Manuscript submitted to *The Econometrics Journal*, pp. 1–32.

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

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

[†] *Economics Department, Sciences Po 28 rue des Saint-Pères 75005 Paris, France.*
E-mail: clement.dechaisemartin@sciencespo.fr

[‡] *CREST-ENSAE 5 avenue Henry le Chatelier 91120 Palaiseau, France.*
E-mail: xavier.dhaultfoeuille@ensae.fr

Summary
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 shown 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. We use those alternative estimators to revisit Wolfers (2006a).

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

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 D'Haultfœuille (2021a) 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:

$$\text{DID} = Y_{s,2} - Y_{s,1} - (Y_{n,2} - Y_{n,1})\,. \tag{1.1}$$

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)):

$$\begin{aligned} E\left[\text{DID}\right] =& E\left[Y_{s,2} - Y_{s,1} - (Y_{n,2} - Y_{n,1})\right] \\ =& E\left[Y_{s,2}(1) - Y_{s,1}(0) - (Y_{n,2}(0) - Y_{n,1}(0))\right] \\ =& E\left[Y_{s,2}(1) - Y_{s,2}(0)\right] + E\left[Y_{s,2}(0) - Y_{s,1}(0)\right] - E\left[Y_{n,2}(0) - Y_{n,1}(0)\right] \\ =& E\left[Y_{s,2}(1) - Y_{s,2}(0)\right], \end{aligned} \tag{1.2}$$

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]

Motivated by the fact that in the two-groups and two-periods design described above, DID is equal to the treatment coefficient in a TWFE regression, researchers have also estimated TWFE regressions in more complicated designs with many groups and periods, variation in treatment timing, treatments switching on and off, and/or non-binary treatments. Recent research has shown that in those more complicated designs, TWFE estimators are unbiased for an ATE if parallel trends holds, and if another assumption is satisfied: the treatment effect should 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 methods 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. We then give an overview of the papers that have proposed alternative estimators robust

[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. This restriction is not of essence to derive Equation (1.2) below, but it is of essence for some of the other results we cover, as noted later in the paper. We relax it in Section 3.2.

[2] Pre-trends tests come with caveats unveiled by a recent literature, see Kahn-Lang and Lang (2020), Bilinski and Hatfield (2018), and Roth (2021). Similarly, recent papers have proposed relaxations of the parallel trends assumption (see, e.g., Manski and Pepper, 2018; Rambachan and Roth, 2019; Freyaldenhoven et al., 2019). Though we allude to it in Section 3.2, this literature is mostly beyond the scope of this survey. See Roth et al. (2022) for a review.

to heterogeneous treatment effects. Finally, we revisit Wolfers (2006a), a famous TWFE application, in light of the recent literature discussed in this survey. As a word of caution, note that this literature is very recent, so several of the papers we review are still working papers, which have not been through the peer-review process yet.

Table 2 in the conclusion summarizes the heterogeneity-robust estimators available to applied researchers, depending on their research design. 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. Finally, the Stata code for our re-analysis of Wolfers (2006a), where several of the estimators discussed in this survey are computed, is available on the journal's website.

2. TWFE REGRESSIONS WITH HETEROGENEOUS TREATMENT EFFECTS

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

We consider a panel of G groups observed at T periods, respectively indexed by the placeholders g and t, which can refer to any group or time period. Typically, groups are geographical entities gathering many observations, but a group could also just be a single individual or firm. Let $\widehat{\beta}_{fe}$ denote the coefficient of $D_{g,t}$, the treatment in group g at period t, in an OLS regression of $Y_{g,t}$, the outcome of group g at period t, on group fixed effects, period fixed effects, and $D_{g,t}$:

$$Y_{g,t} = \widehat{\alpha}_g + \widehat{\gamma}_t + \widehat{\beta}_{fe} D_{g,t} + \epsilon_{g,t}, \tag{2.1}$$

where $\epsilon_{g,t}$ denotes the regression residual. We assume that the regression is unweighted, but it is sometimes weighted by $N_{g,t}$, the population of group g at period t. The results discussed below also apply to this weighted regression, see de Chaisemartin and D'Haultfœuille (2020).[3]

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

$$E\left[\widehat{\beta}_{fe}\right] = E\left[\sum_{(g,t):D_{g,t}\neq 0} W_{g,t} TE_{g,t}\right]. \tag{2.2}$$

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 $D_{g,t}$.[4] The $W_{g,t}$ are weights summing to 1, that are proportional to and of the same sign as

$$D_{g,t} - D_{g,.} - D_{.,t} + D_{.,.}, \tag{2.3}$$

[3] The regression could also be estimated using more disaggregated outcome data. For instance, groups may be US counties, and one may estimate the regression using individual-level outcome measures, assigning group membership based on county of residence. This disaggregated regression is equivalent to the aggregated regression in (2.1), provided $Y_{g,t}$ is defined as the average outcome of individuals in cell (g,t), and the aggregated regression is weighted by the number of individuals in cell (g,t). Accordingly, the results below also apply to disaggregated regressions, see de Chaisemartin and D'Haultfœuille (2020).

[4] de Chaisemartin and D'Haultfœuille (2020) derive Equation (2.2) assuming that groups' potential outcomes only depend on their current treatment, not on their past treatments. With dynamic effects, Equation (2.2) 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.

where $D_{g,.}$ is the average treatment of group g across periods, $D_{.,t}$ is the average treatment at period t across groups, and $D_{.,.}$ is the average treatment across groups and periods.

Equations (2.2) and (2.3) have two important consequences. First, $W_{g,t}$ is in general not equal to one divided by the number of treated (g,t) cells, so $\widehat{\beta}_{fe}$ may be biased for the average treatment effect across those cells, the ATT. A special case where $W_{g,t}$ is equal to one divided by the number of treated (g,t) cells, and where $\widehat{\beta}_{fe}$ is therefore unbiased for the ATT is when (i) the design is staggered, meaning that groups' treatment can only increase over time and can change at most once;[5] (ii) the treatment is binary; and (iii) there is no variation in treatment timing: all treated groups start receiving the treatment at the same date. 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 $D_{g,t} - D_{g,.} - D_{.,t} + D_{.,.}$ 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).[6] Unfortunately, this no-correlation condition is often implausible. To see this, note that $D_{g,t} - D_{g,.} - D_{.,t} + D_{.,.}$ is decreasing in $D_{g,.}$, meaning that $\widehat{\beta}_{fe}$ downweights the treatment effect of groups with the highest average treatment from period 1 to T. However, groups with the largest and lowest average treatment may have systematically different treatment effects. Similarly, $D_{g,t} - D_{g,.} - D_{.,t} + D_{.,.}$ is decreasing in $D_{.,t}$, 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. In staggered adoption designs, $D_{.,t}$ is increasing in t so the weights are decreasing in t. If the treatment effect is also monotonically increasing or decreasing in t, this no-correlation condition will fail. This no-correlation condition is partly testable, if one observes a proxy variable $P_{g,t}$ that is likely to be correlated with $TE_{g,t}$. Then, one can just test if $D_{g,t} - D_{g,.} - D_{.,t} + D_{.,.}$ is correlated with $P_{g,t}$.

Second, and perhaps more worryingly, Equation (2.3) 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 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

[5] Together, (i) and (ii) imply that groups can only switch from untreated to treated, and may do so at different points in time. This is probably the definition of a staggered design many people have in mind. (i) extends the definition of a staggered design to non-binary treatments.

[6] 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 (2.2) that $E\left[\widehat{\beta}_{fe}\right] = \delta$. However, constant effect is most often an implausible assumption.

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.

Inasmuch as "no-sign-reversal" is a desirable property, it becomes interesting to understand when $\widehat{\beta}_{fe}$ may satisfy it. Equation (2.3) 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 (2.2) 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 (2.3) 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 (2.3) 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 ($D_{g,t} \geq D_{g,.}$ for all gs such that $D_{g,t} \neq 0$). This condition is likely to often fail.

The `twowayfeweights` Stata (see de Chaisemartin et al., 2019) and R (see Zhang and de Chaisemartin, 2021) commands compute the weights $W_{g,t}$ in (2.2). The basic syntax of the Stata command is:

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

A decomposition similar to (2.2) can be obtained for TWFE regressions with control variables, and for $\widehat{\beta}_{fd}$, the treatment's coefficient in a regression of the outcome's first difference on the treatment's first difference and period fixed effects. de Chaisemartin and D'Haultfœuille (2020) also derive decompositions similar to (2.2), for $\widehat{\beta}_{fe}$ and $\widehat{\beta}_{fd}$, 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.

de Chaisemartin and D'Haultfœuille (2020) use the `twowayfeweights` Stata command to revisit Gentzkow et al. (2011). The authors regress the change in turnout in county g between two elections on the change of the county's number of newspapers and state-year fixed effects. They find that $\widehat{\beta}_{fd} = 0.0026$ (s.e. = 0.0009): one more newspaper increases turnout by 0.26 percentage points. Using the `twowayfeweights` Stata package, de Chaisemartin and D'Haultfœuille (2020) find that under parallel trends, $\widehat{\beta}_{fd}$ estimates a weighted sum of the effects of newspapers on turnout in 10,077 county×election cells, where 5,472 effects are weighted positively while 4,605 are weighted negatively, and where negative weights sum to -1.43. Accordingly, $\widehat{\beta}_{fd}$ is far from estimating a convex combination of effects. The weights are negatively correlated with the election year: $\widehat{\beta}_{fd}$ is more likely to upweight newspapers' effects in early elections, and to downweight or weight negatively newspapers' effects in late elections. This may lead $\widehat{\beta}_{fd}$ to be biased if newspapers' effects change over time. Similar results apply to $\widehat{\beta}_{fe}$: more than half of the weights attached to that coefficient are negative, and negative weights sum to -0.53.

