This is very useful when there is no never-treated group: in that case, the cohort-and-period specific effects $TE_{c,c+\ell}$ can still be estimated, for every $c \geq 2$ and $\ell \geq 0$ such that $\ell + c \leq U$, where U is the last period when at least one group is still untreated. Even when there are never-treated groups, the not-yet-treated is a larger control group, so this second set of estimators may be more precise than the first. Note that in staggered adoption designs with a binary treatment, the DID_M estimator proposed by de Chaisemartin and D'Haultfœuille (2020) also compares switchers to not-yet-treated, and is unbiased for switchers' instantaneous treatment effect.

Callaway and Sant'Anna (2021) also propose estimators relying on a conditional parallel trends assumption. Rather than assuming that all groups have the same evolution of their never-treated outcome, conditional parallel trends requires that groups with the same value of some time-invariant covariates have the same evolution of their never-treated outcome, an often weaker assumption. In the minimum wage example, if the proportion of college graduates at the start

of the panel is conditioned upon, conditional parallel trends assumes that in the absence of a minimum wage, counties with the same proportion of college graduates would have experienced the same employment evolution, rather than assuming that all counties would have experienced the same employment evolution.

Finally, Callaway and Sant'Anna (2021) propose placebo estimators to test the parallel trends assumption underlying their estimators. Their placebo estimators are robust to heterogeneous effects, unlike the coefficients $\widehat{\beta}_\ell$ for $\ell \leq -2$ from the event-study regression in (12).

The estimators proposed by Callaway and Sant'Anna (2021) are computed by the `csdid` Stata command (see Rios-Avila et al., 2021), and by the `did` R command (see Sant'Anna and Callaway, 2021). The basic syntax of the Stata command is

```
csdid outcome, time(timeid) gvar(cohort)
```

where `cohort` is a variable equal to the period when a group starts receiving the treatment.

The estimators proposed by Sun and Abraham (2021)

Sun and Abraham (2021) also propose DID estimators of the cohort-and-period specific effects $TE_{c,c+\ell}$ that only rely on the parallel trends assumption in Equation (14), and that are robust to heterogeneous treatment effects. Their estimators either use the never-treated groups as controls, or the last-treated groups if there are no never-treated. With the former control group, their estimators of the $TE_{c,c+\ell}$ parameters are identical to those proposed by Callaway and Sant'Anna (2021) with the same control group. Operationally, they show that their estimators can be computed via a simple linear regression, which may reduce computing time. Unlike Callaway and Sant'Anna (2021), they do not propose estimators relying on a conditional parallel trends assumption, and they also do not propose estimators using the not-yet-treated as controls.

Their estimators are computed by the `eventstudyinteract` Stata command (see Sun, 2021). Its basic syntax is

```
eventstudyinteract outcome {rel_time_list}, absorb(i.groupid i.timeid)
cohort(first_treatment) control_cohort(controlgroup)
```

where `rel_time_list` is the list of relative-time indicators $1\{F_g = t - \ell\}$ one would include in the dynamic TWFE regression in (12), `first_treatment` is a variable equal to the period when group g got treated for the first time, and `controlgroup` is an indicator for the control group observations (e.g.: the never treated).

The estimators proposed by Borusyak et al. (2021)

Borusyak et al. (2021) have proposed estimators that may be more efficient than those in Callaway and Sant'Anna (2021) and Sun and Abraham (2021), under some assumptions. Their estimators can be obtained by running a TWFE regression of the outcome on group and time

