Difference in Differences and Staggered Adoption Designs

Jaehyun Choi, Seth Freedman, Jennifer Mangano, Coady Wing

October 10, 2026

Abstract

The Difference in Differences (DID) design is a workhorse method of modern empirical research, widely used to estimate causal effects of policies and interventions. However, recent developments in the econometrics of DID estimators show that the standard two-way fixed effects (TWFE) estimator can misrepresent causal relationships when treatment timing is staggered or effects are heterogeneous. This chapter establishes a foundation for understanding more complex designs by developing the classic two-group, two-period setup. The chapter then examines two alternative modern approaches—the stacked DID estimator and an imputation-based estimator. These developments highlight the importance of separating estimation from aggregation (both over time and across units) to clarify what is being estimated. The chapter situates these innovations within a unified conceptual framework and provides applied researchers with practical insights for implementing credible, interpretable, and robust DID analyses.

> It is one of the happy incidents of the federal system that a single courageous State may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country.
>
> *Louis D. Brandeis*

1 — Introduction

The Difference in Differences (DID) design is the workhorse method of modern quantitative social science research, supporting causal inference under clear assumptions and with data requirements that are manageable in a diverse array of settings. A huge number of studies exploit changes across geographic areas to understand the way government policies and market conditions affect outcomes related to employment, earnings, education, health, housing, crime, charitable contributions, and more. Empirical strategies built around DID estimators are the most common way researchers analyze data generated by the federalist laboratory. Brandeis would be proud.

However, federalism is not the only complement to the DID design. The design also pairs well with surveillance-style data infrastructure, raising the value of repeated cross-sectional surveys, longitudinal surveys, research versions of administrative databases, and record linkage projects that combine information from multiple sources. For example, repeated cross-sectional surveys such as the Current Population Survey and large administrative datasets such as unemployment insurance records allow researchers to observe outcomes before and after policy changes across many jurisdictions.

In its most fundamental form, the DID is a comparison of the change in an outcome over two time periods in two groups. The treatment group starts out untreated and is exposed to treatment in the second period. The control group is untreated in both periods. The DID is simply the difference between the pre-post change in the treatment group and the pre-post change in the control group. Under the often plausible assumptions (SUTVA, common trends and no-anticipation), the DID is a consistent estimator of the average treatment effect in the treatment group.

The basic design is easy to understand. But researchers often face more complicated situations with more than two groups and more than two time periods. Exposure dates may be staggered so that the post period for one group is the pre-period for other groups.

Additionally, DID applications often raise practical complications. Statistical inference can be challenging due to clustering, which may arise from spatial dependence and serial

correlation within groups. In other settings, outcomes may be bounded, treatments may be multi-valued, or policies may be implemented and later rescinded. While these issues are important, this chapter focuses on challenges arising from staggered adoption and treatment effect heterogeneity.

Until recently, researchers analyzed data from these "generalized" forms of the DID using a two way fixed effects (TWFE) — a regression of outcomes on group fixed effects, time period fixed effects, and a treatment variable (Meyer, Viscusi, and Durbin 1995). The two way fixed effects regression is easy to work with and it seems to embody the common trends and no-anticipation assumptions required by the basic DID while also allowing for the messier complexity encountered in many applications with multiple groups and periods, dependent error structures, continuous and multi-valued treatments, and more.

Recent work by Goodman-Bacon (2018) and De Chaisemartin and d'Haultfoeuille (2020) revealed threats to the validity of the TWFE estimator that were not previously well understood. When treatment effects are heterogeneous and time varying, the standard TWFE estimator does not correspond to a sensible average of underlying causal effects. A key issue is that the standard TWFE estimator combines variation from multiple underlying DID-like comparisons, including comparisons of late vs early adopters which can violate the common trends assumption if treatment effects are time varying. These results tipped off a flurry of innovations. There is now a collection of new econometric estimators designed to identify and estimate coherent causal effects from staggered adoption DID designs without the problems involved in the TWFE estimator.

Methodological work in this space will likely continue in part because the DID design is such a central empirical strategy in the modern social sciences. Efforts to understand the strengths and weaknesses of alternative analytic strategies related to the design have high value. New methods that make the design more robust and convincing become widely used in top journals very quickly. It is already true that reviewers at top journals in fields like economics, public policy, and health policy expect studies using a DID design to account for challenges with staggered adoption using new methods if appropriate.

To clarify the key issues in modern DID research, the chapter uses a common framework to develop the DID design and its extensions, beginning with the most basic 2×2 design as a foundation for understanding recent debates.

The chapter is organized as follows. Section 2 develops the core DID framework, introduces the key identifying assumptions, and works through causal identification in the simple two-group, two-period case, with California's paid family leave program as a running empirical example (Rossin-Slater, Ruhm, and Waldfogel 2013). Section 3 turns to staggered adoption designs, presenting the stacked DID estimator of Wing et al. (2023) and the

fixed effect imputation estimator of Borusyak, Jaravel, and Spiess (2024); other approaches to the staggered adoption problem are reviewed in Wing et al. (2024). Section 4 discusses statistical inference, including cluster-robust standard errors and the wild cluster bootstrap. Section 5 concludes.

2 — Framework and Assumptions

2.1. Separating Estimation from Aggregation

The goal of a DID design is to estimate the *average treatment effect on the treated* (ATT), or the average causal impact of a policy on the units that actually received it. To illustrate this, consider the following example that will be used throughout the chapter. California adopted the nation's first universal paid family leave (PFL) program in July 2004, providing up to six weeks of partial wage replacement to workers taking leave to care for a newborn or seriously ill family member (Rossin-Slater, Ruhm, and Waldfogel 2013). The ATT in this setting is the causal effect of PFL on leave-taking among California mothers, or in other words, by how much did the probability of taking maternity or family leave increase for CA mothers *because of* the program? Rossin-Slater, Ruhm, and Waldfogel 2013 estimate that PFL increased the probability of taking maternity leave by approximately 6.3 percentage points. Answering this question requires a counterfactual. In this case, what would leave rates have looked like in California after 2004 if the state had not adopted PFL? The DID design answers this question by using trends in a comparison group to approximate that counterfactual. The rest of this section formalizes that logic.

In this section, we define notation and present the key assumptions of a DID research design. We begin with the case in which treatment adoption is not "staggered." In other words, all treated units adopt treatment in the same time period. However, we intentionally set up the notation to emphasize that *treatment effects* may vary over time. Later, we allow for staggered treatment and show why the traditional TWFE regression model leads to biased estimates if treatment is staggered *and* treatment effects do in fact vary over time.

Historically, DID papers primarily estimate a regression of the form:

$$Y_{s,t} = \beta_0 + \beta_1 Treat_s + \beta_2 Post_t + \beta_3 (Treat_s \times Post_t) + \varepsilon_{s,t}$$

Where $Y_{s,t}$ is the outcome for unit s at time period t, $Treat_s$ is an indicator for units that ever adopt the treatment, and $Post_t$ is an indicator for post treatment time periods. In the PFL example, $Y_{s,t}$ is the probability of being on maternity leave for mothers in state s in year t, $Treat_s = 1$ for California, and $Post_t = 1$ for years 2005 and later. The coeffi-

cient β_3 on the interaction between $Treat_s$ and $Post_t$ is the parameter of interest. In the PFL context, it estimates the ATT, the increase in leave-taking among California mothers attributable to the program. Often, especially when treatment is staggered, $Treat_s$ and $Post_t$ are replaced by unit and time fixed effects. However, unless treatment effects are constant both over time and across units, the estimate of β_3 is averaging, potentially, a great deal of heterogeneity. Researchers often supplement this basic DID regression with an "event study" regression to trace out how the effects vary over time (more detail on this point is provided in Section 2.4). The central insight of the modern DID literature is that there is not necessarily a single, unique treatment effect. Treatment effects may vary across units and over time; how to summarize that heterogeneity is a key substantive and methodological decision and a recurring theme in this literature. Rather than forcing all of this variation into a single regression coefficient, the modern DID literature encourages separating the full process into two steps:

1. Estimate well-defined average causal effects separately for each treatment cohort and each time period.
2. Aggregate to summary numbers across groups or time periods depending on the particular context or research question.

This can be seen in the stacked DID framework (Wing et al. 2023) and the fixed effect imputation framework (Borusyak, Jaravel, and Spiess 2024) that we outline below. It can also be found in methods from Callaway and P. H. C. Sant'Anna 2021 and others. Throughout our notation we emphasize this point by beginning with a model that allows for heterogeneity across both treatment cohorts and time, and pointing out where and how researchers may want to aggregate these effects.

