Two-Way Fixed Effects and Difference-in-Differences Estimators with Heterogeneous Treatment Effects and Imperfect Parallel Trends

Clément de Chaisemartin* Xavier D'Haultfœuille†

June 21, 2023

Textbook in progress, comments welcome.

*Economics Department, Sciences Po, clement.dechaisemartin@sciencespo.fr
†CREST-ENSAE, xavier.dhaultfoeuille@ensae.fr

Contents

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 (2021) 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, environmental sciences, and epidemiology.

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

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

DID relies on a parallel trends assumption: in the absence of the treatment, both groups would have experienced the same outcome evolution. Specifically, for every $g \in \{s, n\}$ and $t \in \{1, 2\}$, let $Y_{g,t}(0)$ and $Y_{g,t}(1)$ denote the potential outcomes in group g at period t without and with treatment, respectively (Neyman et al., 1990; Rubin, 1974) . 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)):

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

Where does the last equality come from?

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 the two-groups and two-periods design described above, DID is equal to the treatment coefficient in a TWFE regression. Motivated by this fact, 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, confident that there as well, TWFE was giving them an estimation method that only relied on a partly testable parallel trends assumption. Two recent strands of literature have shattered that confidence.

First, a recent strand of literature has shown that in more complicated designs than the one with two groups and two periods above, TWFE estimators are no longer equivalent to simple DID estimators, and no longer only rely on a parallel trends assumption. For those estimators to be unbiased for the treatment's effect, that 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 the quantitative social sciences relies on an often-implausible assumption has spurred a flurry of methodological papers. Some of them have diagnosed this issue and analyzed its origins. Other papers have proposed alternative estimators relying on parallel

trends conditions, like TWFE estimators, but robust to heterogeneous effects, unlike TWFE estimators. Hereafter, those alternative estimators are referred to as heterogeneity-robust DID estimators.

Second, in a recent paper, Roth (2022) has shown that tests of the parallel trends assumption often lack statistical power, and may fail to detect differential trends between treated and control locations that are often large enough to account for a significant share of the policy's estimated effect. This realization has spurred a growing interest among practitioners for a second strand of literature, that has proposed alternative estimation methods relying on weaker assumptions than parallel trends. Examples include estimators relying on a conditional parallel trends assumption (see, e.g., Abadie, 2005), estimators assuming bounded differential trends (see, e.g., Manski and Pepper, 2018; Rambachan and Roth, 2023), estimators assuming a factor model with interactive fixed effects (see, e.g., Bai, 2003) and synthetic control estimators (see, e.g., Abadie et al., 2010), and estimators assuming grouped patterns of heterogeneity (see,e.g., Bonhomme and Manresa, 2015).

This textbook aims to provide an overview of these two strands of literature, as well as other panel data methods routinely used for causal inference by practitionners, such as panel Bartik regressions (see, e.g., Goldsmith-Pinkham et al., 2020), or sequential randomized experiments (see, e.g., Bojinov et al., 2021). When available, the Stata and R commands implementing the diagnostics tools and alternative estimators discussed in this textbook are referenced, and the basic syntax of the Stata command is provided. We refer the reader to the commands' help files for further details on their syntax.

2 Data, notation, and assumptions

2.1 Group-level panel data

We seek to estimate the effect of a treatment on an outcome. For that purpose, we use a panel of G groups observed at T periods, respectively indexed by g and t. Typically, groups are geographical entities, like states, counties or municipalities, but a group could also just be

a single individual or firm. The group-level panel data may be constructed by aggregating an individual-level panel or repeated cross-section data set at the (g,t) level, defining groups, say, as individuals' county of birth. The group-level panel data may also be constructed from a single cross-section dataset, with cohort of birth playing the role of the time variable.[1] The estimators discussed below are not weighted by $N_{g,t}$, the population of cell (g,t). This is just to reduce notational complexity: studying weighted estimators is a mechanical extension. Throughout the textbook, we also assume that the group-level panel dataset is balanced: the outcome and treatment of each group is observed at every period. Again, this is mainly to reduce notational complexity.[2]

2.2 Treatment and potential outcomes

Treatment. When the treatment is assigned at the (g,t) level, as is the case when the treatment is a county-level law or regulation, like the minimum wage, $D_{g,t}$ denotes the treatment of group g at period t. When the treatment varies within (g,t) cells, a so-called fuzzy design (see de Chaisemartin and D'Haultfœuille, 2018) that may arise when groups are geographical entities and individuals or firms within the same cell may not all have the same treatment, $D_{g,t}$ denotes the average treatment in cell (g,t). Then, let $\mathcal{D}_t$ be the set of values $D_{g,t}$ can take at period t (i.e.: its support), let $\boldsymbol{D}_g = (D_{g,1},...,D_{g,T})$ be a $1 \times T$ vector stacking the treatments of group g from period 1 to T, and let $\boldsymbol{D} = (\boldsymbol{D}_1,...,\boldsymbol{D}_G)$ be a vector stacking the treatments of all groups at every period. $\boldsymbol{D}$ is referred to as the design of a study.

Potential outcomes. For all $(d_1,...,d_T) \in \mathcal{D}_1 \times ... \times \mathcal{D}_T$, let $Y_{g,t}(d_1,...,d_T)$ denote the potential outcome of group g at t if $(D_{g,1},...,D_{g,T}) = (d_1,...,d_T)$, and let $Y_{g,t} = Y_{g,t}(\boldsymbol{D}_g)$ denote the observed outcome of g at t. This dynamic potential outcome framework follows Robins (1986). It explicitly allows groups' outcome at t to depend on their past and future treatments.

[1] Some of the commands implementing the estimators reviewed in this textbook can be used with data at a more disaggregated level than the (g,t) level, see the commands' help files for further details.

[2] Some of the commands implementing the estimators reviewed in this textbook can be used with an imbalanced panel of groups, see the commands' help files for further details.

Sources of randomness. In most of this textbook, the study design $\boldsymbol{D}$ is implicitly conditioned upon. This means that our assumptions are made conditional on the design, and our results hold conditional on the design. Concretely, whenever there is an $E[X]$ below, it should actually be understood as $E[X|\boldsymbol{D}]$. Leaving this conditioning implicit greatly alleviates the notational burden. Conditioning on the design does not affect very much the results below that concern estimators' expectations. If $E\left[\widehat{\theta}\middle|\boldsymbol{D}\right] = \theta_0$, then by the law of iterated expectations, $E\left[\widehat{\theta}\right] = E[\theta_0]$: if $\widehat{\theta}$ is conditionally unbiased for θ_0, then $\widehat{\theta}$ is unconditionally unbiased for $E[\theta_0]$. Conditioning on the design does affect estimators' variances: $E\left[V\left[\widehat{\theta}\middle|\boldsymbol{D}\right]\right] \neq V\left[\widehat{\theta}\right]$, but estimators' variances is not a central issue in this textbook. Conditional on the design, groups' potential outcomes are the only source of randomness left. Probabilistic statements below are with respect to (wrt) their joint probability distribution. In the causal inference literature, there are three main ways of rationalizing why the data can be thought of as realizations as random variables: the units we observe are a random sample from a larger population; the treatment is randomly assigned to units; units' outcomes are affected by random shocks. This textbook adopts the third approach.

2.3 Identifying assumptions

No anticipation. Throughout the textbook, we maintain a no-anticipation assumption.

Assumption 1 *(No Anticipation) For all g and $(d_1, ..., d_T) \in \mathcal{D}_1 \times ... \times \mathcal{D}_T$, $Y_{g,t}(d_1, ..., d_T) = Y_{g,t}(d_1, ..., d_t)$.*

Assumption 1 requires that a group's current outcome do not depend on its future treatments, the so-called no-anticipation hypothesis. Abbring and Van den Berg (2003) have discussed it in duration models, and Malani and Reif (2015), Botosaru and Gutierrez (2018), Callaway and Sant'Anna (2021), and Sun and Abraham (2021) have discussed it in DID models. Assumption 1 is often testable, as we will see later.

No dynamic effects. Some of our results will also rely on the assumption that there are no dynamic effects.

Assumption 2 *(No Dynamic Effects) For all g and $(d_1, ..., d_t) \in \mathcal{D}_1 \times ... \times \mathcal{D}_t$, $Y_{g,t}(d_1, ..., d_t) = Y_{g,t}(d_t)$.*

Assumption 2 requires that a group's current outcome do not depend on its past treatments, the so-called no-dynamic-effects or no carry-over-effects hypothesis. Under Assumption 2 and with a binary treatment, each cell (g, t) as two potential outcomes: $Y_{g,t}(0)$ if g is untreated at t, and $Y_{g,t}(1)$ if g is treated at t. Then, we are back to the standard Neyman-Rubin model of potential outcomes.

Parallel trends. Recent papers have considered treatment effect estimation with panel data and random treatment timing (Athey and Imbens, 2022; Roth and Sant'Anna, 2021) or sequential randomization (Bojinov et al., 2021). Instead, in most of this textbook we will assume that groups are on parallel trends, which is typically weaker than assuming that treatment is randomly assigned. Our baseline parallel trends assumption is as follows. For all t, let $\mathbf{0}_t$ denote a vector of t zeros. $Y_{g,t}(\mathbf{0}_t)$ is the outcome of group g at t if it remains untreated from period 1 to t, hereafter referred to as the never-treated outcome of g at t.

Assumption 3 *(Parallel trends for the never-treated outcome) For all $t \geq 2$, $E\left[Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})\right]$ does not vary across g.*

Assumption 3 requires that every group experiences the same expected evolution of its never-treated potential outcome. Under Assumption 2, Assumption 3 reduces to

Assumption 4 *(Parallel trends for the untreated outcome) For all $t \geq 2$, $E\left[Y_{g,t}(0) - Y_{g,t-1}(0)\right]$ does not vary across g.*

Assumption 4 is the standard parallel trends assumption in classical DID models (see, e.g., Abadie, 2005). Thus, Assumption 3 is a generalization of this standard parallel trends assumption to potential outcome models allowing for dynamic effects. Assumption 3 was for instance considered by Callaway and Sant'Anna (2021) and Sun and Abraham (2021).

Independent groups.

Assumption 5 *(Independent groups) The vectors $(Y_{g,t}(d_1, ..., d_t))_{(d_1,...,d_t)\in\mathcal{D}_1\times...\times\mathcal{D}_t, 1\leq t\leq T}$ are mutually independent.*

Assumption 5 requires that the potential outcomes of different groups be independent, a commonly-made assumption in DID analysis, where standard errors are usually clustered at the group level, to account for the potential serial correlation of a group's potential outcomes over time (see Bertrand et al., 2004). Assumption 5 rules out common shocks affecting several groups. Thus, such common shocks are implicitly conditioned upon in our analysis.

3 TWFE estimators with heterogeneous treatment effects.

This section gives a general decomposition of TWFE estimators with heterogeneous treatment effects.

TWFE estimator. Let $\widehat{\beta}^{fe}$ denote the sample 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 $D_{g,t}$:

$$Y_{g,t} = \sum_{g'=1}^{G} \widehat{\alpha}_{g'} 1\{g = g'\} + \sum_{t'=1}^{T} \widehat{\gamma}_{t'} 1\{t = t'\} + \widehat{\beta}^{fe} D_{g,t} + \hat{\epsilon}_{g,t}, \tag{3}$$

where $\hat{\epsilon}_{g,t}$ denotes the regression residual. For instance, Gentzkow et al. (2011) use an 1868-1928 US county-level panel data set to test a conjecture in De Tocqueville (1850), that newspapers encourage citizens to participate more in democratic institutions. For that purpose, they let $Y_{g,t}$ denote the turnout rate in county g and the presidential election that took place in year t, they let $D_{g,t}$ denote the number of newspapers in county g and year t, and they run a regression closely related to (3), described in details below.

No dynamic effects. Throughout this section, we maintain Assumption 2 and assume that the treatment has no dynamic effects. This is consistent with the regression in (3), where the current treatment $D_{g,t}$ is one of the dependent variables, but the lagged treatments $D_{g,t-1}$, $D_{g,t-2}$ etc. are not part of the dependent variables. In Gentzkow et al. (2011), (3) implicitly assumes that the number of newspapers available in county g in previous elections no longer affects the turnout rate in election-year t. That turnout only depends on the number of newspapers available at t.

Target parameter. Under Assumption 2, for all (g,t) such that $D_{g,t} \neq 0$, let

$$\text{TE}_{g,t} = \frac{E\left[Y_{g,t}(D_{g,t}) - Y_{g,t}(0)\right]}{D_{g,t}}.$$

Interpret $\text{TE}_{g,t}$, in general and in the context of Gentzkow et al. (2011).

$\text{TE}_{g,t}$ denotes the expected effect in cell (g,t) of moving the treatment from 0 to $D_{g,t}$, scaled by $D_{g,t}$. In other words, $\text{TE}_{g,t}$ is the slope of (g,t)'s potential outcome function, from 0 to its actual treatment $D_{g,t}$. In Gentzkow et al. (2011), $\text{TE}_{g,t}$ is the difference between the actual turnout rate in county g and year t and its counterfactual turnout rate without any newspaper, divided by its number of newspapers. Thus, $\text{TE}_{g,t}$ can be interpreted as an effect per newspaper. If the treatment is binary, for all (g,t) such that $D_{g,t} = 1$, $\text{TE}_{g,t} = E\left[Y_{g,t}(1) - Y_{g,t}(0)\right]$, the ATE in cell (g,t). Let N_1 denote the number of cells such that $D_{g,t} \neq 0$, namely the number of treated (g,t) cells. A natural target parameter is

$$\text{ATT} = \frac{1}{N_1} \sum_{(g,t):D_{g,t}\neq 0} \text{TE}_{g,t},$$

the average treatment effect across treated cells (ATT). In Gentzkow et al. (2011), ATT is the average turnout increase produced per newspaper, across all county×election-year with at least one newspaper.

Theorem 1 in de Chaisemartin and D'Haultfœuille (2020) Theorem 1 in de Chaisemartin and D'Haultfœuille (2020), restated below, shows that under a parallel trends assumption on the potential outcome without treatment $Y_{g,t}(0)$, $\widehat{\beta}^{fe}$ is unbiased for a weighted sum of the $\text{TE}_{g,t}$s, that may differ from ATT. Let

$$W_{g,t} = \frac{\hat{u}_{g,t} D_{g,t}}{\frac{1}{N_1} \sum_{(g',t'):D_{g',t'}\neq 0} \hat{u}_{g',t'} D_{g',t'}},$$

where $\hat{u}_{g,t}$ denotes the sample residual from a regression of $D_{g,t}$ on group and period fixed effects. In Gentzkow et al. (2011), $\hat{u}_{g,t}$ is the residual from a regression of the number of newspapers in county g and year t on county and year fixed effects.

Theorem 1 *If Assumptions 1, 2, and 4 hold,*

$$E\left[\widehat{\beta}^{fe}\right] = \frac{1}{N_1} \sum_{(g,t):D_{g,t}\neq 0} W_{g,t}\, TE_{g,t}. \tag{4}$$

Interpret Theorem 1, in general and in the context of Gentzkow et al. (2011).

Theorem 1 says that $\widehat{\beta}^{fe}$ is unbiased for a weighted sum of the treatment effects $\text{TE}_{g,t}$, across all treated (g,t) cells, and where the treatment effect of cell (g,t) receives a weight equal to $\frac{W_{g,t}}{N_1}$. In Gentzkow et al. (2011), $\widehat{\beta}^{fe}$ is unbiased for a weighted sum of the effects of newspapers on turnout across all county×year cells with at least one newspaper, and where the effect of newspapers in cell (g,t) receives a weight equal to $\frac{W_{g,t}}{N_1}$.

What is the value of $\sum_{(g,t):D_{g,t}\neq 0} \frac{W_{g,t}}{N_1}$?

It directly follows from the definition of $W_{g,t}$ that $\sum_{(g,t):D_{g,t}\neq 0} \frac{W_{g,t}}{N_1} = 1$. Therefore, $\widehat{\beta}^{fe}$ is unbiased for a weighted sum of the treatment effects $\text{TE}_{g,t}$, across all treated (g,t) cells, with weights summing to one.

Proof of Theorem 1 It follows from Assumption 4 that $E\left[Y_{g,t}(0) - Y_{g,1}(0)\right]$ is constant across g. Then, let $\gamma_t = E\left[Y_{g,t}(0) - Y_{g,1}(0)\right]$, and let $\alpha_g = E\left[Y_{g,1}(0)\right]$. One has that

$$E\left[Y_{g,t}(0)\right] = E\left[Y_{g,1}(0)\right] + E\left[Y_{g,t}(0) - Y_{g,1}(0)\right] = \alpha_g + \gamma_t. \tag{5}$$

Moreover,

$$\begin{aligned} E\left[Y_{g,t}\right] =& E\left[Y_{g,t}(0)\right] + E\left[Y_{g,t}(D_{g,t}) - Y_{g,t}(0)\right] \\ =& E\left[Y_{g,t}(0)\right] + D_{g,t} E\left[Y_{g,t}(D_{g,t}) - Y_{g,t}(0)\right] / D_{g,t} \\ =& E\left[Y_{g,t}(0)\right] + D_{g,t}\text{TE}_{g,t}, \end{aligned} \tag{6}$$

with the convention that $0/0 = 0$. Then, it follows from the Frisch-Waugh-Lovell theorem that

$$\widehat{\beta}^{fe} = \frac{\sum_{g,t} \hat{u}_{g,t} Y_{g,t}}{\sum_{g,t} \hat{u}_{g,t}^2} = \frac{\sum_{g,t} \hat{u}_{g,t} Y_{g,t}}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}}, \tag{7}$$

where the second equality follows from the regression decomposition of $D_{g,t}$ wrt to group and time fixed effects $(1\{g = g'\})_{g' \in \{1,\ldots,G\}}$ and $(1\{t = t'\})_{t' \in \{1,\ldots,T\}}$:

$$D_{g,t} = \sum_{g'=1}^{G} 1\{g = g'\}\widehat{\theta}_{g'} + \sum_{t'=1}^{T} 1\{t = t'\}\widehat{\theta}_{t'} + \hat{u}_{g,t}, \tag{8}$$

and the fact that by the first-order conditions attached to an OLS regression, $\hat{u}_{g,t}$ is uncorrelated to all the group and time fixed effects: for all g',

$$\sum_{g,t} \hat{u}_{g,t} 1\{g = g'\} = 0 \Leftrightarrow \sum_{t=1}^{T} \hat{u}_{g',t} = 0, \tag{9}$$

and for all t',

$$\sum_{g,t} \hat{u}_{g,t} 1\{t = t'\} = 0 \Leftrightarrow \sum_{g=1}^{G} \hat{u}_{g,t'} = 0. \tag{10}$$

Finally,

$$\begin{aligned}
E\left[\widehat{\beta}^{fe}\right] =& \frac{\sum_{g,t} \hat{u}_{g,t} E\left[Y_{g,t}\right]}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} \\
=& \frac{\sum_{g,t} \hat{u}_{g,t} (E\left[Y_{g,t}(0)\right] + D_{g,t}\mathrm{TE}_{g,t})}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} \\
=& \frac{\sum_{g,t} \hat{u}_{g,t} \alpha_g}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} + \frac{\sum_{g,t} \hat{u}_{g,t} \gamma_t}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} + \frac{\sum_{g,t} \hat{u}_{g,t} D_{g,t} \mathrm{TE}_{g,t}}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} \\
=& \frac{\sum_{g=1}^{G} \alpha_g \sum_{t=1}^{T} \hat{u}_{g,t}}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} + \frac{\sum_{t=1}^{T} \gamma_t \sum_{g=1}^{G} \hat{u}_{g,t}}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} + \frac{\sum_{g,t} \hat{u}_{g,t} D_{g,t} \mathrm{TE}_{g,t}}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} \\
=& \frac{\sum_{(g,t):D_{g,t} \neq 0} \hat{u}_{g,t} D_{g,t} \mathrm{TE}_{g,t}}{\sum_{(g,t):D_{g,t} \neq 0} \hat{u}_{g,t} D_{g,t}} \\
=& \frac{1}{N_1} \sum_{(g,t):D_{g,t} \neq 0} W_{g,t} \mathrm{TE}_{g,t}.
\end{aligned}$$

The first equality follows from (7) and the fact the design is conditioned upon, the second equality follows from (6), the third equality follows from (5), the fifth equality follows from (9) and (10), the last equality follows from the definition of $W_{g,t}$ **QED.**

According to Theorem 1, do we have that $\widehat{\beta}^{fe}$ is biased or unbiased for the ATT?

$\widehat{\beta}^{fe}$ may be biased for the ATT. One can show that

$$\hat{u}_{g,t} = D_{g,t} - D_{g,.} - D_{.,t} + D_{.,.}, \tag{11}$$

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. Therefore, while the average of $W_{g,t}$ across all treated (g,t) is equal to one, $W_{g,t}$ may not be equal to one for all treated (g,t). Thus, $\widehat{\beta}^{fe}$ may give a weight larger than $\frac{1}{N_1}$ to the treatment effect of some (g,t) cells, and a weight lower than $\frac{1}{N_1}$ to the treatment effect of other cells. Then, $\widehat{\beta}^{fe}$ may be biased for the ATT, if the treatment effects of cells down- and up-weighted by $\widehat{\beta}^{fe}$ differ.

Try to find an assumption such that adding that assumption to those in Theorem 1, one gets that $\widehat{\beta}^{fe}$ is unbiased for the ATT.

$\widehat{\beta}^{fe}$ is unbiased for the average treatment effect on the treated if the weights are uncorrelated with cells treatment effects. $E\left[\widehat{\beta}^{fe}\right] = \text{ATT}$ if on top of the assumptions underlying Theorem 1, one is ready to further assume that

$$\frac{1}{N_1} \sum_{(g,t):D_{g,t}\neq 0} (W_{g,t} - 1)(\text{TE}_{g,t} - \text{ATT}) = 0, \tag{12}$$

meaning that $W_{g,t}$ is uncorrelated with $\text{TE}_{g,t}$. Then, the treatment effects that are up- and down-weighted by $\widehat{\beta}^{fe}$ do not systematically differ, and one can show that $\widehat{\beta}^{fe}$ is unbiased for ATT (see Corollary 2 in de Chaisemartin and D'Haultfœuille, 2020).

Do you think that the no-correlation condition in (12) is often likely to hold?

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. For instance, in a Roy selection model, groups with the largest average treatment may be those with the largest treatment effect. Similarly, $D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot}$ is decreasing in $D_{\cdot,t}$, and the treatment effects at time periods with the highest average treatment may also systematically differ from the treatment effects at time periods where the average treatment is lower.

Is (12) testable? If not, can you think of a way to suggestively test it?

Testing (12) would require observing $W_{g,t}$ and $TE_{g,t}$, so as to compute their correlation. However, $W_{g,t}$ is observed, but $TE_{g,t}$ is not. Still, (12) can be suggestively tested, if one observes a proxy variable $P_{g,t}$ likely to be correlated with $\text{TE}_{g,t}$. Then, one can test if $W_{g,t}$ and $P_{g,t}$ are correlated.

$\widehat{\beta}^{fe}$ is unbiased for the average treatment effect on the treated if the treatment effect is constant. A special case of this "no-correlation" condition is if the treatment effect is constant, i.e. $\text{TE}_{g,t} = \delta$ for all (g,t). Then, it directly follows from Equation (4) and the fact that $\frac{1}{N_1}\sum_{(g,t):D_{g,t}\neq 0} W_{g,t} = 1$ that $E\left[\widehat{\beta}^{fe}\right] = \delta$. However, constant effect is often an implausible assumption.

Are the weights $W_{g,t}/N_1$ positive for all (g,t)?

$\widehat{\beta}^{fe}$ may not satisfy the no-sign reversal property. Equation (11) 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 the weights $W_{g,t}$ are negative: when all those weights are positive, $\widehat{\beta}^{fe}$ does satisfy the no-sign reversal property.

No-sign reversal is connected to the economic concept of Pareto efficiency. 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.

Stata and R commands to compute the weights attached to any TWFE regression. The `twowayfeweights` Stata (see de Chaisemartin, D'Haultfœuille and Deeb, 2019) and R (see Zhang and de Chaisemartin, 2021) commands compute the weights $W_{g,t}/N_1$ in (4). The basic syntax of the Stata command is:

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

Decompositions of other TWFE estimators. de Chaisemartin and D'Haultfœuille (2020) show that $\widehat{\beta}^{fd}$, the treatment's coefficient in a regression of the outcome's first difference on the treatment's first difference and period fixed effects, can also be decomposed as a weighted sum of $TE_{g,t}$ under Assumptions 1, 2, and 4, with weights that differ from those in (4) but that also sum to one and that may also be negative. This implies that under constant treatment effects, the expectations of $\widehat{\beta}^{fe}$ and $\widehat{\beta}^{fd}$ are equal. Accordingly, if the two coefficients signifi-

cantly differ, under Assumptions 1, 2, and 4 one can reject the null that the treatment effect is constant. de Chaisemartin and D'Haultfœuille (2020) also derive a decomposition similar to (4) for TWFE regressions with control variables. Finally, they also derive decompositions similar to (4), 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.

Application. de Chaisemartin and D'Haultfœuille (2020) use the `twowayfeweights` Stata command to revisit Gentzkow et al. (2011). The authors do not compute $\widehat{\beta}^{fe}$. Instead, they 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, thus computing $\widehat{\beta}^{fd}$. 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-year 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. Finally, $\widehat{\beta}^{fd} - \widehat{\beta}^{fe}$ is significantly different from 0 (t-stat=2.86): under Assumptions 1, 2, and 4 one can reject the null that the effect of newspapers on turnout is constant across counties and over time.

Related results in other papers. The decomposition in (4) 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): to our knowledge, those are the first decompositions of TWFE regression coefficients as weighted sums of treatment effects. 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 (4).[3]

Next steps. Theorem 1 in de Chaisemartin and D'Haultfœuille (2020) is a general result, that applies to any research design. It shows that under the standard parallel trends assumption invoked to justify DID estimators, TWFE estimators may not be robust to heterogeneous treatment effects. The result relies on a no-dynamic-effect assumption that is rather strong, but that assumption is motivated by the TWFE specification considered in this section, and intuitively, it seems likely that relaxing this assumption would not yield more positive results for TWFE regressions. In Gentzkow et al. (2011), we find that TWFE regressions estimate highly non-convex combinations of treatment effects. This shows that the problem can be serious, but is it always serious? And when it is, is it possible to propose estimators relying on a parallel trends assumption, like TWFE estimators, but robust to heterogeneous effects? The answers to those questions are going to be design-specific, so we will now consider various commonly found research designs. In each design, we will assess if TWFE estimators are robust to heterogeneous effects. When they are not, we will try to understand intuitively where their lack of robustness stems from, in order to propose robust estimators.

4 Designs with a binary treatment, and no variation in treatment timing.

Throughout this section, we assume that the design is a classical DID design, with a binary treatment and no variation in treatment timing, as in, say, Card and Krueger (1994), who use the fact that New Jersey increased its minimum wage on April 1st 1992 while the neighboring state of Pennsylvania kept its minimum wage unchanged, to study the effect of the minimum wage on fast-food employment.

[3]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.

Design 1 *(Binary treatment, no variation in treatment timing)* $D_{g,t} = 1\{t \geq F\}T_g$, *with* $F \geq 2$, $T_g \in \{0,1\}$ *for all* g, *and* $\min_g T_g = 0$ *and* $\max_g T_g = 1$.

T_g is an indicator equal to 1 for treatment groups (e.g. New Jersey fast-foods in Card and Krueger, 1994), that all become treated at period F (e.g. April 1st 1992 in Card and Krueger, 1994), while T_g is equal to 0 for control groups that never become treated (e.g. Pennsylvania fast-foods in Card and Krueger, 1994). We require that $F \geq 2$, and that there is at least one treatment and one control group, as otherwise the TWFE regression is not identified. One can directly verify from the data whether the conditions in Design 1 are met. Let $G_1 = \sum_{g=1}^{G} T_g$ denote the number of treated groups, and let $G_0 = G - G_1$ denote the number of control groups. Let $T_1 = T - F + 1$ denote the number of treated periods, and let $T_0 = T - T_1$ denote the number of control periods. One has $N_1 = G_1 T_1$: the number of treated cells is equal to the number of treated groups times the number of treated periods.