fixed effects, and fixed effects for every treated (g,t) cell. To be concrete, if the data has 50 groups, 10 time periods, and 100 treated (g,t) cells, the regression has a constant and 158 fixed effects (49 for groups, 9 for time periods, and 100 for the treated (g,t) cells). Under the assumptions of the Gauss-Markov theorem, the coefficients from this regression are the linear estimators of the population coefficients with the lowest variance. But under parallel trends, the population coefficient on the fixed effect for treated cell (g,t) is actually equal to $TE_{g,t}$, the ATE in cell (g,t), so the estimators in Borusyak et al. (2021) are the linear estimators of those ATEs with the lowest variance. With estimators of $TE_{g,t}$ in hand, one can estimate $TE_{c,c+\ell}$ as the average of all the $TE_{g,t}$s such that group g started receiving the treatment at period c and $t = c + \ell$. Again, Gauss-Markov ensures that this estimator is the best linear estimator of $TE_{c,c+\ell}$. As the estimators in Callaway and Sant'Anna (2021) and Sun and Abraham (2021) are also linear estimators, those in Borusyak et al. (2021) have a lower variance. Liu et al. (2021) and Gardner (2021) have independently proposed the same estimators as Borusyak et al. (2021),[12] but the result showing that this estimator is efficient under the assumptions of the Gauss-Markov theorem only appears in Borusyak et al. (2021). Note also that Wooldridge (2021) proposes an estimation strategy connected, and in some cases numerically equivalent, to that of Borusyak et al. (2021).

The estimators proposed by Borusyak et al. (2021) are computed by the `did_imputation` Stata command (see Borusyak, 2021) and by the `didimputation` R command (see Butts, 2021). The basic syntax of the Stata command is:

`did_imputation outcome groupid timeid first_treatment`,

where `first_treatment` is a variable equal to the period when group g first got treated.

Understanding the differences between those estimators

When deciding whether to use the estimator in Borusyak et al. (2021), or that in Callaway and Sant'Anna (2021) or Sun and Abraham (2021), there are a number of things one may want to keep in mind. First, the efficiency result in Borusyak et al. (2021) holds under the assumptions of the Gauss-Markov theorem. Those require, among other things, that the never treated potential outcome $Y_{g,t}(\mathbf{0}_t)$ be independent, both across groups and over time. It is often implausible that the potential outcomes of the same group are uncorrelated over time. With serial correlation, it is no longer guaranteed that the estimators in Borusyak et al. (2021) will always be more efficient than those in Callaway and Sant'Anna (2021) and Sun and Abraham (2021), even though one can probably still expect efficiency gains when serial correlation is moderate.

Perhaps more importantly, the estimators in Borusyak et al. (2021) may be more biased than those in Callaway and Sant'Anna (2021) or Sun and Abraham (2021) when parallel trends does

[12]Before that, Gobillon and Magnac (2016) have proposed a similar strategy to estimate treatment effects under a factor model.

not exactly hold. Borusyak et al. (2021) do not provide a closed-form of their estimators, but one can show that with only one treated group s, which starts to receive the treatment at period t_s, their estimator of that group's effect at $t_s + \ell$ is equal to

$$Y_{s,t_s+\ell} - \frac{1}{t_s - 1} \sum_{k=1}^{t_s-1} Y_{s,k} - \frac{1}{G-1} \sum_{g \neq s} \left(Y_{g,t_s+\ell} - \frac{1}{t_s - 1} \sum_{k=1}^{t_s-1} Y_{g,k} \right).$$

The estimators in Callaway and Sant'Anna (2021) and Sun and Abraham (2021) use groups' $t_s - 1$ outcome, the last period before s gets treated, as the baseline outcome. The estimator in Borusyak et al. (2021) instead uses the average outcome from period 1 to $t_s - 1$ as the baseline, which is why it may be more precise. However, the estimator in Borusyak et al. (2021) may also be more biased if parallel trends does not exactly hold. Roth (2019) shows that with monotonic differential trends, leveraging earlier pre-treatment periods increases the bias of a DID estimator, since one makes comparisons from earlier periods. This suggests that there may be a bias-variance trade-off between the estimators of Borusyak et al. (2021) and those of Callaway and Sant'Anna (2021) and Sun and Abraham (2021) when parallel trends does not exactly hold. Studying this trade-off may be an interesting avenue for future research.