The decomposition in (2.2) 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 another result related to that in Equation (2.2).[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

$$\widehat{\beta}_{fe} = \sum_{g \neq g', t < t'} v_{g,g',t,t'} \mathrm{DID}_{g,g',t,t'}, \tag{2.4}$$

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 the $DID_{g,g',t,t'}$s in Equation (2.4) 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 (2.2) originate from this second type of DIDs.

To see that, let us consider a simple example, first introduced by Borusyak and Jaravel

[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.

(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 example, Equation (2.4) reduces to

$$\widehat{\beta}_{fe} = (\mathrm{DID}_{e,\ell,1,2} + \mathrm{DID}_{\ell,e,2,3})/2, \tag{2.5}$$

with

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

$\mathrm{DID}_{e,\ell,1,2}$ 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. $\mathrm{DID}_{e,\ell,1,2}$ is similar to the DID estimator in Equation (1.1), and under parallel trends it is unbiased for the treatment effect in group e at period 2:

$$E\left[\mathrm{DID}_{e,\ell,1,2}\right] = E\left[TE_{e,2}\right]. \tag{2.6}$$

$\mathrm{DID}_{\ell,e,2,3}$, 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,

$$E\left[\mathrm{DID}_{\ell,e,2,3}\right] = E\left[TE_{\ell,3} - TE_{e,3} + TE_{e,2}\right], \tag{2.7}$$

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 (2.5), (2.6), and (2.7) that

$$E\left[\widehat{\beta}_{fe}\right] = E\left[1/2TE_{\ell,3} + TE_{e,2} - 1/2TE_{e,3}\right]. \tag{2.8}$$

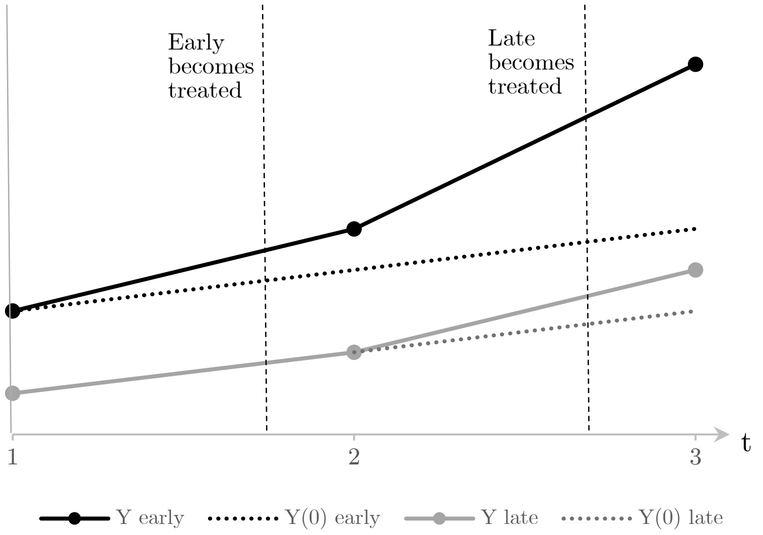
In this simple example, Equation (2.2) reduces to (2.8). The right-hand side of Equation (2.8) is a weighted sum of three ATEs where one ATE receives a negative weight. As the previous derivation shows, this negative weight comes from the fact $\widehat{\beta}_{fe}$ leverages $\mathrm{DID}_{\ell,e,2,3}$, a DID comparing a group switching from untreated to treated to a group treated at both periods.

To make things more concrete, Figure 1 below shows the actual and counterfactual outcome evolution, in a numerical example with three periods and an early and a late treated group. All treatment effects are positive: the actual outcomes, on the solid lines, are always above the counterfactual outcomes on the dashed lines. However, $\widehat{\beta}_{fe}$ is negative. $\widehat{\beta}_{fe}$ is the simple average of the DID comparing the early- to the late-treated group from period one to two, which is positive, and of the DID comparing the late- to the early-treated group from period two to three, which is negative, and larger in absolute

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

value than the first DID. The reason why the second DID is negative is that the treatment effect of the early-treated group increases substantially from period two to three, so this group's outcome increases more than that of the late-treated group.

Figure 1. A numerical example with three periods, an early and a late treated group



If one is ready to assume that the treatment effect does not change over time, $TE_{e,3} = TE_{e,2}$, and (2.7) simplifies to

$$E\left[\text{DID}_{\ell,e,2,3}\right] = E\left[TE_{\ell,3}\right]. \tag{2.9}$$

Then, the negative weight in (2.7) 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 (2.4) 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 (2.4) now indicates that $\widehat{\beta}_{fe}$ assigns a weight equal to 1/6 to DIDs comparing a switcher to a group treated at both periods. On the other hand, all the weights in (2.2) 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 (2.2) are positive, but the sum of the weights in (2.4) on DIDs comparing a switcher to a group treated at both periods is equal to 0.06. Beyond these examples, one can show that having DIDs comparing a switcher to a group treated at both periods in (2.4) is necessary but not sufficient to have negative weights in (2.2). Similarly, the sum of the weights on DIDs comparing a switcher to a group treated at both periods in (2.4) is always larger than the absolute value of the sum of the negative weights in (2.2). The reason why Equation (2.4) “overestimates” the negative weights in (2.2) 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 (2.4) putting less weight on DIDs using a group treated at both periods as the control group.[10]

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 (2.4), the weights assigned to them, as well as the sum of the weights on $\text{DID}_{g,g',t,t'}$s using a group treated at both periods as the control group. The basic syntax of the `bacondecomp` Stata command is:

```
bacondecomp outcome treatment, ddetail
```

2.2.2. “Forbidden comparisons” when the design is not staggered or treatment is not binary When the treatment is not staggered 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. In fact, with two groups m and ℓ and two periods, one can show that

$$\widehat{\beta}_{fe} = \frac{Y_{m,2} - Y_{m,1} - (Y_{\ell,2} - Y_{\ell,1})}{D_{m,2} - D_{m,1} - (D_{\ell,2} - D_{\ell,1})}, \tag{2.12}$$

where the right hand side of the previous display is the Wald-DID estimator studied by de Chaisemartin and D’Haultfœuille (2018). The Wald-DID compares the outcome evolution of groups m and ℓ, and scales that comparison by the differential evolution of m’s and ℓ’s treatments. de Chaisemartin and D’Haultfœuille (2018) show that the Wald-DID may not estimate a convex combination of effects, unless the treatment effect is constant over time and is the same in groups m and ℓ. This second requirement was not present in the binary and staggered case. In that case, we have seen before that if the treatment effect is constant over time, $\widehat{\beta}_{fe}$ estimates a convex combination of effects, even if the treatment effect varies between groups.

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

$$\text{DID}_{\ell,e,t_1,t_2} = \text{DID}_{\ell,n,t_0,t_2} - \text{DID}_{e,n,t_0,t_2} + \text{DID}_{e,\ell,t_0,t_1}. \tag{2.10}$$

Then, it follows from Equation (2.10) that

$$\begin{aligned}&v_{\ell,e,t_1,t_2}\text{DID}_{\ell,e,t_1,t_2} + v_{e,n,t_0,t_2}\text{DID}_{e,n,t_0,t_2}\\ =&(v_{\ell,e,t_1,t_2} - \underline{v})\text{DID}_{\ell,e,t_1,t_2} + \underline{v}\text{DID}_{\ell,n,t_0,t_2} + \underline{v}\text{DID}_{e,\ell,t_0,t_1} + (v_{e,n,t_0,t_2} - \underline{v})\text{DID}_{e,n,t_0,t_2}.\end{aligned} \tag{2.11}$$

Plugging Equation (2.11) into Equation (2.4) will yield a different decomposition of $\widehat{\beta}_{fe}$ as a weighted average of DIDs. But the weight on DIDs using a group treated at both periods as the control group is equal to v_{ℓ,e,t_1,t_2} in the left-hand-side of Equation (2.11), 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 (2.4) on DIDs using a group treated at both periods as the control group.