4.1 In Design 1, $\widehat{\beta}^{fe}$ is unbiased for ATT.

Applying Theorem 1 in Design 1. In Design 1, it follows from (11) that for all (g,t),

$$\hat{u}_{g,t} = D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot} = 1\{t \geq F\}T_g - \frac{T_1}{T}T_g - \frac{G_1}{G}1\{t \geq F\} + \frac{G_1 \times T_1}{G \times T}. \tag{13}$$

Therefore, $W_{g,t}$ is constant across all (g,t) such that $D_{g,t} = 1$. As the average of $W_{g,t}$ across those (g,t)s is equal to 1, $W_{g,t} = 1$ for all those (g,t)s. Then, it directly follows from Theorem 1 that under Assumptions 1, 2, and 4, $\widehat{\beta}^{fe}$ is unbiased for the ATT:

$$E\left[\widehat{\beta}^{fe}\right] = \text{ATT}.$$

In Design 1, $\widehat{\beta}^{fe}$ is a simple DID estimator. Notice that

$$1 - \frac{T_1}{T} - \frac{G_1}{G} + \frac{G_1 \times T_1}{G \times T} = \frac{T_0}{T} - \frac{G_1 \times T_0}{G \times T} = \frac{G_0 \times T_0}{G \times T}. \tag{14}$$

Then, plugging (13) and (14) into (7),

$$
\begin{aligned}
&\widehat{\beta}^{fe} \\
=&\frac{\sum_{g,t} \hat{u}_{g,t} Y_{g,t}}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} \\
=&\frac{\sum_{g,t} \left(1\{t \geq F\} T_g - \frac{T_1}{T} T_g - \frac{G_1}{G} 1\{t \geq F\} + \frac{G_1 \times T_1}{G \times T}\right) Y_{g,t}}{\sum_{(g,t):D_{g,t}=1} \left(1 - \frac{T_1}{T} - \frac{G_1}{G} + \frac{G_1 \times T_1}{G \times T}\right)} \\
=&\frac{1}{G_1 T_1 \frac{G_0 T_0}{GT}} \left(\frac{G_0 T_0}{GT} \sum_{g:T_g=1, t \geq F} Y_{g,t} - \frac{G_0 T_1}{GT} \sum_{g:T_g=1, t<F} Y_{g,t} - \frac{G_1 T_0}{GT} \sum_{g:T_g=0, t \geq F} Y_{g,t} + \frac{G_1 T_1}{GT} \sum_{g:T_g=0, t<F} Y_{g,t} \right) \\
=&\frac{1}{G_1 T_1} \sum_{g:T_g=1, t \geq F} Y_{g,t} - \frac{1}{G_1 T_0} \sum_{g:T_g=1, t<F} Y_{g,t} - \frac{1}{G_0 T_1} \sum_{g:T_g=0, t \geq F} Y_{g,t} + \frac{1}{G_0 T_0} \sum_{g:T_g=0, t<F} Y_{g,t}. \qquad (15)
\end{aligned}
$$

(15) shows that in designs with a binary treatment and no variation in treatment timing, $\widehat{\beta}^{fe}$ is a simple DID estimator comparing the average outcome evolution, before and after F, in treatment and control groups. This DID estimator could also have been obtained from a simpler regression of $Y_{g,t}$ on the treatment-group indicator T_g, the post-treatment indicator $1\{t \geq F\}$, and $D_{g,t}$, the interaction of the two indicators. This shows that in Design 1, adding all group and period fixed effects yields the same treatment coefficient as just having the treatment-group and post-treatment indicators.

Using (15) **to show that $\widehat{\beta}^{fe}$ is unbiased for the ATT.** In a simple design like Design 1, there are more direct ways of showing that $\widehat{\beta}^{fe}$ is unbiased for the ATT than invoking Theorem 1. One can for instance leverage (15). To avoid repeating twice the same result, we will prove it under Assumption 3, and without ruling out dynamic effects. For all t, let $\mathbf{1}_t$ denote a vector of t ones. For all (g,t) such that $T_g = 1$ and $t \geq F$, let

$$\text{TE}^{\text{dyn}}_{g,t} = E\left[Y_{g,t}(\mathbf{0}_{F-1}, \mathbf{1}_{t-F+1}) - Y_{g,t}(\mathbf{0}_t)\right]$$

denote the expected effect in cell (g,t) of having been treated rather than untreated for $t-F+1$ periods, from F to t. A natural target parameter is

$$\text{ATT}_{\text{dyn}} = \frac{1}{G_1 T_1} \sum_{(g,t):D_{g,t}=1} \text{TE}^{\text{dyn}}_{g,t},$$

the average treatment effect across treated cells (ATT).

Theorem 2 *In Design 1, if Assumptions 1 and 3 hold,*

$$E\left[\widehat{\beta}^{fe}\right] = ATT_{dyn}. \tag{16}$$

Theorem 2 shows that in Design 1, $\widehat{\beta}^{fe}$ remains unbiased for the ATT if one allows for dynamic effects. Doing so just slightly changes the interpretation of the ATT, which now becomes an average effect of having been treated for $t - F + 1$ periods, across all treated (g, t) cells. On average, those cells have been treated for $\frac{T-F}{2} + 1$ periods, a number that may help interpret ATT_{dyn}.

Proof of Theorem 2 It follows from Assumption 3 that $E\left[Y_{g,t}(\mathbf{0}_t) - Y_{g,1}(\mathbf{0}_1)\right]$ is constant across g. Then, let $\gamma_t = E\left[Y_{g,t}(\mathbf{0}_t) - Y_{g,1}(\mathbf{0}_1)\right]$, and let $\alpha_g = E\left[Y_{g,1}(\mathbf{0}_1)\right]$. One has that

$$E\left[Y_{g,t}(\mathbf{0}_t)\right] = \alpha_g + \gamma_t. \tag{17}$$

Moreover,

$$E\left[Y_{g,t}\right] = E\left[Y_{g,t}(\mathbf{0}_t)\right] + D_{g,t}E\left[Y_{g,t} - Y_{g,t}(\mathbf{0}_t)\right] = E\left[Y_{g,t}(\mathbf{0}_t)\right] + D_{g,t}\text{TE}^{\text{dyn}}_{g,t}. \tag{18}$$

Then,

$$\begin{aligned}
&E\left[\widehat{\beta}^{fe}\right] \\
=&\frac{1}{G_1T_1}\sum_{g:T_g=1,t\geq F} E\left[Y_{g,t}\right] - \frac{1}{G_1T_0}\sum_{g:T_g=1,t<F} E\left[Y_{g,t}\right] - \frac{1}{G_0T_1}\sum_{g:T_g=0,t\geq F} E\left[Y_{g,t}\right] + \frac{1}{G_0T_0}\sum_{g:T_g=0,t<F} E\left[Y_{g,t}\right] \\
=&\text{ATT}_{\text{dyn}} \\
+&\frac{1}{G_1T_1}\sum_{g:T_g=1,t\geq F} (\alpha_g + \gamma_t) - \frac{1}{G_1T_0}\sum_{g:T_g=1,t<F} (\alpha_g + \gamma_t) - \frac{1}{G_0T_1}\sum_{g:T_g=0,t\geq F} (\alpha_g + \gamma_t) + \frac{1}{G_0T_0}\sum_{g:T_g=0,t<F} (\alpha_g + \gamma_t) \\
=&\text{ATT}_{\text{dyn}}.
\end{aligned}$$

The first equality follows from (15) and the fact the design is conditioned upon. The second equality follows from (18), (17), Design 1, and the definition of ATT_{dyn}. The third equality follows after some algebra **QED.**

4.2 Estimating dynamic effects and testing the parallel trends assumption.

TWFE event-study regression in designs with a binary treatment and no variation in treatment timing. In Design 1, researchers have often estimated the following regression:

$$Y_{g,t} = \widehat{\alpha}_0 + \widehat{\alpha}_1 T_g + \sum_{t' \neq F-1} \widehat{\gamma}_{t'} 1\{t = t'\} + \sum_{t' \neq F-1} \widehat{\beta}^{fe}_{t'} 1\{t = t'\} T_g + \hat{\epsilon}_{g,t}. \tag{19}$$

(19) is often referred to as a TWFE event-study (ES) regression, and researchers often plot its coefficients $(\widehat{\beta}^{fe}_t)_{t \neq F-1}$ on a so-called ES graph, with $(t - F + 1)$, the relative time to treatment onset for the treated groups,[4] on the $x-$axis. As the regression is saturated in $T_g \times$ time fixed effects, it is easy to show that for all $t \neq F - 1$,

$$\widehat{\beta}^{fe}_t = \frac{1}{G_1} \sum_{g:T_g=1} (Y_{g,t} - Y_{g,F-1}) - \frac{1}{G_0} \sum_{g:T_g=0} (Y_{g,t} - Y_{g,F-1}), \tag{20}$$

a simple DID comparing the $F - 1$ to t outcome evolution in treatment and control groups. All DIDs are relative to $F - 1$, the period prior to the treatment onset for the treated groups, and the omitted time period in (19). Accordingly, for $t \geq F$, the DID in (20) goes from the past to the future, but for $t \leq F-2$, the DID goes from the future to the past. For $\ell \in \{-1, ..., -F+2\}$, $\widehat{\beta}^{fe}_{F-1+\ell}$ is often referred to as a placebo estimator. Intuitively, it assesses if before the treatment onset, treatment and control groups were on parallel trends. As we will see below, this placebo estimator can be used to formally test Assumptions 1 and 3.

Dynamic treatment effects: target parameters. For any $\ell \in \{1, ..., T - F + 1\}$, let

$$\text{ATT}_\ell = \frac{1}{G_1} \sum_{g:T_g=1} E\left[Y_{g,F-1+\ell}(\mathbf{0}_{F-1}, \mathbf{1}_\ell) - Y_{g,F-1+\ell}(\mathbf{0}_{F-1+\ell})\right].$$

At period $F - 1 + \ell$, treated groups have been treated for ℓ periods, so ATT_ℓ is the average effect of having been treated for ℓ periods, across all treated groups and at period $F - 1 + \ell$. Can $\ell \mapsto \text{ATT}_\ell$ be used to determine if past treatments affect the outcome? For instance, if

[4] Researchers usually rather define relative time as $(t - F)$, with treatment onset corresponding to relative time 0 rather than to relative time 1. We prefer to define relative time as $(t - F + 1)$, to ensure that the ES graph is normalized at 0, and that effects and placebos are shown symetrically around 0, with the first effect at 1 and the first placebo at -1.

$ATT_2 > ATT_1 > 0$, can we conclude that the first treatment lag has an effect on the current outcome?

Without any restriction on treatment effect heterogeneity, $\ell \mapsto \text{ATT}_\ell$ cannot be used to determine if past treatments affect the outcome. For instance, if $ATT_2 > ATT_1 > 0$, it may be tempting to conclude that being treated for two periods has a larger effect than being treated for one period, thus implying that the first treatment lag has a positive effect on the current outcome. However, one may have that $ATT_2 > ATT_1 > 0$ even if the first treatment lag has no effect on the outcome, if the effect of the current treatment on the outcome is larger at period $F+1$ than at period F. Mathematically,

$$\begin{aligned}
&\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F+1}(\mathbf{0}_{F-1},1,1) - Y_{g,F+1}(\mathbf{0}_{F+1})\right] \\
=&\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F+1}(\mathbf{0}_{F-1},1,0) - Y_{g,F+1}(\mathbf{0}_{F+1})\right] + \frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F+1}(\mathbf{0}_{F-1},1,1) - Y_{g,F+1}(\mathbf{0}_{F-1},1,0)\right],
\end{aligned}$$

so

$$\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F+1}(\mathbf{0}_{F-1},1,1) - Y_{g,F+1}(\mathbf{0}_{F+1})\right] > \frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F}(\mathbf{0}_{F-1},1) - Y_{g,F}(\mathbf{0}_F)\right]$$

implies that either

$$\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F+1}(\mathbf{0}_{F-1},1,0) - Y_{g,F+1}(\mathbf{0}_{F+1})\right] > 0,$$

meaning that the first treatment lag has a positive effect on the current outcome, or

$$\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F+1}(\mathbf{0}_{F-1},1,1) - Y_{g,F+1}(\mathbf{0}_{F-1},1,0)\right] > \frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F}(\mathbf{0}_{F-1},1) - Y_{g,F}(\mathbf{0}_F)\right],$$

meaning that the effect of the current treatment is larger at $F+1$ than at F. If one assumes that the effects of the current treatment and of its lags are additively separable, meaning that the current treatment and its lags are neither complements or substitutes, and constant over time, then

$$\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F+1}(\mathbf{0}_{F-1},1,1) - Y_{g,F+1}(\mathbf{0}_{F-1},1,0)\right] = \frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F}(\mathbf{0}_{F-1},1) - Y_{g,F}(\mathbf{0}_F)\right],$$

so $ATT_2 > ATT_1 > 0$ implies that the first-treatment lag has an effect on the current outcome. More generally, it is only if one assumes additively separable and constant effects over time that $\ell \mapsto \text{ATT}_\ell$ can be used to test for the presence of dynamic effects, and to separately estimate the effect of the current treatment and of its lags on the outcome.

Using the TWFE event-study regression to estimate dynamic effects and test Assumptions 1 and 3.

Theorem 3 *In Design 1, if Assumptions 1 and 3 hold, then for all* $\ell \in \{1, ..., T - F + 1\}$,

$$E\left[\widehat{\beta}^{fe}_{F-1+\ell}\right] = ATT_\ell. \tag{21}$$

and if $F \geq 3$, *for all* $\ell \in \{-1, ..., -F + 2\}$

$$E\left[\widehat{\beta}^{fe}_{F-1+\ell}\right] = 0. \tag{22}$$

Equation (21) shows that in Design 1, for all $\ell \in \{1, ..., T - F + 1\}$, $\widehat{\beta}^{fe}_{F-1+\ell}$ is unbiased for ATT_ℓ.

Based on Theorem 3, how can one test Assumptions 1 and 3?

Equation (22) shows that if $F \geq 3$, Assumptions 1 and 3 are testable, as they imply that $E\left[\widehat{\beta}^{fe}_{F-1+\ell}\right] = 0$ for all $\ell \in \{-1, ..., -F+2\}$. Therefore, if we can reject the null that $E\left[\widehat{\beta}^{fe}_{F-1+\ell}\right] = 0$ for all $\ell \in \{-1, ..., -F + 2\}$, we can reject the null that Assumptions 1 and 3 both hold.

If $F \geq 4$, $E\left[\widehat{\beta}^{fe}_{F-1+\ell}\right] = 0$ for all $\ell \in \{-1, ..., -F+2\}$ is a null hypothesis on a vector of dimension $F - 2 > 1$. Then, would testing separately that $E\left[\widehat{\beta}^{fe}_{F-1+\ell}\right] = 0$ for all $\ell \in \{-1, ..., -F+2\}$ yield a valid test of Assumptions 1 and 3?

No. This would give rise to a multiple hypothesis testing problem. To account for that, one can

adjust p-values for multiple testing, using, say, a Bonferroni adjustment, but this may lead to a test with low power. Or one can run an F-test that $E\left[\widehat{\beta}^{fe}_{F-1+\ell}\right]=0$ for all $\ell \in \{-1,...,-F+2\}$.

Proof of Theorem 3 For $\ell \in \{1,...,T-F+1\}$,

$$
\begin{aligned}
&E\left[\widehat{\beta}^{fe}_{F-1+\ell}\right] \\
=&\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F-1+\ell}-Y_{g,F-1}\right]-\frac{1}{G_0}\sum_{g:T_g=0} E\left[Y_{g,F-1+\ell}-Y_{g,F-1}\right] \\
=&\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F-1+\ell}(\mathbf{0}_{F-1},\mathbf{1}_\ell)-Y_{g,F-1}(\mathbf{0}_{F-1})\right]-\frac{1}{G_0}\sum_{g:T_g=0} E\left[Y_{g,F-1+\ell}(\mathbf{0}_{F-1+\ell})-Y_{g,F-1}(\mathbf{0}_{F-1})\right] \\
=&\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F-1+\ell}(\mathbf{0}_{F-1},\mathbf{1}_\ell)-Y_{g,F-1+\ell}(\mathbf{0}_{F-1+\ell})\right] \\
+&\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F-1+\ell}(\mathbf{0}_{F-1+\ell})-Y_{g,F-1}(\mathbf{0}_{F-1})\right]-\frac{1}{G_0}\sum_{g:T_g=0} E\left[Y_{g,F-1+\ell}(\mathbf{0}_{F-1+\ell})-Y_{g,F-1}(\mathbf{0}_{F-1})\right] \\
=&\mathrm{ATT}_\ell.
\end{aligned}
$$

The first equality follows from (20) and the fact the design is conditioned upon, the second equality follows from Design 1, the third equality follows adding and subtracting $Y_{g,F-1+\ell}(\mathbf{0}_{F-1+\ell})$, the fourth equality follows from Assumption 3 and the definition of ATT_ℓ. This proves (21). Then, if $F \geq 3$, for $\ell \in \{-1,...,-F+2\}$,

$$
\begin{aligned}
&E\left[\widehat{\beta}^{fe}_{F-1+\ell}\right] \\
=&\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F-1+\ell}-Y_{g,F-1}\right]-\frac{1}{G_0}\sum_{g:T_g=0} E\left[Y_{g,F-1+\ell}-Y_{g,F-1}\right] \\
=&\frac{1}{G_1}\sum_{g:T_g=1} E\left[Y_{g,F-1+\ell}(\mathbf{0}_{F-1+\ell})-Y_{g,F-1}(\mathbf{0}_{F-1})\right]-\frac{1}{G_0}\sum_{g:T_g=0} E\left[Y_{g,F-1+\ell}(\mathbf{0}_{F-1+\ell})-Y_{g,F-1}(\mathbf{0}_{F-1})\right] \\
=&0.
\end{aligned}
$$

The first equality follows from (20) and the fact the design is conditioned upon, the second equality follows from Design 1, the third equality follows from Assumption 3. This proves (22) **QED.**

Efficient estimation of ATT_ℓ. One can show that under Assumptions 1 and 3,

$$
\widehat{\beta}^{b,l,g}_{F-1+\ell} \equiv \frac{1}{G_1}\sum_{g:T_g=1}\left(Y_{g,F-1+\ell}-\frac{1}{F-1}\sum_{t=1}^{F-1} Y_{g,t}\right)-\frac{1}{G_0}\sum_{g:T_g=0}\left(Y_{g,F-1+\ell}-\frac{1}{F-1}\sum_{t=1}^{F-1} Y_{g,t}\right) \quad (23)
$$

is also unbiased for ATT_ℓ. The estimator of ATT_ℓ proposed by Borusyak et al. (2021), Liu et al. (2021), and Gardner (2021) for more general binary and staggered designs reduces to $\widehat{\beta}^{b,l,g}_{F-1+\ell}$ in Design 1, hence its subscript. $\widehat{\beta}^{fe}_{F-1+\ell}$ use groups' $F-1$ outcome, the last period before treatment onset, as the baseline outcome. On the other hand, $\widehat{\beta}^{b,l,g}_{F-1+\ell}$ uses their average outcome from period 1 to $F-1$. As $\widehat{\beta}^{b,l,g}_{F-1+\ell}$ uses more data than $\widehat{\beta}^{fe}_{F-1+\ell}$, one may expect the former estimator to be more precise than the latter. Actually, whether this is the case depends on the data generating process. If

$$Y_{g,t}(\mathbf{0}_t) = \alpha_g + \gamma_t + \varepsilon_{g,t}, \tag{24}$$

with $\varepsilon_{g,t}$ independent and identically distributed (iid) across both g and t, then Borusyak et al. (2021) show that $\widehat{\beta}^{b,l,g}_{F-1+\ell}$ is the best linear unbiased estimator (BLUE) of ATT_ℓ. Thus,

$$V\left[\widehat{\beta}^{b,l,g}_{F-1+\ell}\right] \leq V\left[\widehat{\beta}^{fe}_{F-1+\ell}\right], \tag{25}$$

as $\widehat{\beta}^{fe}_{F-1+\ell}$ is also a linear unbiased estimator of ATT_ℓ. However, this result relies on a very restrictive assumption, namely that within group g the errors $\varepsilon_{g,t}$ are uncorrelated over time. This rules out within-group serial correlations that are likely to be present in many empirical settings, and that Bertrand et al. (2004) recommend accounting for when conducting inference in DID and TWFE studies. Moreover, (25) can reverse if one relaxes this assumption. Instead of assuming independent errors, assume that in each group errors follow a random walk: $\varepsilon_{g,t} = \varepsilon_{g,t-1} + u_{g,t}$, with $u_{g,t}$ iid. This means that in each group, errors are very strongly positively correlated over time. Then, Harmon (2022) shows that $\widehat{\beta}^{fe}_{F-1+\ell}$ is the BLUE estimator of ATT_ℓ, thus implying that

$$V\left[\widehat{\beta}^{b,l,g}_{F-1+\ell}\right] \geq V\left[\widehat{\beta}^{fe}_{F-1+\ell}\right]. \tag{26}$$

Some intuition for (26) goes as follows. Assume that $F = 3$. Then, $\widehat{\beta}^{b,l,g}_{3}$ is a difference between independent averages of

$$Y_{g,3} - \frac{1}{2}(Y_{g,2} + Y_{g,1}) = \gamma_3 - \frac{1}{2}(\gamma_2 + \gamma_1) + \varepsilon_{g,3} - \frac{1}{2}(\varepsilon_{g,2} + \varepsilon_{g,1}) = \gamma_3 - \frac{1}{2}(\gamma_2 + \gamma_1) + u_{g,3} + \frac{1}{2}u_{g,2},$$

where the first equality follows from (24) and the second from the random walk assumption. On the other hand, $\widehat{\beta}^{fe}_{3}$ is a difference between independent averages of

$$Y_{g,3} - Y_{g,2} = \gamma_3 - \gamma_2 + u_{g,3}.$$

Then, it is easy to see that $V\left[\widehat{\beta}_3^{b,l,g}\right] > V\left[\widehat{\beta}_3^{fe}\right]$. At the end of the day, we recommend using $\widehat{\beta}_{F-1+\ell}^{fe}$ rather than $\widehat{\beta}_{F-1+\ell}^{b,l,g}$. Under Assumptions 1 and 3, the two estimators should be close, so using one or the other should not make a large difference.[5] But $\widehat{\beta}_{F-1+\ell}^{b,l,g}$ is more biased than $\widehat{\beta}_{F-1+\ell}^{fe}$, when Assumption 3 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 (2022) notes that leveraging earlier pre-treatment periods increases the bias of a DID estimator, since one makes comparisons from earlier periods. On the other hand, if Assumption 1 fails due to anticipation effects arising a few periods before F, $\widehat{\beta}_{F-1+\ell}^{b,l,g}$ may be less biased than $\widehat{\beta}_{F-1+\ell}^{fe}$. However, violations of Assumptions 3 and 1 may not be equally problematic. Often times, $\widehat{\beta}_{F-1+\ell}^{b,l,g}$ and $\widehat{\beta}_{F-1+\ell}^{fe}$ can be immunized against anticipation effects, by redefining F as the date when the treatment was announced. On the other hand, it is often harder to immunize those estimators against violations of Assumption 3.

Estimation under a weaker parallel trends assumption The parallel trends condition in Assumption 3 is strong, as it requires that all groups experience parallel trends. Actually, Theorems 2 and 3 still hold under the following weaker parallel trends assumption: for all $t \geq 2$,

$$E\left[\frac{1}{G_1}\sum_{g:T_g=1}\left(Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})\right)\right] = E\left[\frac{1}{G_0}\sum_{g:T_g=0}\left(Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})\right)\right], \tag{27}$$

meaning that the average expected evolution of the never-treated outcome is the same across treated and control groups.

First- and long-difference placebos. For $\ell \in \{-1, ..., -F+2\}$, instead of $\widehat{\beta}_{F-1+\ell}^{fe}$ one may use

$$\widehat{\beta}_{F-1+\ell}^{fd} = \frac{1}{G_1}\sum_{g:T_g=1}\left(Y_{g,F+\ell-1} - Y_{g,F+\ell}\right) - \frac{1}{G_0}\sum_{g:T_g=0}\left(Y_{g,F+\ell-1} - Y_{g,F+\ell}\right) \tag{28}$$

as another placebo estimator, that should also not significantly differ from zero under Assumptions 1 and 3. $\widehat{\beta}_{F-2}^{fe}$ and $\widehat{\beta}_{F-2}^{fd}$ are equal, but $\widehat{\beta}_{F-1+\ell}^{fe}$ and $\widehat{\beta}_{F-1+\ell}^{fd}$ differ for $\ell \leq -2$. $\widehat{\beta}_{F-1+\ell}^{fe}$ is a "long-difference" placebo, that compares treatment- and control-groups' outcome evolution over several periods before the treatment onset. $\widehat{\beta}_{F-1+\ell}^{fd}$ instead is a "first-difference" placebo, that

[5] Under Assumptions 1 and 3, both estimators are unbiased for ATT_ℓ so if they significantly differ that implies that Assumption 1 or 3 must fail.

compares treatment- and control-groups' outcome evolution over consecutive periods before the treatment onset.

Intuitively, explain the pros and cons of using $\widehat{\beta}^{fd}_{F-1+\ell}$ rather than $\widehat{\beta}^{fe}_{F-1+\ell}$ to test Assumptions 1 and 3.

If treatment and control groups follow different linear trends, differential trends will be larger, and easier to detect, over several periods than over two consecutive periods. Then, the long-difference placebos may lead to a more powerful test of Assumptions 1 and 3. On the other hand, the first-difference placebos may be useful to specifically test Assumption 1, the no-anticipation assumption. Assume for instance that those placebos are insignificant, except between $F-2$ and $F-1$. This may suggest that Assumption 3 holds, but Assumption 1 fails: at period $F-1$, treatment groups are already affected by their upcoming treatment in the next period. Then, one may just recompute the estimators defined above, redefining F as $F-1$, or as the date when the treatment was announced.

4.3 Issues with parallel trends tests.

Tests of parallel trends are often underpowered in published economics papers. Roth (2022) evaluates the power of parallel trends tests in a sample of 12 papers published in the American Economic Review, American Economic Journal: Applied Economics, and American Economic Journal: Economic Policy between 2014 and June 2018, that contain the phrase "event study", whose data is publicly available, and that estimated Regression (19). For each paper, he collects the event-study coefficients $(\widehat{\beta}^{fe}_{F-1+\ell})_{\ell\in\{-F+2,...,-1,1,...,T-F+1\}}$ and their estimated variance-covariance matrix $\widehat{\Sigma}$, and then runs the following simulations. In each simulation draw, he generates coefficients $(\widehat{\beta}^{s,fe}_{F-1+\ell})_{\ell\in\{-F+2,...,-1,1,...,T-F+1\}}$ from a normal distribution with variance-covariance matrix $\widehat{\Sigma}$, and where

$$E(\widehat{\beta}^{s,fe}_{F-1+\ell}) = \gamma|\ell| + 1\{\ell \geq 1\}\widehat{\beta}^{fe}_{F-1+\ell},$$

for some real number $\gamma \neq 0$. Interpret the DGP in those simulations. Does the parallel trends assumption hold? What is the value of ATT_ℓ?

For $\ell \in \{-F+2, ..., -1\}$, $E(\widehat{\beta}^{s,fe}_{F-1+\ell}) = \gamma|\ell| \neq 0$, so parallel trends fails in this DGP. Treated and control groups experience their own linear trends, and the difference between the linear trends of treated and control groups is equal to γ. ATT_ℓ is equal to $\widehat{\beta}^{fe}_{F-1+\ell}$, the actual value of the estimated ATT_ℓ in each paper. In each simulation draw, Roth mimicks a pre-trends test, which is rejected if there is at least one $\ell \in \{-F+2, ..., -1\}$ such that $|\widehat{\beta}^{s,fe}_{F-1+\ell}| > 1.96\widehat{\sigma}_{F-1+\ell}$, where $\widehat{\sigma}_{F-1+\ell}$ is the estimated standard error of $\widehat{\beta}^{fe}_{F-1+\ell}$. He evaluates the power of the pre-trends tests across many values of γ, until finding the value $\gamma_{0.5}$ such that the pre-trends test is rejected 50% of the time. In each of the 12 papers he considers, $\gamma_{0.5}$ represents the differential linear trends between treatment and control groups that have 50% chances of being detected by the researcher. Finally, he evaluates $\frac{1}{T-F+1}\sum_{\ell=1}^{T-F+1}\gamma_{0.5}|\ell|$ and compares that quantity to $\frac{1}{T-F+1}\sum_{\ell=1}^{T-F+1}\widehat{\beta}^{fe}_{F-1+\ell}$. What is the purpose of that comparison?