2.2. Notation

In many DID applications, the data may contain individual observations (households, workers, patients, etc.) nested within groups (states, firms, hospitals, etc.). In such settings, treatment is often assigned at the group level, while outcomes are observed for each individual. Throughout the chapter, $i = 1, \ldots, N$ indexes individual units and $s = 1, \ldots, S$ indexes groups. Calendar time periods are indexed by $t = T_{min}, \ldots, T_{max}$. In the PFL example, i would be individual mothers, s would be states, and t would be survey years.[1] Treatment (the passage of a PFL law) is determined at the group-time level. In most of the

1. The index s is chosen because the Brandeis quote in the epigraph suggests that the natural unit of analysis is a state.

chapter, the analysis focuses on the case where treatment is an absorbing state[2], where once a group is exposed to treatment, it remains exposed forever. Let A_s represent the calendar time period when group s is first exposed to treatment. Some groups may never adopt the treatment during the study period. For these never-treated groups, $A_s = \infty$.

The causal identification argument is most easily stated using the concept of *potential outcomes* (Rubin 1974). The core idea is that each unit has a potential outcome under each possible treatment state, only one of which is actually observed.

Let $Y_{i,s,t}(0)$ represent the *potential outcome* of unit i from group s in time period t in a scenario where unit $i's$ group never adopts treatment. In contrast, $Y_{i,s,t}(a)$ represents the outcome that the same unit would experience in the same time period if her group had *initiated treatment for the first time* in period $t = a$. In this formulation, each person has a separate potential outcome associated with each possible year of adoption.[3] For any individual person, only one potential outcome is actually realized. The other potential outcomes are counterfactual. Let $\beta_{i,s,t}(a) = Y_{i,s,t}(a) - Y_{i,s,t}(0)$ represent the causal effect of adopting treatment for the first time in period a on the outcomes that unit i experiences in time period t. Then the realized outcome is $Y^{obs}_{i,s,t} = Y_{i,s,t}(0) + \sum_a \beta_{i,s,t}(a) \times 1(A_s = a)$. Writing the realized outcome this way emphasizes that the outcomes we observe in the data are equal to the untreated outcome plus the causal effect of treatment if the unit actually adopted treatment.

This notation accommodates both simple DID designs with only a single treatment adoption event as well as staggered adoption designs in which groups adopt treatment at varied times. In the simple DID case, we have a collection of units that are exposed to treatment some specified time a^*, and a collection of units that are never exposed. In this case the treatment group consists of units with $A_s = a^*$ and the control group consists of all groups with $A_s = \infty$. The observed outcome is $Y^{obs}_{i,s,t} = Y_{i,s,t}(0) + \beta_{i,s,t}(a^*)1(A_s = a^*)$ and there is only one DID comparison to consider for any given time period. In contrast, the staggered adoption case consists of collection of adoption events, each defining a treatment group, *clean* control group (one not yet exposed to treatment), post period, and causal effect.

2.3. Assumptions

Causal inference in the DID design depends on three key assumptions: SUTVA, common trends, and no-anticipation. This subsection lays out the formal conditions for a DID design

2. The absorbing state assumption rules out treatment reversals and policy repeals. In settings where treatment can be switched on and off, a more general framework is needed.

3. For example, in a study covering the years 2000 to 2003, person i from group s could have adopted treatment for the first time in periods $a \in \{2000, 2001, 2002, 2003, \infty\}$. For any given calendar time period, the person has five potential outcomes: $(Y_{i,s,t}(0), Y_{i,s,t}(2000), Y_{i,s,t}(2001), Y_{i,s,t}(2002), Y_{i,s,t}(2003))$.

to identify causal effects. These conditions are rarely exactly satisfied in practice, and that is to be expected. Exploring the extent to which they hold — and the implications when they do not — is an important and constructive part of applied DID research.

DID Assumptions

A1. **SUTVA.** For all $i \in \{1, \ldots, N\}$,

$$Y_i(A_{s(1)}, A_{s(2)}, \ldots, A_{s(N)}) = Y_i(A_{s(i)}).$$

A2. **No-Anticipation.** For all units i and for all periods $t' < a$,

$$E[Y_{i,s,t'}(a) - Y_{i,s,t'}(0) \mid A_{s(i)} = a] = 0.$$

A3. **Common Trends.** For treatment group $A_s = a$ and control group $A_s > a + \kappa$,

$$E[Y_{i,s,a+e}(0) \mid A_s = a] - E[Y_{i,s,a-1}(0) \mid A_s = a] =$$
$$E[Y_{i,s,a+e}(0) \mid A_s > a + \kappa] - E[Y_{i,s,a-1}(0) \mid A_s = a + \kappa].$$

Assumption A1 (SUTVA) requires that treatment response is individualistic so that unit $i's$ response depends on its own treatment status and not on the treatment assignments of other units. In practice, the SUTVA condition may fail because of spillover effects. For example, in a study of the effects of sugar sweetened beverage taxes on sales of sweetened beverages, it is plausible that taxes levied in one state may affect sales in other states through cross-border shopping effects. A standard DID estimator does not recover a causal effect if the control group includes states that respond to the treatment adopted in the treated state. In this case, one might assume that SUTVA holds for states that do not share a border and exclude the border states from the set of clean controls.

Assumption A2 (No-Anticipation) requires that the average causal effect of adopting treatment in period a is equal to zero in all periods prior to a. In other words, units scheduled to receive treatment at period a do not adjust their outcomes in anticipation of treatment. The no-anticipation condition restriction may fail because of reverse causality, where policy adoption may be a direct response to historical patterns in the outcome, creating the appearance of "anticipation effects." For example, a city that experiences rising crime might install CCTV cameras the following year. In this case, the trend in crime rate prior to installation is not independent of treatment. Sometimes this type of reverse causality is called "policy endogeneity." The no-anticipation assumption may also be violated because of "announcement effects." For instance, if a state government announces in 2025 that it

will expand Medicaid in 2026, hospitals and households may begin adjusting their behavior immediately. Providers may change hiring decisions and patients may postpone care until coverage begins.

Assumption A3 (Common Trends) requires that, in the absence of treatment, the treated group and the control group would have followed the same trend in outcomes in the post-treatment period. The left-hand side of the equation is the counterfactual trend in the treatment group and the right-hand side is the observed trend in the control group. The assumption implies that these two post-period trends are equal. The assumption is not directly testable because the post-period trend in untreated outcomes is an unknown counterfactual in the treatment group.

Importantly, the treated and control groups are allowed to have different levels of the outcome. The assumption only requires that, in the absence of treatment, the trajectory of the treated group would have been parallel to that of the control group. As long as the difference between the treated group and the control group remains constant during the post-treatment period, the starting levels can be different. The common trends do not need to be linear trends. Changes over time that trace out non-linear or "wiggly" paths satisfy common trends as long the treated and control groups move in sync over time in the absence of treatment.

Violations of common trends may occur when there is some unmeasured covariate that affects outcomes and evolves differently across states. For instance, if treated and control states enact other policies — beyond the policy of interest — at different times, these concurrent policy changes may induce diverging trends unrelated to the treatment under study. This concern is especially acute when the legislation of interest is multi-component. For instance, if a reform bundles together several policy changes, it may be difficult to attribute any observed effect to a specific component. Common trends might also fail if time-varying aggregate shocks have heterogeneous effects on individual units. Suppose a nationwide recession is concentrated in the technology sector. It is possible that the same national shock may have a bigger effect on unemployment in states with a large technology sector. Differential trends that arise because of heterogeneous responses to aggregate shocks create problems for DID estimators. These are also a motivation for synthetic control estimators, which is discussed in a future chapter.

2.4. Causal Identification in a Simple DID

Even in the simple two-group DID case, treatment effects need not be homogeneous. There are at least two sources of heterogeneity to keep in mind: event-study and cross-unit. **Event-study heterogeneity** describes how the impact of treatment may vary with the

amount of time elapsed since the intervention. **Cross-unit heterogeneity** explains that even holding event time fixed, treatment effects may differ across units. Under assumptions A1–A3, the DID estimator averages these individual-level effects across all units in the treated group, recovering the ATT.

To see how a simple two-group DID comparison identifies a causal effect, consider a context in which there is only one adoption group that first receives treatment in period $A_s = a$. The goal is to show that, under assumptions A1–A3, a simple pairwise DID comparison recovers $ATT(a, a+e)$, the average treatment effect on the treated group, evaluated e periods after initial exposure. The DID comparison is built around changes in outcomes between a pre-treatment period $a-1$ and a post-treatment period $a+e$.