Another difference between these approaches is that Borusyak et al. (2021) impose parallel trends for every group and between every pair of consecutive time periods.[13] Callaway and Sant'Anna (2021), on the other hand, impose a weaker parallel trends assumption: from period c onwards, cohort c must be on the same trend as the never-treated groups, but before that cohort c may have been on a different trend. The assumption in Callaway and Sant'Anna (2021) is the minimal assumption ensuring that all the $TE_{c,c+\ell}$ can be unbiasedly estimated, but it is not testable. We refer the reader to Marcus and Sant'Anna (2021) and Borusyak et al. (2021) for detailed discussions of the differences between parallel trends assumptions. Here, we just want to emphasize that the parallel trends assumption in Callaway and Sant'Anna (2021) is conditional on the design: which groups are required to be on parallel trends at which dates depends on groups' realized treatments. Whether the design should or should not be conditioned upon depends on the type of conclusions one would like to draw. If one would like to draw conclusions specific to the realized treatments (e.g. estimating the ATE across the (g,t) cells that actually got treated), the design should be conditioned upon. If, on the other hand, one would like to draw conclusions that extend beyond the specific values of the realized treatments (e.g. estimating the ATE across the (g,t) cells that could have been treated, weighted by their treatment probabilities), then the design should not be conditioned upon.

Beyond those perhaps slightly abstract considerations, whether the design is conditioned upon or not has an important practical implication. If the design is not conditioned upon, the randomness arising from groups' treatments needs to be accounted for when performing inference. Then,

[13] de Chaisemartin and D'Haultfœuille (2020) and Sun and Abraham (2021) also impose that assumption.

standard errors should be clustered at least at the group level, as recommended by Bertrand et al. (2004). Clustering at the group level accounts for the fact that the treatment is assigned at the (g,t) level, and that groups' treatments exhibit a strong serial correlation: in a staggered design, $D_{g,t} = D_{g,t+1}$, except at the time period where g gets treated. If the design is conditioned upon, standard errors no longer necessarily have to be clustered at the level at which the treatment is assigned. Which standard errors one should use becomes an open question. To answer it, the analyst needs to determine the type of outcome shocks they would like their conclusions to be robust to, while bearing in mind that their conclusions will be conditional to the other shocks that their standard errors do not account for (see, e.g., Deeb and de Chaisemartin, 2019, for a discussion of this issue in the context of randomized controlled trials).

3.2.2 ... When the treatment is not binary or the design is not staggered.

de Chaisemartin and D'Haultfœuille (2021*a*) propose treatment effect estimators robust to heterogeneous and dynamic treatment effects and which can be used even if the treatment is not binary or the design is not staggered. In their survey of 26 highly cited 2015-2019 AER papers using a TWFE regression, they find that only 19% have a binary treatment and a staggered adoption design, so being able to accommodate more general designs is important.

For simplicity, we describe those estimators when the treatment is binary but non-staggered, and refer the reader to de Chaisemartin and D'Haultfœuille (2021*a*) for details as to how the approach below can be adapted to non-binary treatments. When the treatment is binary and non-staggered, groups untreated at $t = 1$ can still be aggregated into cohorts, according to the date at which they first get treated. Then, for every $c \geq 2$ and $\ell \geq 0$ such that $\ell + c \leq T$, de Chaisemartin and D'Haultfœuille (2021*a*) let N_c be the number of groups in cohort c and define

$$FTT_{c,c+\ell} = \frac{1}{N_c} \sum_{g \in c} \left(Y(\mathbf{0}_{c-1}, 1, D_{g,c+1}, ..., D_{g,c+\ell}) - Y(\mathbf{0}_{c+\ell}) \right),$$

the average effect, across all groups in cohort c, of having been treated for the first time ℓ periods ago, relative to having never been treated. Notice that in the non-staggered case, a group treated at c may revert to being untreated after that, and $FTT_{c,c+\ell}$ is defined conditional on groups' treatment trajectories from $c+1$ to $c+\ell$, $(D_{g,c+1}, ..., D_{g,c+\ell})$. de Chaisemartin and D'Haultfœuille (2021*a*) show that under the parallel trends assumption in Equation (14), $FTT_{c,c+\ell}$ can be unbiasedly estimated by a DID comparing the $c-1$ to $c+\ell$ outcome evolution of groups in cohort c and groups untreated from period 1 to $c+\ell$.