$\frac{1}{T-F+1}\sum_{\ell=1}^{T-F+1}\gamma_{0.5}|\ell|$ is the bias in the event-study estimate of the ATT, under differential linear trends that have 50% chances of being detected. It is interesting to compare the magnitude of this potential bias to the magnitude of the actual event-study estimate of the ATT, to assess if differential trends that have high chances of not being detected by the researcher can account for a large share of the estimated treatment effect.[6] Results are compelling. His appendix Figure D1, reproduced below, shows that in 7 papers out of 12, under differential linear trends that have 50% chances of being detected, the bias in the event-study estimate of the ATT, in green, is no

[6] Perhaps it would have made even more sense to compare $\frac{1}{T-F+1}\sum_{\ell=1}^{T-F+1}\gamma_{0.5}|\ell|$ to $\frac{1}{T-F+1}\sum_{\ell=1}^{T-F+1}\gamma_{0.5}|\ell| + \frac{1}{T-F+1}\sum_{\ell=1}^{T-F+1}\widehat{\beta}^{fe}_{F-1+\ell}$, the expectation of the estimated ATT in the simulations.

smaller than a half of the actual estimate of the ATT, in blue. In other words, in 7 papers out of 12, the authors had 50 chances of failing to detect differential trends large enough to account for at least a half of their estimated ATT. Based on his findings, Roth (2022) recommends that practitioners run simulations similar to his, and provides the R package `pretrends` for that purpose. Thus, researchers can assess the power of pre-trends tests in their application, and whether they could fail to detect differential trends large enough to account for a substantial fraction of their estimated ATT.

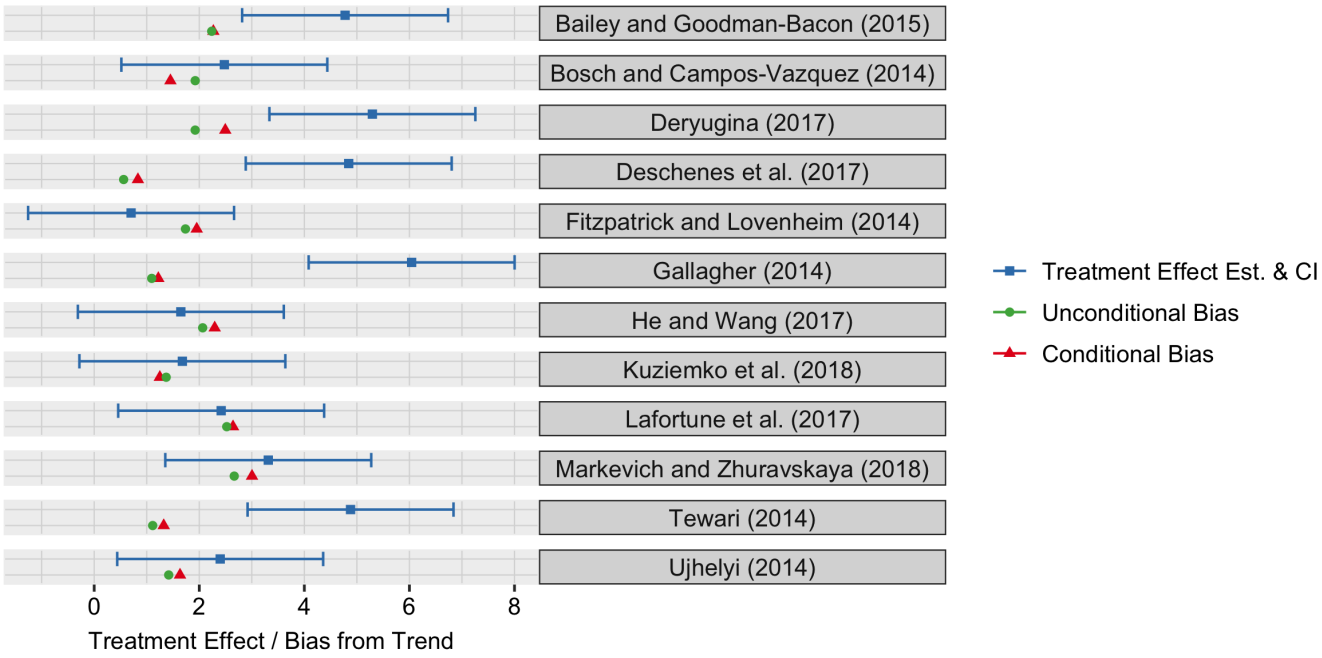


Figure 1: Power of pre-trends tests in 12 published economics papers: reproduction of Figure D1 in Roth (2022).

Tests of parallel trends may lead to a pre-testing problem. Parallel trends tests are often used as a way to decide whether the analysis should be continued, or which specification should be reported: researchers may add control variables, add group-specific linear trends, change the definition of their control group, etc., until the parallel-trends test is not rejected. This means that often times, the vector of estimated effects we observe $(\widehat{\beta}^{fe}_{F-1+\ell})_{\ell\in\{1,...,T-F+1\}}$ is conditional on values of the pre-trends coefficients $(\widehat{\beta}^{fe}_{F-1+\ell})_{\ell\in\{-F+2,...,-1\}}$ such that the pre-trends test is not rejected. Let Pub be an indicator equal to 1 when that event is realized, where Pub

stands for publishable. If $(\widehat{\beta}^{fe}_{F-1+\ell})_{\ell\in\{1,...,T-F+1\}}$ and $(\widehat{\beta}^{fe}_{F-1+\ell})_{\ell\in\{-F+2,...,-1\}}$ are not independent, we may have that for $\ell \in \{1, ..., T-F+1\}$

$$E(\widehat{\beta}^{fe}_{F-1+\ell}|\text{Pub}=1) \neq E(\widehat{\beta}^{fe}_{F-1+\ell}),$$

so such pre-testing could lead to a bias of $\widehat{\beta}^{fe}_{F-1+\ell}$, on top of the potential bias that may come from differential trends. Reassuringly, Proposition 1 in Roth (2022) shows that when trends are parallel, this additional bias is equal to zero: under parallel trends, testing for pre-trends and estimating the treatment effect only if the pre-trends test is not rejected does not lead to a bias. However, Proposition 2 therein shows that if trends are not parallel, differential trends widen over time, and the estimates of the pre-trends and actual effects are positively correlated and homoscedastic, then pre-testing leads to a bias which goes in the same direction as the bias coming from differential trends, thus exacerbating it. In practice, Figure D1 in Roth (2022), reproduced above, shows that this bias exacerbation phenomenon is modest in the 12 papers he reviews: $E(\widehat{\beta}^{fe}_{F-1+\ell}|\text{Pub}=1)$, in red, is in most cases fairly close to $E(\widehat{\beta}^{fe}_{F-1+\ell})$, in green. This may be due to the fact that under (24), if in each group errors follow a random walk: $\varepsilon_{g,t} = \varepsilon_{g,t-1} + u_{g,t}$, then it is easy to show that $(\widehat{\beta}^{fe}_{F-1+\ell})_{\ell\in\{1,...,T-F+1\}}$ and $(\widehat{\beta}^{fe}_{F-1+\ell})_{\ell\in\{-F+2,...,-1\}}$ are independent. Of course, the random walk assumption is rather strong, but Figure D1 in Roth (2022) suggests that in real-life examples, this assumption may not be too far from being satisfied, hence the modest bias exacerbation from pre-testing.

4.4 Estimating heterogeneous treatment effects.

Estimating the distribution of treatment effects? Under Assumption 1 and the strong parallel-trends condition in Assumption 3, one can unbiasedly estimate the group-level treatment effects

$$\text{TE}^{\text{dyn}}_{g,F-1+\ell} = E\left[Y_{g,F-1+\ell}(\mathbf{0}_{F-1}, \mathbf{1}_{\ell}) - Y_{g,F-1+\ell}(\mathbf{0}_{F-1+\ell})\right],$$

for instance using

$$\widehat{\text{TE}}^{\text{dyn}}_{g,F-1+\ell} = Y_{g,F-1+\ell} - Y_{g,F-1} - \frac{1}{G_0}\sum_{g':T_{g'}=0}(Y_{g',F-1+\ell} - Y_{g',F-1}).$$

Then, one may consider using the estimators $(\widehat{\text{TE}}^{\text{dyn}}_{g,F-1+\ell})_{g:T_g=1}$ to assess if the effect of having been treated for ℓ periods varies across groups, and to estimate the distribution of the effects $(\text{TE}^{\text{dyn}}_{g,F-1+\ell})_{g:T_g=1}$. Under the independent groups framework in Assumption 5, an estimator may be consistent if it averages outcomes from a number of groups that goes to infinity when $G \to \infty$. For any value of G, $Y_{g,F-1+\ell} - Y_{g,F-1}$ averages the outcomes of only one group, so $\widehat{\text{TE}}^{\text{dyn}}_{g,F-1+\ell}$ is not consistent. As the estimators $(\widehat{\text{TE}}^{\text{dyn}}_{g,F-1+\ell})_{g:T_g=1}$ are not consistent, naively using them to estimate the distribution of treatment effects across groups would be misleading: one would first need to deconvolute those estimators. There is a vast literature in statistics on deconvolution. It has been successfully used in some subfields of the causal inference literature, for instance for meta-analyses. To our knowledge, deconvolution has never been used to recover the distribution of group-specific treatment effects in DID studies. This is an interesting avenue for future research.

Testing for heterogeneous treatment effects. Another possibility to test for heterogeneous effects is to regress $\widehat{\text{TE}}^{\text{dyn}}_{g,F-1+\ell}$ on group-level covariates, and assess whether the covariates significantly predict those estimated effects (see Muris and Wacker, 2022; de Chaisemartin and Lei, 2021, for similar proposals in more complicated designs). As the $\widehat{\text{TE}}^{\text{dyn}}_{g,F-1+\ell}$ are estimated, it may seem that inference would need to account for that first-step estimation, which may be achieved by bootstrapping the estimation procedure. While this is true in more complicated designs than in Design 1, that is not the case in Design 1. Indeed, note that for a fixed value of ℓ, letting $X_.$ denote the average of the covariate X_g, the coefficient of X_g in a regression of $\widehat{\text{TE}}^{\text{dyn}}_{g,F-1+\ell}$ is merely equal to

$$\frac{\sum_{g:T_g=1}(X_g - X_.)(Y_{g,F-1+\ell} - Y_{g,F-1})}{\sum_{g:T_g=1}(X_g - X_.)^2},$$

the coefficient of X_g in a regression of $Y_{g,F-1+\ell} - Y_{g,F-1}$ on X_g among treated groups, which can readily be estimated without any first-step estimation. That one can test for heterogeneous treatment effects without control groups might be counter-intuitive. This is due to the fact that under the strong parallel-trends condition in Assumption 3, treated groups with high and low values of X_g experience the same counterfactual trends without treatment, so any systematic difference in their outcome evolutions must come from heterogeneous treatment effects. Interestingly this shows that in Design 1, control groups are not necessary to test for heterogeneous

treatment effects, and therefore to test the null of no treatment effect. To our knowledge, these very simple results have not been noted elsewhere. Rather than regressing $\widehat{\text{TE}}^{\text{dyn}}_{g,F-1+\ell}$ on X_g, or equivalently $Y_{g,F-1+\ell} - Y_{g,F-1}$ on X_g, the way empirical researchers test for heterogeneous treatment effects is by running a TWFE regression of $Y_{g,t}$ on group fixed effects, time fixed effects, $D_{g,t}$, and $D_{g,t}X_g$. Strikingly, and as shown by de Chaisemartin and d'Haultfoeuille (2022*b*), the coefficients on $D_{g,t}$ and $D_{g,t}X_g$ in this regression are not robust to heterogeneous treatment effects, and may even suffer from a contamination problem that we will discuss in more details in Section 10. For now, we just provide the Stata code below, which shows that even in Design 1, this issue may be present.

```
drop _all
set obs 1000
gen g=floor((_n-1)/2)+1
gen t=_n-2*floor((_n-1)/2)
gen d=(g>=251&t==2)
gen y=uniform()
gen x=uniform()
gen d_x=d*x
twowayfeweights y g t d, type(feTR) other_treatments(d_x)
```

4.5 Application: Benzarti and Carloni (2019).

In July of 2009, France reduced its VAT on sit-down restaurants from 19.6 to 5.5 percent. Benzarti and Carloni (2019) analyze the impact of this change using the regression in (19), with the control group defined as other market services firms that were not affected by the VAT change. Figure 2 shows one of the event-study graphs in their paper, with the outcome defined as the log of firms' profits. Would you say that Figure 2 convincingly shows that the reform increased restaurants' profits?

Figure 2 shows evidence of a violation of Assumptions 1 and 3, with a significant positive pre-trend coefficient the year before the reform. However, the estimated effects are much larger than the pre-trend coefficient, and lie far outside of the pre-trends' confidence intervals, so it does not seem that pre-trends can account for the entirety of the estimated effects. In our opinion, Figure 2 convincingly shows that the reform increased restaurants' profits. According to the point estimates, the reform increased restaurants' profits by around 20% the year of the reform, by 30% the following year, by 25% two years after, and by 15% three years after.

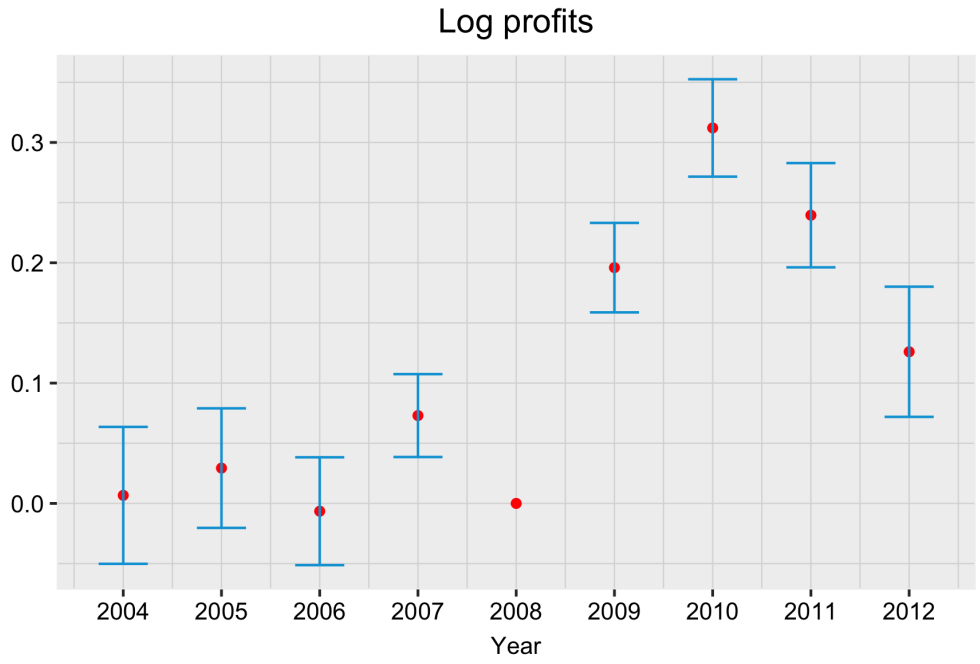


Figure 2: Event-study estimates of the effect of the VAT cut on firms' profits.

5 Imperfect parallel trends: relaxations of the parallel trends assumption.

5.1 Parallel trends with covariates

Maintained assumptions and data configuration. Throughout this section, we assume that Assumption 2 holds: the treatment does not have dynamic effects. In most of this section,

we also assume that the data contains only two time periods: $F = T = 2$. This restrictive set-up is sufficient to present the main ways covariates can be introduced in TWFE and DID analysis.

Conditional parallel trends assumptions. Let X_g denote a vector of time-invariant covariates. In this section, the design and the covariates $(X_g)_{g\in\{1,\ldots,G\}}$ are implicitly conditioned upon, so we can treat the covariates as non-stochastic. Introducing covariates allows us to replace Assumption 4 by the following condition.

Assumption 6 *(Conditional parallel trends) There exists a function $\gamma : x \mapsto \gamma(x)$ such that $E\left[Y_{g,2}(0) - Y_{g,1}(0)\right] = \gamma(X_g)$.*

Interpret Assumption 6.

Assumption 6 is a conditional parallel-trends assumption. It implies that if two groups g_1 and g_2 are such that $X_{g_1} = X_{g_2}$, then g_1 and g_2 have the same expected evolution of their untreated outcome:

$$X_{g_1} = X_{g_2} \Rightarrow E\left[Y_{g_1,2}(0) - Y_{g_1,1}(0)\right] = E\left[Y_{g_2,2}(0) - Y_{g_2,1}(0)\right].$$

But under Assumption 6, groups with different values of their covariates may have different expected evolutions of their untreated outcomes. Thus, Assumption 6 may be more plausible than Assumption 4. Assumption 6 has for instance been considered by Heckman et al. (1997), Blundell et al. (2004), Abadie (2005), and Sant'Anna and Zhao (2020). One could also strengthen Assumption 6, assuming that $\gamma : x \mapsto \gamma(x)$ is linear:

Assumption 7 *(Conditional parallel trends, with a linear functional form) There exists a real number γ_2 and a vector γ_X of same dimension as X_g such that $E\left[Y_{g,2}(0) - Y_{g,1}(0)\right] = \gamma_2 + X_g'\gamma_X$.*

Time-varying covariates. The estimators proposed below can be used with time-varying covariates $X_{g,t}$, letting $X_g = X_{g,2} - X_{g,1}$. This changes the interpretation of Assumption 6

above, which now requires that

$$E\left[Y_{g,2}(0) - Y_{g,1}(0)\right] = \gamma(X_{g,2} - X_{g,1}),$$

a parallel trends assumption conditional on groups' covariates evolution, rather than conditional on the level of their covariates. With time-varying covariates, whether parallel trends is more plausible conditional on $X_{g,1}$ or conditional on $X_{g,2} - X_{g,1}$ will depend on the context, but it is important to note that the two assumptions have different interpretations and lead to different estimators.

Estimating the ATT under Assumption 7: TWFE with controls. Under Assumption 7, propose a TWFE regression to estimate the ATT.

In designs with a binary treatment and no variation in treatment timing, we have seen that $\widehat{\beta}^{fe}$ is unbiased for the ATT under Assumption 4. Then, it would seem natural to assume that in the same designs, the treatment coefficient $\widehat{\beta}^{fe}_X$ in regression (29) below is unbiased for the ATT under Assumption 7:

$$Y_{g,t} = \sum_{g'=1}^{G} \widehat{\alpha}_{g'} 1\{g = g'\} + \widehat{\gamma}_2 1\{t = 2\} + X_g' \widehat{\gamma}_X 1\{t = 2\} + \widehat{\beta}^{fe}_X D_{g,t} + \hat{\epsilon}_{g,t}, \tag{29}$$

With $T = 2$, the regressions in (29) and (3) are very similar, except that the one in (29) also controls for the interaction of the covariates X_g and the period-two indicator. This allows groups' untreated outcome trends to depend linearly on X_g, as assumed in Assumption 7. Assume that the treatment effect is constant: there is a real number δ such that $Y_{g,t}(1) - Y_{g,t}(0) = \delta$ for all (g, t). Then,

$$\begin{aligned}
E(Y_{g,t}) =& E(Y_{g,t}(0) + D_{g,t}(Y_{g,t}(1) - Y_{g,t}(0))) \\
=& E(Y_{g,t}(0)) + \delta D_{g,t} \\
=& E(Y_{g,1}(0)) + 1\{t = 2\} E(Y_{g,2}(0) - Y_{g,1}(0)) + \delta D_{g,t} \\
=& \sum_{g'=1}^{G} E(Y_{g',1}(0)) 1\{g = g'\} + \gamma_2 1\{t = 2\} + X_g' \gamma_X 1\{t = 2\} + \delta D_{g,t},
\end{aligned}$$

where the second equality follows from the constant effect assumption, and the fourth follows from Assumption 7. Therefore, under the constant effect assumption and Assumption 7, the regression in (29) is correctly specified and one can show that $\widehat{\beta}_X^{fe}$ is unbiased for the constant effect δ. But with heterogeneous treatment effects, it follows from Theorem S4 in the Web Appendix of de Chaisemartin and D'Haultfœuille (2020) that even in designs with a binary treatment and no variation in treatment timing, $\widehat{\beta}_X^{fe}$ may not be unbiased for the ATT: it estimates a weighted sum of the $TE_{g,t}$s, potentially with some negative weights. With control variables, Theorem S4 in the Web Appendix of de Chaisemartin and D'Haultfœuille (2020) shows that the weights in the decomposition of $\widehat{\beta}_X^{fe}$ come from the residual from a regression of $D_{g,t}$ on group fixed effects, a period-two fixed effect, and the interaction of the covariates X_g and the period-two fixed effect. With a single binary covariate X_g, this regression is saturated: it is actually equivalent to a regression of $D_{g,t}$ on X_g, a period-two fixed effect, and the interaction of X_g and the period-two fixed effect. This is a standard 2×2 DID regression which models exactly the conditional mean of $D_{g,t}$ given t and X_g so the regression never predicts values of $D_{g,t}$ above 1 and cannot yield negative residuals. But if X_g is not binary, the regression determining the weights is not saturated anymore, it could be misspecified for the conditional mean of $D_{g,t}$ given t and X_g and yield predicted values above 1, and negative residuals.

Estimating the ATT under Assumption 7: linear outcome regression. To estimate the ATT under Assumption 7, Heckman et al. (1997) propose the following procedure. First, one estimates the regression below, restricting the sample to groups such that $T_g = 0$:

$$Y_{g,2} - Y_{g,1} = \widehat{\gamma}_2^{or} + X_g' \widehat{\gamma}_X^{or} + \hat{\epsilon}_{g,t}. \tag{30}$$

Then, propose an unbiased estimator of the ATT.

As $\widehat{\gamma}_2^{or}$ and $\widehat{\gamma}_X^{or}$ are unbiased for γ_2 and γ_X, one can show that

$$\text{DID}_{X,lin-or} \equiv \frac{1}{G_1} \sum_{g:T_g=1} (Y_{g,2} - Y_{g,1} - (\widehat{\gamma}_2^{or} + X_g' \widehat{\gamma}_X^{or}))$$

is unbiased for the ATT. Here is some intuition. Under Assumption 7, groups' outcome evolution without treatment is a linear function of their covariates. We estimate the coefficients in that linear function by regressing control groups' outcome evolution on their covariates. We then compute treated groups' predicted outcome evolution without treatment, based on that regression. Finally, DID_X subtracts from treated groups' actual outcome evolution their predicted outcome evolution without treatment, to recover their treatment effect.

Estimating the ATT under Assumption 6: non-parametric outcome regression. Under Assumption 6, Heckman et al. (1997) propose a similar procedure as that we just discussed to estimate the ATT, except that the fist step has to rely on a non-parametric regression. One first estimates a non-parametric regression of $Y_{g,2} - Y_{g,1}$ on X_g, restricting the sample to groups such that $T_g = 0$. Non-parametric regression models for instance include kernel regressions, or series estimators where $Y_{g,2} - Y_{g,1}$ is regressed on a polynomial in X_g, whose degree goes to infinity when the sample size increases. Once the non-parametric regression has been estimated, one can compute the predicted outcome evolution without treatment of all treated groups, based on their X_g, which we denote $\widehat{\gamma}(X_g)$. Finally, one can use

$$\text{DID}_{X,np-or} \equiv \frac{1}{G_1} \sum_{g:T_g=1} (Y_{g,2} - Y_{g,1} - \widehat{\gamma}(X_g))$$

to estimate the ATT.

Estimating the ATT under Assumption 6: propensity-score reweighting. Under Assumption 6, one can also use propensity-score reweighting to estimate the ATT. In a first step, one regresses the treatment group indicator T_g on X_g, using either a non-parametric regression model, or a parametric model for binary outcome variables such as a logit or a probit. Let $\widehat{p}(X_g)$ denote group's g predicted probability to be a treatment group according to this regression, its so-called propensity score. Then, Abadie (2005) proposes to use

$$\text{DID}_{X,ps} \equiv \frac{1}{G_1} \sum_{g:T_g=1} (Y_{g,2} - Y_{g,1}) - \frac{1}{G_0} \sum_{g:T_g=0} \frac{\widehat{p}(X_g)}{1-\widehat{p}(X_g)} \frac{G_0}{G_1} (Y_{g,2} - Y_{g,1})$$

to estimate the ATT. Intuitively, $\text{DID}_{X,ps}$ compares the outcome evolution of treated and control groups, after reweighting control groups. As $x \mapsto x/(1-x)$ is increasing in x on $(0,1)$, the reweighting gives more weight to control groups that have a larger value of $\widehat{p}(X_g)$, namely to

control groups who, based on their X_g, have a larger predicted probability of being treated. Thus, the reweighting gives more weight to control groups that "look like" treatment groups. Actually, one can formally show that the reweighting ensures that the distribution of X_g is the same in the treatment group and in the reweighted control group, which is why $\text{DID}_{X,ps}$ is consistent for the ATT.

Estimating the ATT under Assumption 6: doubly-robust estimator. When one wants to control for several, potentially continuous covariates, using non-parametric regressions to estimate $\gamma(X_g)$ or groups' propensity scores may yield estimators with poor finite-sample properties, owing to the so-called curse of dimensionality. At the same time, $\text{DID}_{X,lin-or}$ is inconsistent if Assumption 6 holds but $\gamma(X_g)$ is not linear. And $\text{DID}_{X,ps}$ with a parametric logit or probit model is also inconsistent if the propensity score does not follow the chosen parametric model. Then, Sant'Anna and Zhao (2020) have proposed to use a so-called doubly-robust estimator, that combines outcome regression and propensity-score reweighting. Specifically, one of their estimators of the ATT is

$$\text{DID}_{X,dr} \equiv \frac{1}{G_1} \sum_{g:T_g=1} (Y_{g,2} - Y_{g,1} - (\widehat{\gamma}_2^{or} + X_g' \widehat{\gamma}_X^{or})) - \frac{1}{G_0} \sum_{g:T_g=0} \frac{\widehat{p}(X_g)}{1-\widehat{p}(X_g)} \frac{G_0}{G_1} (Y_{g,2} - Y_{g,1} - (\widehat{\gamma}_2^{or} + X_g' \widehat{\gamma}_X^{or})),$$

where $(\widehat{\gamma}_2^{or}, \widehat{\gamma}_X^{or})$ are the coefficients from the linear regression in (30), and where $\widehat{p}(X_g)$ comes from a parametric (e.g. probit or logit) regression of T_g on X_g. Sant'Anna and Zhao (2020) show that $\text{DID}_{X,dr}$ is consistent for the ATT if either $\gamma(X_g)$ is linear, or the parametric model for the propensity score is correctly specified. It is only if $\gamma(X_g)$ is non linear and the parametric model for the propensity score is incorrectly specified that $\text{DID}_{X,dr}$ is inconsistent.

Estimation with group-specific linear trends. In this paragraph, we assume that $T = F = 3$: the data contains three time periods, and treated groups become treated at period 3. We consider the following assumption.

Assumption 8 *(Common deviations from linear trends) For all* $t \geq 2$, $E\left[Y_{g,t}(0) - Y_{g,t-1}(0)\right] = \gamma_t + \lambda_g$.

Assumption 8 allows groups to experience group-specific linear trends, but requires that between each pair of consecutive periods, all groups have the same expected deviation from their linear

trend. Under Assumption 8, Mora and Reggio (2019) propose

$$\text{DID}_{X,tr-lin} \equiv \frac{1}{G_1} \sum_{g:T_g=1} (Y_{g,3} - Y_{g,2} - (Y_{g,2} - Y_{g,1})) - \frac{1}{G_0} \sum_{g:T_g=0} (Y_{g,3} - Y_{g,2} - (Y_{g,2} - Y_{g,1}))$$

to estimate the ATT. Intuitively, $\text{DID}_{X,tr-lin}$ compares second-differences of the outcome in the treatment and in the control group.