The control group consists of units with $A_s > a+e$, which have not yet been exposed to treatment by period $a+e$. This is a minimal requirement for a *clean* control group. In practice, researchers often impose additional restrictions, such as limiting to never-treated units ($A_s = \infty$) or baseline covariate matching, to strengthen the credibility of assumptions A1–A3.

The simple DID statistic, denoted $\Delta_{DID(a,e)}$, is the difference in mean outcome changes between the treated and control groups:

$$\Delta_{DID(a,e)} = E[Y_{i,s,a+e} - Y_{i,s,a-1} \mid A_s = a] - E[Y_{i,s,a+e} - Y_{i,s,a-1} \mid A_s > a+e]$$

This quantity is stated in terms of observed outcomes. The derivation below shows that, under A1–A3, $\Delta_{DID(a,e)} = ATT(a, a+e)$.

$$\begin{aligned}
\Delta_{DID(a,e)} &= E\left[Y_{i,s,a+e} - Y_{i,s,a-1} | A_s = a\right] - E\left[Y_{i,s,a+e} - Y_{i,s,a-1} | A_s > a+e\right] \\
&= E\left[Y_{i,s,a+e}(a) - Y_{i,s,a-1}(a) | A_s = a\right] - E\left[Y_{i,s,a+e}(0) - Y_{i,s,a-1}(0) | A_s > a+e\right] \\
&= E\left[Y_{i,s,a+e}(a) - Y_{i,s,a-1}(0) | A_s = a\right] - E\left[Y_{i,s,a+e}(0) - Y_{i,s,a-1}(0) | A_s > a+e\right] \\
&= E\left[\beta_{i,s,a+e}(a) | A_s = a\right] \\
&\quad - \left\{E\left[Y_{i,s,a+e}(0) - Y_{i,s,a-1}(0) | A_s = a\right] - E\left[Y_{i,s,a+e}(0) - Y_{i,s,a-1}(0) | A_s > a+e\right]\right\} \\
&= E\left[\beta_{i,s,a+e}(a) | A_s = a\right] \\
&= ATT(a, a+e)
\end{aligned}$$

Line 1. The first line defines $\Delta_{DID(a,e)}$ in terms of observed outcomes: the pre-to-post change in the treated group minus the pre-to-post change in the control group. This is the quantity that can be computed directly from data, without any assumptions.

Line 2. The second equality imposes SUTVA (A1), which allows observed outcomes to be replaced with individualistic potential outcomes. In the treated group ($A_s = a$), observed outcomes equal $Y_{i,s,t}(a)$, the potential outcome under adoption at time a. In the control group ($A_s > a + e$), observed outcomes equal $Y_{i,s,t}(0)$, the potential outcome under no treatment.

Line 3. The third equality imposes no anticipation (A2). Since the pre-period $a - 1$ occurs before treatment begins at a, units in the treated group have not yet been affected: $Y_{i,s,a-1}(a) = Y_{i,s,a-1}(0)$. The pre-period treated potential outcome can therefore be replaced by the untreated potential outcome.

Line 4. To obtain this separation, substitute $Y_{i,s,a+e}(a) = Y_{i,s,a+e}(0) + \beta_{i,s,a+e}(a)$ into the treated group's post-period outcome. After substituting and regrouping, the expression splits into (i) the average unit-level treatment effect $E[\beta_{i,s,a+e}(a) \mid A_s = a]$, and (ii) the term in braces, which is the difference in counterfactual untreated trends between the treated and control groups.

Line 5. The term in braces is equal to zero under common trends (A3): in the absence of treatment, the treated and control groups would have followed the same trend. Setting it to zero leaves $E[\beta_{i,s,a+e}(a) \mid A_s = a]$.

Line 6. The remaining term is the group-time ATT, as defined above: $ATT(a, a + e) = E[Y_{i,s,a+e}(a) - Y_{i,s,a+e}(0) \mid A_s = a]$. The DID comparison identifies this causal effect under assumptions A1–A3.

Expressing the DID as differences in means makes the identification argument transparent. The same quantity also has a clean regression representation. To see why, note that the two-by-two OLS regression below is a *saturated* model, because it has as many parameters as data points. The model includes a group indicator ($Treat_s$), a period indicator ($Post_t$), and their interaction. These three regressors, together with the intercept, provide four parameters that correspond exactly to the four group-period cells (treated–pre, treated–post, control–pre, control–post). Because the model is saturated, OLS fits the mean outcome in each cell exactly. As a result, the estimated coefficient on the interaction term, β_3, equals the difference in the pre-to-post outcome change for the treated group relative to the control group. This quantity is precisely the DID statistic $\Delta_{DID(a,e)}$. Restricting the sample to units in the adoption group $A_s = a$ and the control group with $A_s > a + e$, and to periods $t = a - 1$ and $t = a + e$, the regression is:

$$
\begin{aligned}
Y_{st} &= \beta_0 + \beta_1 1(A_s = a) + \beta_2 1(t = a + e) + \beta_3 1(A_s = a) \times 1(t = a + e) + u_{st} \\
&= \beta_0 + \beta_1 Treat_s + \beta_2 Post_t + \beta_3 Treat_s \times Post_t + u_{st}
\end{aligned}
$$

Each coefficient has a clear interpretation. β_0 is the mean outcome for the control group in the pre-period. β_1 captures the pre-treatment difference in mean outcomes between the treated and control groups, which is allowed to be non-zero because DID requires only parallel trends, not identical levels. β_2 captures the pre-to-post change in the control group: the common trend. β_3 is the additional pre-to-post change in the treated group relative to the control group: the DID estimate. The four predicted cell means are: β_0 (control, pre), $\beta_0 + \beta_2$ (control, post), $\beta_0 + \beta_1$ (treated, pre), and $\beta_0 + \beta_1 + \beta_2 + \beta_3$ (treated, post). Computing the difference-in-differences as (treated post $-$ treated pre) $-$ (control post $-$ control pre) yields β_3. In the PFL example, β_3 estimates how much more leave-taking increased in California relative to the comparison group after 2004. As shown above, $\beta_3 = \Delta_{DID(a,e)} = ATT(a, a + e)$ under the DID assumptions.

The DID can also be represented as a coefficient from a TWFE regression fitted to the same data:

$$
Y_{i,s,t} = \delta_{fe} D_{s,t} + \alpha_s + \gamma_t + m_{i,s,t}
$$

Because the sample is restricted to the time periods $a - 1$ and $a + e$ for treated units and control units, the time fixed effect (γ_t) absorbs the $Post_t$ variable and the group fixed effect (α_s) absorbs the $Treat_s$ variable. $D_{st} = 1[t \geq A_s]$ is a dummy indicating that treatment is active, which is equivalent to the interaction term. In this set up, the coefficient on treatment exposure is equivalent to the DID: $\delta_{fe} = \beta_3 = \Delta_{DID(a,e)}$.

Figure 1 provides a graphical illustration of the 2×2 DID design and its identifying logic. The two solid lines show the observed mean outcomes for the treatment and control groups across the two periods. The key object is the dashed blue line: it represents what would have happened to the treatment group in the absence of treatment. Under the parallel trends assumption, the treatment group would have followed the same trend as the control group. The dashed line is therefore constructed by taking the treatment group's pre-period level and adding the control group's pre-to-post change. The gap between the observed post-period outcome for the treatment group and this counterfactual gives $\hat{\beta}_{DiD}$—the DID estimate.