In a staggered design, one has

$$FTT_{c,c+\ell} = \frac{1}{N_c} \sum_{g \in c} \left(Y(\mathbf{0}_{c-1}, 1, 1, ..., 1) - Y(\mathbf{0}_{c+\ell}) \right),$$

so $FTT_{c,c+\ell}$ is just the cumulative effect of having been treated for $\ell + 1$ periods. Outside

of staggered designs, $FTT_{c,c+\ell}$ may be harder to interpret, as it may bundle together many different treatment effects. For instance, if some of the groups that get treated for the first time at period c remain treated at $c+1$ while others go back to being untreated at $c+1$, $FTT_{c,c+1}$ aggregates together $E(Y_{g,c+1}(\mathbf{0}_{c-1},1,1)-Y_{g,c+1}(\mathbf{0}_{c+1}))$, the effect of two treatment periods for the first set of groups, and $E(Y_{g,c+1}(\mathbf{0}_{c-1},1,0)-Y_{g,c+1}(\mathbf{0}_{c+1}))$, the effect of having been treated for one period one period ago for the second set of groups. One could try to separately estimate the effects of all the treatment trajectories observed in the data, as Callaway and Sant'Anna (2021) successfully do in the staggered case. But in the staggered case, there can be at most $T+1$ treatment trajectories, while in the non-staggered case there may be up to 2^T of them. Trying to separately estimate the effects of all trajectories may often yield noisy estimates.

Instead, de Chaisemartin and D'Haultfœuille (2021*a*) show that the $FTT_{c,c+\ell}$ parameters can be aggregated into an economically interpretable parameter, namely the cost-benefit ratio a planner would use to conduct a cost-benefit analysis comparing the actual treatments of groups that were initially untreated to the counterfactual *status quo* scenario where they would have remained untreated throughout the panel. In other words, the planner seeks to assess whether the treatment changes that occurred over the panel were beneficial. de Chaisemartin and D'Haultfœuille (2021*a*) show that this cost-benefit ratio is equal to a weighted sum, across c and ℓ, of the $FTT_{c,c+\ell}$ parameters. Accordingly, this cost-benefit ratio can simply be estimated by a weighted sum of DIDs. Importantly, this cost-benefit ratio can also be interpreted as some average of the effect produced by a one-unit increase of the treatment. By symmetry, de Chaisemartin and D'Haultfœuille (2021*a*) show that the cost-benefit ratio comparing the actual treatments of groups that were initially treated to the counterfactual *status quo* scenario where they would have remained treated throughout the panel can also be estimated by a weighted sum of DIDs.

The estimators in de Chaisemartin and D'Haultfœuille (2021*a*) can be used with a binary treatment switching on and off, with a discrete treatment, or with a continuous and staggered treatment (groups start getting treated at different dates, with differing intensities, but once a group gets treated its treatment intensity never changes). They can, of course, also be used with a binary and staggered treatment. Without covariates in the estimation, they are then equivalent to the estimators proposed by Callaway and Sant'Anna (2021) using the not-yet-treated as controls. With covariates, the estimators in Callaway and Sant'Anna (2021) and de Chaisemartin and D'Haultfœuille (2021*a*) differ: Callaway and Sant'Anna (2021) account for covariates non-parametrically, and de Chaisemartin and D'Haultfœuille (2021*a*) do so linearly. Accounting for covariates linearly relies on stronger assumptions, but for instance allows one to have group-specific linear-trends in the estimation.

The estimators proposed by de Chaisemartin and D'Haultfœuille (2021*a*) are computed by the `did_multiplegt` Stata and R commands. To compute those estimators rather than those proposed in de Chaisemartin and D'Haultfœuille (2020), the Stata command's basic syntax is:

```
did_multiplegt outcome groupid timeid treatment, robust_dynamic dynamic(#)
average_effect
```

where `dynamic(#)` specifies the horizon over which effects of a first treatment switch have to be estimated.