When to include covariates in the estimation? Introducing covariates, one can work under a conditional rather than unconditional parallel trends assumption. If pre-trend coefficients without control variables are precisely estimated and not significantly different from zero, there may not be a compelling reason to control for covariates. An issue with conditionally-valid identifying assumptions is that different researchers may disagree on the variables one should control for, and different sets of controls may lead to different results. This has generated vigorous controversies in the matching (see LaLonde, 1986; Dehejia and Wahba, 1999; Smith and Todd, 2005) and teachers' value-added (see Chetty et al., 2014; Rothstein, 2017) literature. If placebo tests are significant or imprecise without controls, there may be a more compelling argument to control for some covariates, if pre-trends are insignificant and precisely estimated with covariates. Note that the course of action we recommend here involves some pre-testing. This may lead to a bias when pre-tests lack power, even though the results in Roth (2022) suggest that this bias may be small relative to the bias that could arise from differential trends.

Stata commands to compute DID estimators with covariates. $\text{DID}_{X,lin-or}$, $\text{DID}_{X,np-or}$, $\text{DID}_{X,ps}$, and $\text{DID}_{X,dr}$ are computed by the `drdid` Stata (see Rios-Avila, Sant'Anna and Naqvi, 2021) and R (see Sant'Anna and Zhao, 2022) commands. The basic syntax of the Stata command is:

`drdid outcome [controls_var_names], ivar(groupid) time(timeid) treatment(var_name).`

With time-varying covariates, the control variables inputted to the command have to be defined as $X_g = X_{g,2} - X_{g,1}$ at every date. If the user inputs $X_{g,t}$, the command controls for $X_{g,1}$. $\text{DID}_{X,lin-or}$, $\text{DID}_{X,np-or}$, and $\text{DID}_{X,tr-lin}$ are computed by the `did_multiplegt` Stata (see de Chaisemartin, D'Haultfœuille and Guyonvarch, 2019) and R (see Zhang and de Chaisemartin, 2020) commands. To compute $\text{DID}_{X,lin-or}$, the syntax of the Stata command is:

`did_multiplegt outcome groupid timeid treatment, controls(var_names)`.
With time-invariant controls, the control variables inputted to the command have to be defined as $X_g \times t$. If the user defines the controls as X_g, the controls are dropped from the estimation. To compute $\text{DID}_{X,np-or}$, the syntax is:

`did_multiplegt outcome groupid timeid treatment, trends_non_param(var_names)`.
Only discrete variables coarser than the group variable can be inputted to `trends_non_param`. To compute $\text{DID}_{X,tr-lin}$, the syntax is:

`did_multiplegt outcome groupid timeid treatment, trends_lin(groupid)`.
Note that both `did_multiplegt` and `drdid` can be used with more than two time periods.

5.2 Stationary differential trends.

Stationary differential trends. In view of the low power of parallel-trends tests in economics articles documented by Roth (2022), Rambachan and Roth (2023) propose to relax the parallel trends assumption. To present their idea, let us assume that Assumption 2 holds, and that $F = T = 3$. The authors propose to replace the parallel trends condition in (27) by the following weaker condition:

Assumption 9 *(Stationary differential trends) There is a positive real number M such that*

$$\begin{aligned}&\left|E\left[\frac{1}{G_1}\sum_{g:T_g=1}(Y_{g,3}(0)-Y_{g,2}(0))\right]-E\left[\frac{1}{G_0}\sum_{g:T_g=0}(Y_{g,3}(0)-Y_{g,2}(0))\right]\right|\\ \leq& M\left|E\left[\frac{1}{G_1}\sum_{g:T_g=1}(Y_{g,2}(0)-Y_{g,1}(0))\right]-E\left[\frac{1}{G_0}\sum_{g:T_g=0}(Y_{g,2}(0)-Y_{g,1}(0))\right]\right|. \qquad (31)\end{aligned}$$

Assumption 9 allows treated and control groups to experience differential trends, but requires that their period-2-to-3 differential trend be bounded in absolute value by some constant M times their period-1-to-2 differential trend. In other words, differential trends cannot vary "too much" from period to period, where M indexes by how much differential trends can differ from period 2 to 3 and from period 1 to 2. Note that with $M = 0$, Assumption 9 is equivalent to parallel-trends from period 2 to 3. Similarly, if

$$\left|E\left[\frac{1}{G_1}\sum_{g:T_g=1}(Y_{g,2}(0)-Y_{g,1}(0))\right]-E\left[\frac{1}{G_0}\sum_{g:T_g=0}(Y_{g,2}(0)-Y_{g,1}(0))\right]\right|=0,$$

then Assumption 9 implies parallel trends from period 2 to 3, irrespective of the value of M. Note that Rambachan and Roth (2023) do not propose a name for Assumption 9: "stationary differential trends" is our own interpretation of their assumption, highlighting the fact that it requires differential trends to not change too much over time.

Connection with bounded differential trends. Assumption 9 is related to, but different from, the bounded differential trends assumption that had been previously proposed by Manski and Pepper (2018): there is a positive real number M such that

$$\left| E\left[\frac{1}{G_1}\sum_{g:T_g=1}(Y_{g,3}(0)-Y_{g,2}(0))\right] - E\left[\frac{1}{G_0}\sum_{g:T_g=0}(Y_{g,3}(0)-Y_{g,2}(0))\right]\right| \leq M.$$

The key difference between the two assumptions is that under Assumption 9, one can leverage pre-trends estimates to infer the differential trends that treatment and control groups would have experienced from period two to three, even in the absence of the treatment.

Partial identification of the ATT under Assumption 9. Under Assumption 9, and denoting by β_t^{fe} the expectations of the coefficients $\widehat{\beta}_t^{fe}$ in Regression (19), one can show that

$$\beta_1^{fe} - M|\beta_{-1}^{fe}| \leq \text{ATT} \leq \beta_1^{fe} + M|\beta_{-1}^{fe}|.$$

Given M, the lower and upper bounds in the previous display can respectively be estimated by $\widehat{\beta}_1^{fe} - M|\widehat{\beta}_{-1}^{fe}|$ and $\widehat{\beta}_1^{fe} + M|\widehat{\beta}_{-1}^{fe}|$. Which condition should hold to have that 0 does not belong to the interval $\left[\widehat{\beta}_1^{fe} - M|\widehat{\beta}_{-1}^{fe}|, \widehat{\beta}_1^{fe} + M|\widehat{\beta}_{-1}^{fe}|\right]$?

One should have that $|\widehat{\beta}_1^{fe}| > M|\widehat{\beta}_{-1}^{fe}|$. Whenever $|\widehat{\beta}_1^{fe}| \leq M|\widehat{\beta}_{-1}^{fe}|$, 0 is included between the lower and upper bounds for the ATT, so we cannot reject ATT $= 0$. Of course, $0 \notin \left[\widehat{\beta}_1^{fe} - M|\widehat{\beta}_{-1}^{fe}|, \widehat{\beta}_1^{fe} + M|\widehat{\beta}_{-1}^{fe}|\right]$ is not sufficient to reject ATT $= 0$: one also needs to take into account the sampling error in the estimation of the bounds. To use the bounds to construct a confidence interval for ATT, Rambachan and Roth (2023) leverage results from the moment inequality literature (see Andrews et al., 2019).

Sensitivity analysis. Practitioners may not have a good sense of which value of M they should choose. Rather than recommending a particular value, Rambachan and Roth (2023) recommend that they conduct the following sensitivity analysis. Assume that $\widehat{\beta}_1^{fe}$ is strictly positive and significantly different from zero. Under parallel trends, researchers would conclude that the treatment has a positive effect. But how robust is that conclusion to violations of parallel trends of a similar order magnitude as that observed from period 1 to 2, before the treatment onset? To answer that question, Rambachan and Roth (2023) propose to use M^*, the lowest value of M such that 0 belongs to the confidence interval of ATT. If $M^* = 5$, that means that even under differential trends five times larger from period 2 to 3 than from period 1 to 2, one can still conclude that the treatment had a positive effect: the researcher's conclusion is very robust to allowing for differential trends. On the other hand, $M^* = 0.2$ means that differential trends fives times smaller from period 2 to 3 than from period 1 to 2 are enough for the researcher's conclusion to break down, thus suggesting that results are not robust to plausible differential trends.

Generalization to multiple time periods. With more than three time periods, if $F \geq 3$ Assumption 9 can be generalized as follows: for all $t \geq F$, there is a positive real number M such that

$$\begin{aligned}
&\left| E\left[\frac{1}{G_1}\sum_{g:T_g=1}(Y_{g,t}(0) - Y_{g,t-1}(0))\right] - E\left[\frac{1}{G_0}\sum_{g:T_g=0}(Y_{g,t}(0) - Y_{g,t-1}(0))\right]\right| \\
\leq & M \max_{t' \in \{2,\ldots,F-1\}} \left| E\left[\frac{1}{G_1}\sum_{g:T_g=1}(Y_{g,t'}(0) - Y_{g,t'-1}(0))\right] - E\left[\frac{1}{G_0}\sum_{g:T_g=0}(Y_{g,t'}(0) - Y_{g,t'-1}(0))\right]\right|.
\end{aligned}$$

Fixing M, would you argue that the plausibility of the assumption above is: increasing or decreasing in $(T-F)/(F-2)$?

The assumption above states that the $T - F$ differential trends between pairs of consecutive post-treatment periods should all be below the highest differential trend between the $F-2$ pairs of consecutive pre-treatment periods. If, say $(T-F) = (F-2) = 5$, then we require that 5 post-

treatment differential trends are all bounded by the max of 5 pre-treatment differential trends. But if $(T-F)=50$ and $(F-2)=2$, we require that 50 post-treatment differential trends are all bounded by the max of 2 pre-treatment differential trends. For a fixed M, the assumption is more plausible in the former than in the latter case. In other words, the interpretation of M^* may depend on $(T-F)/(F-2)$. If $(T-F)/(F-2)=1$, $M^* \geq 1$ may be enough to consider that the results are robust. But if $(T-F)/(F-2)=25$, $M^* \geq 1$ may not be enough to consider that the results are robust.

Stata and R commands to compute M^* and assess the robustness of one's finding to plausible violations of the parallel trends assumption. M^* as well as the confidence intervals for ATT under varying values of M are computed by the `honestdid` Stata (see Bravo et al., 2022) and R (see Rambachan, 2022) commands. The basic syntax of the Stata command is:

`honestdid, numpre(#) b('beta') vcov('sigma')`,

where `#` is the number of pre-trends coefficients that have been estimated, `'beta'` contains the event-study coefficients, and `'sigma'` contains their variance-covariance matrix. Note that `honestdid` can also compute values of M^* and confidence intervals for ATT under other relaxations of parallel trends than that in Assumpion 9. In this course, we choose to present one of the relaxations of parallel trends proposed by Roth (2022), but they propose other relaxations.

Application 1: Benzarti and Carloni (2019). Based on Figure 2, do you expect that $M^* > 1$ or $M^* < 1$ in Benzarti and Carloni (2019)?

In Figure 2, there is a significant pre-trend coefficient, but the estimated effects are much larger than the pre-trends. Consistent with that, Figure 3 below, a reproduction of the left panel of Figure 5 in Rambachan and Roth (2023), shows that the estimated treatment effect in 2009, at the time of the reform, is fairly robust to violations of parallel trends: M^* is just below 2 for that parameter. The right panel of Figure 5 in Rambachan and Roth (2023) shows that the

estimated average treatment effect is a bit less robust: for that parameter, M^* is just below 1.

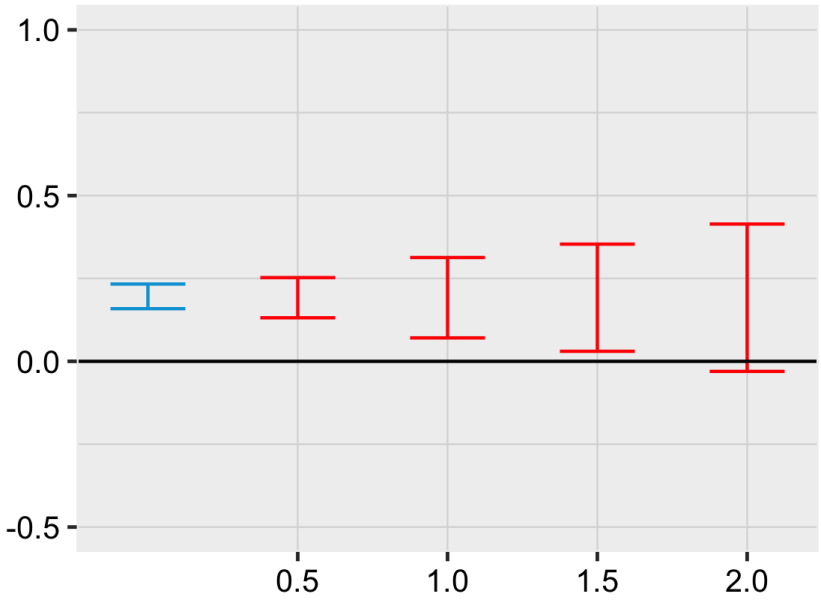


Figure 3: Confidence intervals for the effect of the VAT cut on firms' profits in 2009, under Assumption 9 and varying M.

Application 2: Lovenheim and Willén (2019). Roth (2019) also review Lovenheim and Willén (2019), who study the impact of state-level public sector duty-to-bargain (DTB) laws in the US, which mandated that school districts bargain in good faith with teachers' unions, on the wages of people who were students around the time that these laws were passed. The event-study graph for males' wages is shown in Figure 4 below. Based on Figure 4, do you think that the results in Lovenheim and Willén (2019) are robust to violations of parallel trends?

The estimated effects on Figure 4 are statistically significant while the pre-trends estimates are not, but the two sets of estimates are of a similar order of magnitude, and the confidence intervals attached to the pre-trends often contain the estimated effects. Consistent with that, Rambachan and Roth (2023) find that $M^* = 0.01$: minimal violations of parallel trends are

sufficient to overturn the paper's finding. Overall, those examples make it clear that while applications where the estimated effects are of a different magnitude than the estimated pre-trends can survive the sensitivity analysis proposed by Rambachan and Roth (2023), applications where the two sets of estimates are more comparable may not. In such cases, the data does not rule out the possibility that pre-trends account for most or all of the estimated effects.

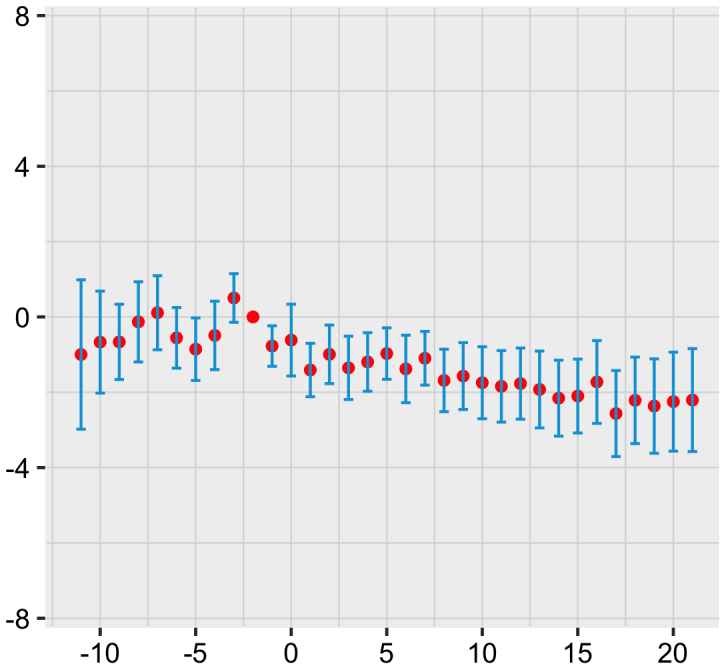


Figure 4: Event-study estimates of the effect of duty-to-bargain laws on males' wages.

5.3 Factor models and synthetic controls.

5.4 Grouped Patterns of Heterogeneity.

6 Binary and staggered designs.

Throughout this section, we assume that the treatment is binary and staggered, meaning that groups can switch into the treatment at heterogeneous dates and cannot switch out of the

treatment:

Design 2 *(Binary and staggered design)* $D_{g,t} = 1\{t \geq F_g\}$, *with* $\min_{g:F_g>1} F_g < \max_g F_g$.

F_g is the first date at which group g becomes treated, and group g remains treated thereafter. If g never becomes treated over the study period, we let $F_g = T+1$. F_g may be equal to 1, meaning that group g is always treated. $\min_{g:F_g>1} F_g < \max_g F_g$ requires that among groups that are untreated at period 1, not all groups get treated at the same period. If that condition fails, the two-way fixed effects regression in (3) is not identified. The only difference between Designs 1 and 2 is that in Design 2, the date when treated groups get treated can vary across groups: there can be variation in treatment timing. Let us illustrate those notation with a concrete example. Between 1968 and 1988, 29 US states adopted a unilateral divorce law (UDL), allowing one spouse to terminate the marriage without the consent of the other. Wolfers (2006), building upon Friedberg (1998), studies the effects of those laws on divorce rates. The UDL treatment satisfies Design 2: the treatment is binary, states adopt UDL laws at different dates, and they never repeal those laws. Then, F_g denotes the year when state g adopts a UDL.

6.1 TWFE regressions may not be robust to heterogeneous effects in binary and staggered designs.

6.1.1 $\widehat{\beta}^{fe}$ may be biased for the ATT and may not estimate a convex combination of effects.

Applying Theorem 1 in Design 2. First, note that in binary and staggered designs, it is easy to show that Theorem 1 still holds if the treatment has dynamic effects. Then, one just has that $\widehat{\beta}^{fe}$ identifies a weighted sum of effects of having been treated for $t - F_g + 1$ periods across all treated (g,t) cells, with the same weights as in Theorem 1. It follows from (11) and the definition of Design 2 that in binary and staggered designs, for all (g,t) such that $D_{g,t} = 1$,

$$\hat{u}_{g,t} = 1 - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot}. \tag{32}$$

$W_{g,t}$ is not constant across all (g,t) such that $D_{g,t} = 1$, so $\widehat{\beta}^{fe}$ may be biased for the ATT.

Are all the weights $W_{g,t}$ necessarily positive?

(32) also implies that some of the weights $W_{g,t}$ may be negative, if there are (g,t)s such that $1 + D_{.,.} < D_{g,.} + D_{.,t}$.

Which groups are the most likely to be such that some of the weights $W_{g,t}$ are negative for some t?

Groups whose average treatment $D_{g,.}$ is the highest. In Design 2, $D_{g,.} = (T - F_g + 1)/T$, so group whose average treatment is the highest are those for which F_g is the lowest, namely the groups that become treated early. Always treated groups are such that $D_{g,.} = 1$, so for them $\hat{u}_{g,t} = D_{.,.} - D_{.,t}$. As $D_{.,t}$ is weakly increasing in t in Design 2, $D_{.,T} > D_{.,.}$, so if there are always treated groups, their treatment effect at the last period is always weighted negatively by $\widehat{\beta}^{fe}$.

Which time periods are the most likely to be such that some of the weights $W_{g,t}$ are negative for some g?

The last time periods of the panel, because $D_{.,t}$ is weakly increasing in t in Design 2. To our knowledge, Borusyak and Jaravel (2017) were the first to note that $\widehat{\beta}^{fe}$ is more likely to assign a negative weight to treatment effects at the last periods of the panel in binary and staggered designs. This has led Jakiela (2021) to propose to estimate the TWFE regression after dropping the last periods of the data from the estimation, to mitigate or eliminate the negative weights. One could also drop the always-treated groups, if there are any.

When are all the weights $W_{g,t}$ likely to be positive?

All the weights are positive if and only if $D_{g,.} + D_{.,t} \leq 1 + D_{.,.}$ for all (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 period where most groups are treated. For instance, if a large proportion of groups are never treated, it is likely that $\widehat{\beta}^{fe}$ estimates a convex combination of effects, thus implying that $\widehat{\beta}^{fe}$ satisfies the no-sign reversal property. Even then, $\widehat{\beta}^{fe}$ may still be biased for the ATT.

6.1.2 The origin of the negative weights in binary and staggered designs.

The Goodman-Bacon decomposition. Goodman-Bacon (2021) shows that in Design 2,

$$\widehat{\beta}^{fe} = \sum_{g \neq g', t < t'} v_{g,g',t,t'} \text{DID}_{g,g',t,t'}, \tag{33}$$

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.[7] Some of the $DID_{g,g',t,t'}$s in Equation (33) compare a group switching from untreated to treated between t and 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 (4) originate from this second type of DIDs.

Forbidden comparisons. To see that, let us consider a simple example, first introduced by Borusyak and Jaravel (2017),[8] 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, the Goodman-Bacon decomposition reduces to

$$\widehat{\beta}^{fe} = (\text{DID}_{e,\ell,1,2} + \text{DID}_{\ell,e,2,3})/2, \tag{34}$$

[7] 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.

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

with

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

$\text{DID}_{e,\ell,1,2}$ compares the period-1-to-2 outcome evolution of group e, that becomes treated at period 2, to the outcome evolution of group ℓ that is untreated at both periods. $\text{DID}_{e,\ell,1,2}$ is similar to the DID estimator in Equation (1), and under no-anticipation and parallel trends assumptions it is unbiased for the treatment effect in group e at period 2:

$$E\left[\text{DID}_{e,\ell,1,2}\right] = E\left[TE_{e,2}\right]. \tag{35}$$

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

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

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

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

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

$$E\left[\text{DID}_{\ell,e,2,3}\right] = E\left[TE_{\ell,3} - TE_{e,3} + TE_{e,2}\right], \tag{36}$$

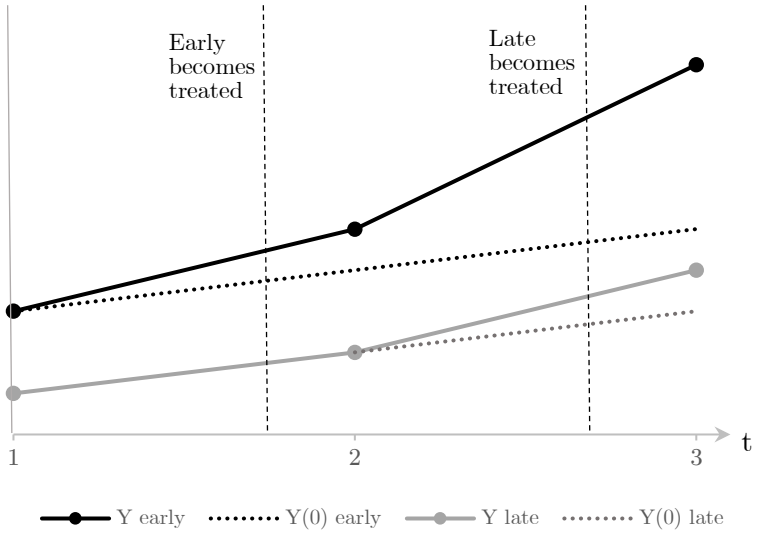
where $E\left[Y_{e,3}(0) - Y_{e,2}(0)\right]$ and $E\left[Y_{\ell,3}(0) - Y_{\ell,2}(0)\right]$ cancel out under the parallel trends assumption. Finally, it follows from Equations (34), (35), and (36) that

$$E\left[\widehat{\beta}^{fe}\right] = E\left[1/2TE_{\ell,3} + TE_{e,2} - 1/2TE_{e,3}\right]. \tag{37}$$

In this simple example, the decomposition of $\widehat{\beta}^{fe}$ in Theorem 1 in de Chaisemartin and D'Haultfœuille (2020) reduces to (37). The right-hand side of Equation (37) is a weighted sum of three treatment effects where one effect receives a negative weight. As the previous derivation shows, this negative weight comes from the fact $\widehat{\beta}^{fe}$ leverages $\text{DID}_{\ell,e,2,3}$, a DID comparing a group switching from untreated to treated to a group treated at both periods.

Numerical example. To make things more concrete, Figure 5 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 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 5: Example with three periods, an early and a late treated group



Find an assumption on the treatment effects $TE_{g,t}$ such that under that supplementary assumption, the negative weight in the decomposition of $\widehat{\beta}^{fe}$ in 37 disappears.

$\widehat{\beta}^{fe}$ estimates a convex combination of effects if treatment effects do not change over time. If one is ready to assume that the treatment effect does not change over time, $TE_{e,3} = TE_{e,2}$, and (36) simplifies to

$$E\left[\text{DID}_{\ell,e,2,3}\right] = E\left[TE_{\ell,3}\right]. \tag{38}$$

Then, the negative weight in (36) 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. This rules dynamic treatment effects. Even assuming away dynamic effects, this rules out the possibility that the treatment effect may change over time.