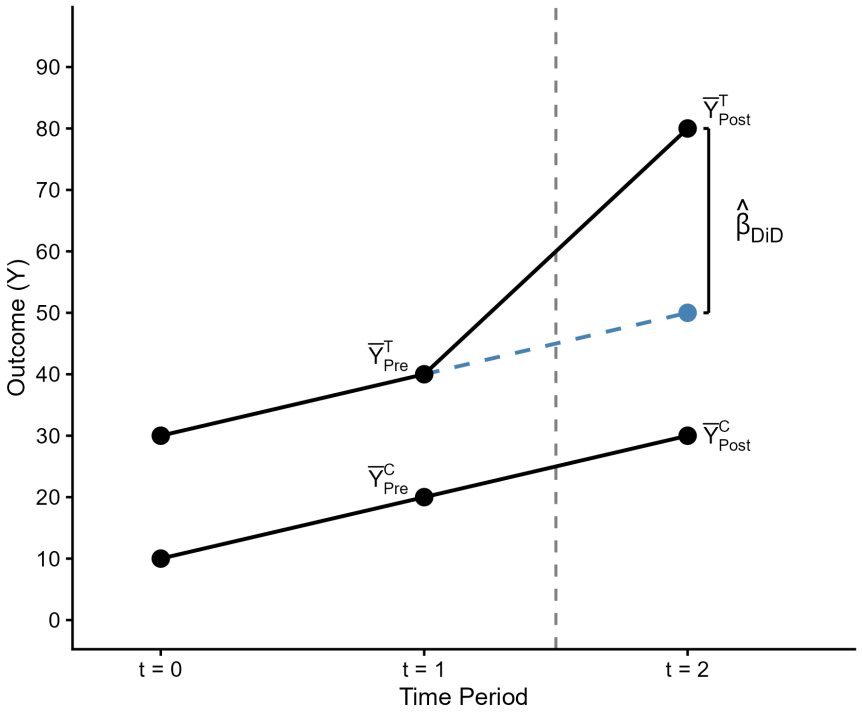


Figure 1 — The 2×2 difference-in-differences design

Note: The solid lines show observed mean outcomes for the treatment and control groups across three periods. In the pre-treatment periods ($t = 0$ and $t = 1$), both groups trend in parallel. The vertical dashed grey line marks the onset of treatment between $t = 1$ and $t = 2$. The dashed blue line is the counterfactual: under parallel trends, the treatment group would have continued on the same trajectory as the control group. The filled blue circle marks the implied counterfactual outcome $\hat{Y}^{T}_{Post}(0)$. The bracket shows $\hat{\beta}_{DiD} = \bar{Y}^{T}_{Post} - \hat{Y}^{T}_{Post}(0)$.

2.5. Event Study Regression

The two-period framework in Section 2.4 is analytically convenient, but in practice researchers typically observe many time periods. When treatment effects evolve over time after adoption, collapsing the data to a single pre-period and a single post period obscures potentially important information. An event study regression addresses this by estimating a separate DID comparison for each event time e. This serves two purposes: it traces the full time path of treatment effects, and it provides a diagnostic for the common trends assumption by examining whether treated and control groups followed parallel trends prior to treatment.

In the TWFE framework, the $ATT(a, a+e)$ for the set of units treated at time a can be estimated for all time periods by fitting the following regression:

$$Y_{s,t} = \sum_{h=-\kappa_{pre};h\neq -1}^{\kappa_{post}} [\theta_h(D_{s,a} \times 1[e=h])] + \alpha_s + \gamma_t + m_{s,t}$$

The summation term is a set of interactions between a dummy for being a treated unit and each time period relative to treatment. The reference year, $e = -1$, is the excluded term from this set of interactions. This setup implies that the θ_h terms trace out each of the simple DID's between period -1 and period $a + e$.

The coefficients for negative event times ($e < 0$) provide a diagnostic for the *common trends* and *no-anticipation* assumptions. If the common trends assumption holds, the pre-treatment coefficients should hover around zero or be statistically indistinguishable from zero. Failing to reject a zero effect in the pre-treatment period does not necessarily prove that the common trends assumption holds. It merely suggests that we do not detect a difference. However, if any of the pre-treatment coefficients are statistically different from zero, this is strong evidence that the treated and control groups followed different paths before treatment, violating the common trends assumption. In addition, statistically significant coefficients in the periods immediately preceding treatment may also indicate violations of the no-anticipation assumption, since treated units appear to respond before the policy is actually implemented.

The coefficients for positive event times ($e > 0$) trace the dynamic treatment effects over time. They describe how the average treatment effect evolves after adoption—whether it rises, fades, or persists. In settings where treatment effects accumulate gradually or adapt to behavioral adjustments, these dynamics are substantively important for interpreting policy impact.

To ground the event-study framework in a concrete setting, the chapter turns to the study by Rossin-Slater, Ruhm, and Waldfogel (2013), "The Effects of California's Paid Family Leave Program on Mothers' Leave-Taking and Subsequent Labor Market Outcomes." The authors use March Current Population Survey (CPS) data from 1999 to 2010 and a difference-in-differences design to estimate how California's first-in-the-nation Paid Family Leave (PFL) program, implemented in 2004, affected mothers' leave-taking following childbirth and subsequent labor market outcomes. Figure 2 reports a replication of the paper's leave-taking results.

Interpreting an event-study plot is straightforward. Event time e is on the x-axis, and the estimated coefficient θ_h is on the y-axis. The pre-treatment coefficients provide a visual check on the common trends assumption, while the post-treatment coefficients describe the trajectory of treatment effects over time. By convention, the coefficient at $e = -1$

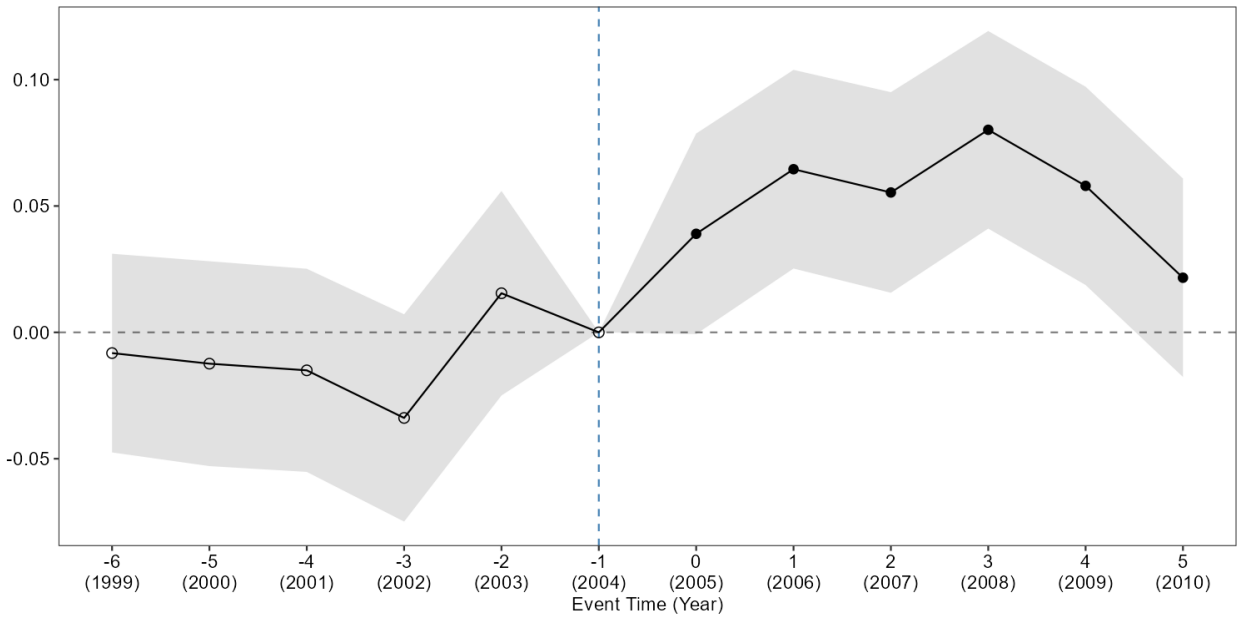


Figure 2 — Event-study estimates of the effect of California's Paid Family Leave program on family leave-taking

Note: The sample is California mothers of infants (treatment) vs. California mothers of children aged 5–17 (comparison), conditional on working a positive number of hours (TOT sample). Open circles: pre-treatment periods; filled circles: post-treatment periods. Dashed vertical line: CA PFL effective (July 2004). Shaded band: 95% CI. Data: IPUMS CPS ASEC, 1999–2010 (Rossin-Slater, Ruhm, and Waldfogel 2013).

(the year before treatment) is normalized to zero, so all other coefficients are interpreted relative to that baseline.

Each point shows the estimated treatment effect for a given year relative to PFL adoption. In other words, it shows how much more likely California mothers of infants were to take family leave compared with California mothers whose youngest child was aged 5–17. Because the coefficient for $e = -1$ is normalized to zero, all estimates are interpreted relative to the year immediately preceding treatment. The pre-treatment estimates (open circles, 1999–2004) hover near zero, providing suggestive evidence consistent with the common trends assumption. Before PFL, mothers of infants and mother of older children followed similar trajectories. After PFL's implementation in 2004, the estimates rise sharply and remain elevated in subsequent years. Mothers of infants became roughly 5–9 percentage points more likely to take family leave relative to the comparison group, suggesting that the program substantially increased leave-taking among new mothers (Rossin-Slater, Ruhm, and Waldfogel 2013).

As discussed above, our notation emphasizes identifying separate DID comparisons for each time period (both before and after treatment) relative to a fixed reference year, $a - 1$. Sometimes researchers would like to generate a one-number summary of the average treat-

ment effect across the whole post period. In this framework, one constructs this estimate by simply averaging the θ_h coefficient estimates from the post-period years, $0, .., \kappa_{post}$. Because this is just a linear combination of coefficients (the sum of the post-period coefficients divided by the number of post-period years), standard errors are straightforward to estimate by using built in commands such as `lincom` in STATA or the `hypothesis()` function in R's `marginaleffects` package.