To see that with a non-binary or non-staggered treatment $\widehat{\beta}_{fe}$ may not estimate a convex combination of effects even if the treatment effect is constant over time, let us consider a simple example. Assume that group m goes from 0 to 2 units of treatment from period 1 to 2, while group ℓ goes from 0 to 1 unit. Then, the denominator of the Wald-DID is equal to $2-0-(1-0)=1$, so

$$\widehat{\beta}_{fe}=Y_{m,2}-Y_{m,1}-\left(Y_{\ell,2}-Y_{\ell,1}\right).$$

To simplify, let us also assume that in both groups, potential outcomes are linear in the number of treatment units, with slopes that are constant over time but may differ for groups m and ℓ:

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

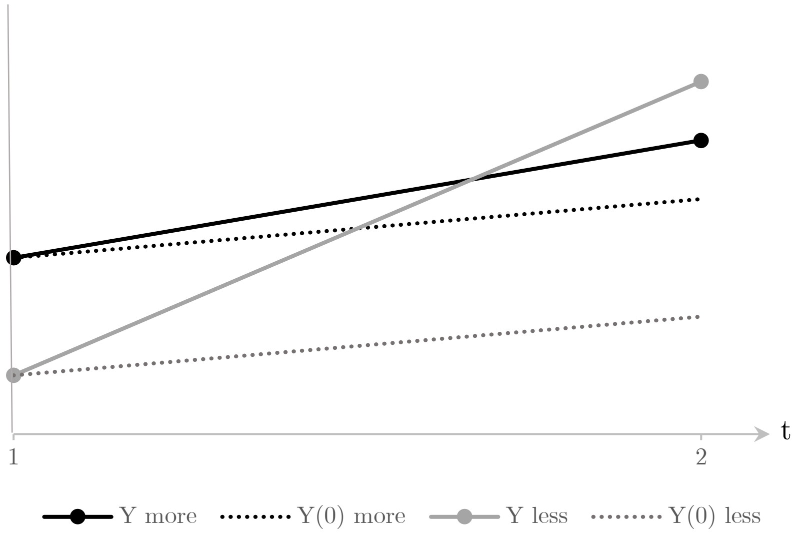
Then, under parallel trends,

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

a weighted sum of m and ℓ's treatment effects, where group ℓ's effect is weighted negatively. Intuitively, group ℓ is also treated at period two, and $\widehat{\beta}_{fe}$, which uses ℓ as a control group, subtracts its treatment effect out. This example also shows that $\widehat{\beta}_{fe}$ may fail to identify a convex combination of effects, even without variation in treatment timing: here, both m and ℓ start getting treated at period 2.

To make things more concrete, Figure 2 below shows the actual and counterfactual outcome evolution, in a numerical example with two periods, a group whose treatment increases more, from 0 to 2 units, and a group whose treatment increases less, from 0 to 1 unit. All treatment effects are positive: the actual outcomes, on the solid lines, are always above the counterfactual outcomes on the dashed lines. However, $\widehat{\beta}_{fe}$, which is equal to the DID comparing the more- and the less-treated groups from period one to two, is negative. The reason why this DID is negative is that the treatment effect, per treatment unit, of the less-treated group is more than twice larger than the treatment effect of the more-treated group. Accordingly, the outcome of the less-treated group increases more, despite the fact that this group receives a twice smaller treatment dose in period 2.

Figure 2. A numerical example with two periods, a more- and a less-treated group



2.3. Decomposition results for other TWFE regression coefficients

2.3.1. Dynamic TWFE regressions In staggered designs with a binary treatment, Sun and Abraham (2021) consider event-study regressions:

$$Y_{g,t} = \widehat{\gamma}_g + \widehat{\lambda}_t + \sum_{\ell=-K,\ell\neq -1}^{L} \widehat{\beta}_\ell 1\{F_g = t-\ell\} + \varepsilon_{g,t}, \tag{2.13}$$

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. Researchers have sometimes estimated a variant of this regression, where the first and last indicators $1\{F_g = t+K\}$ and $1\{F_g = t-L\}$ are respectively replaced by an indicator for being at least K periods away from adoption ($1\{F_g \geq t+K\}$) and an indicator for having adopted at least L periods ago ($1\{F_g \leq t-L\}$). Such endpoint binning is for instance recommended by Schmidheiny and Siegloch (2020): without it, the regression implicitly assumes that the treatment no longer has any effect after L periods. Instead, with endpoint binning the regression assumes that that the treatment effect is constant after L periods, a more plausible assumption.

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

$$E\left[\widehat{\beta}_\ell\right] = E\left[\sum_g w_{g,\ell} TE_g(\ell) + \sum_{\ell'\neq\ell}\sum_g w_{g,\ell'} TE_g(\ell')\right], \tag{2.14}$$

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 (2.14) 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 (2.2), 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 (2.14): 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, (2.13) 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 (2.13), `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.

Event-study regressions can only be used in staggered designs with a binary treatment. In more complicated designs where the treatment is not binary or a group's treatment can increase or decrease multiple times, some researchers have estimated TWFE regressions of the outcome on the treatment and its first K lags, the so-called distributed-lag regression. Other researchers have estimated a panel-data version of the local-projection method proposed by Jordà (2005) for time-series data: $Y_{g,t+\ell}$ is regressed on group and period FEs and $D_{g,t}$, for $\ell\in\{0,...,K\}$. de Chaisemartin and D'Haultfœuille (2021a) show that those regressions suffer from similar issues as the event-study regression: under parallel trends, the distributed-lag and local-projection regressions may produce biased estimates of the

[11] Equation (2.14) 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.

treatment's instantaneous and dynamic effects, if effects are heterogeneous across groups and over time. In particular, they do not satisfy the no-sign reversal property: one could have that the treatment's instantaneous and dynamic effects are positive in every (g,t) cell, but the expectations of those regression coefficients are negative. de Chaisemartin and D'Haultfœuille (2021a) also show that the panel-data version of the local-projection method may yield biased estimates even if effects are homogeneous.

2.3.2. 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 (2021b) show that in those regressions, 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 HETEROGENEITY-ROBUST DID ESTIMATORS

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. We start by reviewing estimators ruling out dynamic effects, i.e. that assume that a group's current outcome only depends on its current treatment, before reviewing estimators that allow dynamic effects. In complicated designs, say with a continuous treatment that changes often, allowing for dynamic effects comes with a number of costs: it may result in imprecise estimators, and may complicate the interpretation of the estimated effects. Then, one may want to carefully evaluate if past treatments are indeed likely to affect the current outcome.

3.1. Estimators ruling out dynamic effects

With a binary treatment, de Chaisemartin and D'Haultfœuille (2020) propose to use the DID_M estimator. With two time periods, DID_M is merely a weighted average of

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

and of

$$\mathrm{DID}_{-} = \frac{1}{N_{1,1}} \sum_{g: D_{g,1}=1, D_{g,2}=1} (Y_{g,2} - Y_{g,1}) - \frac{1}{N_{1,0}} \sum_{g: D_{g,1}=1, D_{g,2}=0} (Y_{g,2} - Y_{g,1}),$$

where for all $(d_1, d_2) \in \{0,1\}^2$, N_{d_1,d_2} denotes the number of groups such that $D_{g,1} = d_1$ and $D_{g,2} = d_2$.[12] DID_+ is a DID comparing the period-one-to-two outcome evolution of groups going from untreated to treated, the "switchers in", and of groups untreated at both dates. It is similar to the DID estimator in Equation (1.1), and it is unbiased for the treatment effect of the switching-in groups at period 2, under a parallel trends assumption on the untreated outcome $Y_{g,t}(0)$. DID_- is a DID comparing the period-one-to-two outcome evolution of groups treated at both dates, and of groups going from treated to untreated, the "switchers out". DID_- is also similar to the DID estimator in Equation (1.1), switching "treatment" and "non-treatment". Then, one can show that DID_- is unbiased for the treatment effect of the switching-out groups at period 2, under a parallel trends assumption on the treated outcome $Y_{g,t}(1)$.

The $\mathrm{DID}_{\mathrm{M}}$ estimator can easily be extended to applications with more than two time periods. For each pair of consecutive time periods, one can compute a $\mathrm{DID}_{+,t}$ estimator comparing groups going from untreated to treated from $t-1$ to t to groups untreated at both dates, and a $\mathrm{DID}_{-,t}$ estimator comparing groups treated at $t-1$ and t to groups going from treated to untreated from $t-1$ to t. Then, one averages the $\mathrm{DID}_{+,t}$ and $\mathrm{DID}_{-,t}$ estimators across t. de Chaisemartin and D'Haultfœuille (2020) show that the resulting estimator is unbiased for the average treatment effect across all switching (g,t) cells, namely cells such that $D_{g,t} \neq D_{g,t-1}$. They also propose placebo estimators to test the parallel trends assumptions underlying $\mathrm{DID}_{\mathrm{M}}$. The placebos compare the outcome trends of switchers and non-switchers, before the switchers switch.

With more than two time periods, the $\mathrm{DID}_{\mathrm{M}}$ estimator may be biased if the treatment has dynamic effects. For instance, to infer the counterfactual trend that groups going from untreated to treated from $t-1$ to t would have experienced without that switch, $\mathrm{DID}_{+,t}$ uses as controls all groups untreated at $t-1$ and t. However, some of those groups may have been treated, say, at $t-2$. If the treatment has dynamic effects, this past treatment may affect their period $t-1$-to-t outcome evolution, thus making them potentially invalid controls. Note that if the treatment is binary and staggered, such situations cannot arise: groups untreated at $t-1$ and t have been untreated all along. Accordingly, $\mathrm{DID}_{\mathrm{M}}$ is robust to dynamic effects in binary and staggered designs.