4 Conclusion, and avenues for future research

The literature reviewed in this survey has shown that TWFE regressions may not always estimate a convex combination of treatment effects. In such cases, it may be hard to give them a causal interpretation, as TWFE coefficients could for instance be of a different sign than every unit's treatment effect. The literature so far has mostly focused on providing alternative estimators for the case with a binary treatment and staggered adoption. Heterogeneity-robust DID estimators that can be used in more complicated designs are scarce, while many applications where TWFE regressions have been used either do not have a staggered design, or do not have a binary treatment. Developing more estimators that can be used in such designs is a promising avenue for future research. This can often be done by building upon the insights gained from studying the binary-and-staggered case. For instance, the estimators proposed by de Chaisemartin and D'Haultfœuille (2021*a*) build upon those proposed by Callaway and Sant'Anna (2021) for the binary-and-staggered case. Heterogeneity-robust DID estimators are particularly needed to study the effects of continuous treatments. The estimators proposed by de Chaisemartin and D'Haultfœuille (2021*a*) and Callaway et al. (2021) can accommodate continuous treatments, but only if the design is staggered. Many continuous treatments, such as precipitations or trade tariffs, do not follow a staggered design: they can change multiple times and can both go up or down. We hope that the whirlwind of DID working papers shall continue, till heterogeneity-robust DID estimators are as widely applicable as TWFE regressions. It is also important to stress that at this stage, it is still unclear whether researchers should systematically abandon TWFE estimators: those estimators sometimes estimate a convex combination of effects, and they often have a lower variance than the heterogeneity-robust estimators reviewed in the previous section. A comparison of the mean-squared error of TWFE and heterogeneity-robust DID estimators in a broad set of applications is another promising avenue for future research.

References

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

Athey, S. and Imbens, G. W. (2022), 'Design-based analysis in difference-in-differences settings with staggered adoption', *Journal of Econometrics* **226**, 62–79.

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.

Bilinski, A. and Hatfield, L. A. (2018), Nothing to see here? non-inferiority approaches to parallel trends and other model assumptions. arXiv preprint arXiv:1805.03273.

Bojinov, I., Rambachan, A. and Shephard, N. (2021), 'Panel experiments and dynamic causal effects: A finite population perspective', *Quantitative Economics* **12**, 1171–1196.

Borusyak, K. (2021), 'DID_IMPUTATION: Stata module to perform treatment effect estimation and pre-trend testing in event studies'.
URL: *https://ideas.repec.org/c/boc/bocode/s458957.html*

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

Borusyak, K., Jaravel, X. and Spiess, J. (2021), Revisiting event study designs: Robust and efficient estimation. Working Paper.

Butts, K. (2021), 'didimputation: Imputation Estimator from Borusyak, Jaravel, and Spiess (2021) in R'.
URL: *https://cran.r-project.org/web/packages/didimputation/index.html*

Callaway, B., Goodman-Bacon, A. and Sant'Anna, P. H. (2021), Difference-in-differences with a continuous treatment. arXiv preprint arXiv:2107.02637.

Callaway, B. and Sant'Anna, P. H. (2021), 'Difference-in-differences with multiple time periods', *Journal of Econometrics* **225**, 200–230.

Chernozhukov, V., Fernández-Val, I., Hahn, J. and Newey, W. (2013), 'Average and quantile effects in nonseparable panel models', *Econometrica* **81**(2), 535–580.

de Chaisemartin, C. and D'Haultfœuille, 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. and D'Haultfœuille, X. (2020), 'Two-way fixed effects estimators with heterogeneous treatment effects', *American Economic Review* **110**(9), 2964–2996.

de Chaisemartin, C. and D'Haultfœuille, X. (2021*a*), Difference-in-differences estimators of intertemporal treatment effects. arXiv preprint arXiv:2007.04267.

de Chaisemartin, C. and D'Haultfœuille, X. (2021*b*), Two-way fixed effects regressions with several treatments. arXiv preprint arXiv:2012.10077.

de Chaisemartin, C., D'Haultfœuille, X. and Deeb, A. (2019), 'twowayfeweights: Estimation of the Weights Attached to the Two-Way Fixed Effects Regressions in Stata'.
URL: *https://ideas.repec.org/c/boc/bocode/s458611.html*

de Chaisemartin, C., D'Haultfœuille, X. and Guyonvarch, Y. (2019), 'did_multiplegt: DID Estimation with Multiple Groups and Periods in Stata'.
URL: *https://ideas.repec.org/c/boc/bocode/s458643.html*

Deeb, A. and de Chaisemartin, C. (2019), 'Clustering and external validity in randomized controlled trials', *arXiv preprint arXiv:1912.01052* .