2.6. Bad DID Comparisons

The TWFE specification is appealing because it encodes the common trends assumption clearly. The a_s allow for time-invariant differences across groups. The b_t allow for time varying effects that affect all units in the same way. The specification also appears to scale well to accommodate multiple groups and time periods. Indeed, the TWFE regression has been widely used to analyze data from staggered adoption designs in which treatment adoption occurs at different times for different units. Specifically, one can fit the same regression of outcomes on group fixed effects, time fixed effects, and a treatment dummy to datasets that include multiple groups and multiple time periods with staggered treatment adoption.

The problem arises when treatment effects are not constant across adoption cohorts or over event time. In that case, it is worth asking first: what is the parameter of interest? In principle, a researcher might want to know the ATT for a specific cohort at a specific event time, $ATT(a, a+e)$, or an average of these across adoption cohorts. The conventional TWFE estimator does not deliver any of these coherent targets.

There are two distinct problems with applying the conventional TWFE estimator to staggered adoption data. The first is **event-study heterogeneity**, in which treatment effects may vary over event time within a cohort, but the TWFE estimator implicitly averages across the post period and imposes constant effects. The second is **cross-unit heterogeneity**, in which different adoption cohorts may experience different average treatment effects, yet TWFE pools them all together. Goodman-Bacon (2018) and De Chaisemartin and d'Haultfoeuille (2020) showed that the TWFE coefficient in the staggered adoption setting can be expressed as a weighted average of many simple 2×2 DID comparisons:

$$\delta_{fe}^{staggered} = \sum_{(g,h,t)} \omega_{g,h,t} \, \Delta_{DID}^{(g,h,t)},$$

where $\Delta_{DID}^{(g,h,t)}$ denotes the DID comparison between groups g and h at time t, and the weights $\omega_{g,h,t}$ depend on group sizes and treatment timing. Larger groups receive more weight, reflecting efficiency, or comparisons based on more observations are estimated

more precisely. The deeper problem is that some of these comparisons lack a causal interpretation when treatment effects vary over time.

To see why, consider the simple case with three groups: an early adopter k, a late adopter ℓ, and a never-treated group U. The TWFE estimator draws on four distinct 2×2 DID comparisons. Two are *clean*: early adopters versus never-treated (k vs. U) and late adopters versus never-treated (ℓ vs. U). A third compares early adopters to late adopters during the period after k adopts but before ℓ adopts using the not-yet-treated group as a valid control. The fourth comparison is the problematic one: it compares late adopters to early adopters during periods when k is *already* treated. Here the early adopters serve as the control group, but their outcomes already reflect the treatment they received. If treatment effects change over event time, this comparison subtracts the evolving effects of k's treatment and no longer identifies a coherent causal parameter.

The mechanism behind this problem becomes clearer if we focus on a single late–early comparison. Consider a two-group, two-period DID comparison built around a late treatment group that adopts treatment in period $A_s = a_2$, and an early treatment group with $A_s = a_1$ with $a_1 < a_2 - 1$. The simple DID comparison is $\Delta^{a_1,a_2}_{DID_{(a,e)}} = E\left[Y_{i,s,a+e} - Y_{i,s,a-1} | A_s = a_2\right] - E\left[Y_{i,s,a+e} - Y_{i,s,a-1} | A_s = a_1\right]$. To understand what this particular DID comparison reveals, we proceed using the standard technique of replacing observed outcomes with potential outcomes and imposing the identifying assumptions.

$$\Delta^{a_1,a_2}_{DID_{(a,e)}} = E\left[Y_{i,s,a_2+e} - Y_{i,s,a_2-1} \mid A_s = a_2\right] - E\left[Y_{i,s,a_2+e} - Y_{i,s,a_2-1} \mid A_s = a_1\right]$$

Observed outcome DID: late adopters minus early adopters, differenced over time.

$$= E\left[Y_{i,s,a_2+e}(a_2) - Y_{i,s,a_2-1}(a_2) \mid A_s = a_2\right] - E\left[Y_{i,s,a_2+e}(a_1) - Y_{i,s,a_2-1}(a_1) \mid A_s = a_1\right]$$

Potential outcomes *Replace observed outcomes with potential outcomes. Both arms use treated potential outcomes.*

$$= E\left[Y_{i,s,a_2+e}(a_2) - Y_{i,s,a_2-1}(0) \mid A_s = a_2\right] - E\left[Y_{i,s,a_2+e}(a_1) - Y_{i,s,a_2-1}(a_1) \mid A_s = a_1\right]$$

No anticipation *Pre-period outcome for late adopters set to $Y(0)$; early adopter group unchanged.*

$$\begin{aligned} &= E\left[\beta_{i,s,a_2+e}(a_2) \mid A_s = a_2\right] - E\left[\beta_{i,s,a_2+e}(a_1) - \beta_{i,s,a_2-1}(a_1) \mid A_s = a_1\right] \\ &\quad + \left\{E\left[Y_{i,s,a_2+e}(0) - Y_{i,s,a_2-1}(0) \mid A_s = a_2\right] - E\left[Y_{i,s,a_2+e}(0) - Y_{i,s,a_2-1}(0) \mid A_s = a_1\right]\right\} \end{aligned}$$

Decompose treated outcomes as $\beta + Y(0)$. Three components: (i) ATT for late adopters, (ii) change in early adopter treatment effect, (iii) untreated trend difference. Term (iii) vanishes under parallel trends.

$$= ATT(a_2,\, a_2+e) - \{ATT(a_1,\, a_2+e) - ATT(a_1,\, a_2-1)\}$$

Parallel trends *Untreated trend term cancels. Estimand equals late adopters' ATT minus change in early adopters' ATT. Bias exists unless early adopter effects are constant.*

$$\begin{aligned}
\Delta^{a_1,a_2}_{DID_{(a,e)}} &= E\left[Y_{i,s,a_2+e} - Y_{i,s,a_2-1} | A_s = a_2\right] - E\left[Y_{i,s,a_2+e} - Y_{i,s,a_2-1} | A_s = a_1\right] \\
&= E\left[Y_{i,s,a_2+e}(a_2) - Y_{i,s,a_2-1}(a_2) | A_s = a_2\right] - E\left[Y_{i,s,a_2+e}(a_1) - Y_{i,s,a_2-1}(a_1) | A_s = a_1\right] \\
&= E\left[Y_{i,s,a_2+e}(a_2) - Y_{i,s,a_2-1}(0) | A_s = a_2\right] - E\left[Y_{i,s,a_2+e}(a_1) - Y_{i,s,a_2-1}(a_1) | A_s = a_1\right] \\
&= E\left[\beta_{i,s,a_2+e}(a_2) | A_s = a_2\right] - E\left[\beta_{i,s,a_2+e}(a_1) - \beta_{i,s,a_2-1}(a_1) | A_s = a_1\right] \\
&+ \{E\left[Y_{i,s,a_2+e}(0) - Y_{i,s,a_2-1}(0) | A_s = a_2\right] - E\left[Y_{i,s,a_2+e}(0) - Y_{i,s,a_2-1}(0) | A_s = a_1\right]\} \\
&= ATT(a_2, a_2+e) - \{ATT(a_1, a_2+e) - ATT(a_1, a_2-1)\}
\end{aligned}$$

The logic of the derivation is very similar to the conventional 2 x 2 DID case considered above. The second equality replaces observed outcomes with potential outcomes. But notice that in the late-early comparison case, we replace observed outcomes with "treated outcomes" in both the treated arm and the control arm of the estimator. In the late adopter treated arm ($A_s = a_2$), observed outcomes are replaced with $Y_{i,s,t}(a_2)$. In the early adopter control arm ($A_s = a_1$), observed outcomes are replaced with $Y_{i,s,t}(a_1)$. In the third equality, we impose the no anticipation assumption to set the pre-period outcome to $Y_{i,s,t}(0)$ for the late adopter group. There is no change to the early adopter control group because both periods $a_2 - 1$ and $a_2 + e$ occur after period a_1.

The fourth equality decomposes the treated outcomes and re-arranges. For the late adopter group, we re-write $Y_{i,s,a_2+e}(a_2) = \beta_{i,s,a_2+e}(a_2) + Y_{i,s,a_2+e}(0)$. For the early adopter group, we re-write the outcomes for both time periods. In period a_2+e, we have $Y_{i,s,a_2+e}(a_1) = \beta_{i,s,a_2+e}(a_1) + Y_{i,s,a_2+e}(0)$. In period $a_2 - 1$, we have $Y_{i,s,a_2-1}(a_1) = \beta_{i,s,a_2-1}(a_1) + Y_{i,s,a_2-1}(0)$. Re-grouping the terms shows that the DID comparison equals the sum of three components: (i) the average treatment effect on the late adopters, (ii) the change in the treatment effect for the early adopters over time, and (iii) the difference in untreated outcome trends between the early and late adopters. The third term cancels out under the common trends

assumption.