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

[12] Implicitly, this definition of DID_+ and DID_- assumes that all groups have the same sizes. The $\mathrm{DID}_{\mathrm{M}}$ estimator can easily be extended to instances where groups have heterogeneous sizes, see de Chaisemartin and D'Haultfœuille (2020).

The DID_{M} estimator is computed by the `did_multiplegt` Stata (see de Chaisemartin et al., 2019) and R (see Zhang and de Chaisemartin, 2020) commands. The basic syntax of the Stata command is:

```
did_multiplegt outcome groupid timeid treatment
```

de Chaisemartin and D'Haultfœuille (2020) compute the DID_{M} estimator in the Gentzkow et al. (2011) example mentioned above, that studies the effect of newspapers on turnout in US presidential elections. de Chaisemartin and D'Haultfœuille (2020) find that $\text{DID}_{\text{M}} = 0.0043$ (s.e. $= 0.0014$), meaning that one more newspaper increases turnout by 0.43 percentage point. DID_{M} is 66% larger than, and significantly different from, $\widehat{\beta}_{fd}$, the estimator reported by Gentzkow et al. (2011).

de Chaisemartin et al. (2022) extend the DID_{M} estimator to continuous treatments. To simplify, we present their estimators in the case with two time periods, though they readily extend to the case with more periods. de Chaisemartin et al. (2022) assume that from period one to two, the treatment of some units, hereafter referred to as the movers, changes. They also assume that the treatment of other units, hereafter referred to as the stayers, does not change. This assumption is likely to be met when the treatment is say, trade tariffs: tariffs' reforms rarely apply to all products, so it is likely that tariffs of at least some products stay constant over time. On the other hand, this assumption is unlikely to be met when the treatment is say, precipitations: geographical units never experience the exact same precipitations over two consecutive years.

Under the assumption that there are some stayers, the estimator proposed by de Chaisemartin et al. (2022) compares the outcome evolution of movers and stayers, with the same period-one treatment. With a continuous treatment, such comparisons can either be achieved by reweigthing stayers by propensity score weights, or by adjusting movers' outcome change using a nonparametric regression of the outcome change on the period-one treatment among the stayers. Under parallel trends assumptions, the corresponding estimands identify a weighted average of the effect, across all movers, of moving their treatment from its period-one to its period-two value, scaled by the difference between these two values. This effect is a weighted average of the slopes of movers' potential outcome function, between their period-one and period-two treatments.

The estimators in de Chaisemartin et al. (2022) can be extended to the case where there are no stayers, provided there are quasi-stayers, meaning units whose treatment barely changes from period one to two. Alternatively, one could also use the estimator proposed by Graham and Powell (2012), which compares the outcome evolution of movers and quasi stayers, but without conditioning on units' period-one treatment. Their estimator relies on a linear treatment effect assumption, unlike those in de Chaisemartin et al. (2022). When there are no true stayers, both estimators require choosing a bandwidth, namely the lowest treatment change below which a unit can be considered as a quasi-stayer. Neither de Chaisemartin et al. (2022) or Graham and Powell (2012) derive an "optimal" bandwidth, so for now bandwidth choice is left to the discretion of the researcher. If the data has at least three periods, one could also use the correlated-random-coefficient estimator proposed by Chamberlain (1992). While it allows for some treatment effect heterogeneity, that estimator relies on a linear treatment effect assumption, like the estimator in Graham and Powell (2012).

de Chaisemartin et al. (2022) show that after some relabelling, some of their estimators are equivalent or nearly equivalent to estimators that had been previously proposed by de Chaisemartin and D'Haultfœuille (2018), Abadie (2005), and Callaway and Sant'Anna (2021). This implies that their estimators can be computed, up to small tweaks, by the companion software for those papers. We refer the reader to de Chaisemartin et al. (2022) for a precise description of how their estimators can be computed using existing software.

3.2. Estimators allowing for dynamic effects when the treatment is binary and the design is staggered.

For any $t \in \{1, ..., T\}$, let $\mathbf{0}_t$ (resp. $\mathbf{1}_t$) denote a vectors of t zeros (resp. ones). With dynamic effects, group g's outcome at time t is allowed to depend on her past treatments. For any $(d_1, ..., d_t)$, let $Y_{g,t}(d_1, ..., d_t)$ denote group g's potential outcome at period t with treatments $(d_1, ..., d_t)$ from period 1 to t.[13] In particular, $Y_{g,t}(\mathbf{0}_t)$ is group g's outcome without ever being treated from period 1 to t. With dynamic effects, Callaway and Sant'Anna (2021) and Sun and Abraham (2021) have proposed to replace the parallel trends assumption on $Y_{g,t}(0)$ by a parallel trends assumption on $Y_{g,t}(\mathbf{0}_t)$: for all $g \neq g'$ and $t \geq 2$,

$$E\left[Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})\right] = E\left[Y_{g',t}(\mathbf{0}_t) - Y_{g',t-1}(\mathbf{0}_{t-1})\right]. \tag{3.1}$$

We now review the estimators 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 (3.1).

3.2.1. 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. For all c and t, and for all $\ell \in \{0, ..., t\}$ let $\overline{Y}_{c,t}$ denote the average outcome at period t across groups belonging to cohort c, and let $\overline{Y}_{n,t}$ denote the average outcome at period t across groups that remain untreated from period 1 to T, hereafter referred to as the never-treated groups, assuming for now that such groups exist. Callaway and Sant'Anna (2021) define their parameters of interest as

$$TE_{c,c+\ell} = E\left[\overline{Y}_{c,c+\ell}(\mathbf{0}_{c-1}, \mathbf{1}_{\ell+1}) - \overline{Y}_{c,c+\ell}(\mathbf{0}_{c+\ell})\right],$$

the average effect of having been treated for $\ell + 1$ periods in the cohort that started receiving the treatment at period c, for every $c \in \{2, ..., T\}$ and $\ell \geq 0$ such that $\ell + c \leq T$. To estimate, say, $TE_{c,c}$, Callaway and Sant'Anna (2021) propose

$$\widehat{\text{DID}}_{c,0} = \overline{Y}_{c,c} - \overline{Y}_{c,c-1} - \left(\overline{Y}_{n,c} - \overline{Y}_{n,c-1}\right),$$

a DID estimator comparing the period $c-1$-to-c outcome evolution in cohort c and in the never-treated groups n. $\widehat{\text{DID}}_{c,0}$ is unbiased for $TE_{c,c}$:

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

[13] This notation implicitly rules out anticipation effects: the outcome cannot depend on a group's future treatment.

where the last equality follows from Equation (3.1). More generally, to estimate $TE_{c,c+\ell}$, Callaway and Sant'Anna (2021) propose

$$\overline{\text{DID}}_{c,\ell} = \overline{Y}_{c,c+\ell} - \overline{Y}_{c,c-1} - \left(\overline{Y}_{n,c+\ell} - \overline{Y}_{n,c-1}\right),$$

a DID estimator comparing the period-$c-1$-to-$c+\ell$ outcome evolution in cohort c and in the never-treated groups n.

Callaway and Sant'Anna (2021) extend those baseline estimators in various directions. First, they propose more aggregated estimators, such as DID_ℓ, a weighted average of the $\overline{\text{DID}}_{c,\ell}$ estimators across all cohorts reaching ℓ periods after their first treatment before the end of the panel. Second, they propose estimators similar to those above, but that use the not-yet-treated instead of the never-treated as controls. For instance, all groups not yet treated at period c can be used as control groups in the definition of $\overline{\text{DID}}_{c,0}$. This is very useful when there is no never-treated group: in that case, the 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, one may worry that such groups are less comparable to groups that get treated at some point, and researchers sometimes prefer to discard them and only leverage variation in treatment timing. Finally, even when one is fine with keeping the never-treated groups, the not-yet-treated is a larger control group, and may lead to more precise estimators. Note that in staggered adoption designs with a binary treatment, the DID_M estimator proposed by de Chaisemartin and D'Haultfœuille (2020) also uses the not-yet-treated as controls, and is identical to the DID_0 estimator of the instantaneous treatment effect using the not-yet-treated as controls in Callaway and Sant'Anna (2021). Third, Callaway and Sant'Anna (2021) also propose estimators relying on a conditional parallel trends assumption. Fourth, they suggest placebo estimators to test the parallel trends assumptions underlying their estimators. These placebos are robust to heterogeneous effects, unlike the coefficients $\widehat{\beta}_\ell$ for $\ell \leq -2$ from the event-study regression in (2.13).

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 equal to the period when a group starts receiving the treatment.

3.2.2. 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 (3.1), 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 event-study regression in (2.13), `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).

3.2.3. The estimators proposed by Borusyak et al. (2021), Gardner (2021), and Liu et al. (2021) Borusyak et al. (2021), Gardner (2021), and Liu 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. We start by reviewing Borusyak et al. (2021), before discussing the connection between their results and those in Gardner (2021) and Liu et al. (2021). The estimators in Borusyak et al. (2021) 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.