The Goodman-Bacon decomposition cannot be used to assess if $\widehat{\beta}^{fe}$ estimates a convex combination of effects. The decomposition in Equation (33) 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 Goodman-Bacon decomposition 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 the decomposition in Theorem 1 in de Chaisemartin and D'Haultfœuille (2020) 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 Theorem 1 in de Chaisemartin and D'Haultfœuille (2020) are positive, but the sum of the weights in the Goodman-Bacon decomposition 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 the Goodman-Bacon decomposition is necessary

but not sufficient to have negative weights in Theorem 1 in de Chaisemartin and D'Haultfœuille (2020). Similarly, the sum of the weights on DIDs comparing a switcher to a group treated at both periods in the Goodman-Bacon decomposition is always larger than the absolute value of the sum of the negative weights in Theorem 1 in de Chaisemartin and D'Haultfœuille (2020). The reason why the Goodman-Bacon decomposition "overestimates" the negative weights in Theorem 1 in de Chaisemartin and D'Haultfœuille (2020) 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 the Goodman-Bacon decomposition, putting less weight on DIDs using a group treated at both periods as the control group.[9]

Stata and R commands to compute the weights in the Goodman-Bacon decomposition. 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 (33), 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
```

[9]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

$$\text{DID}_{\ell,e,t_1,t_2} = \text{DID}_{\ell,n,t_0,t_2} - \text{DID}_{e,n,t_0,t_2} + \text{DID}_{e,\ell,t_0,t_1}. \tag{39}$$

Then, it follows from Equation (39) that

$$\begin{aligned} &v_{\ell,e,t_1,t_2}\text{DID}_{\ell,e,t_1,t_2} + v_{e,n,t_0,t_2}\text{DID}_{e,n,t_0,t_2} \\ =&(v_{\ell,e,t_1,t_2} - \underline{v})\text{DID}_{\ell,e,t_1,t_2} + \underline{v}\text{DID}_{\ell,n,t_0,t_2} + \underline{v}\text{DID}_{e,\ell,t_0,t_1} + (v_{e,n,t_0,t_2} - \underline{v})\text{DID}_{e,n,t_0,t_2}. \end{aligned} \tag{40}$$

Plugging Equation (40) into Equation (33) 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 (40), 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 (33) on DIDs using a group treated at both periods as the control group.

6.1.3 *TWFE event-study regressions are also not robust to heterogeneous effects, and may suffer from a contamination bias.*

TWFE event-study regressions. In Design 2, to estimate dynamic effects and test the no-anticipation and parallel-trends assumptions, researchers have often estimated the following TWFE event-study regression:

$$Y_{g,t} = \sum_{g'=1}^{G} \widehat{\alpha}_{g'} 1\{g = g'\} + \sum_{t'=1}^{T} \widehat{\gamma}_{t'} 1\{t = t'\} + \sum_{\ell=-K, \ell \neq 0}^{L} \widehat{\beta}_{\ell}^{fe} 1\{t = F_g - 1 + \ell\} + \hat{\epsilon}_{g,t}. \quad (41)$$

In words, the outcome is regressed on group and period fixed effects, and relative-time indicators $1\{t = F_g - 1 + \ell\}$ equal to 1 if at t, group g has been treated for ℓ periods. For $\ell \geq 1$, $\widehat{\beta}_{\ell}^{fe}$ is supposed to estimate the cumulative effect of ℓ treatment periods. For $\ell \leq -1$, $\widehat{\beta}_{\ell}^{fe}$ 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 $1\{t = F_g - 1 - K\}$ and $1\{t = F_g - 1 + L\}$ are respectively replaced by an indicator for being at least K periods away from the period before adoption ($1\{t \leq F_g - 1 - K\}$) and an indicator for having been treated for at least L periods ago ($1\{t \geq F_g - 1 + L\}$). Such endpoint binning is for instance recommended by Schmidheiny and Siegloch (2020): without it, even under constant effect the regression implicitly assumes that the treatment no longer has any effect after L periods. Instead, with endpoint binning the regression assumes that the treatment effect is constant after L periods, a more plausible assumption.

TWFE event-study regressions are not robust to heterogeneous effects, and may suffer from a contamination bias. Remember that for any integer $k \geq 1$, $\mathbf{1}_k$ denotes a vector of k ones. For all g such that $F_g \leq T$, and for $\ell \in \{1, ..., T - F_g + 1\}$, let

$$TE_{g,\ell} = E\left[Y_{g,F_g-1+\ell}(\mathbf{0}_{F_g-1}, \mathbf{1}_\ell) - Y_{g,F_g-1+\ell}(\mathbf{0}_{F_g-1+\ell})\right].$$

Interpret $TE_{g,\ell}$.

$TE_{g,\ell}$ is the expected effect, in group g and a period $F_g - 1 + \ell$, of having been treated rather than untreated from period F_g to $F_g - 1 + \ell$, namely for ℓ periods. The result below follows from Proposition 3 in Sun and Abraham (2021):

Theorem 4 *In Design 2, under Assumptions 1 and 3, for $\ell \in \{1, ..., L\}$,*

$$E\left[\widehat{\beta}_{\ell}^{fe}\right] = E\left[\sum_{g:F_g-1+\ell\leq T} w_{g,\ell} TE_{g,\ell} + \sum_{\ell'\neq\ell} \sum_{g:F_g-1+\ell'\leq T} w_{g,\ell'} TE_{g,\ell'}\right], \tag{42}$$

where $w_{g,\ell}$ and $w_{g,\ell'}$ are weights such that $\sum_{g:F_g-1+\ell\leq T} w_{g,\ell} = 1$ and $\sum_{g:F_g-1+\ell'\leq T} w_{g,\ell'} = 0$ for every ℓ'.[10]

The first summation in the right-hand side of Equation (42) is a weighted sum across groups of the cumulative effect of ℓ treatment periods, with weights summing to 1 but that may be negative. This first summation resembles that in the decomposition of $\widehat{\beta}^{fe}$ in Theorem 1 in de Chaisemartin and D'Haultfœuille (2020), and it implies that $\widehat{\beta}_{\ell}^{fe}$ may be biased if the cumulative effect of ℓ treatment periods varies across groups. The second summation is a weighted sum, across $\ell' \neq \ell$ and groups, of the cumulative effect of ℓ' treatment periods in group g, with weights summing to 0. This second summation was not present in the decomposition of $\widehat{\beta}^{fe}$. Importantly, its presence implies that $\widehat{\beta}_{\ell}^{fe}$, which is supposed to estimate the cumulative effect of ℓ treatment periods, may in fact be contaminated by the effects of ℓ' treatment periods. As $\sum_{g:F_g-1+\ell'\leq T} w_{g,\ell'} = 0$ 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.

If treatment effects are heterogeneous, TWFE event-study regressions cannot be used to test the no-anticipation and parallel-trends assumptions. For $\ell \leq -1$, and without making Assumptions 1 and 3, Sun and Abraham (2021) show that $\widehat{\beta}_{\ell}^{fe}$ estimates the sum of two terms. As intended, the first term measures differential 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 (42): a weighted sum, across $\ell' \geq 1$

[10] Equation (42) 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 in Theorem 4.

and groups, of the cumulative effect of ℓ' treatment periods in group g, with weights summing to zero. Due to the presence of this second term, the expectation of $\widehat{\beta}_{\ell}^{fe}$ 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, the TWFE event-study regression in (41) cannot be used to test for parallel trends.

Stata command to compute the weights attached to any TWFE event-study regression. The `eventstudyweights` Stata command (see Sun, 2020) computes the weights attached to TWFE 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 $1\{t = F_g - 1 + \ell\}$ included in (41), `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.

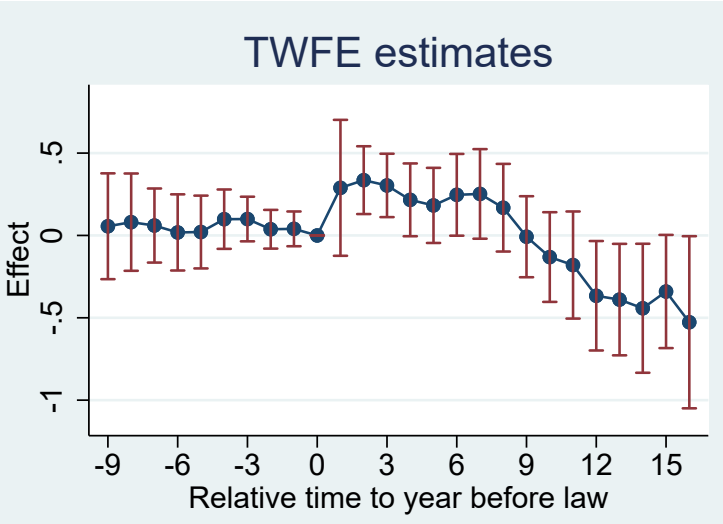
Application. Wolfers (2006) studies the effects of UDLs on divorce rates, using a yearly-panel of US states, and using the staggered adoption of UDLs in 29 states between 1968 and 1988. The author estimates a version of the event-study regression in (41), to estimate the effect of having been exposed to a UDL for up to 16 years. Figure 6 below shows those estimated effects, according to the event-study regression in (41), with $L = 16$, $K = 9$, and endpoint binning. According to this regression, UDLs increase the divorce rate for 8 years. After 12 years, the effect becomes significantly negative. The placebo estimates are small, and individually and jointly insignificant (F-test p-value=0.863). Those results are consistent with those in Column (1) of Table 2 of Wolfers (2006).[11] One can follow Sun and Abraham (2021), and compute

[11]The event-study regression in Figure 6 and that in Wolfers (2006) differ on two dimensions: Wolfers (2006) 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 adoption year and the year after that, one indicator for the two following years, etc.). Results seem fairly robust to those specification choices.

the weights attached to $\widehat{\beta}_1^{fe}$ in this event-study regression.[12] As shown in Equation (42), this coefficient can be decomposed as the sum of two terms. The first term is a weighted sum of effects of having been exposed to a UDL for one year, across 29 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 effects in states passing a law early, and downweights effects in states passing a law late. Accordingly, this first term may differ from ATT_1, the average effect of having been exposed to a UDL for one year across all states, if the effect of having been exposed to a UDL for one year varies between early- and late-adopting states, but it at least estimates a convex combination of effects. The second term is a weighted sum of being exposed to a UDL for more than one year. 29 effects of having been exposed to a UDL for two years 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 been exposed to a UDL for three years 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 been exposed to a UDL for four, five, ..., 15, and more than 16 years 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' effects vary across states, that second term may not be equal to zero, thus further biasing $\widehat{\beta}_1^{fe}$ with respect to its target parameter ATT_1. However, those contamination weights are not very large, so this bias is likely to be small. Overall, $\widehat{\beta}_1^{fe}$ seems fairly robust to heterogeneous treatment effects, presumably because there is a large share of never-treated groups in this application (around 40%). A similar analysis of the weights attached to the other event-study coefficients $\widehat{\beta}_\ell^{fe}$ for $\ell \geq 2$ suggest that they too are fairly robust to heterogeneous treatment effects.

[12] 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.

Figure 6: Effects of Unilateral Divorce Laws, using the data in Wolfers (2006): event-study regression



6.2 Heterogeneity-robust DID estimators in binary and staggered designs.

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

Target parameters. In Design 2, groups can be aggregated into cohorts that start receiving the treatment at the same period. Let $\mathcal{C} = \{c \in \{2, ..., T\} : \exists g : F_g = c\}$ denote the set of dates at which at least one group adopts the treatment. $\mathcal{C}$ is the set of all adoption cohorts. Let $\underline{c} = \min \mathcal{C}$ denote the earliest adoption cohort. For all $c \in \mathcal{C}$ and $t \in \{1, ..., T\}$, 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) and Sun and Abraham (2021) define a first set of parameters of interest as

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

for all $c \in \mathcal{C}$ and $\ell \in \{1, ..., T - c + 1\}$ ($T - c + 1$ is the number of periods for which cohort c has been treated at period T). $TE_{c,\ell}$ is the average effect of having been treated for ℓ periods in the cohort that started receiving the treatment at period c. Callaway and Sant'Anna (2021) and Sun

and Abraham (2021) also consider more aggregated parameters. Let N_c denote the number of groups in adoption-cohort c. Notice that $T-\underline{c}+1$ is the number of periods for which the earliest adoption cohort has been treated at period T. Then, for $\ell \leq T-\underline{c}+1$, let $N_\ell = \sum_{c:c-1+\ell\leq T} N_c$ denote the number of groups reaching ℓ treatment periods before period T, and let

$$\text{ATT}_\ell = \sum_{c:c-1+\ell\leq T} \frac{N_c}{N_\ell} TE_{c,\ell}.$$

Interpret ATT_ℓ.

ATT_ℓ is the average effect of having been treated for ℓ periods, across all groups reaching ℓ treatment periods before period T. ATT_ℓ generalizes the parameter ATT_ℓ defined in Section 4, to binary and staggered designs with variation in treatment timing. In that section, we have seen that without any restriction on treatment effect heterogeneity, $\ell \mapsto \text{ATT}_\ell$ cannot be used to determine if past treatments affect the outcome, and to disentangle the effects of the current treatment and of its lags on the outcome. This remains true here. But in binary and staggered designs, a new difficulty arises when it comes to interpreting $\ell \mapsto \text{ATT}_\ell$, which was not present in Section 4. Even without dynamic and time-varying effects, could we have that $\text{ATT}_\ell \neq \text{ATT}_{\ell'}$ in a binary and staggered design?

Yes, because for $\ell \neq \ell'$, ATT_ℓ and $\text{ATT}_{\ell'}$ do not apply to the same groups, as fewer and fewer groups reach ℓ treatment periods before T as ℓ increases. Thus, variations in ATT_ℓ across ℓ can come from dynamic treatment effects, time-varying effects, and compositional changes if treatment effects vary across groups.

Unbiased estimators. To estimate, say, $TE_{c,1}$, Callaway and Sant'Anna (2021) and Sun and Abraham (2021) propose

$$\widehat{\beta}^{cs,sa}_{c,1} = \overline{Y}_{c,c} - \overline{Y}_{c,c-1} - \left(\overline{Y}_{n,c} - \overline{Y}_{n,c-1}\right),$$

a DID estimator comparing the period $c-1$-to-c outcome evolution in cohort c and in the never-treated groups n. More generally, to estimate $TE_{c,\ell}$, Callaway and Sant'Anna (2021) and Sun and Abraham (2021) propose

$$\widehat{\beta}^{cs,sa}_{c,\ell} = \overline{Y}_{c,c-1+\ell} - \overline{Y}_{c,c-1} - \left(\overline{Y}_{n,c-1+\ell} - \overline{Y}_{n,c-1}\right),$$

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

Theorem 5 *In Design 2, under Assumptions 1 and 3, for* $c \in \{2, ..., T\}$ *and* $\ell \in \{1, ..., T-c+1\}$,

$$E\left[\widehat{\beta}^{cs,sa}_{c,\ell}\right] = TE_{c,\ell}. \tag{43}$$

Theorem 5 shows that $\widehat{\beta}^{cs,sa}_{c,\ell}$ is unbiased for a well-defined treatment effect parameter, $TE_{c,\ell}$, under Assumptions 1 and 3 alone, even if the treatment effect is heterogeneous, across groups or over time. Intuitively, why is it that unlike $\widehat{\beta}^{fe}$, $\widehat{\beta}^{cs,sa}_{c,\ell}$ is robust to heterogeneous treatment effects?

We saw that $\widehat{\beta}^{fe}$ is not robust to heterogeneous treatment effects, because it leverages DIDs comparing a group going from untreated to treated to a group treated at both periods. $\widehat{\beta}^{cs,sa}_{c,\ell}$ does not leverage such comparisons, as it compares groups going from untreated to treated to groups untreated at both periods. Based on Theorem 5, propose an unbiased estimator of ATT_ℓ.

It directly follows from Theorem 5 that

$$\widehat{\beta}_{\ell}^{cs,sa} \equiv \sum_{c:c-1+\ell \leq T} \frac{N_c}{N_\ell} \widehat{\beta}_{c,\ell}^{cs,sa}$$

is unbiased for ATT_ℓ.

Proof of Theorem 5

$$\begin{aligned}
&E\left[\widehat{\beta}_{c,\ell}^{cs,sa}\right] \\
=&E\left[\overline{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1}, \mathbf{1}_\ell) - \overline{Y}_{c,c-1}(\mathbf{0}_{c-1}) - \left(\overline{Y}_{n,c-1+\ell}(\mathbf{0}_{c-1+\ell}) - \overline{Y}_{n,c-1}(\mathbf{0}_{c-1})\right)\right] \\
=&E\left[\overline{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1}, \mathbf{1}_\ell) - \overline{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1+\ell})\right] \\
+&E\left[\overline{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1+\ell}) - \overline{Y}_{c,c-1}(\mathbf{0}_{c-1}) - \left(\overline{Y}_{n,c-1+\ell}(\mathbf{0}_{c-1+\ell}) - \overline{Y}_{n,c-1}(\mathbf{0}_{c-1})\right)\right] \\
=&E\left[\overline{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1}, \mathbf{1}_\ell) - \overline{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1+\ell})\right].
\end{aligned}$$

The first equality follows from the definitions of $\overline{Y}_{c,t}$ and $\overline{Y}_{n,t}$, Design 2, and Assumption 1. The second equality follows from adding and subtracting $\overline{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1+\ell})$. The third equality follows from Assumption 3 **QED.**

Pre-trends tests of Assumptions 1 and 3. For $c \in \{3, ..., T\}$ and $\ell \in \{1, ..., c-2\}$, let

$$\widehat{\beta}_{c,-\ell}^{cs,sa} = \overline{Y}_{c,c-1-\ell} - \overline{Y}_{c,c-1} - \left(\overline{Y}_{n,c-1-\ell} - \overline{Y}_{n,c-1}\right)$$

be a placebo DID estimator comparing the outcome evolution in cohort c and in the never-treated groups, from period $c-1$ to $c-1+\ell$, namely over ℓ periods before cohort c got treated. $\widehat{\beta}_{c,-\ell}^{cs,sa}$ exactly mimicks $\widehat{\beta}_{c,\ell}^{cs,sa}$, the estimator of the effect of having been treated for ℓ periods in cohort c. To mimick $\widehat{\beta}_{\ell}^{cs,sa}$, one may then use

$$\widehat{\beta}_{-\ell}^{cs,sa} \equiv \sum_{c:c-1+\ell \leq T, c-1-\ell \geq 1} \frac{N_c}{N_\ell^{\text{pl}}} \widehat{\beta}_{c,-\ell}^{cs,sa},$$

where $N_\ell^{\text{pl}} = \sum_{c:c-1+\ell \leq T, c-1-\ell \geq 1} N_c$, for any ℓ such that $N_\ell^{\text{pl}} > 0$. $\widehat{\beta}_{-\ell}^{cs,sa}$ almost perfectly mimicks $\widehat{\beta}_{\ell}^{cs,sa}$, except that groups such that $c-1-\ell < 1$ cannot be included in the placebo, because their outcome evolution over ℓ periods before they adopt the treatment is not observed. One can show that under Assumptions 1 and 3, $E\left[\widehat{\beta}_{-\ell}^{cs,sa}\right] = 0$, so one can reject Assumptions 1 and 3 if $\widehat{\beta}_{-\ell}^{cs,sa}$ is significantly different from zero. Note that this test of Assumptions 1 and 3 is robust to heterogeneous treatment effects, unlike the test based on the TWFE event-study regression in (41).

Sensitivity analysis under stationary differential trends. The approach proposed by Rambachan and Roth (2023) in Design 1 to bound ATT_ℓ and derive a confidence interval for it under a stationary differential trends assumption may also be used in Design 2, using the estimated effects $\widehat{\beta}_{\ell}^{cs,sa}$, the placebos $\widehat{\beta}_{-\ell}^{cs,sa}$, and their variance-covariance matrix as inputs to the procedure. Perhaps the only slight caveat is that because $\widehat{\beta}_{-\ell}^{cs,sa}$ does not apply to the exact same groups as $\widehat{\beta}_{\ell}^{cs,sa}$, the stationary parallel trends assumption underlying the procedure is a little bit less appealing here: the differential trends experienced by groups included in $\widehat{\beta}_{-\ell}^{cs,sa}$ may differ from the differential trends experienced by groups included in $\widehat{\beta}_{\ell}^{cs,sa}$.

Extensions. Callaway and Sant'Anna (2021) extend their estimators in a number of important directions. First, 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 $\widehat{\beta}_{c,1}^{cs,sa}$. This is very useful when there is no never-treated group: in that case, the effects $TE_{c,\ell}$ 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. Without never-treated groups, Sun and Abraham (2021) propose to use the last treated cohort as the control group, but this may result in imprecise estimators when that cohort is small. 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), which we will discuss later, also uses the not-yet-treated as controls, and is identical to the estimator of ATT_1 using the not-yet-treated as controls in Callaway and Sant'Anna (2021). Second, Callaway and Sant'Anna (2021) also propose estimators relying on a conditional parallel trends assumption, which extend the DID estimators with covariates reviewed in Section 5.1 to binary and staggered designs.

Stata and R commands to compute the estimators proposed by Callaway and Sant'Anna (2021). The estimators proposed by Callaway and Sant'Anna (2021) are computed by the `csdid` Stata command (see Rios-Avila, Sant'Anna and Callaway, 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.

Stata command to compute the estimators proposed by Sun and Abraham (2021). The estimators proposed by Sun and Abraham (2021) 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 (41), `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).

6.2.2 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 differ from those in Callaway and Sant'Anna (2021) and Sun and Abraham (2021). 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).

TWFE model allowing for arbitrarily heterogeneous effects. 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). To estimate $TE_{g,t}$, one can use the coefficient for treated cell (g,t) in this regression. Then, to estimate $TE_{c,\ell}$, one can use the average of all the $TE_{g,t}$s such that group g started receiving the treatment at period c and $t = c-1+\ell$. One can show that the resulting estimators are unbiased under Assumptions

1 and 3, and are therefore robust to heterogeneous treatment effects. Intuitively, why is it that those estimators are robust to heterogeneous treatment effects?

Because the TWFE regression in Borusyak et al. (2021) does not impose any restriction on treatment effect heterogeneity, as it has one coefficient per treated (g,t) cell. Instead of including fixed effects for group, period, and treated cells, Wooldridge (2021) shows that a TWFE regression of the outcome on cohort, period, and cohort$\times$ time-since-adoption fixed effects yields estimators of the effects $TE_{c,\ell}$ that are numerically equivalent to those of Borusyak et al. (2021). Another 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 (g,t) cells, and using that regression to predict the counterfactual untreated outcome of treated cells. Estimates of the treatment effect of those cells are then merely obtained by subtracting their counterfactual to their actual outcome. This imputation method is often computationally faster than the first two methods to obtain the estimators described above. It also readily generalizes to more complicated specifications, such as triple-differences, or models allowing for group-specific linear trends. 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 `did_imputation` 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 Borusyak et al. (2021),[13] but the result showing that the resulting estimators are efficient under (24) with iid errors, which we discuss below, only appears in Borusyak et al. (2021).

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

Non-linear models. Wooldridge (2022) proposes to use the strategy he proposed in Wooldridge (2021) in non-linear models such as logit or probit regressions. In a binary and staggered design, to estimate, say, a two-way fixed effects probit model robust to heterogeneous effects, one can include a set of cohort, period, and cohort× time-since-adoption fixed effects. Including cohort rather than group fixed effects is important, and not only for computational reasons, in non-linear models: including group fixed effects could lead to an incidental parameter problem (Neyman and Scott, 1948), which could severely bias the parameter estimates if T, the number of time periods of the panel, is low. If the number of groups is large relative to the number of cohorts, the strategy proposed by Wooldridge (2022) will not be subject to this incidental parameter problem. Note that in non-linear models, the regression with a full set of cohort, period, and cohort× time-since-adoption fixed effects is not equivalent to the imputation strategy where one would estimate the model with cohort and period fixed effects in the sample of untreated (g,t) cells, and use this model to predict the counterfactual latent outcome of treated cells.

6.2.3 Understanding the differences between those estimators

Which estimator is more efficient depends on the serial correlation over time of groups' potential outcomes. Under (24), and if errors are independent and identically distributed across g and t, Borusyak et al. (2021) show that their estimators are the BLUE of $TE_{c,\ell}$. In Section 4, we have seen that in binary designs without variation in treatment timing, the estimator of ATT_ℓ proposed by Borusyak et al. (2021) reduces to $\widehat{\beta}^{b,l,g}_{F-1+\ell}$, while that proposed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021) reduces to the event-study coefficient $\widehat{\beta}^{fe}_{F-1+\ell}$. Thus, in those designs the efficiency ranking of the two estimators reverses if one instead assumes that the errors in (24) follow a random walk. The results in Harmon (2022) show that this remains true in binary and staggered designs. He shows that the estimator of ATT_1 proposed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021) is the BLUE under (24), if errors follow a random walk. For $\ell > 1$, Harmon (2022) shows that neither the estimators of Callaway and Sant'Anna (2021) and Sun and Abraham (2021) nor those of Borusyak et al. (2021) are the BLUE under his random-walk-errors assumption. The BLUE of $TE_{c,\ell}$ is

$$\widehat{\beta}^{h,b}_{c,\ell} = \sum_{k=1}^{\ell} \left(\overline{Y}_{c,c-1+k} - \overline{Y}_{c,c-1+k-1} - \left(\overline{Y}_{nyt,c-1+k} - \overline{Y}_{nyt,c-1+k-1} \right) \right),$$

where for all t, $\overline{Y}_{nyt,t}$ denotes the average outcome of groups not-yet-treated at t. With the not-yet treated as the control group, the estimator of Callaway and Sant'Anna (2021) compares the $c-1$ to $c-1+\ell$ outcome evolution of cohort c and groups not-yet-treated at $c-1+\ell$. Instead, $\widehat{\beta}^{h,b}_{c,\ell}$ is a "chained" DID estimator, comparing the $c-1$ to c outcome evolution of cohort c and groups not-yet-treated at c, and adding to it a comparison of the c to $c+1$ outcome evolution of cohort c and groups not-yet-treated at $c+1$, ..., and adding to it a comparison of the $c-1+\ell-1$ to $c-1+\ell$ outcome evolution of cohort c and groups not-yet-treated at $c-1+\ell$. This chained DID estimator has also been proposed by Bellégo et al. (2023), as a way to estimate long-run treatment effects with an imbalanced panel, hence its "h, b" subscript. Bellégo et al. (2023) have also coined the chained DID terminology.

The estimators of Borusyak et al. (2021) may be more biased than those of Callaway and Sant'Anna (2021) and Sun and Abraham (2021) under differential trends that widen over time. In binary designs without variation in treatment timing, we have seen that the estimators of Borusyak et al. (2021) are more biased than those of Callaway and Sant'Anna (2021) and Sun and Abraham (2021) if Assumption 3 is violated with differential trends that widen over time, and less biased if Assumption 1 is violated due to anticipation effects arising a few periods before the treatment onset. We have also argued that violations of Assumptions 3 and 1 may not be equally problematic, as estimators can often be immunized against anticipation effects. The estimators of Borusyak et al. (2021) and those of Callaway and Sant'Anna (2021) and Sun and Abraham (2021) probably still have the same pros and cons in binary and staggered designs, though the fact that the estimators of Borusyak et al. (2021) do not have a simple closed-form expression in those designs makes it hard to ascertain. Simulations that would compare the bias of those estimators under violations of Assumptions 3 and 1 in binary and staggered designs would be useful.

The estimators of Callaway and Sant'Anna (2021) and Sun and Abraham (2021) are more amenable to the estimation approach under stationary differential trends in Rambachan and Roth (2023). As discussed above, one can construct placebo estimators that closely mimick the actual treatment effect estimators of Callaway and Sant'Anna (2021) and Sun and Abraham (2021), by comparing the outcome evolutions of (almost) the same

groups over the same number of periods, before the treatment onset. This makes those estimators amenable to the estimation approach under stationary differential trends proposed by Rambachan and Roth (2023). Building a placebo that would similarly mimick the estimator proposed by Borusyak et al. (2021) is not feasible, because that estimator leverages all pre-treatment periods to construct its baseline, as shown in Design 1.

6.2.4 Application

Figure 7 below shows estimates of the average effect of having been exposed to a UDL for ℓ years, for $\ell \in \{1, ..., 16\}$, using the data from Wolfers (2006), and according to four estimation methods. The figure also shows placebo estimates testing the parallel trends assumption up to 10 years before the adoption of UDLs, according to the same four estimation methods. The top-left panel of the figure replicates the TWFE event-study coefficients shown in Figure 6 above.

Estimators of Sun and Abraham (2021). The top-right panel of Figure 7 shows the estimators proposed by Sun and Abraham (2021), computed using the `eventstudyinteract` Stata command. The estimated effects are very similar to the TWFE event-study coefficients 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 by the analysis of the weights attached to this regression we conducted above. Interestingly, the confidence intervals are, if anything, slightly wider in the top-left than in the top-centre panel of Figure 7, 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.

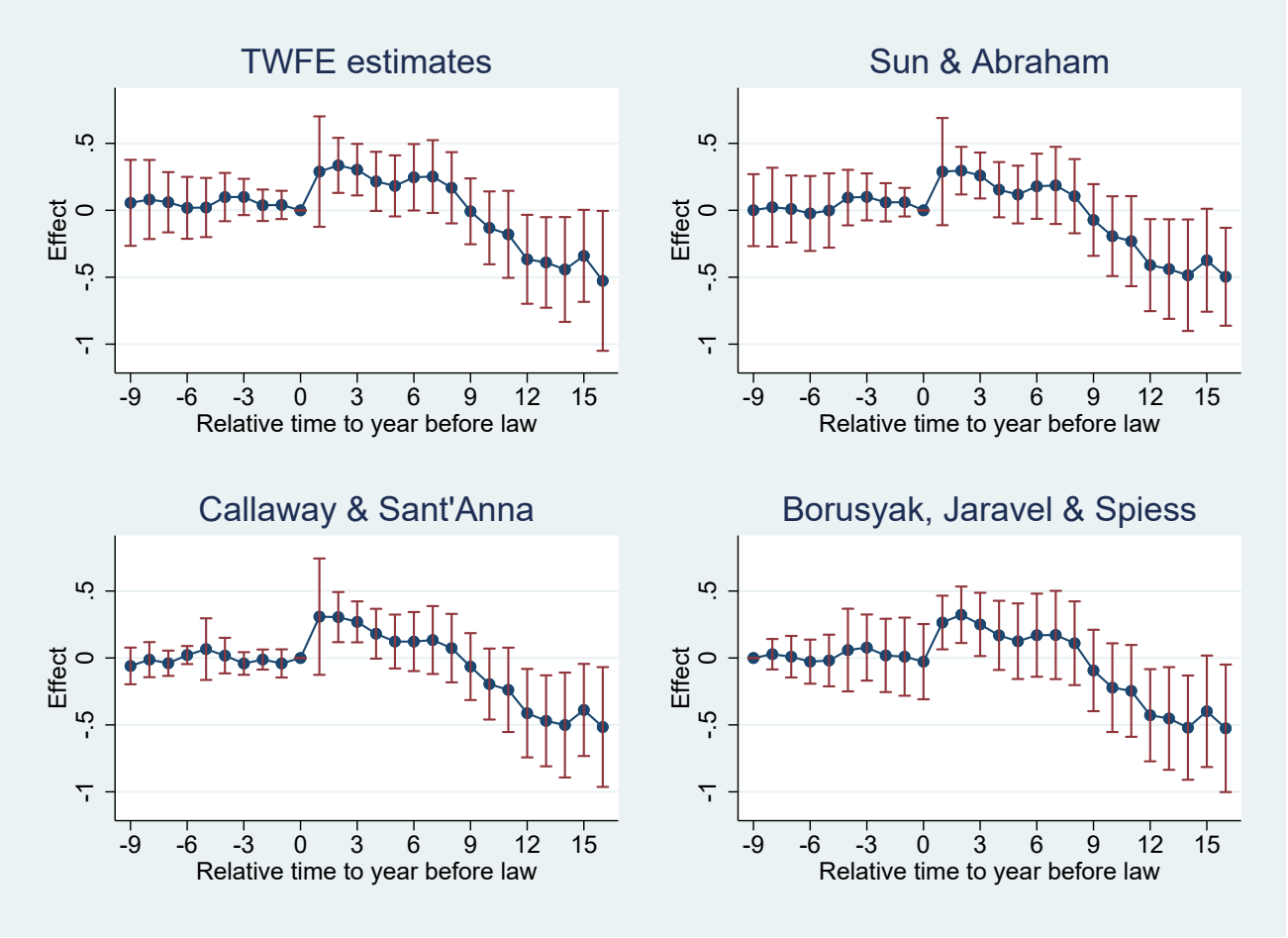
Estimators of Callaway and Sant'Anna (2021). The bottom-left panel of Figure 7 shows the estimators proposed by Callaway and Sant'Anna (2021), computed using the `csdid` Stata command, using the "not-yet-treated" states as the control group. The estimated effects are very similar to those in the top-centre 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. On the other hand, `eventstudyinteract` computes long-difference placebos.

Estimators of Borusyak et al. (2021). The bottom-right panel of Figure 7 shows the estimators proposed by Borusyak et al. (2021), computed using the `did_imputation` command. The effects are very similar to those found with the previous estimators. The confidence interval of the effect of having been exposed to a UDL for one year is much tighter in the bottom-right panel than in all other panels: for that effect, the estimator proposed by Borusyak et al. (2021) does lead to a large precision gain. However, the opposite can hold when one considers other effects. For instance, the confidence interval of the effect of having been exposed to a UDL for three years 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).[14] 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, and uses the leads' coefficients as the placebos. 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-right panel, instead of at $t = -1$ in the other panels.

[14] 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 7 are significant: it computes a joint nullity test, but for more disaggregated placebos.

Figure 7: Effects of Unilateral Divorce Laws, using the data in Wolfers (2006)