The final line shows that the late-early DID comparison identifies a composite of two terms. The first term is the group-time ATT for the late adopters, evaluated in period $a_2 + e$. The second term is the change in the group-time ATT for the early adopters across periods $a_2 - 1$ and $a_2 + e$. This shows that the late–early DID comparison identifies the late adopters' ATT only if the treatment effects for the early adopters remain constant over the comparison window. This requirement contrasts with the event-study framework developed above, which is designed precisely to estimate how treatment effects evolve over event time. The late-early comparison imposes the opposite requirement—that the early adopter's treatment effects remain constant—because any evolution in those effects is subtracted from the late adopter's DID estimate, introducing bias.

Table 1 — Observed outcomes in the numerical example. States A, B, and C adopt treatment in periods 2, 3, and never, respectively. Panel (a) assumes constant treatment effects; Panel (b) allows A's effect to grow over event time.

	Period		
State	$t = 1$	$t = 2$	$t = 3$
Panel (a): Constant effects, $ATT(A) = 2$, $ATT(B) = 4$			
A (early adopter)	1	3	3
B (late adopter)	1	1	5
C (never adopter)	1	1	1
Panel (b): Growing effects, $ATT(A,2) = 2$, $ATT(A,3) = 6$, $ATT(B) = 4$			
A (early adopter)	1	3	7
B (late adopter)	1	1	5
C (never adopter)	1	1	1

The intuition behind this result becomes clearer in a simple numerical example. Consider three states over three periods ($t = 1, 2, 3$): state A adopts in period 2, B adopts in period 3, C never adopts, and $Y_{s,t}(0) = 1$ throughout. With $ATT(A) = 2$ in periods 2 and 3 (constant) and $ATT(B) = 4$ in period 3, observed outcomes are $Y_A = (1, 3, 3)$, $Y_B = (1, 1, 5)$, $Y_C = (1, 1, 1)$ (Panel (a) of Table 1). The four DID comparisons yield: (i) A vs. C, periods 1–2: $\Delta = 2$; (ii) B vs. C, periods 1–3: $\Delta = 4$; (iii) A vs. B, periods 1–2: $\Delta = 2$; (iv) B vs. A, periods 2–3: $\Delta = (5 - 1) - (3 - 3) = 4$. All four recover the true ATT and the TWFE coefficient is a clean weighted average.

Now suppose A's effect grows: $ATT(A, 2) = 2$ and $ATT(A, 3) = 6$, giving $Y_{A,3} = 7$ (Panel (b)). Comparisons (i)–(iii) are unchanged. But comparison (iv) now yields $(5 - 1) -$

$(7-3)=0$—severely biased. The derivation above explains exactly why: $ATT(a_2, a_2+e) - [ATT(a_1, a_2+e) - ATT(a_1, a_2-1)] = 4-(6-2)=0$. The growth in A's treatment effect over event time ($6-2=4$) is subtracted from B's true ATT of 4, reducing the estimate to zero.

The event-study estimates above come from a single-state policy setting: California is the sole policy-adopting state, and the leave-taking analysis compares mothers of infants with an internal comparison group before and after the program took effect in 2004. California, however, was only the first adopter. Since 2004, additional states have implemented PFL programs over a series of years, creating a multi-cohort, staggered-adoption setting in which a conventional TWFE regression can combine clean DID comparisons with late-versus-early comparisons that become contaminated when treatment effects evolve over time. Table 2 lists the states with mandatory PFL programs and the year each took effect.

Section 3 extends the chapter's running PFL example from this California-only design to the multi-state setting and presents two estimators designed to address this problem: the stacked DID estimator, which restricts attention to valid comparisons, and the fixed effect imputation estimator, which constructs counterfactual outcomes using untreated observations.

3 — Handling Staggered Adoption Designs

The recent literature contains a growing set of estimators for staggered adoption DID designs. Broadly speaking, the new estimators share two common goals: (i) to ensure that "bad (late vs early adopters) DID comparisons" are not included in the analysis of the staggered design, and (ii) to construct an interpretable aggregation of the underlying heterogeneous causal effects. The target parameter in this setting is a weighted average of the cohort-specific $ATT(a, a+e)$ introduced in Section 2—averaged across adoption cohorts and, often, across post-treatment event times. The new estimators are designed to recover consistent versions of this parameter; the conventional TWFE estimator, as shown in Section 2, may not. In this section, we examine two specific techniques for estimating causal effects from a staggered adoption design. The first method is the weighted stacked DID estimator developed in Wing et al. (2023). The second method is the fixed effect imputation estimator developed in Borusyak, Jaravel, and Spiess (2024), hereafter BJS.

Several other estimators address the same problem. Callaway and P. H. Sant'Anna (2020) aggregate cohort-specific ATTs using a not-yet-treated or never-treated comparison group. Sun and Abraham (2020) propose interaction-weighted estimators based on cohort-by-event-time interactions, free of the contamination that affects standard TWFE event-study

Table 2 — States with mandatory Paid Family Leave programs, by effective year

Subexperiment	Starting Year	Ending Year	Treated States	Control States
2004	1999	2009	CA	Never Adopters + RI, NY, DC, WA, MA, CT, OR, CO, DE, ME, MN, MD, VA
2009	2004	2014	NJ	Never Adopters + NY, DC, WA, MA, CT, OR, CO, DE, ME, MN, MD, VA
2014	2009	2019	RI	Never Adopter + CA, DC, WA, MA, CT, OR, CO, DE, ME, MN, MD, VA
2018	2013	2023	NY	Never Adopter + CA, NJ, CO, DE, ME, MN, MD, VA
2020	2015	2025	DC, WA	Never Adopter + CA, NJ, RI, DE, ME, MN, MD, VA
2021	2016	2025	MA	Never Adopter + CA, NJ, RI, DE, ME, MN, MD, VA
2022	2017	2025	CT	Never Adopter + CA, NJ, RI, DE, ME, MN, MD, VA
2023	2018	2025	OR	Never Adopter + CA, NJ, RI, DE, ME, MN, MD, VA
2024	2019	2025	CO	Never Adopter + CA, NJ, RI, NY, DE, ME, MN, MD, VA
Passed but not yet implemented				DE, ME, MN, MD, VA
Never Adopters: AL, AK, AZ, AR, FL, GA, HI, ID, IL, IN, IA, KS, KY, LA, MI, MS, MO, MT, NE, NV, NH, NM, NC, ND, OH, OK, PA, SC, SD, TN, TX, UT, VT, WV, WI, WY				

coefficients. Gardner (2022) proposes a two-stage estimator that removes group and time fixed effects—estimated from untreated observations—and then regresses the residualized outcomes on treatment status.

3.1. Stacked DID

As discussed in Section 2, the standard TWFE model pools many 2×2 comparisons when treatments are adopted at different times. When treatment effects differ across cohorts or change over event time, this pooling creates problems: the TWFE coefficient includes "bad" comparisons between newly treated units and those already treated earlier. Because already-treated units are following trajectories shaped by their own earlier treatment, using them as controls contaminates the comparison whenever treatment effects evolve over time.

The Stacked DID method, formalized by Wing et al. (2023), offers a transparent and intuitive way to address the forbidden comparison problem in staggered adoption settings. The method is conceptually simple, easy to implement without special packages, and appealing because it allows researchers to control exactly which comparisons are being made. Throughout this section, we illustrate the method using the staggered adoption of paid family leave (PFL) programs across U.S. states, extending the California example from Section 2.

The core idea of Stacked DID is to build a separate dataset for each "sub-experiment," which is defined around each treatment cohort—in other words, any unit treated at the same point in time. Each sub-experiment dataset includes observations on two groups:

- **Treated group:** units first treated in a specific adoption year a;
- **Control group:** units that remain untreated during the entire event window around that adoption.

The key point is that every sub-experiment satisfies the same structure—a pre-period, a treatment adoption period, and a post-period—with the same event-time window applied to each. Once these sub-experiments are defined, they are simply "stacked" into one large dataset, on which a simple pooled regression can be estimated to obtain the DID estimate. In the rest of this section, we provide step-by-step guidance on carrying out a stacked-DID regression.