A second, numerically equivalent way of computing the estimators in Borusyak et al. (2021) amounts to fitting a regression of the outcome on group and time fixed effects in the sample of untreated observations, and using that regression to predict the counterfactual outcome of treated observations. Estimates of the treatment effect of those observations are then merely obtained by substracting their counterfactual to their actual outcome. This imputation method is computationally faster than the first. It also readily generalizes to more complicated specifications, such as triple-differences, or models allowing for group-specific linear trends. Using this representation of their estimator, Borusyak et al. (2021) show that it can also be used to estimate the effect of a binary and non-staggered treatment, if that treatment does not have dynamic effects. This imputation method is the one used by the `did_imputation` Stata command (see Borusyak, 2021) and by the `didimputation` R command (see Butts, 2021) to compute the estimators proposed by Borusyak et al. (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.

Before Borusyak et al. (2021), Liu et al. (2021) and Gardner (2021) have proposed the same imputation method as in Borusyak et al. (2021),[14] but the result showing that the

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

resulting estimators are efficient under the assumptions of the Gauss-Markov theorem only appears in Borusyak et al. (2021). Note that Wooldridge (2021) has also proposed an estimation strategy connected, and in some cases numerically equivalent, to that of Borusyak et al. (2021).

3.2.4. Understanding the differences between those estimators Under parallel trends, the estimators in Borusyak et al. (2021) may offer precision gains with respect to those in Callaway and Sant'Anna (2021) or Sun and Abraham (2021), under the assumptions of the Gauss-Markov theorem. Those require, among other things, that the never treated potential outcomes $Y_{g,t}(\mathbf{0}_t)$ be independent of each other, both across groups and over time. It is, of course, 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), but simulations in Borusyak et al. (2021) suggest that one can still expect efficiency gains with moderate serial correlation.

If trends are not exactly parallel, the estimators in Borusyak et al. (2021) may be more or less biased than those in Callaway and Sant'Anna (2021) or Sun and Abraham (2021) depending on the nature of the violation of parallel trends. 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 treatment effect at $t_s + \ell$ is

$$Y_{s,t_s+\ell} - \frac{1}{t_s-1}\sum_{k=1}^{t_s-1} Y_{s,k} - \frac{1}{G-1}\sum_{g\neq s}\left(Y_{g,t_s+\ell} - \frac{1}{t_s-1}\sum_{k=1}^{t_s-1} Y_{g,k}\right), \tag{3.2}$$

while the estimator in Callaway and Sant'Anna (2021) and Sun and Abraham (2021) is

$$Y_{s,t_s+\ell} - Y_{s,t_s-1} - \frac{1}{G-1}\sum_{g\neq s}\left(Y_{g,t_s+\ell} - Y_{g,t_s-1}\right). \tag{3.3}$$

Equation (3.3) shows that the estimator 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, while Equation (3.2) shows that the estimator in Borusyak et al. (2021) instead uses the average outcome from period 1 to $t_s - 1$ as the baseline. This is why the latter estimator is often more precise. However, it is also more biased, when parallel trends does not exactly hold and the discrepancy between groups' trends gets larger over longer horizons, as would for instance happen when there are group-specific linear trends. In such instances, Roth (2021) notes that leveraging earlier pre-treatment periods increases the bias of a DID estimator, since one makes comparisons from earlier periods. If, on the other hand, parallel trends fails due to anticipation effects arising a few periods before t_s, Equations (3.2) and (3.3) imply that the estimator in Borusyak et al. (2021) is less biased than that in Callaway and Sant'Anna (2021) and Sun and Abraham (2021). However, these two types of violations of parallel trends may not be equally problematic. Often times, both estimators can be immunized against anticipation effects, by redefining t_s as the date when the treatment was announced. On the other hand, it is often harder to immunize them against differential trends widening over time (see de Chaisemartin and D'Haultfœuille, 2021a, for further discussion). Beyond the simple example we consider here, deriving a closed-form expression of the estimators in

Borusyak et al. (2021) is not straightforward. Whether the conclusions we derive in this simple example carry through to more complicated designs is thus an open question.

If one views parallel trends as a reasonable first-order approximation rather than an assumption that holds exactly, it may make sense to investigate how sensitive one's findings are to violations of parallel trends. To do so, one may for instance implement the partial identification approach in Manski and Pepper (2018) or Rambachan and Roth (2019). The latter approach assumes that parallel trends do not hold exactly, and that the magnitude of placebo estimators is informative as to the magnitude of the bias in the actual estimators caused by differential trends. The estimators proposed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021) may be more amenable to the approach in Rambachan and Roth (2019) than the estimators proposed by Borusyak et al. (2021). Consider again the same simple example as above. For any $\ell \leq t_s - 2$, one can construct the following placebo estimator:

$$Y_{s,t_s-1} - Y_{s,t_s-\ell-2} - \frac{1}{G-1}\sum_{g\neq s}\left(Y_{g,t_s-1} - Y_{g,t_s-\ell-2}\right). \tag{3.4}$$

This placebo compares the treated and control groups' outcome evolution, from period $t_s - \ell - 2$ to $t_s - 1$, namely over $\ell + 1$ periods before group s got treated. It exactly mimicks the estimator of group s's treatment effect at period $t_s + \ell$ proposed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021), which compares the same groups, over the same number of periods. Accordingly, the magnitude of that placebo may indeed be informative as to the magnitude of the bias of the estimator in Equation (3.3), as requested by Rambachan and Roth (2019). Building a placebo that would similarly mimick the estimator proposed by Borusyak et al. (2021) is not feasible, precisely because that estimator leverages all pre-treatment periods to construct its baseline. See de Chaisemartin and D'Haultfœuille (2021a) for more discussion of the advantages of having placebos that mimick actual estimators.

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.[15] 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 conditional on the design: which groups are required to be on parallel trends at which dates depends on groups' realized treatments. It is also not testable. We refer the reader to Marcus and Sant'Anna (2021) and Borusyak et al. (2021) for further discussion on the differences between parallel trends assumptions.

Overall, whether the estimators in Borusyak et al. (2021) should be preferred to those in Callaway and Sant'Anna (2021) and Sun and Abraham (2021) may depend on one's degree of confidence in the parallel trends assumption, on the type of violations of this assumption that seems more likely to arise in the application at hand, on whether it is possible to immunize the estimators against anticipation effects by redefining the treatment date as the announcement date, and on one's willingness to undertake a sensitivity analysis such as the one proposed by Rambachan and Roth (2019). Note also that if

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

the estimators proposed by Borusyak et al. (2021), Callaway and Sant'Anna (2021), and Sun and Abraham (2021) are significantly different, this implies that the parallel trends assumption, at least the "strong version" of this assumption imposed by Borusyak et al. (2021) and Sun and Abraham (2021), must be violated.

3.3. Estimators allowing for dynamic effects when the treatment is not binary or the design is not staggered.

de Chaisemartin and D'Haultfœuille (2021a) propose treatment effect estimators robust to heterogeneous and dynamic treatment effects and that 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 4 have a binary treatment and a staggered design, so being able to accommodate more general designs is important. The paper's main idea is to propose a generalization of the event-study approach to such designs, by defining the event as the period where a group's treatment changes for the first time. With a binary-and-staggered treatment, the event per this definition is the period where a group gets treated, so this definition extends the standard one to general designs.

More specifically, de Chaisemartin and D'Haultfœuille (2021a) start by showing that for any group g whose treatment changed for the first time at period F_g, the instantaneous and dynamic effects of that change can be unbiasedly estimated. Let

$$\delta_{g,\ell} = E(Y_{g,F_g+\ell} - Y_{g,F_g+\ell}(D_{g,1},...,D_{g,1}))$$

be the expected difference between group g's actual outcome at $F_g+\ell$ and the counterfactual "status quo" outcome it would have obtained if its treatment had remained equal to its period-one value from period one to $F_g+\ell$. Let $N^c_{g,\ell}$ denote the number of groups whose treatment has not changed yet at $F_g+\ell$, and with the same treatment as g at period one. de Chaisemartin and D'Haultfœuille (2021a) show that

$$\text{DID}_{g,\ell} = Y_{g,F_g+\ell} - Y_{g,F_g-1} - \frac{1}{N^c_{g,\ell}} \sum_{g':D_{g',1}=D_{g,1},F_{g'}>F_g+\ell} (Y_{g',F_g+\ell} - Y_{g',F_g-1}),$$

a DID estimator comparing the F_g-1-to-$F_g+\ell$ outcome evolution between group g and groups whose treatment has not changed yet at $F_g+\ell$ and with the same treatment as g at period one, is unbiased for $\delta_{g,\ell}$ under parallel trends assumptions. To test those parallel trends assumptions, they propose placebo estimators comparing the outcome trends of switchers and non-switchers before the switchers switch.