Note: This figure shows the estimated effects of Unilateral Divorce Laws on the divorce rate and placebo estimates, using the data in Wolfers (2006) and four estimation methods. In the top-left panel, we show estimated effects per the event-study regression in (41), with $L = 15$, $K = 10$, and endpoint binning. In the top-right (resp. bottom-left, bottom-right) panel, we show estimated effects per the `eventstudyinteract` (resp. `csdid`, `did_imputation`) Stata commands. 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.

6.3 Heterogeneity-robust synthetic control estimators in binary and staggered designs.

7 Heterogeneous adoption designs.

Set-up. Throughout this section, we assume that $T = 2$, and treatment follows an heterogeneous adoption design:[15] groups are untreated at period one, and receive a positive dose of treatment at period two, with some variation across groups:

Design 3 *(Heterogeneous adoption design)* $D_{g,1} = 0$*,* $D_{g,2} \geq 0$*, and* $\min_g D_{g,2} < \max_g D_{g,2}$*.*

We also assume that Assumption 2 holds, which is without loss of generality because $T = 2$ and groups are untreated at period 1.

Examples. Enikolopov et al. (2011) study the effect of NTV, an independent TV channel introduced in 1996 in Russia, on voting behavior, using region-level voting outcomes for the 1995 and 1999 elections. After 1996, NTV's coverage rate is heterogeneous across regions: while a large fraction of the population receives it in urbanized regions, a smaller fraction receives it in more rural regions. Their treatment is $D_{g,t}$, the proportion of the population having access to NTV in region g and election-year t. By definition, $D_{g,1} = 0$. Moreover, $D_{g,2} \geq 0$ and $\min_g D_{g,2} < \max_g D_{g,2}$, so the conditions in Design 3 are met. de Chaisemartin (2011) studies the effect of varenicline, a new smoking cessation treatment introduced in February 2007 in France. Among French smoking cessation clinics, some started prescribing it to a large percentage of their patients, while others only prescribed it to a small percentage. This also gives rise to an heterogeneous adoption design. By definition, $D_{g,1} = 0$: in every clinic 0% of patients receive varenicline prior to its introduction. Moreover, $D_{g,2} \geq 0$ and $\min_g D_{g,2} < \max_g D_{g,2}$: after varenicline's introduction, the proportion of patients receiving it varies across clinics.

Fuzzy designs. As the two examples above show, heterogeneous adoption designs are often (though not always) fuzzy designs, where $D_{g,t}$ is the average treatment of the individuals or

[15] To our knowledge, the terminology "heterogeneous adoption design" was coined by de Chaisemartin and D'Haultfœuille (2020), see Assumption S1 therein.

firms in cell (g,t), that may not all have the same treatment. For instance, in Enikolopov et al. (2011) some individuals in region g have access to the NTV channel in period two, while other individuals do not, and $D_{g,2}$ is the proportion of individuals that have access in region g. In fuzzy designs, our (g,t)-level potential outcome notation $Y_{g,t}(d)$ assumes that the outcome of g at t can only depend on the proportion of treated units in g at t, not on the identities of the treated units. Actually, the results below still hold if potential outcomes depend on the identities of the treated units. Letting $Y_{i,g,t}(0)$, $Y_{i,g,t}(1)$, and $Y_{i,g,t}$ denote the untreated, treated, and observed outcomes of unit i in group g at t, letting $N_{1,g,t}$ denote the number of treated units in group g at t, letting $TE_{g,t} = \frac{1}{N_{1,g,t}} \sum_{i:D_{i,g,t}=1} E[Y_{i,g,t}(1) - Y_{i,g,t}(0)]$ denote the ATT across all the treated units in g at t, and letting $Y_{g,t}(0)$ and $Y_{g,t}$ denote the average untreated and observed outcomes across the $N_{g,t}$ units in group g at t, we have:

$$\begin{aligned}
E[Y_{g,t}] =& E\left[\frac{1}{N_{g,t}}\sum_{i=1}^{N_{g,t}} Y_{i,g,t}\right] \\
=& E\left[\frac{1}{N_{g,t}}\sum_{i=1}^{N_{g,t}} (Y_{i,g,t}(0) + D_{i,g,t}(Y_{i,g,t}(1) - Y_{i,g,t}(0)))\right] \\
=& E[Y_{g,t}(0)] + E\left[\frac{N_{1,g,t}}{N_{g,t}}\frac{1}{N_{1,g,t}}\sum_{i:D_{i,g,t}=1} (Y_{i,g,t}(1) - Y_{i,g,t}(0))\right] \\
=& E[Y_{g,t}(0)] + D_{g,t}TE_{g,t}, \qquad (44)
\end{aligned}$$

where the last equality follows from the fact the design is conditioned upon. Therefore, (6) still holds in a fuzzy design, without assuming that the potential outcomes of cell (g,t) can only depend on the proportion of treated units in g at t, after a slight change of our definition of $TE_{g,t}$. Then, one can for instance show that Theorem 1 holds in fuzzy designs, as de Chaisemartin and D'Haultfœuille (2020) do in the proof of their theorem S1. Similarly, in fuzzy designs the results we show below do not assume that the potential outcomes of cell (g,t) can only depend on the proportion of treated units in g at t.

7.1 Decomposition of $\widehat{\beta}^{fe}$ in heterogeneous adoption designs.

Applying Theorem 1 in Design 3. In Design 3,

$$\hat{u}_{g,2} = D_{g,2} - D_{g,2}/2 - D_{.,2} + D_{.,2}/2 = 1/2(D_{g,2} - D_{.,2}). \qquad (45)$$

Then, it directly follows from Theorem 1 that[16]

$$E\left[\widehat{\beta}^{fe}\right] = \frac{1}{N_1} \sum_{g:D_{g,2}>0} \frac{D_{g,2}(D_{g,2} - D_{.,2})}{\frac{1}{N_1}\sum_{g':D_{g',2}>0} D_{g',2}(D_{g',2} - D_{.,2})} \text{TE}_{g,2}. \tag{46}$$

The weights vary across groups, so $\widehat{\beta}^{fe}$ may not be unbiased for the ATT. Are all the weights in (46) positive?

Not necessarily. If there is a g such that $0 < D_{g,2} < D_{.,2}$, meaning that g's period-two treatment is strictly positive but below the average period-two treatment, then its period-two treatment effect is weighted negatively. Conversely, it is only if $D_{.,2} \leq \min_{g:D_{g,2}>0} D_{g,2}$ that $\widehat{\beta}^{fe}$ estimates a convex combination of effects. When is it likely that $D_{.,2} \leq \min_{g:D_{g,2}>0} D_{g,2}$?

If there are many groups untreated at period 2. For any set A let $\#A$ denote the number of elements of that set, i.e. its cardinality. Let $G_0 = \#\{g : D_{g,2} > 0\}$ denote the number of untreated groups at period 2, and let $G_1 = G - G_0$ denote the number of treated groups.

$$D_{.,2} = \frac{1}{G}\sum_{g=1}^{G} D_{g,2} = \frac{1}{G}\sum_{g:D_{g,2}>0} D_{g,2} = \left(1 - \frac{G_0}{G}\right)\frac{1}{G_1}\sum_{g:D_{g,2}>0} D_{g,2}.$$

$\frac{1}{G_1}\sum_{g:D_{g,2}>0} D_{g,2}$ is necessarily larger than $\min_{g:D_{g,2}>0} D_{g,2}$, so the only way one can have that $D_{.,2} < \min_{g:D_{g,2}>0} D_{g,2}$ is if the proportion of untreated groups $\frac{G_0}{G}$ is "large enough", relative to the dispersion of $D_{g,2}$ among treated groups:

$$\frac{G_0}{G} \geq 1 - \frac{\min_{g:D_{g,2}>0} D_{g,2}}{\frac{1}{G_1}\sum_{g:D_{g,2}>0} D_{g,2}}.$$

[16]Proposition S1 in de Chaisemartin and D'Haultfœuille (2020) is equivalent to Equation (46), when the heterogeneous adoption design arises due to the heterogeneous adoption of an individual-level binary treatment across groups at period 2.

7.2 The origin of the negative weights in heterogeneous adoption designs.

Intuition for the negative weights in (46)**.** When $T = 2$, $\widehat{\beta}^{fe}$ is algebraically equivalent to $\widehat{\beta}^{fd}$, the first difference coefficient discussed in Section 3. Moreover, as $D_{g,1} = 0$, the first-differenced treatment is equal to $D_{g,2}$. Thus, $\widehat{\beta}^{fe}$ is algebraically equivalent to the coefficient on $D_{g,2}$ in a regression of $Y_{g,2} - Y_{g,1}$ on a constant and $D_{g,2}$. Then,

$$\widehat{\beta}^{fe} = \frac{\sum_{g=1}^{G}(D_{g,2} - D_{.,2})(Y_{g,2} - Y_{g,1})}{\sum_{g=1}^{G}(D_{g,2} - D_{.,2})^2}.$$

Intuitively, groups such that $D_{g,2} - D_{.,2} > 0$ are used as "treatment groups" by $\widehat{\beta}^{fe}$: their outcome evolution is weighted positively. On the other hand, groups such that $D_{g,2} - D_{.,2} < 0$ are used as "control groups": their outcome evolution is weighted negatively. In Design 3, $Y_{g,2} - Y_{g,1} = Y_{g,2}(0) - Y_{g,1}(0) + D_{g,2}TE_{g,2}$. The effect of the period-two treatment enters with a positive sign in the outcome evolution of control groups, so it gets weighted negatively by $\widehat{\beta}^{fe}$.

Forbidden comparisons in heterogeneous adoption designs. In heterogeneous adoption designs, the negative weights come from the fact $\widehat{\beta}^{fe}$ may leverage another type of forbidden comparison: $\widehat{\beta}^{fe}$ may compare the outcome evolution of a group m that is treated more at period two, to the outcome evolution of a group ℓ that is treated less. In fact, with two groups m and ℓ and two periods, one can show that

$$\widehat{\beta}^{fe} = \frac{Y_{m,2} - Y_{m,1} - (Y_{\ell,2} - Y_{\ell,1})}{D_{m,2} - D_{\ell,2}}. \tag{47}$$

The right hand side of (47) is a version of the Wald-DID estimator studied by de Chaisemartin and D'Haultfœuille (2018). de Chaisemartin and D'Haultfœuille (2018) have shown that this estimator does not always identify a convex combination of treatment effects. To see that, assume that group m receives two units of treatment at period 2, while group ℓ receives one unit. Then,

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

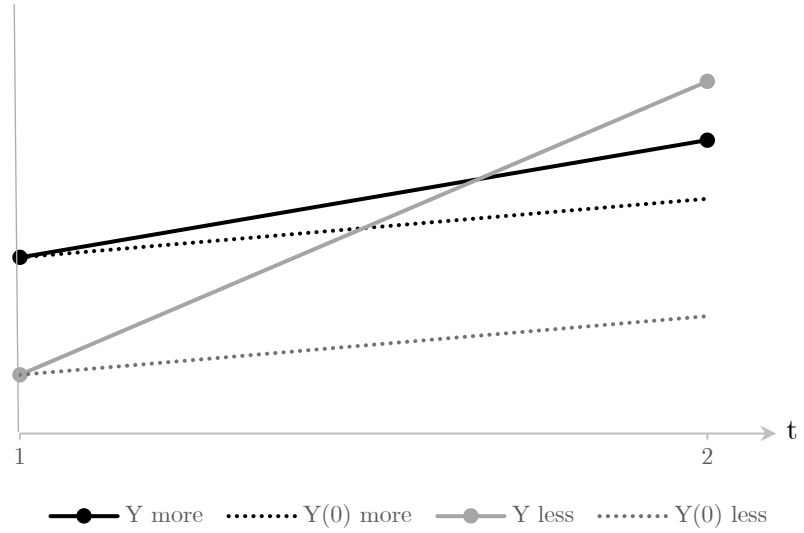
Then, under parallel trends,

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

where the last equality follows from the parallel trends assumption. The right-hand-side of the previous display is a weighted sum of m and ℓ's treatment effects, with weights summing to one, and 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. In the binary and staggered case, we have seen that $\widehat{\beta}^{fe}$ may not be to robust to time-varying treatment effects, but 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. The above example shows that in heterogeneous adoption designs, the opposite is true: $\widehat{\beta}^{fe}$ may not be to robust to heterogeneous effects across groups, but it is robust to time-varying effects if effects do not vary across groups. 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.

Numerical example. To make things more concrete, Figure 8 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 8: Example with two periods, and a more- and a less-treated group



7.3 Testing whether $\widehat{\beta}^{fe}$ is robust to heterogeneous treatment effects.

The results presented till the end of our study of heterogeneous adoption designs can be found in de Chaisemartin, D'Haultfoeuille and Gurgand (2022).

Independent and identically distributed groups. Till the end of our study of heterogeneous adoption designs, we will deviate from the conceptual framework we adopted so far. We will assume that the groups we observe are an independent and identically distributed (iid) sample, drawn from an infinite super-population of groups. When possible, we try to avoid resorting to this modelling framework, because we find it quite abstract, and because introducing it is not always necessary. Introducing it here is not strictly speaking necessary, but doing so will considerably reduce the notational burden. In that framework, we can drop the g subscript as groups are identically distributed. With identically distributed groups, Assumption 4 may seem to hold automatically, but this is because the conditioning on the design is left implicit in Assumption 4, so we need to make that conditioning explicit. With iid groups and under Assumption 5, Assumption 4 reduces to Assumption 10 below.

Assumption 10 *(Strong exogeneity for the untreated outcome) There is a real number* μ_0 *such*

that $E\left[Y_2(0) - Y_1(0)|D_2\right] = \mu_0$.

Assumption 10 requires that groups' untreated outcome evolution be mean independent of their period-two treatment, which is very similar to a strong exogeneity assumption in panel data models. This shows that parallel trends and strong exogeneity are, essentially, two different ways of expressing the same idea. With iid groups drawn from a super-population of groups, we assume that we have an heterogeneous adoption design in the super-population:

Design 3 *(Heterogeneous adoption design)* $D_1 = 0$*,* $D_2 \geq 0$*, and* $V(D_2) > 0$.

The last condition ensures that in the superpopulation, not all groups have the same period-two treatment. Note that this ensures that $\mathbb{P}(D_2 > 0) > 0$.

Target parameters. We consider two target parameters:

$$\begin{aligned}\text{ATT} \equiv& E\left(TE_2|D_2 > 0\right)\\ \text{ATT}_w \equiv& E\left[\frac{D_2}{E[D_2|D_2 > 0]}TE_2\middle|D_2 > 0\right].\end{aligned}$$

ATT is the average of the slopes of treated groups' potential outcome functions between 0 and their actual treatments, a parameter which generalizes the average treatment effect on the treated to settings with a non-binary treatment. ATT is just a super-population version of the ATT parameter we considered in previous sections. Note that by the law of iterated expectations,

$$E\left(TE_2|D_2 > 0\right) = E\left(E(TE_2|D_2)|D_2 > 0\right): \tag{48}$$

ATT is a weighted average of the period-two conditional ATEs (CATEs) $E(TE_2|D_2 = d_2)$, the ATE across groups in the superpopulation with $D_2 = d_2$, where $E(TE_2|D_2 = d_2)$ receives a weight proportional to $\mathbb{P}(D_2 = d_2)$, or proportional to $f_{D_2}(d_2)$, the density of D_2 at d_2, if D_2 is continuously distributed. ATT_w is a weighted average of treated groups' slopes, where groups with a larger period-two treatment receive more weight. We also consider ATT_w, because estimating ATT may sometimes be more difficult than estimating ATT_w. When there are treated groups with a value of D_2 close to zero, the denominator of TE_2 is close to zero for those groups. Then, estimators of ATT may suffer from a small-denominator problem, which could substantially increase their variance, and even affect their convergence rate (see Graham and

Powell, 2012; Sasaki and Ura, 2021; de Chaisemartin, D'Haultfoeuille, Pasquier and Vazquez-Bare, 2022, for similar issues in related contexts). On the other hand, it follows from the definition of TE_2 that

$$\text{ATT}_w = E\left[\frac{D_2}{E[D_2|D_2>0]}\frac{Y_2(D_2)-Y_2(0)}{D_2}\middle|D_2>0\right] = \frac{E[Y_2(D_2)-Y_2(0)|D_2>0]}{E[D_2|D_2>0]}, \tag{49}$$

so estimators of ATT_w are not affected by a small-denominator problem, even if there are treated groups with a value of D_2 close to zero.

Asymptotic decomposition of $\widehat{\beta}^{fe}$. In this modified framework, the result in (46) still holds, but it will be convenient to consider an asymptotic version of that result, as stated below. Under some technical conditions, one can show that under the strong exogeneity condition in Assumption 10, when the number of groups G goes to infinity,

$$\widehat{\beta}^{fe} \xrightarrow{\mathbb{P}} E\left(\frac{(D_2-E(D_2))D_2}{E\left((D_2-E(D_2))D_2|D_2>0\right)}E(TE_2|D_2)\middle|D_2>0\right). \tag{50}$$

(50) says that the plim of $\widehat{\beta}^{fe}$ is a weighted sum of the CATEs $E(TE_2|D_2=d_2)$, across all treated groups in the super population, where $E(TE_2|D_2=d_2)$ receives a weight proportional to $(d_2-E(D_2))d_2\mathbb{P}(D_2=d_2)$ (or proportional to $(d_2-E(D_2))d_2f_{D_2}(d_2)$ if D_2 is continuously distributed). As $(d_2-E(D_2))d_2 \neq 1$, (48) and (50) imply that $\widehat{\beta}^{fe}$ may not be consistent for ATT. Find a sufficient condition to have that the right-hand-side of (50) is equal to ATT.

If

$$1\{D_2>0\}E(TE_2|D_2) = 1\{D_2>0\}E\left(TE_2|D_2>0\right), \tag{51}$$

then

$$\begin{aligned}
&E\left(\frac{(D_2-E(D_2))D_2}{E\left((D_2-E(D_2))D_2|D_2>0\right)}E(TE_2|D_2)\middle|D_2>0\right)\\
=&E\left(TE_2|D_2>0\right)E\left(\frac{(D_2-E(D_2))D_2}{E\left((D_2-E(D_2))D_2|D_2>0\right)}\middle|D_2>0\right)\\
=&E\left(TE_2|D_2>0\right).
\end{aligned}$$

Intuitively, if treated groups with different period-two treatments have the same CATE, then the CATEs and the weights in (50) are uncorrelated, and $\widehat{\beta}^{fe}$ converges towards the ATT, a result similar to that we derived under (12) in Section 3. Note that (51) requires that the CATEs are constant across treated groups, but it does not require that untreated and treated groups have the same CATEs.

Testing the condition under which $\widehat{\beta}^{fe}$ is consistent for the ATT. If (51) holds, then $\widehat{\beta}^{fe}$ is consistent for the ATT. It turns out that (51) has a testable implication, and is fully testable in heterogeneous adoption designs with stayers or quasi-stayers.

Design 3' *(Heterogeneous adoption design with stayers or quasi-stayers) The conditions in Design 3 hold, and either of the two conditions below is satisfied:*

1. $\mathbb{P}(D_2 = 0) > 0$.
2. *D_2 is continuously distributed on $\mathbb{R}_+$ with density $d_2 \mapsto f_{D_2}(d_2)$ with respect to the Lebesgue measure such that $f_{D_2}(0) > 0$ and $d \mapsto f_{D_2}(d)$ is continuous.*

Point 1 of the condition in Design 3' holds when there are groups whose period-two treatment is equal to zero. Hereafter, those groups are referred to as stayers. Point 2 holds when D_2 is continuously distributed on $\mathbb{R}_+$, thus implying that there are no stayers ($\mathbb{P}(D_2 = 0) = 0$), but D_2 has a continuous density that is strictly positive at 0, thus implying that there are groups whose period-two treatment is "very close" to zero: for any $\delta > 0$, $\mathbb{P}(0 < D_2 < \delta) > 0$. Hereafter, those groups are referred to as quasi stayers.

Theorem 6 *Suppose that Assumption 10 holds.*

1. *In Design 3, if (51) holds, then there exist real numbers α_0 and α_1 such that $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$.*
2. *In Design 3', if there exist real numbers α_0 and α_1 such that $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$, then (51) holds.*

Is it possible to test whether there exist real numbers α_0 and α_1 such that $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$?

Yes, this amounts to testing whether the linear regression function of $Y_2 - Y_1$ on a constant and D_2 is equal to the CEF of $Y_2 - Y_1$ given D_2. To test whether $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$, one can for instance regress $Y_2 - Y_1$ on a high-order polynomial in D_2, and test whether the coefficients on the terms of degree higher than one are all equal to zero. A more principled way of conducting the test is to use a non-parametric regression technique to estimate $E(Y_2 - Y_1|D_2)$, and then compare the fit of the non-parametric model and of the linear regression of $Y_2 - Y_1$ on a constant and D_2 (see, e.g., Hardle and Mammen, 1993), though this non-parametric implementation of the test requires choosing tuning parameters.

If one rejects the null that there exist real numbers α_0 and α_1 such that $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$, can we reject (51), the sufficient condition under which $\widehat{\beta}^{fe}$ is consistent for the ATT?

Point 1 of Theorem 6 shows that (51) implies that there exist real numbers α_0 and α_1 such that $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$. By contraposition, if there does not exist real numbers α_0 and α_1 such that $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$, then (51) cannot hold. So if the test is rejected, (51) is rejected, and $\widehat{\beta}^{fe}$ may not be consistent for the ATT.

If one does not reject the null that there exist real numbers α_0 and α_1 such that $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$, can we claim that $\widehat{\beta}^{fe}$ is consistent for the ATT?

If there are stayers or quasi stayers, there is an "if and only if" relationship between (51) and $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$. Therefore, if there exist real numbers α_0 and α_1 such that $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$, $\widehat{\beta}^{fe}$ is consistent for the ATT. Then, the following estimation rule may seem reasonable: one uses $\widehat{\beta}^{fe}$ if a test of $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$ is not rejected, and one uses one of the heterogeneity-robust estimators proposed below if the test is rejected.

However, such pre-testing could lead to bias and size distortion if the power of the test of $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$ is low. Assessing the magnitude of the bias and size distortion that could arise from such pre-testing in simulations tailored to actual applications is a promising area for future research. If there are no stayers or quasi-stayers, we no longer have an "if and only if" relationship between (51) and $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$: $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$ could hold while (51) fails. Therefore, even if the test of $E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2$ is not rejected, $\widehat{\beta}^{fe}$ may not be consistent for the ATT outside of Design 3'.

Proof of Theorem 6. In Design 3, if (52) holds,

$$\begin{aligned}
E(Y_2 - Y_1|D_2) =& E(Y_2(D_2) - Y_1(0)|D_2) \\
=& E(Y_2(0) - Y_1(0)|D_2) + E(Y_2(D_2) - Y_2(0)|D_2) \\
=& E(Y_2(0) - Y_1(0)|D_2) + 1\{D_2 > 0\} E(Y_2(D_2) - Y_2(0)|D_2) \\
=& E(Y_2(0) - Y_1(0)|D_2) + 1\{D_2 > 0\} D_2 E((Y_2(D_2) - Y_2(0))/D_2|D_2) \\
=& \mu_0 + D_2 1\{D_2 > 0\} E(\text{TE}_2|D_2) \qquad (52) \\
=& \mu_0 + D_2 1\{D_2 > 0\} E(\text{TE}_2|D_2 > 0) \\
=& \mu_0 + D_2 E(\text{TE}_2|D_2 > 0),
\end{aligned}$$

where the last-but-two equality follows from Assumption 10 and the definition of TE_2, and the last-but-one equality follows from (52). This proves Point 1 of Theorem 6, with $\alpha_0 = \mu_0$, and $\alpha_1 = E(\text{TE}_2|D_2 > 0)$.

Then, assume that there exist real numbers α_0 and α_1 such that

$$E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2 = \alpha_0 + \alpha_1 D_2 1\{D_2 > 0\}. \qquad (53)$$

0 belongs to the support of D_2 in Design 3',[17] so $E(Y_2 - Y_1|D_2 = 0)$ is well defined. Equating (52) and (53) at $D_2 = 0$ yields $\alpha_0 = \mu_0$. Then, equating (52) and (53) implies that

$$D_2 1\{D_2 > 0\} E(\text{TE}_2|D_2) = \alpha_1 D_2 1\{D_2 > 0\},$$

and dividing the previous display by D_2 yields

$$1\{D_2 > 0\} E(\text{TE}_2|D_2) = \alpha_1 1\{D_2 > 0\}.$$

[17] Actually, for Point 2 of Theorem 6 to hold, it is enough to assume that 0 belongs to the support of D_2, which Design 3' is a special case of.

Taking the expectation of the previous display, dividing by $\mathbb{P}(D_2 > 0) > 0$ and using the law of iterated expectations yields $\alpha_1 = \text{ATT}$. This proves Point 2 of Theorem 6 **QED.**

7.4 Heterogeneity-robust DID estimators...

7.4.1 ... In designs with stayers or quasi-stayers.

In this section, we restrict attention to heterogeneous adoption designs with stayers or quasi-stayers, defined in Design 3'. Our target parameter is ATT_w, as heterogeneity-robust estimators of ATT may be very noisy with quasi-stayers, owing to the small-denominator problem mentioned above.

Identification of ATT_w with stayers or quasi-stayers.

Theorem 7 *Suppose that we are in Design 3' and Assumption 10 holds. Then,*

$$ATT_w = \frac{E[Y_2 - Y_1|D_2 > 0] - E[Y_2 - Y_1|D_2 = 0]}{E[D_2|D_2 > 0]}. \tag{54}$$

Theorem 7 shows that with stayers or quasi-stayers, ATT_w is identified by an estimand comparing the outcome evolution of treated and untreated groups, and scaling that comparison by the average treatment of treated groups. This estimand is a version of the Wald-DID estimand studied by de Chaisemartin and D'Haultfœuille (2018). It is robust to heterogeneous treatment effects, because it uses as controls groups that are untreated at both periods, a special case of the more general recommendation of de Chaisemartin and D'Haultfœuille (2018) of using as controls groups whose treatment is not changing over time.

Proof of Theorem 6. In Design 3', $E[Y_2 - Y_1|D_2 = 0]$ is well defined.

$$\begin{aligned}
&E[Y_2 - Y_1|D_2 > 0] - E[Y_2 - Y_1|D_2 = 0]\\
=&E[Y_2(D_2) - Y_1(0)|D_2 > 0] - E[Y_2(0) - Y_1(0)|D_2 = 0]\\
=&E[Y_2(D_2) - Y_2(0)|D_2 > 0] + E[Y_2(0) - Y_1(0)|D_2 > 0] - E[Y_2(0) - Y_1(0)|D_2 = 0]\\
=&E[Y_2(D_2) - Y_2(0)|D_2 > 0].
\end{aligned}$$

The first equality follows from Design 3. The third equality follows from Assumption 10. The result follows from the previous display and (49) **QED.**

Estimation of ATT$_w$ with stayers. In Design 3', if $\mathbb{P}(D_2 = 0) > 0$ it follows from Theorem 7 that with iid groups,

$$\widehat{\beta}^{het,s} = \frac{\frac{1}{G_1}\sum_{g:D_{g,2}>0}(Y_{g,2} - Y_{g,1}) - \frac{1}{G_0}\sum_{g:D_{g,2}=0}(Y_{g,2} - Y_{g,1})}{\frac{1}{G_1}\sum_{g:D_{g,2}>0} D_{g,2}}$$

converges towards ATT$_w$ when the number of groups tends to infinity. To compute $\widehat{\beta}^{het,s}$, one can merely run a 2SLS regression of $Y_{g,2} - Y_{g,1}$ on a constant and $D_{g,2}$, using $1\{D_{g,2} > 0\}$ as the instrument for $D_{g,2}$.

Estimation of ATT$_w$ without stayers but with quasi-stayers. In Design 3', if $\mathbb{P}(D_2 = 0) = 0$, estimating $E[Y_2 - Y_1|D_2 = 0]$ is not straightforward, as there are no stayers. We propose to use observations with D_2 lower than some bandwidth h to estimate $E[Y_2 - Y_1|D_2 = 0]$, by running a regression of $Y_2 - Y_1$ on D_2 in that subsample, and using the regression's intercept to estimate $E[Y_2 - Y_1|D_2 = 0]$. Intuitively, as the bandwidth h increases, the bias of the estimator of ATT$_w$ increases, as it uses groups that received a higher treatment dose to infer groups' counterfactual outcome trend without treatment. At the same time, as h increases, the variance of the estimator of ATT$_w$ decreases, as it estimates groups' counterfactual outcome trend without treatment out of a larger sample. This suggests that there might exist an optimal bandwidth, that trades off the estimator's bias and variance optimally. de Chaisemartin, D'Haultfoeuille and Gurgand (2022) derive a bandwidth minimizing an asymptotic approximation of the mean-squared error of the resulting estimator of ATT$_w$, in the spirit of the work of Imbens and Kalyanaraman (2012) for regression discontinuity designs. Letting $\sigma^2(d) = V(Y_2 - Y_1|D_2 = d)$ and $f(d) = E[Y_2 - Y_1|D_2 = d]$, they show that the optimal bandwidth is equal to

$$h^* = \left[\frac{144\sigma^2(0)}{Gf''(0)^2 f_{D_2}(0)}\right]^{1/5}.$$

Estimating h^* requires estimating $\sigma^2(0)$, $f''(0)$, and $f_{D_2}(0)$. $\sigma^2(0)$ may be estimated as the difference between the intercept in a local linear regression of $(Y_{g,2} - Y_{g,1})^2$ on $D_{g,2}$ and the square of the intercept in a local linear regression of $Y_{g,2} - Y_{g,1}$ on $D_{g,2}$. $f''(0)$ may be estimated as the coefficient on $D_{g,2}^2$ in a local quadratic regression of $Y_{g,2} - Y_{g,1}$ on $D_{g,2}$ and $D_{g,2}^2$. Finally, $f_{D_2}(0)$ may be estimated using a kernel density estimator. In each of those non-parametric estimations, the bandwidth may be selected using cross-validation. Then, to estimate ATT$_w$,

de Chaisemartin, D'Haultfoeuille and Gurgand (2022) propose to use

$$\widehat{\beta}^{het,qs} = \frac{\frac{1}{G}\sum_{g=1}^{G}(Y_{g,2} - Y_{g,1}) - \widehat{\mu}_{\widehat{h}^*}}{\frac{1}{G}\sum_{g=1}^{G} D_{g,2}},$$

with $\widehat{\mu}_{\widehat{h}^*}$ the intercept in the local linear regression of $Y_{g,2} - Y_{g,1}$ on $D_{g,2}$ for groups such that $D_{g,2} \leq \widehat{h}^*$.

Inference on ATT$_w$ without stayers but with quasi-stayers. Let

$$h_c^* = \left[\frac{144\sigma^2(0)}{f''(0)^2 f_{D_2}(0)}\right]^{1/5},$$

and let $\overline{D}_2 = \frac{1}{G}\sum_{g=1}^{G} D_{g,2}$. In view of the literature on similar estimators (see e.g. Calonico et al., 2014), de Chaisemartin, D'Haultfoeuille and Gurgand (2022) conjecture that conditional on the design, under regularity conditions,

$$G^{2/5}\left(\widehat{\beta}^{het,qs}_{\widehat{h}^*} - \text{AST}_w\right) \xrightarrow{d} N\left(\frac{f''(0)h_c^*}{12\overline{D}_2}, \frac{4\sigma^2(0)}{h_c^*\overline{D}_2^2}\right).$$

Thus, $\widehat{\beta}^{het,qs}_{\widehat{h}^*}$ converges towards AST$_w$ at the non-parametric $G^{2/5}$ rate, rather than at the standard $G^{1/2}$ parametric rate. This shows that while it is possible to propose estimators robust to heterogeneous effects in heterogeneous adoption designs without stayers, doing so comes with a cost in terms of statistical precision. Moreover, the asymptotic distribution of $G^{2/5}(\widehat{\beta}^{het,qs}_{\widehat{h}^*} - \text{AST}_w)$ has a first-order bias, which has to be estimated to construct valid confidence intervals, in the spirit of the work of Calonico et al. (2014) for regression discontinuity designs. One can follow the strategy used to estimate h^* to obtain consistent estimators

$$\widehat{B} \equiv \frac{\widehat{f''}(0)\widehat{h}_c^*}{12\overline{D}_2}$$

and

$$\widehat{V} \equiv \frac{4\widehat{\sigma}^2(0)}{\widehat{h}_c^*\overline{D}_2^2}$$

of the asymptotic bias and variance of $\widehat{\beta}^{het,qs}_{\widehat{h}^*}$. Then, letting q_x denote the quantile of order x of a standard normal, de Chaisemartin, D'Haultfoeuille and Gurgand (2022) conjecture that

$$\left[\widehat{\beta}^{het,qs}_{\widehat{h}^*} - \frac{\widehat{B}}{G^{2/5}} - q_{1-\alpha/2}\sqrt{\frac{\widehat{V}}{G^{4/5}}}, \widehat{\beta}^{het,qs}_{\widehat{h}^*} - \frac{\widehat{B}}{G^{2/5}} + q_{1-\alpha/2}\sqrt{\frac{\widehat{V}}{G^{4/5}}}\right]$$

is a confidence interval for AST_w with asymptotic coverage $1-\alpha$. Note that the estimator's estimated asymptotic variance $\widehat{V}$ is just the asymptotic variance of

$$\widehat{\beta}^{np,qs}_{\widehat{h}^*} \equiv \frac{\widehat{\mu}_h}{\frac{1}{G}\sum_{g=1}^{G} D_{g,2}}.$$

While it is asymptotically valid to not take into account the asymptotic variance of

$$\widehat{\beta}^{p,qs}_{\widehat{h}^*} \equiv \frac{\frac{1}{G}\sum_{g=1}^{G}(Y_{g,2}-Y_{g,1})}{\frac{1}{G}\sum_{g=1}^{G} D_{g,2}},$$

as $\widehat{\beta}^{p,qs}_{\widehat{h}^*}$ converges faster than $\widehat{\beta}^{np,qs}_{\widehat{h}^*}$, doing so may result in poor confidence interval coverage in finite samples. Then, letting $\widehat{\sigma}^2$ denote the sample variance of $Y_{g,2}-Y_{g,1}$, de Chaisemartin, D'Haultfoeuille and Gurgand (2022) finally propose to use

$$CI_{1-\alpha} \equiv \left[\widehat{\beta}^{het,qs}_{\widehat{h}^*} - \frac{\widehat{B}}{G^{2/5}} - q_{1-\alpha/2}\sqrt{\frac{\widehat{V}}{G^{4/5}} + \frac{\widehat{\sigma}^2}{G\overline{D}_2^2}}, \widehat{\beta}^{het,qs}_{\widehat{h}^*} - \frac{\widehat{B}}{G^{2/5}} + q_{1-\alpha/2}\sqrt{\frac{\widehat{V}}{G^{4/5}} + \frac{\widehat{\sigma}^2}{G\overline{D}_2^2}}\right]$$

as a $1-\alpha$ level confidence interval for ATT_w.[18]

7.4.2 ... In designs without stayers or quasi-stayers.

Identifying assumptions.

Assumption 11 *(Linear treatment effect model) There exists a random variable Δ_2 such that*

$$Y_2(d) = Y_2(0) + \Delta_2 d. \tag{55}$$

Δ_2 is a random variable, thus allowing for heterogeneous treatment effects across groups, though the treatment effect is assumed to be linear. Under Assumption 11, we define our target parameter as $\text{ATE} \equiv E[\Delta_2]$. We identify ATE under a parametric assumption.

Assumption 12 *There exist a known integer K, K known functions $f_1(d)$, ..., and $f_K(d)$ and K unknown real numbers δ_1, ..., δ_K such that*

$$E[\Delta_2|D_2=d] = \sum_{k=0}^{K} \delta_k f_k(d). \tag{56}$$