3.1.1 Constructing the Stack

The heart of the stacked-DID methodology is in constructing the stacked data. We outline the steps here, along with some important researcher decisions:

Constructing the Stack

1. Start with a long-form panel (e.g., state $\times$ year) that contains outcomes and treatment adoption dates.
2. Choose an event window, typically defined by κ_{pre} pre-treatment years and κ_{post} post-treatment years (e.g., three years before and three years after adoption).
3. Trim the data so that only adoption events with full pre- and post-periods are included. This ensures compositional balance—each sub-experiment has the same time window.
4. Define clean controls for each adoption year, or units that remain untreated through the entire event window. They may be "never treated," "not yet treated," or both.
5. Create sub-experiment datasets. For each valid adoption year a, keep only the treated and clean control units observed in the selected event window.
6. Relabel time in each sub-experiment into "event time," where $e = 0$ represents the year of adoption, $e < 0$ pre-treatment years, and $e > 0$ post-treatment years.
7. Stack all sub-experiments vertically into one dataset. Each observation is now indexed by unit $\times$ sub-experiment $\times$ event time.

To illustrate what the stacked dataset looks like, consider two treatment cohorts: early adopters who first adopt in year a_1 and late adopters who first adopt in year $a_2 > a_1$, with $\kappa_{pre} = \kappa_{post} = 2$. Sub-experiment 1 is built around the early adopters. It includes the early-adopting units (treated) and any units that remain untreated throughout the window $[a_1 - 2,\ a_1 + 2]$ (clean controls), with time relabeled to event times $e \in \{-2, -1, 0, 1, 2\}$. The late adopters qualify as clean controls in this sub-experiment because they have not yet adopted by $a_1 + 2$.

Sub-experiment 2 is built around the late adopters. It includes the late-adopting units (treated) and units that remain untreated throughout $[a_2 - 2,\ a_2 + 2]$ (clean controls), again relabeled to $e \in \{-2, -1, 0, 1, 2\}$. The early adopters are excluded from this control group because they are already treated by the time this event window begins.

Stacking these two sub-experiments vertically produces a dataset in which each unit–sub-experiment pair is indexed by the same event-time window and no unit is ever used as a control after it has received treatment. This process can be implemented with simple code—by repeatedly filtering, saving, and concatenating data files.

An important researcher decision is choosing the event window, κ_{pre} and κ_{post}. A wider window captures more of the treatment effect trajectory and allows for a clearer assessment of pre-treatment trends, but it restricts which cohorts can be included. For instance, adoption events that lack a full κ_{pre} pre-treatment periods or κ_{post} post-treatment periods must be trimmed in Step 3. A narrower window retains more cohorts but may miss important dynamics or provide limited evidence on pre-trends. In practice, the window should be wide enough to assess parallel trends and capture the bulk of the treatment effect, while remaining feasible given the available data.

In the PFL application, the data cover mothers of infants across all U.S. states from 1999 to 2025 (IPUMS CPS ASEC). We set $\kappa_{pre} = \kappa_{post} = 5$, yielding a symmetric eleven-year event window around each adoption. A symmetric window helps ensure that the same set of cohorts contributes at each event-time bin, making pre- and post-period estimates directly comparable. A cohort is feasible only if its full event window $[a - \kappa_{pre},\, a + \kappa_{post}]$ falls within the data range. With $\kappa_{pre} = \kappa_{post} = 5$ and data spanning 1999–2025, this yields six feasible adoption cohorts: California (2004), New Jersey (2009), Rhode Island (2014), New York (2018), and the District of Columbia and Washington (both 2020). States that adopted later—Massachusetts (2021), Connecticut and Virginia (2022)—are excluded because their post-treatment window extends beyond the data. This restriction ensures that each sub-experiment has a complete and balanced event-time window, avoiding composition changes across event-time bins. DC and Washington share a single sub-experiment since they adopted in the same year. States that never adopted a PFL program within the data window serve as clean controls in every sub-experiment. Earlier adopters are excluded from the control pool in later sub-experiments, so no state ever appears as a control after adopting PFL itself. Table 3 summarizes the six feasible cohorts; Figure 3 displays their adoption years on a timeline.

Table 3 — PFL adoption cohorts included in the stacked DID

Adoption Year	States
2004	CA
2009	NJ
2014	RI
2018	NY
2020	DC, WA

3.1.2 Estimating effects

Once the data are stacked, the researcher fits a regression of the form:

$$Y_{s,a,e} = \alpha_0 + \alpha_1 D_{s,a} + \sum_{h=-\kappa_{pre},\, h\neq -1}^{\kappa_{post}} \left[\lambda_h \, 1[e = h] + \delta_h \left(D_{sa} \times 1[e = h]\right)\right] + v_{s,a,e},$$

where Y_{sae} is the outcome for unit s in sub-experiment a at event time e, $D_{s,a}$ indicates whether the unit is treated in that sub-experiment, and $e = -1$ is the reference (baseline) period just before treatment.

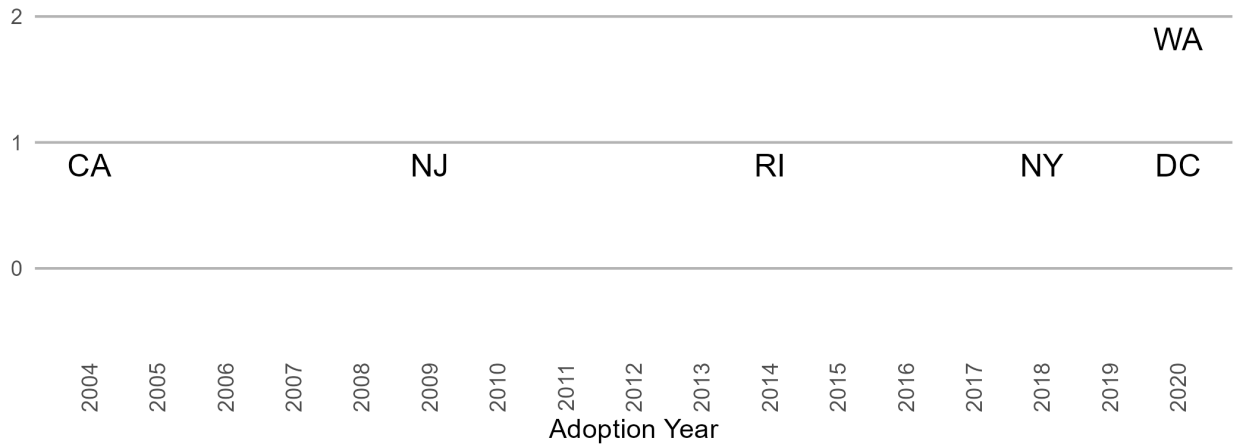


Figure 3 — Timeline of state PFL adoption cohorts

The goal of this regression is for the coefficients δ_h to trace out the average treatment effects by event time—the equivalent of an event study for the staggered adoption setting. In other words, the idea is for these coefficients to represent the simple DID estimate from period $a-1$ to period h, aggregated in some way across treatment cohorts. However, Wing et al. 2023 show that an OLS estimation of these coefficients does not quite lead to a true average treatment effect, but a simple weighted least squares regression does. They show that the following weights should be applied to each observation from unit s and sub-experiment a based on if the unit is in the treatment or control group for that particular sub-experiment:

$$Q_{sa} = \begin{cases} 1, & \text{if treated in sub-experiment } a, \\ \dfrac{N_a^D / N_{\Omega_\kappa}^D}{N_a^C / N_{\Omega_\kappa}^C}, & \text{if control in sub-experiment } a, \end{cases}$$

where N_a^D and N_a^C are the number of treated and control units in sub-experiment a, and $N_{\Omega_\kappa}^D$ and $N_{\Omega_\kappa}^C$ are total number of treated and control units across all sub-experiments.

When these weights are applied, Wing et al. 2023 show that the coefficients from the above event study are equivalent to:

$$\sum_{a \in \Omega_\kappa} ATT(a, a+e) \times \frac{N_a^D}{N_{\Omega_\kappa}^D}.$$

In other words, each coefficient represents the weighted average across sub-experiments

of simple DIDs from period $a-1$ to period $a+e$, where each sub-experiment is weighted based on the share of treatment units in that sub-experiment. So for example, if there are three treatment cohorts, that each have one treated unit, the weighted least squares coefficients represent an average effect in which each sub-experiment has received a weight of $1/3$. Alternatively, if the first treatment cohort has 2 treated units, the second has 1, and the third has 1, the weighted least squares regression coefficients represent an average effect in which the first sub-experiment has received a weight of $2/4 = 1/2$ while the second and third sub-experiments have received a weight of $1/4$.