Then, de Chaisemartin and D'Haultfœuille (2021a) aggregate the $\text{DID}_{g,\ell}$ estimators into an estimator of the effect of having experienced a weakly higher amount of treatment for ℓ periods. For any real number x and $t \in \{1,...,T\}$, let $\boldsymbol{x}_t$ denote a $1\times t$ vector with coordinates equal to x. When the treatment is binary, for groups untreated at period one, $D_{g,1}=0$, so

$$\delta_{g,\ell} = E(Y_{g,F_g+\ell}(\mathbf{0}_{F_g-1},1,D_{g,F_g+1},...,D_{g,F_g+\ell}) - Y_{g,F_g+\ell}(\mathbf{0}_{F_g+\ell})).$$

For groups treated at period one, $D_{g,1}=1$, so

$$-\delta_{g,\ell} = E(Y_{g,F_g+\ell}(\mathbf{1}_{F_g+\ell}) - Y_{g,F_g+\ell}(\mathbf{1}_{F_g-1},0,D_{g,F_g+1},...,D_{g,F_g+\ell})).$$

The right-hand side of the two equations above are effects of having experienced a weakly higher amount of treatment for $\ell + 1$ periods. Accordingly, the $\text{DID}_{g,\ell}$ estimators are aggregated into a DID_ℓ estimator, multiplying by minus one the $\text{DID}_{g,\ell}$ of groups treated at period one. With a non-binary treatment, one can also aggregate the $\text{DID}_{g,\ell}$ to estimate the effect of having experienced a weakly higher amount of treatment for $\ell + 1$ periods.

Ultimately, this approach leads to an event-study graph, with the distance to the first treatment change on the x-axis, the DID_ℓ estimators on the y-axis to the right of zero, and placebo estimators on the y-axis to the left of zero. This event-study graph is useful to test the parallel trends assumption, and to provide reduced-form evidence of whether weakly increasing the treatment for $\ell + 1$ periods increases or decreases the outcome on average. However, interpreting the magnitude of the DID_ℓ estimators might be complicated. For instance, with three periods and three groups such that $(D_{1,1} = 0, D_{1,2} = 4, D_{1,3} = 1)$, $(D_{2,1} = 0, D_{2,2} = 2, D_{2,3} = 3)$, and $(D_{2,1} = 0, D_{2,2} = 0, D_{2,3} = 0)$, DID_1 estimates the average of $E(Y_{1,3}(0,4,1) - Y_{1,3}(0,0,0))$ and $E(Y_{2,3}(0,2,3) - Y_{2,3}(0,0,0))$. Accordingly, DID_1 does not estimate by how much the outcome increases on average when the treatment increases by a given amount for a given number of periods.

To circumvent this important limitation, two strategies can be implemented. First, the reduced-form event-study graph described above can be complemented with a first-stage event-study graph, where the outcome is replaced by the treatment. The estimators on the first-stage graph show the average value of $|D_{g,F_g+\ell} - D_{g,1}|$ across all groups entering in DID_ℓ. In the example above, the first two estimates on the first-stage graph are equal to $1/2(D_{1,2} - D_{1,1} + D_{2,2} - D_{2,1}) = 3$ and $1/2(D_{1,3} - D_{1,1} + D_{2,3} - D_{2,1}) = 2$. This reflects the fact that in this example, DID_1 is an effect produced by increasing the previous and current treatment by 3 and 2 units on average. Second, a weighted average across ℓ of the reduced-form estimators divided by a weighted average across ℓ of the first-stage estimators is unbiased for a parameter with a clear economic interpretation. That parameter may be used to conduct a cost-benefit analysis comparing groups' actual treatments to the status quo scenario where they would have kept all along the same treatment as in period one. In other words, that parameter can be used to determine if the policy changes that took place over the duration of the panel led to a better situation than the one that would have prevailed if no policy change had been undertaken, a natural policy question. Importantly, that parameter can also be interpreted as an average total effect per unit of treatment, where "total effect" refers to the sum of the instantaneous and dynamic effects of a treatment.