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

Flack, E. and Edward (2020), 'bacondecomp: Goodman-Bacon Decomposition in R'.
URL: *https://cran.r-project.org/web/packages/bacondecomp/index.html*

Freyaldenhoven, S., Hansen, C. and Shapiro, J. M. (2019), 'Pre-event trends in the panel event-study design', *American Economic Review* **109**(9), 3307–38.

Gardner, J. (2021), Two-stage differences in differences. Working paper.

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.

Gobillon, L. and Magnac, T. (2016), 'Regional policy evaluation: Interactive fixed effects and synthetic controls', *Review of Economics and Statistics* **98**(3), 535–551.

Goodman-Bacon, A. (2021), 'Difference-in-differences with variation in treatment timing', *Journal of Econometrics* **225**, 254–277.

Goodman-Bacon, A., Goldring, T. and Nichols, A. (2019), 'BACONDECOMP: Stata module to perform a Bacon decomposition of difference-in-differences estimation'.
URL: *https://ideas.repec.org/c/boc/bocode/s458676.html*

Imai, K. and Kim, I. S. (2021), 'On the use of two-way fixed effects regression models for causal inference with panel data', *Political Analysis* **29**(3), 405–415.

Jakiela, P. (2021), Simple diagnostics for two-way fixed effects. arXiv preprint arXiv:2103.13229.

Kahn-Lang, A. and Lang, K. (2020), 'The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications', *Journal of Business & Economic Statistics* **38**(3), 613–620.

Liu, L., Wang, Y. and Xu, Y. (2021), A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. arXiv preprint arXiv:2107.00856.

Marcus, M. and Sant'Anna, P. H. (2021), 'The role of parallel trends in event study settings: An application to environmental economics.', *Journal of the Association of Environmental and Resource Economists* **8**(2), 235–275.

Rambachan, A. and Roth, J. (2019), An honest approach to parallel trends. Working paper.

Rios-Avila, F., Sant'Anna, P. and Callaway, B. (2021), 'Csdid: Stata module for the estimation of difference-in-difference models with multiple time periods'.
URL: *https://EconPapers.repec.org/RePEc:boc:bocode:s458976*

Roth, J. (2019), 'Pre-test with caution: Event-study estimates after testing for parallel trends', *Department of Economics, Harvard University, Unpublished manuscript* .

Roth, J. and Sant'Anna, P. H. (2021), Efficient estimation for staggered rollout designs. arXiv preprint arXiv:2102.01291.

Sant'Anna, P. and Callaway, B. (2021), 'did: Treatment effects with multiple periods and groups in r'.
URL: *https://cran.r-project.org/web/packages/did/index.html*

Stevenson, B. and Wolfers, J. (2006), 'Bargaining in the shadow of the law: Divorce laws and family distress', *The Quarterly Journal of Economics* **121**(1), 267–288.

Sun, L. (2020), 'EVENTSTUDYWEIGHTS: Stata module to estimate the implied weights on the cohort-specific average treatment effects on the treated (CATTs) (event study specifications)'.
URL: *https://ideas.repec.org/c/boc/bocode/s458833.html*

Sun, L. (2021), 'EVENTSTUDYINTERACT: Stata module to implement the interaction weighted estimator for an event study'.
URL: *https://ideas.repec.org/c/boc/bocode/s458978.html*

Sun, L. and Abraham, S. (2021), 'Estimating dynamic treatment effects in event studies with heterogeneous treatment effects', *Journal of Econometrics* **225**, 175–199.

Wooldridge, J. (2021), 'Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators', *Available at SSRN 3906345* .

Zhang, S. and de Chaisemartin, C. (2020), 'did_multiplegt: DID Estimation with Multiple Groups and Periods in R'.
URL: *https://cran.r-project.org/web/packages/DIDmultiplegt/index.html*

Zhang, S. and de Chaisemartin, C. (2021), 'TwowayFEWeights: Estimation of the Weights Attached to the Two-Way Fixed Effects Regressions in R'.
URL: *https://cran.r-project.org/web/packages/TwoWayFEWeights/index.html*