[18]Note that this confidence interval still omits the covariance between $\widehat{\beta}^{np,qs}_{\widehat{h}^*}$ and $\widehat{\beta}^{p,qs}_{\widehat{h}^*}$, that one may also want to take into account.

A leading example where Assumption 12 holds is if $E[\Delta_2|D_2 = d] = \delta_0 + \delta_1 d$, meaning that units with different values of D_2 may have different CATEs, but the relationship between their CATE and their value of D_2 is linear. Alternatively one could assume that $E[\Delta_2|D_2 = d]$ is a polynomial of order 2 or 3 in d. Under Assumptions 11 and 12,

$$\text{ATE} = E\left[\sum_{k=0}^{K} \delta_k f_k(D_2)\right],$$

so identifying $(\delta_0, ..., \delta_K)$ is sufficient to identify ATE.

Identification and estimation of ATE under Assumptions 10, 11, and 12. Under Assumptions 10, 11, and 12,

$$\begin{aligned} E[Y_2 - Y_1|D_2] =& E[Y_2(0) - Y_1(0)|D_2] + D_2 E[\Delta_2|D_2] \\ =& \mu_0 + \sum_{k=0}^{K} \delta_k D_2 f_k(D_2) \end{aligned} \tag{57}$$

It directly follows from (57) that $(\mu, \delta_0, ..., \delta_K)$ are the population coefficients from a regression of $Y_2 - Y_1$ on $(1, D_2 f_0(D_2), D_2 f_1(D_2), ..., D_2 f_K(D_2))$. Therefore, ATE is identified. To estimate ATE, one just needs to regress $Y_{g,2} - Y_{g,1}$ on $(1, D_{g,2} f_0(D_{g,2}), D_{g,2} f_1(D_{g,2}), ..., D_{g,2} f_K(D_{g,2}))$ and then use

$$\widehat{\beta}^{het,ns} \equiv \frac{1}{G} \sum_{g=1}^{G} \left(\sum_{k=0}^{K} \widehat{\delta}_k f_k(D_{g,2}) \right).$$

For instance, under the assumption that $E[\Delta_2|D_2 = d] = \delta_0 + \delta_1 d$, one just needs to regress $Y_{g,2} - Y_{g,1}$ on $(1, D_{g,2}, D_{g,2}^2)$ and then use

$$\widehat{\beta}^{het,ns} \equiv \frac{1}{G} \sum_{g=1}^{G} \left(\widehat{\delta}_0 + \widehat{\delta}_1 D_{g,2} \right)$$

to estimate ATE, where $\widehat{\delta}_0$ and $\widehat{\delta}_1$ respectively denote the coefficients on $D_{g,2}$ and $D_{g,2}^2$ in the regression of $Y_{g,2} - Y_{g,1}$ on $(1, D_{g,2}, D_{g,2}^2)$.

Applicability. $\widehat{\beta}^{het,ns}$ relies on Assumptions 11 and 12, which are strong parametric assumptions. $\widehat{\beta}^{het,ns}$ can be used in designs with stayers or quasi-stayers, though it may be not be very attractive in those designs. Then, $\widehat{\beta}^{het,s}$ or $\widehat{\beta}^{het,qs}$ can be used, and unlike $\widehat{\beta}^{het,ns}$, those estimators do not rely on Assumptions 11 and 12.

7.5 Application to Enikolopov et al. (2011)

Data and design. Enikolopov et al. (2011) study the effect of NTV, an independent TV channel introduced in 1996 in Russia, on voting behavior. After 1996, NTV's coverage rate is heterogeneous across regions: while a large fraction of the population receives it in urbanized regions, a smaller fraction receives it in more rural regions. In every region, at least 25% of the population has access to NTV after 1996. Thus this application is an heterogeneous adoption design, without stayers or quasi-stayers.

Scope of our re-analysis. Our re-analysis is concerned with Table 3 in Enikolopov et al. (2011), where the authors use $\widehat{\beta}^{fe}$ to estimate NTV's effect on five outcomes: the share of the electorate voting for the SPS and Yabloko parties, two opposition parties supported by NTV; the share of the electorate voting for the KPRF and LDPR parties, not supported by NTV; and electoral turnout. To do so, the authors compute $\widehat{\beta}^{fe}$, by regressing those outcomes on region fixed effects, an indicator for the 1999 election, and on the share of the population having access to NTV in each region at the time of the election. They find that $\widehat{\beta}^{fe} = 6.65$ (s.e.= 1.40) for the SPS voting rate, and $\widehat{\beta}^{fe} = 1.84$ (s.e.= 0.76) for the Yabloko voting rate. According to these regressions, increasing the share of the population having access to NTV from 0 to 100% increases the share of votes for the SPS and Yabloko opposition parties by 6.65 and 1.84 percentage points, respectively. $\widehat{\beta}^{fe}$ is statistically insignificant for the remaining three outcomes. Importantly, in our re-analysis we do not consider Table 2 in Enikolopov et al. (2011), where the authors only use the 1999 data and cross-sectional regressions to study the effect of NTV on voting behavior.

$\widehat{\beta}^{fe}$ does not estimate a convex combination of effects. We use the `twowayfeweights` Stata package to compute the weights attached to $\widehat{\beta}^{fe}$ in those five regressions. In 1995, all the weights are equal to zero because NTV does not exist yet. In 1999, 918 weights (47.4%) are strictly positive, while 1,020 (52.6%) are strictly negative. The negative weights sum to -2.26, so $\widehat{\beta}^{fe}$ is very far from estimating a convex combination of effects.

The test of (51) is rejected at the 5% level for three outcomes out of five. In spite of the negative weights attached to it, $\widehat{\beta}^{fe}$ may still consistently estimate the ATT if the condition

in (51) holds. Theorem 6 shows that (51) has a testable implication: there should exist real numbers α_0 and α_1 such that

$$E(Y_2 - Y_1|D_2) = \alpha_0 + \alpha_1 D_2.$$

We test this condition parametrically, by regressing $Y_2 - Y_1$ on a polynomial of order three in D_2, and by testing whether the coefficients on the squared and cubic terms are equal to zero. Those tests are rejected with p-values lower than 0.003 for the Yabloko and LDPR voting rates and for the turnout rate. For the KPRF voting rate, the p-value is equal to 0.06 so the test is marginally not rejected at the 5% level. For the SPS voting rate, the p-value is equal to 0.71 so the test is not rejected.

Estimators slightly more robust to heterogeneous effects are different from $\widehat{\beta}^{fe}$. Assuming that $E[\Delta_2|D_2 = d] = \delta_0 + \delta_1 d$, thus allowing for restricted effects' heterogeneity, already yields very different results from those obtained with $\widehat{\beta}^{fe}$, as shown in Table 1 below. $\widehat{\beta}^{ns}$ is close to $\widehat{\beta}^{fe}$ but not significantly different from zero for the SPS vote outcome. $\widehat{\beta}^{ns}$ is seven times larger than $\widehat{\beta}^{fe}$ and significantly different from zero for the Yabloko vote outcome. $\widehat{\beta}^{ns}$ and $\widehat{\beta}^{fe}$ are both insignificantly different from zero for the KPRF vote outcome. $\widehat{\beta}^{ns}$ is very large, negative, and significant for the LDPR vote and Turnout outcomes, while $\widehat{\beta}^{fe}$ is insignificant for those outcomes. Thus, allowing for heterogeneous effects, even in a fairly restricted way, yields noisy estimates, that often differ from the authors' original estimates. Note that for the LDPR vote and Turnout outcomes, $\widehat{\beta}^{ns}$ is implausibly large, which may indicate a violation of Assumption 10, 11, or 12.

Table 1: The Effects of NTV on Voting Behavior

	SPS vote	Yabloko vote	KPRF vote	LDPR vote	Turnout
$\widehat{\beta}^{fe}$	6.65	1.84	-2.20	1.18	-2.06
	(1.40)	(0.76)	(2.12)	(1.38)	(2.01)
$\widehat{\beta}^{ns}$	5.11	12.78	12.80	-39.18	-28.19
	(6.46)	(4.51)	(10.36)	(7.04)	(9.27)
Observations	1,938	1,938	1,938	1,938	1,938

Take-away from our re-analysis. The conclusions from Table 3 in Enikolopov et al. (2011) strongly rely on the assumption that treatment effects are homogeneous, while this assumption is rejected for several outcomes in that table. Remember that our findings do not apply to the cross-sectional regressions in Table 2 of their paper, where the authors regress 1999 voting outcomes on the 1999 exposure to NTV and some controls. The placebo analysis shown in their Table 4, where they run the same regressions as in their Table 2, but with 1995 instead of 1999 voting outcomes, and find insignificant effects of the 1999 exposure to NTV, suggest that treatment may be as good as randomly assigned in this application, conditional on the variables they control for (which they do not control for in their panel regressions). Thus, cross-sectional regressions with controls may be enough to estimate the treatment effect. Leveraging longitudinal variation as in their Table 3 renders the estimated effects non-robust to heterogeneous effects and may actually be worse than just relying on cross-sectional variation. Note that the estimated treatment effects are fairly different in their Tables 2 and 3: our re-analysis suggests that those in Table 2 may be more plausible.

8 General designs, ruling out dynamic effects.

In this section, we leave the design unrestricted: treatment may be binary or not, may increase over time for some groups but decrease for others, etc. The only assumption we make is that $D_{g,t} \geq 0$: we require that treatment be a positive variable, as is very often the case. We also maintain Assumption 2 throughout, thus ruling out dynamic effects. Of course, allowing for dynamic effects is appealing: one would like to use estimators relying on the weakest possible assumptions. However, as will become clear in the next section, in general designs, 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.

8.1 Decomposition of $\widehat{\beta}^{fe}$ in general designs when $T = 2$.

Assuming that $T = 2$ is sufficient to illustrate the forbidden comparisons that lead $\widehat{\beta}^{fe}$ to not be robust to heterogeneous effects in general designs, above and beyond the forbidden comparisons that already lead $\widehat{\beta}^{fe}$ to not be robust in simpler designs. When $T = 2$,

$$\begin{aligned}\hat{u}_{g,1} =& D_{g,1} - (D_{g,1} + D_{g,2})/2 - D_{.,1} + (D_{.,1} + D_{.,2})/2 = 1/2(D_{g,1} - D_{g,2} - (D_{.,1} - D_{.,2})),\\ \hat{u}_{g,2} =& D_{g,2} - (D_{g,1} + D_{g,2})/2 - D_{.,2} + (D_{.,1} + D_{.,2})/2 = 1/2(D_{g,2} - D_{g,1} - (D_{.,2} - D_{.,1})).\end{aligned} \tag{58}$$

Therefore, $\hat{u}_{g,1} = -\hat{u}_{g,2}$. Then, it directly follows from Theorem 1 that

$$\begin{aligned}&E\left[\widehat{\beta}^{fe}\right]\\ =&\sum_{g=1}^{G}\sum_{t=1}^{2}\frac{(D_{g,2} - D_{g,1} - (D_{.,2} - D_{.,1}))D_{g,t}(1\{t = 2\} - 1\{t = 1\})}{\sum_{g'=1}^{G}\sum_{t'=1}^{2}(D_{g',2} - D_{g',1} - (D_{.,2} - D_{.,1}))D_{g',t'}(1\{t' = 2\} - 1\{t' = 1\})}\text{TE}_{g,t}.\end{aligned} \tag{59}$$

If $D_{g,2} - D_{g,1} \neq D_{.,2} - D_{.,1}$ for all g, and $D_{g,t} > 0$ for all (g, t), which proportion of treatment effects $\text{TE}_{g,t}$ are weighted negatively in (59)?

If $D_{g,2} - D_{g,1} \neq D_{.,2} - D_{.,1}$ for all g, and $D_{g,t} > 0$ for all (g, t), $\widehat{\beta}^{fe}$ identifies a weighted sum of $\text{TE}_{g,t}$s where exactly a half of the weights are negative: for every g, either $\text{TE}_{g,2}$ or $\text{TE}_{g,1}$ is weighted negatively.

8.2 The origin of the negative weights in general designs.

Intuition for the negative weights in (59)**.** When $T = 2$, $\widehat{\beta}^{fe}$ is algebraically equivalent to $\widehat{\beta}^{fd}$, the first difference coefficient discussed in Section 3. Thus,

$$\widehat{\beta}^{fe} = \frac{\sum_{g=1}^{G}(D_{g,2} - D_{g,1} - (D_{.,2} - D_{.,1}))(Y_{g,2} - Y_{g,1})}{\sum_{g=1}^{G}(D_{g,2} - D_{.,2})^2}. \tag{60}$$

Moreover,

$$Y_{g,2} - Y_{g,1} = Y_{g,2}(0) - Y_{g,1}(0) + D_{g,2}\text{TE}_{g,2} - D_{g,1}\text{TE}_{g,1}. \tag{61}$$

Equation (60) shows that groups whose treatment-change from period one to two is larger than the average across groups ($D_{g,2} - D_{g,1} > D_{.,2} - D_{.,1}$) are used as "treatment groups" by $\widehat{\beta}^{fe}$: their outcome evolution is weighted positively. But then, (61) implies that their period-one treatment effect $\text{TE}_{g,1}$ gets weighted negatively. Conversely, Equation (60) shows that groups whose treatment-change from period one to two is lower than the average across groups ($D_{g,2} - D_{g,1} < D_{.,2} - D_{.,1}$) are used as "control groups" by $\widehat{\beta}^{fe}$: their outcome evolution is weighted negatively. But then, (61) implies that their period-two treatment effect $\text{TE}_{g,2}$ gets weighted negatively.

Even assuming constant treatment effects over time, $\widehat{\beta}^{fe}$ may still not estimate a convex combination of effects. If $\text{TE}_{g,2} = \text{TE}_{g,1} \equiv \text{TE}_g$, (59) simplifies to

$$E\left[\widehat{\beta}^{fe}\right] = \sum_{g=1}^{G} \frac{(D_{g,2} - D_{g,1} - (D_{.,2} - D_{.,1}))(D_{g,2} - D_{g,1})}{\sum_{g'=1}^{G}(D_{g',2} - D_{g',1} - (D_{.,2} - D_{.,1}))(D_{g',2} - D_{g',1})} \text{TE}_g. \tag{62}$$

To fix ideas, assume that $D_{.,2} - D_{.,1} > 0$: the average treatment increases from period one to two. Then, if there are groups whose treatment increases from period one to two ($D_{g,2} - D_{g,1} > 0$), but increases less than the average increase in the population ($D_{g,2} - D_{g,1} - (D_{.,2} - D_{.,1}) < 0$), their treatment effect is weighted negatively by $\widehat{\beta}^{fe}$.

Forbidden comparisons in general designs. With two groups m and ℓ and two periods, one can show that

$$\widehat{\beta}^{fe} = \frac{Y_{m,2} - Y_{m,1} - (Y_{\ell,2} - Y_{\ell,1})}{D_{m,2} - D_{m,1} - (D_{\ell,2} - D_{\ell,1})}. \tag{63}$$

The right hand side of (63) is the Wald-DID estimator studied by de Chaisemartin and D'Haultfœuille (2018). This estimator is even less robust to heterogeneous effects in general designs than in the heterogeneous adoption designs we considered in the previous section. To see that, assume that $D_{m,1} = 2$, $D_{m,2} = 4$, $D_{\ell,1} = 1$, and $D_{\ell,2} = 2$: group m receives two units of treatment at period one and four units at period two, and group ℓ receives one unit of treatment at period one, and two at period two. Then, under parallel trends,

$$\begin{aligned} E\left[\widehat{\beta}^{fe}\right] =& E\left[Y_{m,2}(4) - Y_{m,1}(2) - (Y_{\ell,2}(2) - Y_{\ell,1}(1))\right] \\ =& E\left[Y_{m,2}(0) - Y_{m,1}(0) - (Y_{\ell,2}(0) - Y_{\ell,1}(0))\right] + 4\text{TE}_{m,2} - 2\text{TE}_{m,1} - 2\text{TE}_{\ell,2} + \text{TE}_{\ell,1} \\ =& 4\text{TE}_{m,2} - 2\text{TE}_{m,1} - 2\text{TE}_{\ell,2} + \text{TE}_{\ell,1}, \end{aligned} \tag{64}$$

a weighted sum of m and ℓ's treatment effects at periods 1 and 2, with weights summing to one, and where two effects enter with negative weights. Assuming that treatment effects are constant across groups but not over time ($\text{TE}_{m,t} = \text{TE}_{\ell,t} \equiv \text{TE}_t$), the previous display reduces to

$$E\left[\widehat{\beta}^{fe}\right] = 2\text{TE}_2 - \text{TE}_1,$$

so unlike what happened in heterogeneous adoption designs, in general designs $\widehat{\beta}^{fe}$ may not identify a convex combination of effects even if effects are constant across groups but vary over time. Assuming that treatment effects are constant over time but not across groups ($\text{TE}_{g,2} = \text{TE}_{g,1} \equiv \text{TE}_g$), the previous display reduces to

$$E\left[\widehat{\beta}^{fe}\right] = 2\text{TE}_m - \text{TE}_\ell,$$

so unlike what happened in binary and staggered designs, in general designs $\widehat{\beta}^{fe}$ may not identify a convex combination of effects even if effects are constant over time but vary across groups. To our knowledge, the first appearance of a result related to that in (64) is in Lemma 1 of de Chaisemartin (2011), though the connection between the Wald-DID estimator and $\widehat{\beta}^{fe}$ was not noted therein.

8.3 Heterogeneity-robust DID estimators...

8.3.1 ... With a binary or discrete treatment.

In this section, we describe the DID_M estimator proposed by de Chaisemartin and D'Haultfœuille (2020), before comparing it to other heterogeneity-robust estimators than can be used in general designs under the assumption that the treatment does not have dynamic effects.

Target parameters. To simplify, let us first assume that $D_{g,t}$ is binary, and let

$$\text{ATE}_S \equiv \frac{1}{N_s} \sum_{(g,t):D_{g,t} \neq D_{g,t-1}} TE_{g,t},$$

where N_s is the number of (g,t) cells such that $D_{g,t} \neq D_{g,t-1}$, hereafter referred to as the switching cells. ATE_S is the average treatment effect across all the switching cells. Similarly, let

$$\text{ATE}_{S,+} \equiv \frac{1}{N_{s,+}} \sum_{(g,t):D_{g,t} > D_{g,t-1}} TE_{g,t},$$

where $N_{s,+}$ is the number of (g,t) cells such that $D_{g,t} > D_{g,t-1}$, hereafter referred to as the switching-in cells. $\text{ATE}_{S,+}$ is the average treatment effect across all the switching-in (g,t) cells. Finally, let

$$\text{ATE}_{S,-} \equiv \frac{1}{N_{s,-}} \sum_{(g,t):D_{g,t}<D_{g,t-1}} TE_{g,t},$$

where $N_{s,-}$ is the number of (g,t) cells such that $D_{g,t} < D_{g,t-1}$, hereafter referred to as the switching-in cells. $\text{ATE}_{S,-}$ is the average treatment effect across all the switching-out (g,t) cells.

Identifying assumptions. $\text{ATE}_{S,+}$ can be unbiasedly estimated under the following parallel trends assumption.

Assumption 13 *(Parallel trends, from $t-1$ to t and among groups untreated at $t-1$) For all $t \geq 2$, for all $g : D_{g,t-1} = 0$, $E\left[Y_{g,t}(0) - Y_{g,t-1}(0)\right]$ does not vary across g.*

Assumption 13 is a relaxation of Assumption 4, which only requires parallel trends on the untreated outcome, across consecutive time periods and among untreated groups at $t-1$. Similarly, $\text{ATE}_{S,-}$ can be unbiasedly estimated under the following parallel trends assumption.

Assumption 14 *(Parallel trends, from $t-1$ to t and among groups treated at $t-1$) For all $t \geq 2$, for all $g : D_{g,t-1} = 1$, $E\left[Y_{g,t}(1) - Y_{g,t-1}(1)\right]$ does not vary across g.*

Assumption 14 requires that the treated potential outcome of all groups that are treated at $t-1$ follow the same evolution from $t-1$ to t.

Estimators when $T = 2$. To simplify, we start by defining the estimators of ATE_S, $\text{ATE}_{S,+}$ and $\text{ATE}_{S,-}$ proposed by de Chaisemartin and D'Haultfœuille (2020) when $T = 2$. Then, let

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

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

and

$$\text{DID}_\text{M} = \frac{G_{0,1}}{G_{0,1}+G_{1,0}} \text{DID}_+ + \frac{G_{1,0}}{G_{0,1}+G_{1,0}} \text{DID}_-,$$

where for all $(d_1, d_2) \in \{0,1\}^2$, G_{d_1,d_2} denotes the number of groups such that $D_{g,1} = d_1$ and $D_{g,2} = d_2$.

Theorem 8 *1. Under Assumptions 2 and 13,* $E\left[DID_+\right] = ATE_{S,+}$.

2. Under Assumptions 2 and 14, $E\left[DID_-\right] = ATE_{S,-}$.

3. Under Assumptions 2, 13, and 14, $E\left[DID_M\right] = ATE_S$.

Intuitively, DID_+ is a DID comparing the period-one-to-two outcome evolution of switchers in and of groups untreated at both dates. Thus, DID_+ is similar to the simple DID estimator in Equation (1), and Point 1 of Theorem 8 shows under that Assumptions 2 and 13, it is unbiased for the ATE of switchers in. Similarly, DID_- is a DID comparing the period-one-to-two outcome evolution of groups treated at both dates, and of the switchers out. Then, DID_- is also similar to the DID estimator in Equation (1), switching "treatment" and "non-treatment", and Point 2 of Theorem 8 shows under that Assumptions 2 and 14, it is unbiased for the ATE of switchers out. Finally, the DID_M estimator is a weighted average of DID_+ and DID_-, with weights proportional to the numbers of switchers in and out, and Point 3 of Theorem 8 shows that under Assumptions 2, 13, and 14, it is unbiased for the ATE of switchers.

In applications with more than two time periods, how could you extend the DID_M estimator?

Extension to applications with more than two periods. The DID_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 $\text{DID}_{+,t}$ estimator comparing groups going from untreated to treated from $t-1$ to t to groups untreated at both dates, and a $\text{DID}_{-,t}$ 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 $\text{DID}_{+,t}$ across t to form the DID_+ estimator, one averages the $\text{DID}_{-,t}$ across t to form the DID_- estimator, and one averages the DID_+ and DID_- estimators to form the DID_M estimator. de Chaisemartin and D'Haultfœuille (2020) show that the resulting DID_+, DID_- and DID_M estimators are still unbiased for $\text{ATE}_{S,+}$, $\text{ATE}_{S,-}$, and ATE_S, under the same assumptions

as in Theorem 8. As will become clear below, considering pairs of consecutive time periods in isolation is only possible under the assumption that the treatment does not have dynamic effects.

Could you propose a placebo DID_M estimator, to test Assumptions 13 and 14?

Placebo estimators to test Assumptions 13 and 14. de Chaisemartin and D'Haultfœuille (2020) also propose a placebo estimator to test the parallel trends assumptions underlying DID_M. The placebo mimicks the actual estimator. The actual estimator compares the $t-1$-to-t outcome evolution of $t-1$-to-t switchers and stayers. The placebo compares the $t-2$-to-$t-1$ outcome evolution of $t-1$-to-t switchers and stayers, restricting attention to $t-2$-to-$t-1$ stayers. Without that restriction, $t-1$-to-t switchers and stayers could experience different $t-2$-to-$t-1$ outcome trends, not because of violations of Assumptions 13 and 14, but because some of those groups' treatment changes from $t-2$-to-$t-1$. Similarly, one can also construct a placebo comparing the $t-3$-to-$t-2$ outcome evolution of the $t-1$-to-t switchers and stayers, restricting attention to groups whose treatment does not change from $t-3$ to $t-2$. Finally, one can compute placebos separately for the switchers in and for the switchers out, if one wants to test Assumptions 13 and 14 separately.

Is DID_M robust to dynamic effects?

Is DID_M robust to dynamic effects? With more than two time periods, DID_M 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, $\text{DID}_{+,t}$ 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, DID_M is robust to dynamic effects in binary and staggered designs. In fact, in binary and staggered designs DID_M is equivalent to the estimator of ATT_1 using not-yet-treated groups as controls that was proposed by Callaway and Sant'Anna (2021).

Comparison with other estimators. With a binary treatment and no dynamic effects, at least two other estimators robust to heterogeneous treatment effects under parallel trends assumptions have been proposed. The multi-period DID estimator in Imai and Kim (2021) is numerically equivalent to the DID_+ estimator, so it can also be used to estimate switchers-in average treatment effect. The imputation estimator proposed by Borusyak et al. (2021) can be used to unbiasedly estimate the ATT under Assumption 4. With respect to DID_+, the imputation estimator is unbiased for an average treatment effect across a broader set of (g,t) cells, and therefore has a greater degree of external validity. On the other hand, the imputation estimator relies on Assumption 4, a stronger parallel trends assumption than that underlying DID_+. In non-staggered designs, this stronger parallel trends assumption justifies DID estimators that may be less natural than those leveraged by DID_+. Under Assumption 4, any DID comparing the outcome evolution of a group untreated at t' and treated at t to that of a group untreated at t' and at t is unbiased for $TE_{g,t}$, even when t' and t are not consecutive time periods, and even if the control group got treated at some point between t and t'. By contrast, DID_+ only leverages DIDs comparing the outcome evolution of switchers-in and untreated groups between consecutive time periods. To formalize the intuition that Assumption 13 may be more plausible than Assumption 4 in non-staggered designs, de Chaisemartin and D'Haultfoeuille (2022*a*) show that unlike the latter, the former assumption may be compatible with some types of Roy selection models, where groups choose their treatment based on the gains they expect from it, and with some types of Ashenfelter's dip models, where groups may choose to get treated after experiencing a series of negative shocks on their untreated outcome. Finally, note that while the authors do not discuss that extension, the imputation estimator of Borusyak et al. (2021) can be extended to non-binary and non-staggered treatments. Then, it

estimates the average treatment effect across all (g,t) cells such that g is untreated at one period at least, and there is at least one other group g' that is untreated at t. This set of (g,t) cells differs from the switching cells, which could lead the imputation and DID_M estimators, even if their respective parallel trends assumptions are satisfied.

In view of Theorem 8, is it the case that only DIDs comparing a group going from untreated to treated to a group untreated at both dates are robust to heterogeneous treatment effects?

More non-forbidden comparisons, and a guiding principle to extend DID_M to non-binary treatments. Point 2 of Theorem 8 shows that under Assumption 2 and a parallel trends assumption on the *treated* potential outcome, a DID estimator comparing groups treated at both dates to groups switching out of treatment is unbiased for the switchers' out treatment effect. More generally, one can show that under Assumption 2, a comparison of the outcome evolution of a switcher to the outcome evolution of a stayer with the same baseline treatment as the switcher can be used to estimate the switcher's treatment effect, under a parallel trends assumption in the counterfactual where the switcher and the stayer would have kept their period-one treatment. This principle guides our extension of DID_M to discrete treatments.

Extension to non-binary discrete treatments. The DID_M estimator can easily be extended to non-binary treatments taking a finite number of values. Then, it is a weighted average, across d and t, of DIDs comparing the $t-1$ to t outcome evolution of switchers whose treatment goes from d to some other value from $t-1$ to t, and of stayers with a treatment equal to d at both dates, normalized by the average absolute value of the treatment change experienced by switchers, to ensure the estimator can be interpreted as an average effect of increasing treatment by one unit. For instance, in Gentzkow et al. (2011), assume that $T=2$ and $G=4$. Assume also that $D_{1,1}=2$, $D_{1,2}=4$, $D_{2,1}=2$, $D_{2,2}=2$, $D_{3,1}=1$, $D_{3,2}=2$, $D_{4,1}=1$, and $D_{4,2}=2$: county 1 switches from 2 to 4 newspapers, county 2 stays at 2 newspapers, county 3 switches

from 1 to 2 newspapers, and county 4 stays at 1 newspaper. Then,

$$\text{DID}_\text{M} = \frac{1/2(Y_{1,2} - Y_{1,1} - (Y_{2,2} - Y_{2,1})) + 1/2(Y_{3,2} - Y_{3,1} - (Y_{4,2} - Y_{4,1}))}{1/2(D_{1,2} - D_{1,1}) + 1/2(D_{3,2} - D_{3,1})}.$$

With a non-binary treatment, the DID_M estimator relies on the following parallel trends assumption.

Assumption 15 *(Parallel trends over consecutive time periods, conditional on groups' baseline treatment) For all $t \geq 2$, for all d such that $D_{g,t-1} = d$ for at least one g, for all $g : D_{g,t-1} = d$ $E\left[Y_{g,t}(d) - Y_{g,t-1}(d)\right]$ does not vary across g.*

Compare Assumptions 15 and 4. Is one assumption weaker than the other? On which dimensions do the two assumptions differ?

Comparing Assumption 15 to Assumption 4, the usual parallel trends assumption. The two assumptions are non-nested, and there are two main differences between them. First, Assumption 4 requires that all groups be on parallel trends, over the entire duration of the panel. Assumption 15, on the other hand, only requires that groups with the same period-$t-1$ treatments be on parallel trends, from $t-1$ to t. Assumption 15 may then be more plausible: groups with the same treatments in the baseline period may be more similar, and may be more likely to experience parallel trends. Moreover, parallel trends may be more likely to hold over consecutive time periods than over the panel's entire duration. Second, Assumption 4 is a parallel trends assumption in the counterfactual where groups do not receive any treatment, while Assumption 15 is a parallel trends assumption in the counterfactual where groups' treatments do not change from $t-1$ to t. Accordingly, Assumption 4 only restricts one potential outcome, the one without any treatment, while Assumption 15 imposes restrictions on many potential outcomes. Still, Assumption 15 does not impose any restriction on treatment effect heterogeneity, because it restricts only one potential outcome per (g,t) cell, namely $Y_{g,t}(d)$ for (g,t) cells such that $D_{g,t-1} = d$. In particular, Assumption 15 does not require that all groups experience the same evolution of their treatment effect.

Under Assumption 15, $\widehat{\beta}_{fe}$ does not estimate a convex combination of effects. Importantly, there are special cases where: i) Assumptions 15 and 4 are equivalent and ii) the decomposition of $\widehat{\beta}_{fe}$ in Theorem 1 under Assumption 4 has negative weights attached to it. This shows that decomposing $\widehat{\beta}_{fe}$ under Assumption 15 can also lead to negative weights. For instance, with two periods and $D_{g,1} = 0$, Assumptions 4 and 15 are equivalent. As a result, negative weights may arise under Assumption 15, as the example in Section 7.2 demonstrates. In designs where Assumptions 15 and 4 are not equivalent, under Assumption 15 we cannot in general write $\widehat{\beta}_{fe}$ as a function of the design and treatment effects only, and replacing Assumption 4 by Assumption 15 may actually exacerbate the problems of TWFE regressions: in addition to not being robust to heterogeneous treatment effects, TWFE regressions may now be biased even with homogenous treatment effects, see de Chaisemartin and d'Haultfoeuille (2022*b*).

A taxonomy of forbidden and non-forbidden comparisons under Assumption 15. Under Assumption 15, the forbidden DIDs that are not robust to heterogeneous treatment effects, even under Assumption 1, are DIDs where:

1. a switcher is compared to another switcher with the same baseline treatment, as in Section 7.2, as those comparisons rule out heterogeneous effects across groups;

2. a switcher is compared to a stayer with a different baseline treatment, as in Section 6.1.2, as those comparisons rule out heterogeneous effects over time;

3. a switcher is compared to another switcher with a different baseline treatment, as in Section 8.2, as those comparisons rule out heterogeneous effects across groups and over time.

On the other hand, DIDs where a switcher is compared to a stayer with the same baseline treatment are valid under Assumption 15. Note that comparing a switcher to a stayer with a different baseline treatment but with the same endline treatment could be justified, under a parallel trends assumption where groups with the same period-t treatment experience parallel trends from $t-1$ to t in the counterfactual where they have their period-t treatment at both dates. For instance, with a binary treatment one could compare a group going from untreated to treated to a group treated at both dates, as suggested by Kim and Lee (2019). Note that jointly assuming that groups with the same baseline treatment are on parallel trends and groups

with the same endline treatment are on parallel trends may imply some restrictions on treatment effects: this may imply that treatment effects have to follow the same evolution over time in some groups, see de Chaisemartin and d'Haultfoeuille (2022*b*).

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