The estimators proposed by de Chaisemartin and D'Haultfœuille (2021a) 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 placebo(#) longdiff_placebo breps(#) cluster(groupid),
```

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

The estimators in de Chaisemartin and D'Haultfœuille (2021a) 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). The estimators proposed by Callaway et al. (2021) can also accommodate continuous and staggered treatments. For continuous and non-staggered treatments, in their Section 4.3 de Chaisemartin et al. (2022) extend their baseline estimators to allow for dynamic effects. With respect to their baseline estimators, the main difference is that when allowing for dynamic effects, fewer units can be used as controls. Without dynamic effects, at period t, any unit whose treatment has not changed between $t-1$ and t can be used as a valid control. With dynamic effects, only units whose treatments have not changed from period 1 to t can be used as valid controls. Therefore, the need for "stayers" becomes even stronger when allowing for dynamic effects: many units need to keep the same value of the treatment for a large number of time periods. Developing estimators robust to dynamic effects that can be used with a continuous treatment and no stayers has not been done yet and is a promising area for future research.

The estimators in de Chaisemartin and D'Haultfœuille (2021a) 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 (2021a) differ. Callaway and Sant'Anna (2021) consider time-invariant covariates, and assume that trends are parallel once we condition on them. de Chaisemartin and D'Haultfœuille (2021a) instead consider time-varying covariates and assume that trends are parallel once the linear effect of those time-varying covariates on the outcome is accounted for. This for instance allows them to include group-specific linear trends in the estimation. With covariates, the parallel trends conditions in Callaway and Sant'Anna (2021) and de Chaisemartin and D'Haultfœuille (2021a) are not nested, and in principle one could combine both.

Finally, it is worth noting that de Chaisemartin and D'Haultfœuille (2021b) propose estimators for the case with several treatments. They propose both estimators that generalize the DID_{M} estimator in de Chaisemartin and D'Haultfœuille (2020) and rule out dynamic effects, and estimators that generalize those in Callaway and Sant'Anna (2021) and allow for dynamic effects.

4. APPLICATION

In this section, we revisit an application with a binary and staggered treatment, thus allowing us to compute several of the heterogeneity-robust DID estimators reviewed above. Between 1968 and 1988, 29 US states adopted a unilateral divorce law (UDL). Wolfers (2006a), building upon Friedberg (1998), studies the effects of those laws on divorce rates, using a version of the event-study regression in (2.13). We use his data (Wolfers, 2006b) to revisit this question. In what follows, estimates are weighted by states' populations and standard errors are clustered at the state level, as in Wolfers (2006a). As the author estimates UDLs' dynamic effects up to 15 years after adoption, in our replication we focus on heterogeneity-robust DID estimators allowing for dynamic effects, and present the estimated effects over the same horizon. We use Stata for this replication exercise, and the versions of the `twowayfeweights`, `eventstudyinteract`, `csdid`, `did_imputation`, and `did_multiplegt` commands available from the SSC repository at the end of April 2022.

Figure 3 below shows the instantaneous and dynamic effects of passing a UDL, according to six estimation methods. In the top-left panel, we show the estimates from the event-study regression in (2.13), with $L = 15$, $K = 10$, and endpoint binning. According to this regression, UDLs increase the divorce rate on the year when the law is passed and for seven years thereafter. 11 years after those laws are passed, their effect becomes significantly negative. Those effects are consistent with those in Column (1) of Table 2 of Wolfers (2006a). Our event-study regression and that in Wolfers (2006a) differ on two dimensions: Wolfers (2006a) does not include any placebo indicator for pre-adoption periods, and he includes post-adoption indicators for bins of two years (one indicator for the year when the law is passed and the year after that, one indicator for the second and third years after the law is passed, etc.). Results seem fairly robust to those specification choices. The placebo estimates are small, and individually and jointly insignificant (F-test p-value=0.863).

We follow Sun and Abraham (2021), and compute the weights attached to UDLs' instantaneous effect in this event-study regression.[16] As shown in Equation (2.14), this coefficient can be decomposed as the sum of two terms. The first term is a weighted sum of UDL's effects in the year when they are passed, across 27 states, where all effects receive a positive weight. The weights are negatively correlated with the year variable (correlation$=-0.232$), so this first term upweights UDLs' instantaneous effects in states passing a law early, and downweights UDLs' instantaneous effects in states passing a law late. Accordingly, this first term may differ from the average instantaneous effects of UDLs if those effects vary between early- and late-adopting states, but it at least estimates a convex combination of effects. The second term is a weighted sum of UDLs' effects in the years after they are passed. 29 effects of having passed a UDL a year ago enter in that second term. 16 enter with a positive weight, and 13 enter with a negative weight. The positive and negative weights respectively sum to 0.012 and -0.012. 28 effects of having passed a UDL two years ago enter in that second term. 10 effects enter with a positive weight, and 18 enter with a negative weight. The positive and negative weights respectively sum to 0.010 and -0.010. Effects of having passed a UDL three, four, ..., 14, and more than 15 years ago also enter in that second term. In total, the positive and negative weights in that second term respectively sum to around 0.064 and -0.064. If UDLs' dynamic effects vary across states, that second term may not be equal to zero, thus further biasing the estimated instantaneous effect in the event-study regression. However, those contamination weights are not very large, so this bias is likely to be small. Overall, this event-study regression seems fairly robust to heterogeneous treatment effects.

In the top-centre panel of Figure 3, we use the `eventstudyinteract` command to compute the estimators proposed by Sun and Abraham (2021). The estimated effects are very similar to those in the top-left panel. This could either be due to the fact that UDLs effects are not very heterogeneous, or to the fact that the event-study regression is fairly robust to heterogeneous treatment effects, as suggested above. Interestingly, the confidence intervals are, if anything, slightly wider in the top-left than in the top-centre panel of Figure 3, thus showing that heterogeneity-robust DID estimators are not always less precise than TWFE estimators. The placebos are individually insignificant. They are also substantially smaller than the estimated effects of UDLs: it does not seem that violations of parallel trends can fully account for those estimated effects.

In the top-right panel of Figure 3, we use the `csdid` command to compute the estima-

[16] In practice, we use the `twowayfeweights` Stata command, which has an option to compute the correlation between the weights and other variables that we use below.

tors proposed by Callaway and Sant'Anna (2021), using the "not-yet-treated" states as the control group. The estimated effects are very similar to those in the top-centre panel. 19 states never adopt a UDL over the period under consideration, so the group of "never-treated" states used as controls by `eventstudyinteract` is quite large, and accounts for a relatively large fraction of the group of "not-yet-treated" states used as controls by `csdid`. This may explain why in this application, the two commands yield very similar estimates. Using the larger control group of "not-yet-treated" states also does not lead to markedly more precise estimates: the widths of the confidence intervals are similar in the two panels. The placebos produced by `csdid` are small and individually insignificant. The placebos are much smaller in the top-right than in the top-centre panel. This is because `csdid` computes first-difference placebos, comparing the outcome evolution of treated and not-yet treated states, before the treated start receiving the treatment, and between pairs of consecutive periods.[17] On the other hand, `eventstudyinteract` computes long-difference placebos. For instance, the second placebo, shown at $t = -3$ on the graph, compares the outcome evolution of treated and never-treated states, from $F_g - 1$, the period before the treated start getting treated, to $F_g - 3$. See de Chaisemartin and D'Haultfœuille (2021a) for a discussion of the respective advantages of long- and first-difference placebos.

In the bottom-left panel of Figure 3, we use the `did_imputation` command to compute the estimators proposed by Borusyak et al. (2021). The effects are very similar to those found by the previous two estimators. The confidence interval of the instantaneous effect is much tighter in the bottom-left panel than in all other panels: for that treatment effect, the estimator proposed by Borusyak et al. (2021) does lead to a large precision gain. However, the opposite can hold when one considers dynamic effects. For instance, the confidence interval of the effect two years after passing a UDL is more than 50% larger per `did_imputation` than per `csdid`. Accordingly, the estimators proposed by Borusyak et al. (2021) do not always lead to precision gains, relative to those proposed by Sun and Abraham (2021) or Callaway and Sant'Anna (2021). The placebos produced by `did_imputation` are small, individually insignificant, and jointly insignificant (F-test p-value = 0.541).[18] Note that the placebos computed by `did_imputation` are different from those computed by the other commands. Essentially, the command estimates a TWFE regression among all the untreated (g, t), with K leads of the treatment. To be consistent with the other estimations, we run the command with 9 leads. Then, everything is relative to 10 periods prior to treatment, which is why the placebo estimate is set to 0 at $t = -10$ in the bottom-left panel, instead of at $t = -1$ in the other panels.

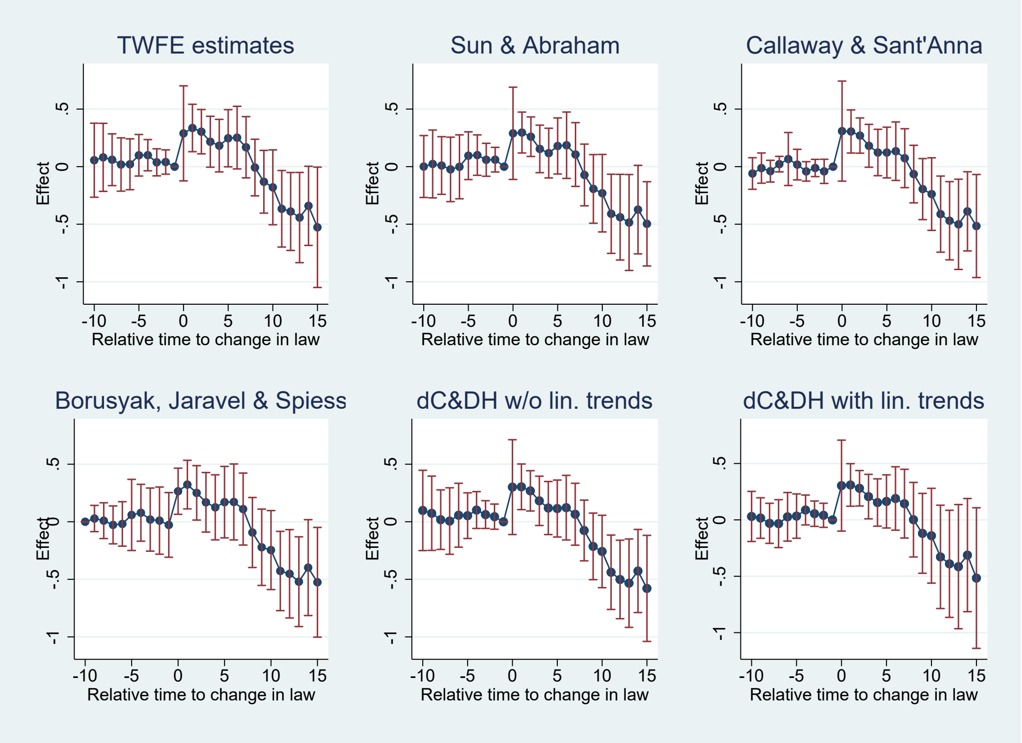
In the bottom-centre panel of Figure 3, we use the `did_multiplegt` command to compute the estimators proposed by de Chaisemartin and D'Haultfœuille (2021a). The resulting estimates are extremely close to those produced by the `csdid` command. The only reason why the two sets of estimates are not identical is that the estimation is weighted by states' population, and the two commands seem to handle weights slightly differently. Without weighting, the two sets of estimates are identical, as expected given that there are no covariates in the estimation and we used `csdid` with the not-yet-treated

[17] `csdid` has an option to compute long-difference placebos, but it returned an error when we used it.

[18] We did not report a joint test that all placebos are equal to 0 based on `eventstudyinteract`: this command does not readily allow to compute this test, as it does not return the covariances between the estimators. Similarly, `csdid` does not allow to jointly test if the placebos in Figure 3 are significant: it computes a joint nullity test, but for more disaggregated placebos.

as controls. The placebos computed by `did_multiplegt` are long-difference placebos, similar to those computed by `eventstudyinteract`, except that `did_multiplegt` uses the not-yet-treated as controls. They are small, and individually and jointly insignificant (F-test p-value = 0.427).

Figure 3. Effects of Unilateral Divorce Laws, using the data in Wolfers (2006a)



Note: This figure shows the estimated effects of Unilateral Divorce Laws on the divorce rate and placebo estimates, using the data in Wolfers (2006a) and six estimation methods. In the top-left panel, we show estimated effects per the event-study regression in (2.13), with $L = 15$, $K = 10$, and endpoint binning. In the top-centre (resp. top-right, bottom-left, bottom-centre) panel, we show estimated effects per the `eventstudyinteract` (resp. `csdid`, `did_imputation`, `did_multiplegt`) Stata command. In the bottom-right panel, we show estimated effects per the `did_multiplegt` Stata command, controlling for state-specific linear trends. All estimations are weighted by states' populations. Standard errors are clustered at the state level. 95% confidence intervals relying on a normal approximation are shown in red.

The estimates discussed so far do not control for state-specific linear trends. Whether such trends should or should not be included to estimate the effect of UDLs has been a debated issue in this literature, with Friedberg (1998) arguing in their favor, and Wolfers (2006a) arguing that they may conflate dynamic effects. The results presented so far already suggest that including state-specific linear trends is unnecessary, as placebos are small and insignificant without them. To confirm that, we run the `did_multiplegt` command again, controlling for state-specific linear trends.[19] The results, displayed in the bottom-right panel of Figure 3, show that results are fairly insensitive to the inclusion of state-linear trends. If anything, adding them makes the estimated long-run effects more noisy. The only argument in favor of state-specific trends is that the placebos are slightly smaller with them, though the difference is most likely insignificant.

Finally, to synthetize our results and obtain a point estimate that can be compared to the results in Wolfers (2006a), we average UDL's effects from the year the law is passed to seven years thereafter. The results are displayed in Table 1. We do not include therein the estimates from the `eventstudyinteract` and `csdid` commands, as one cannot readily obtain the standard error of this average effect from these commands. The results show that according to all estimation methods, UDLs positively affect the divorce rate from the year the law is passed to seven years thereafter. All estimates are fairly similar to each other and point towards an increase of 20%. The estimated standard error is substantially lower using the author's original specification, which is not surprising as it is less flexible than the other estimation methods. The estimated standard error is slightly larger using Borusyak et al. (2021) than the flexible event-study regression or the estimators proposed by de Chaisemartin and D'Haultfœuille (2021a).

Table 1. The short-run effects of Unilateral Divorce Laws

Wolfers (2006a)	0.200 (0.056)
Event-study without binning pairs of years	0.249 (0.106)
Borusyak et al. (2021)	0.198 (0.129)
de Chaisemartin and D'Haultfœuille (2021a), no linear trends	0.185 (0.107)
de Chaisemartin and D'Haultfœuille (2021a), linear trends	0.219 (0.096)

Note: This table shows the estimated effects of Unilateral Divorce Laws on the divorce rate, from 0 to 7 years after adoption, using the data in Wolfers (2006a). The first set of estimates is based on the regression in Column (2) of Table 2 of Wolfers (2006a). The second (resp. third, fourth) set of estimates is based on the results shown in the bottom-left (resp. bottom-centre, bottom-right) panel of Figure 3. All estimations are weighted by states' populations. Standard errors, clustered at the state level, are shown beneath each estimate, between parentheses.

5. 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. Table 2 below summarizes the alternative estimators available to applied researchers, depending on their research design and on whether they

[19] `csdid` does not allow for group-specific trends. `did_imputation` allows in principle for such trends but returned an error when such trends were added. `eventstudyinteract` allows for such trends.

are ready or not to rule out dynamic effects. The table shows that 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 (2021a) build upon those proposed by Callaway and Sant'Anna (2021) for the binary-and-staggered case. 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 under the parallel trends assumption, they may estimate the ATT if the weights attached to them are uncorrelated with the treatment effects $TE_{g,t}$, and they often have a lower variance than the heterogeneity-robust estimators reviewed in the previous section. While there are examples where TWFE and heterogeneity-robust DID estimators are economically and statistically different (see, e.g., the empirical examples in de Chaisemartin and D'Haultfœuille, 2020, 2021a,b; Baker et al., 2022), the previous section also shows a data set where TWFE and heterogeneity-robust DID estimators lead to very similar conclusions. Understanding the circumstances where TWFE and heterogeneity-robust DID estimators are more likely to differ is an important question. We conjecture that differences are likely to be larger in complicated designs (e.g.: a non-binary treatment that can turn on and off multiple times, or several treatments) than in simple designs (e.g.: a single binary and staggered treatment). This conjecture is based on our discussion of Equation (2.3) in Section 3. This is also a pattern we found when computing TWFE and heterogeneity-robust DID estimators in four different data sets, in the empirical examples of this survey and of de Chaisemartin and D'Haultfœuille (2020; 2021a; 2021b). But those examples are not enough to draw general conclusions: a systematic comparison of TWFE and heterogeneity-robust DID estimators in a broad set of applications is in order.

Analyzing estimators' robustness to heterogeneous treatment effects is important, as the assumption that all units are affected in the same way by a treatment is seldom credible. In this survey, we have focused on estimators relying on parallel trends assumptions, but this question is also relevant for other estimators. See for instance Słoczyński (2020) and Blandhol et al. (2022) for instrumental variables estimators with covariates. More closely related to our set-up, the impact of heterogeneous treatment effects in the "group fixed-effects" model of Bonhomme and Manresa (2015) remains to be studied.

Table 2. A summary of available heterogeneity-robust DID estimators

Panel A: Estimators ruling out dynamic effects
Can be used when outcome unaffected by past treatments

Treatment	*Estimators available*	*Stata commands*	*See:*
Binary	de Chaisemartin and D'Haultfœuille (2020)	`did_multiplegt`	3.1
	Imai and Kim (2021)		3.1
	Borusyak et al. (2021)	`did_imputation`	3.2.3
Discrete	de Chaisemartin and D'Haultfœuille (2020)	`did_multiplegt`	3.1
Continuous, with stayers	de Chaisemartin et al. (2022)	See Section 3.1	3.1
Continuous, without stayers	de Chaisemartin et al. (2022)	See Section 3.1	3.1
	Graham and Powell (2012)	`gmm`	3.1
	Chamberlain (1992)	`gmm`	3.1
Several treatments	de Chaisemartin and D'Haultfœuille (2021b)	`did_multiplegt`	3.3

Panel B: Estimators allowing dynamic effects
Can be used when outcome affected by past treatments

Treatment	*Estimators available*	*Stata commands*	*See:*
Binary and staggered	Callaway and Sant'Anna (2021)	`csdid`	3.2.1
	Sun and Abraham (2021)	`eventstudyinteract`	3.2.2
	Borusyak et al. (2021)	`did_imputation`	3.2.3
	de Chaisemartin and D'Haultfœuille (2021a)	`did_multiplegt`	3.3
Binary or discrete, non-staggered	de Chaisemartin and D'Haultfœuille (2021a)	`did_multiplegt`	3.3
Continuous and staggered	de Chaisemartin and D'Haultfœuille (2021a)	`did_multiplegt`	3.3
	Callaway et al. (2021)		3.3
Continuous and non-staggered, with stayers	de Chaisemartin et al. (2022)	See paper	3.3
Continuous and non-staggered, without stayers	No estimator available yet		
Several treatments	de Chaisemartin and D'Haultfœuille (2021b)	`did_multiplegt`	3.3

Note: All the Stata commands have R equivalents with the same name, except `eventstudyinteract` that does not have an R equivalent, and `csdid` whose R equivalent is called `did`. The table's last column indicates the section of the paper where the estimator is described.

ACKNOWLEDGEMENTS

We are very grateful to Francesco Armillei, Kirill Borusyak, Bruno Ferman, Xavier Jaravel, Roberto Ramos, Jonathan Roth, Jann Spiess, Gonzalo Vazquez-Bare, Kaspar Wüthrich, Jaap Abbring (the editor) and two anonymous referees for their helpful comments.

REFERENCES

Abadie, A. (2005, 01). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72*(1), 1–19.

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

Baker, A. C., D. F. Larcker, and C. C. Wang (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics 144*(2), 370–395.

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

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022). When is tsls actually late? NBER working paper 29709.

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

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–1184.

Borusyak, K. (2021, June). DID_IMPUTATION: Stata module to perform treatment effect estimation and pre-trend testing in event studies.

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

Borusyak, K., X. Jaravel, and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. arXiv preprint arXiv:2108.12419.

Butts, K. (2021, August). didimputation: Imputation Estimator from Borusyak, Jaravel, and Spiess (2021) in R.

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

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

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica 60*(3), 567–596.

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

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

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

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

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

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

de Chaisemartin, C., X. D'Haultfœuille, and A. Deeb (2019, February). twowayfeweights: Estimation of the Weights Attached to the Two-Way Fixed Effects Regressions in Stata.

de Chaisemartin, C., X. D'Haultfœuille, and Y. Guyonvarch (2019, May). did_multiplegt: DID Estimation with Multiple Groups and Periods in Stata.

de Chaisemartin, C., X. D'Haultfoeuille, F. Pasquier, and G. Vazquez-Bare (2022). Difference-in-differences estimators of the effect of a continuous treatment. arXiv preprint arXiv:2201.06898.

Flack, E. and Edward (2020, January). bacondecomp: Goodman-Bacon Decomposition in R.

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

Friedberg, L. (1998). Did unilateral divorce raise divorce rates? evidence from panel data. *The American Economic Review 88*(3), 608–627.

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

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

Gobillon, L. and T. Magnac (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., T. Goldring, and A. Nichols (2019, July). BACONDECOMP: Stata module to perform a Bacon decomposition of difference-in-differences estimation.

Graham, B. S. and J. L. Powell (2012). Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica 80*(5), 2105–2152.

Imai, K. and I. S. Kim (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.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American economic review 95*(1), 161–182.

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

Manski, C. F. and J. V. Pepper (2018). How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics 100*(2), 232–244.

Marcus, M. and P. H. Sant'Anna (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 J. Roth (2019). An honest approach to parallel trends. Working paper.

Rios-Avila, F., P. Sant'Anna, and B. Callaway (2021). Csdid: Stata module for the estimation of difference-in-difference models with multiple time periods.

Roth, J. (2021). Pre-test with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights forthcoming.*

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

Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe (2022). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. arXiv preprint arXiv:2201.01194.

Sant'Anna, P. and B. Callaway (2021, December). did: Treatment effects with multiple periods and groups in r.

Schmidheiny, K. and S. Siegloch (2020). On event studies and distributed-lags in two-way fixed effects models: Identification, equivalence, and generalization. ZEW Discussion Paper 20-01.

Słoczyński, T. (2020). When should we (not) interpret linear iv estimands as late? arXiv preprint arXiv:2011.06695.

Stevenson, B. and J. Wolfers (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, September). EVENTSTUDYWEIGHTS: Stata module to estimate the implied weights on the cohort-specific average treatment effects on the treated (CATTs) (event study specifications).

Sun, L. (2021, August). EVENTSTUDYINTERACT: Stata module to implement the interaction weighted estimator for an event study.

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

Wolfers, J. (2006a). Did unilateral divorce laws raise divorce rates? a reconciliation and new results. *American Economic Review 96*(5), 1802–1820.

Wolfers, J. (2006b). Replication data for: Did unilateral divorce laws raise divorce rates? a reconciliation and new results. Technical report, Nashville, TN: American Economic Association [publisher], 2006. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-12-07, https://doi.org/10.3886/E116250V1.

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

Zhang, S. and C. de Chaisemartin (2020, October). did_multiplegt: DID Estimation with Multiple Groups and Periods in R.

Zhang, S. and C. de Chaisemartin (2021, May). TwowayFEWeights: Estimation of the Weights Attached to the Two-Way Fixed Effects Regressions in R.