```
did_multiplegt outcome groupid timeid treatment
```

Application. We 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. We allow for state-specific trends, as the authors do in their first-difference regression. We find that $\text{DID}_\text{M} = 0.0045$ (s.e. $= 0.0011$), meaning that one more newspaper increases turnout by 0.45 percentage point.[19] DID_M is 73% larger than $\widehat{\beta}_{fd}$, the estimator reported by Gentzkow et al. (2011). They also find that the placebo DID_M estimator is equal to -0.010 and is insignificant (s.e. $= 0.0020$): counties that experience and do not experience a change in their number of newspapers from $t-1$ to t do not have significantly different evolutions of their turnout rate from $t-2$ to $t-1$. This lends credibility to the parallel trends assumption underlying DID_M. Alternatively, one could also compute the imputation estimator of Borusyak et al. (2021). Doing so leads to a much lower estimated effect, of 0.014 (s.e. $= 0.005$). The DID_M and imputation estimators apply to non-nested sets of respectively 4,596 and 6,931 (g,t) cells, with only around 2,000 cells in common: this could explain why the estimators differ. By definition, the imputation estimator only applies to groups observed with 0 newspapers at some point over the study period, and the treatment effect of those groups may differ from that of other groups. An alternative explanation for the fact the estimators differ is that they rely on different parallel trends assumptions, and the parallel trends assumption underlying the imputation estimator is rejected. We reestimate the imputation TWFE regression of turnout among (g,t) cells without any newspapers, adding $1\{D_{g,t+1} \geq 0\}$, an indicator for cells that will switch to having at

[19] The numbers shown here differ slightly from those in de Chaisemartin and D'Haultfœuille (2020): in that paper, the authors used the `fuzzydid` command to revisit Gentzkow et al. (2011), which yields very slightly different results.

least one newspaper in the next period, thus yielding a placebo in the spirit of that proposed by Borusyak et al. (2021) for binary and staggered designs. The placebo coefficient is equal to -0.009 and it is significant (s.e. $= 0.004$), thus showing that in counties that will go from 0 newspapers to a strictly positive number of newspapers from t to $t+1$, turnout decreases more from $t-1$ to t than in counties that will stay at 0 newspapers. Similarly, we reestimate the imputation TWFE regression of turnout among (g,t) cells without any newspapers, adding $1\{D_{g,t+2} \geq 0\}$, and again find a significantly negative coefficient (-0.009, s.e.=0.004). Similarly, we reestimate the imputation TWFE regression of turnout among (g,t) cells without any newspapers, adding $1\{D_{g,t+3} \geq 0\}$, and find a marginally insignificant negative coefficient (-0.007, s.e.=0.004). Thus, the imputation estimator may be downward biased by negative pre-trends. As the imputation estimator can compare groups over long horizons rather than over consecutive periods, small period-to-period differential trends may accumulate and finally yield a non-negligible bias.

8.3.2 ... With a continous treatment.

Estimators for treatments continuously distributed at every time period. de Chaisemartin, D'Haultfoeuille, Pasquier and Vazquez-Bare (2022) extend the DID_{M} estimator to treatments that are continuously distributed at every time period. Specifically, they assume that groups' treatments are drawn from a continuous distribution, thus meaning that all groups have a different treatment value. Their approach can also be applied to discrete treatments taking a large number of values, so that many treatment values are such that only a small number of groups have that value of the treatment. Extending the DID_{M} estimator to treatments continuously distributed at every time period is practically important: TWFE regressions have often been used to estimate the effect of trade tariffs (see Fajgelbaum et al., 2020) or precipitations (see Deschênes and Greenstone, 2007), which are continuously distributed at every time period. To simplify, we present their estimators in the case with two time periods, though they readily extend to the case with more periods.

Estimator in designs with stayers. de Chaisemartin, D'Haultfoeuille, Pasquier and Vazquez-Bare (2022) start by assuming that from period one to two, the treatment of some units does

not change: there are some stayers. 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, D'Haultfoeuille, Pasquier and Vazquez-Bare (2022) compares the outcome evolution of switchers and stayers, with the same period-one treatment. With a continuous treatment, a switcher cannot be matched to a stayer with the exact same period-one treatment. Still, comparisons of switchers and stayers with the same period-one treatment can either be achieved by reweigthing stayers by propensity score weights, or by adjusting switchers' outcome change using a nonparametric regression of the outcome change on the period-one treatment among the stayers. This gives rises to conditional DID estimators in the spirit of those we discussed in Section 5.1, but with groups' period-one treatment $D_{g,1}$ as the control variable. Under the following parallel trends assumption:

Assumption 16 *(Parallel trends condition on $D_{g,1}$) There exists a function $\gamma : d \mapsto \gamma(d)$ such that $E\left[Y_{g,2}(D_{g,1}) - Y_{g,1}(D_{g,1})\right] = \gamma(D_{g,1})$,*

one can show that a conditional DID estimator comparing the outcome evolution of switchers and stayers with the same $D_{g,1}$, and normalized by switchers' average treatment change, is consistent for a weighted average of the slopes of switchers' potential outcome function, between their period-one and period-two treatments.

Estimators in designs without stayers but with quasi-stayers. The estimators in de Chaisemartin, D'Haultfoeuille, Pasquier and Vazquez-Bare (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, in the spirit of the estimators proposed by de Chaisemartin, D'Haultfoeuille and Gurgand (2022) for heterogeneous adoption designs. Alternatively, one could also use the estimator proposed by Graham and Powell (2012), which compares the outcome evolution of switchers 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, D'Haultfoeuille, Pasquier and Vazquez-Bare (2022). When there are no true stayers, both esti-

mators require choosing a bandwidth, namely the lowest treatment change below which a unit can be considered as a quasi-stayer. Neither de Chaisemartin, D'Haultfoeuille, Pasquier and Vazquez-Bare (2022) or Graham and Powell (2012) derive an "optimal" bandwidth: determining the optimal bandwidth is more complicated when the treatment is continuously distributed at every time period than in an heterogeneous adoption design.

Stata and R commands to compute estimators in designs with a treatment continuously distributed at every time period. de Chaisemartin, D'Haultfoeuille, Pasquier and Vazquez-Bare (2022) show that after some relabelling, some of their estimators with stayers 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, D'Haultfoeuille, Pasquier and Vazquez-Bare (2022) for a precise description of how their estimators can be computed using existing software. The estimator of Graham and Powell (2012) can be computed by the Stata or R `gmm` command.

8.4 Correlated-random-coefficient estimator.

9 General designs, with dynamic effects.

10 Designs with several treatments, and estimating heterogeneous treatment effects.

11 Designs with randomized treatment timing or sequential randomization.

12 Panel Bartik designs.

References

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

Abadie, A., Diamond, A. and Hainmueller, J. (2010), 'Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program', *Journal of the American statistical Association* **105**(490), 493–505.

Abbring, J. H. and Van den Berg, G. J. (2003), 'The nonparametric identification of treatment effects in duration models', *Econometrica* **71**(5), 1491–1517.

Andrews, I., Roth, J. and Pakes, A. (2019), Inference for linear conditional moment inequalities, Technical report, National Bureau of Economic Research.

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

Bai, J. (2003), 'Inferential theory for factor models of large dimensions', *Econometrica* **71**(1), 135–171.

Bellégo, C., Benatia, D. and Dortet-Bernardet, V. (2023), 'The chained difference-in-differences', *arXiv preprint arXiv:2301.01085* .

Benzarti, Y. and Carloni, D. (2019), 'Who really benefits from consumption tax cuts? evidence from a large vat reform in france', *American Economic Journal: Economic Policy* **11**(1), 38–63.

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

Blundell, R., Dias, M. C., Meghir, C. and Van Reenen, J. (2004), 'Evaluating the employment impact of a mandatory job search program', *Journal of the European economic association* **2**(4), 569–606.

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

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

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

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

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

Botosaru, I. and Gutierrez, F. H. (2018), 'Difference-in-differences when the treatment status is observed in only one period', *Journal of Applied Econometrics* **33**(1), 73–90.

Bravo, M. C., Roth, J. and Rambachan, A. (2022), 'Honestdid: Stata module implementing the honestdid r package'.
URL: *https://EconPapers.repec.org/RePEc:boc:bocode:s459138*

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

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

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014), 'Robust nonparametric confidence intervals for regression-discontinuity designs', *Econometrica* **82**(6), 2295–2326.

Card, D. and Krueger, A. B. (1994), 'Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania', *The American Economic Review* **84**(4), 772–793.

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

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014), 'Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood', *American economic review* **104**(9), 2633–2679.

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

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

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

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

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

de Chaisemartin, C. and D'Haultfoeuille, X. (2022*a*), 'Not all differences-in-differences are equally compatible with outcome-based selection models'.

de Chaisemartin, C. and d'Haultfoeuille, X. (2022*b*), Two-way fixed effects and differences-in-differences estimators with several treatments, Technical report, National Bureau of Economic Research.

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

de Chaisemartin, C., D'Haultfoeuille, X. and Gurgand, M. (2022), 'Two-way fixed effects and differences-in-differences estimators for heterogeneous adoption designs', *Available at SSRN* .

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

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

de Chaisemartin, C. and Lei, Z. (2021), Are bartik regressions always robust to heterogeneous treatment effects? Available at SSRN 3802200.

De Tocqueville, A. (1850), *La démocratie en Amérique*, Pagnerre.

Dehejia, R. H. and Wahba, S. (1999), 'Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs', *Journal of the American statistical Association* **94**(448), 1053–1062.

Deschênes, O. and Greenstone, M. (2007), 'The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather', *American economic review* **97**(1), 354–385.

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

Fajgelbaum, P. D., Goldberg, P. K., Kennedy, P. J. and Khandelwal, A. K. (2020), 'The return to protectionism', *The Quarterly Journal of Economics* **135**(1), 1–55.

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

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

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

Goldsmith-Pinkham, P., Sorkin, I. and Swift, H. (2020), 'Bartik instruments: What, when, why, and how', *American Economic Review* **110**(8), 2586–2624.

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

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

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

Hardle, W. and Mammen, E. (1993), 'Comparing nonparametric versus parametric regression fits', *The Annals of Statistics* pp. 1926–1947.

Harmon, N. A. (2022), 'Difference-in-differences and efficient estimation of treatment effects'.

Heckman, J. J., Ichimura, H. and Todd, P. E. (1997), 'Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme', *The review of economic studies* **64**(4), 605–654.

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

Imbens, G. and Kalyanaraman, K. (2012), 'Optimal bandwidth choice for the regression discontinuity estimator', *The Review of economic studies* **79**(3), 933–959.

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

Kim, K. and Lee, M.-j. (2019), 'Difference in differences in reverse', *Empirical Economics* **57**, 705–725.

LaLonde, R. J. (1986), 'Evaluating the econometric evaluations of training programs with experimental data', *The American economic review* pp. 604–620.

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

Lovenheim, M. F. and Willén, A. (2019), 'The long-run effects of teacher collective bargaining', *American Economic Journal: Economic Policy* **11**(3), 292–324.

Malani, A. and Reif, J. (2015), 'Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform', *Journal of Public Economics* **124**, 1–17.

Manski, C. F. and Pepper, J. V. (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.

Mora, R. and Reggio, I. (2019), 'Alternative diff-in-diffs estimators with several pretreatment periods', *Econometric Reviews* **38**(5), 465–486.

Muris, C. and Wacker, K. (2022), 'Estimating interaction effects with panel data', *arXiv preprint arXiv:2211.01557* .

Neyman, J., Dabrowska, D. M. and Speed, T. P. (1990), 'On the application of probability theory to agricultural experiments. essay on principles. section 9.', *Statistical Science* pp. 465–472.

Neyman, J. and Scott, E. L. (1948), 'Consistent estimates based on partially consistent observations', *Econometrica: Journal of the Econometric Society* pp. 1–32.

Rambachan, A. (2022), 'Robust inference in difference-in-differences and event study designs'.
URL: *https://github.com/asheshrambachan/HonestDiD*

Rambachan, A. and Roth, J. (2023), 'A more credible approach to parallel trends', *Review of Economic Studies* p. rdad018.

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

Rios-Avila, F., Sant'Anna, P. H. and Naqvi, A. (2021), 'DRDID: Stata module for the estimation of Doubly Robust Difference-in-Difference models', Statistical Software Components, Boston College Department of Economics.
URL: *https://ideas.repec.org/c/boc/bocode/s458977.html*

Robins, J. (1986), 'A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect', *Mathematical modelling* **7**(9-12), 1393–1512.

Roth, J. (2022), 'Pretest with caution: Event-study estimates after testing for parallel trends', *American Economic Review: Insights* **4**(3), 305–22.

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

Rothstein, J. (2017), 'Measuring the impacts of teachers: Comment', *American Economic Review* **107**(6), 1656–1684.

Rubin, D. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of Educational Psychology* **66**(5).

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

Sant'Anna, P. H. C. and Zhao, J. (2022), 'DRDID: Doubly Robust Difference-in-Differences Estimators in R'.
URL: *https://cran.r-project.org/web/packages/DRDID/index.html*

Sant'Anna, P. H. and Zhao, J. (2020), 'Doubly robust difference-in-differences estimators', *Journal of Econometrics* **219**(1), 101–122.

Sasaki, Y. and Ura, T. (2021), 'Slow movers in panel data', *arXiv preprint arXiv:2110.12041* .

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

Smith, J. A. and Todd, P. E. (2005), 'Does matching overcome lalonde's critique of nonexperimental estimators?', *Journal of econometrics* **125**(1-2), 305–353.

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

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

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

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

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

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

Wooldridge, J. M. (2022), 'Simple approaches to nonlinear difference-in-differences with panel data', *Available at SSRN 4183726* .

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

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