One important practical consideration concerns statistical inference. Because never-treated and not-yet-treated units may appear as clean controls in multiple sub-experiments, the same unit can contribute observations to more than one sub-experiment in the stacked dataset. As a result, the sub-experiments are not independent, because they may share observations from the same units in their pre- or post-treatment periods. Standard errors should therefore be clustered at the unit level (e.g., the state level) rather than at the unit–sub-experiment level, which would treat repeated appearances of the same unit as independent and understate sampling uncertainty. Wing et al. (2023) show through simulation that unit-level clustering performs well in this setting, provided the number of clusters is not too small.

3.1.3 Aggregation

One advantage of the stacked DID approach is that it takes seriously the heterogeneity of treatment effects across units that adopt the treatment at different times, but provides a relatively simple regression-based approach to obtain an estimate that aggregates these group-specific treatment effects in a coherent causal framework. Again, as discussed above in the non-staggered adoption case, we have continued to allow for heterogeneous treatment effects over event time. If a researcher wants to calculate an average treatment effect across the full post period, the δ_h terms from post period years can be averaged just like in the non-staggered case using the sum of coefficients divided by the number of post period years.

Wing et al. 2023 also provide a variety of corrective weights for different contexts that allow different types of aggregations across groups, including population weighted estimates. Population weighting is often appropriate in policy settings when units differ greatly in size. Without such weights, each state contributes equally to the estimate, so the resulting average reflects the effect for the typical state rather than the typical individual. Weighting by population instead produces an estimate that reflects the average effect

experienced by the typical person. They also provide guidance on applying the weighted stacked estimator to repeated cross section data and to applying propensity score weights in the case that a researcher would like to generate a matched treatment and control group based on pre-treatment observed characteristics.

In the PFL application, averaging the post-period coefficients δ_0 through δ_5 using this formula yields an average post-period ATT of 0.033 (SE$= 0.018$), as reported in Figure 4.

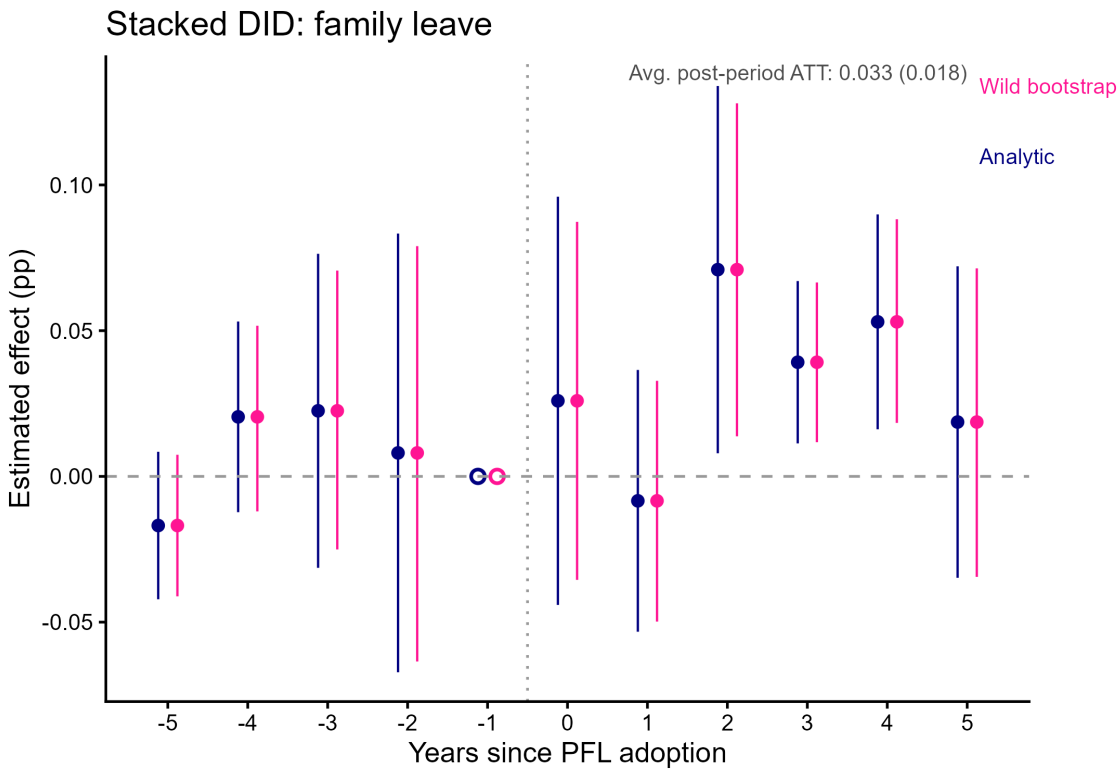


Figure 4 — Stacked DID event-study estimates of the effect of state Paid Family Leave programs on family leave-taking

Note: The sample is mothers of infants (youngest child under age 1) in PFL-adopting states (treatment) vs. mothers of infants in states that never adopted PFL within the data window (control), conditional on working a positive number of hours (TOT sample). The stacked dataset includes six adoption cohorts: California (2004), New Jersey (2009), Rhode Island (2014), New York (2018), and the District of Columbia and Washington (both 2020), each with a symmetric event window of $\kappa_{pre} = \kappa_{post} = 5$ years. Open circle: reference period ($e = -1$). Coefficients represent changes in the probability of family leave-taking; the y-axis is in proportion units (e.g., 0.033 = 3.3 percentage points). Avg. post-period ATT$= 0.033$ (SE$= 0.018$). Standard errors clustered at the state level; 95% confidence intervals shown. Data: IPUMS CPS ASEC, 1999–2025 (Rossin-Slater, Ruhm, and Waldfogel 2013; Wing et al. 2023).

Figure 4 displays the stacked DID event-study estimates for family leave-taking across the six PFL adoption cohorts. The pre-treatment coefficients ($e = -5$ through $e = -2$) are close to zero, supporting the parallel trends assumption: in the years before each state's PFL adoption, mothers of infants in adopting states were on similar trajectories to those

in never-adopting states. After adoption ($e \geq 0$), the point estimates are generally positive, with a dip at $e = 1$.[4] The average post-period ATT is 0.033, or approximately 3.3 percentage points—suggesting that PFL programs increased the probability of family leave-taking among eligible mothers. However, the estimate is imprecise. The 95% confidence interval spans approximately $[-0.2,\ 6.8]$ percentage points, so the effect is not statistically distinguishable from zero. This reflects both the modest number of treated states and the low baseline rate of formal leave-taking in any given reference week.

3.1.4 Practical takeaways

Conceptually, Stacked DID is appealing because it mimics what researchers already do informally: run a clean DID for each adoption group and then average the results. The stacking step simply automates that process into one regression, while preserving transparency and causal validity.

3.2. Fixed Effect Imputation

Borusyak, Jaravel, and Spiess (2024) provides a different approach to DID estimation with staggered treatment timing. Rather than constructing many different sub-experiments with "clean controls" to then be aggregated, this method estimates a model for the untreated potential outcome, $Y_{i,s,t}(0)$, using only untreated observations. This model is then extrapolated to impute $Y_{i,s,t}(0)$ for treated units. This model based imputation is then used as the counterfactual to create the DID for treated units.

This section outlines how to apply the imputation method to the event study regression using our common notation and then discusses aggregation alternatives below. The Borusyak, Jaravel, and Spiess (2024) approach begins with the TWFE model, which we write here for the event study with staggered treatment adoption:

$$Y_{s,t} = \sum_{h=-\kappa_{pre};h\neq-1}^{\kappa_{post}} [\theta_h(D_{s,a} \times 1[e = h])] + \alpha_s + \gamma_t + m_{s,t} \tag{1}$$

Borusyak, Jaravel, and Spiess (2024) set out to retain this familiar TWFE structure while relaxing the restrictive assumption of constant treatment effects across groups applied by the above regression model. They begin by modeling the untreated potential outcome as

4. New Jersey's PFL took effect in July 2009, after the March CPS reference week. The 2009 survey observation for NJ is therefore effectively pre-treatment, which may attenuate the $e = 1$ estimate.

$$Y_{s,t}(0) = \alpha_s + \gamma_t + \epsilon_{s,t} \tag{2}$$

Here notice that unlike the stacked DID, we are maintaining t as calendar time at this point in the modeling.

The key insight of the Borusyak, Jaravel, and Spiess (2024) approach is that under the three key assumptions of the DID identification strategy (SUTVA, common trends, and no-anticipation), it is possible to use this potential outcomes model to construct a counterfactual that is not "contaminated" by previously adopting units. The idea is to restrict the sample to only untreated observations. This will include all observations on units that are never treated during the sample period and observations from treated units, but only in periods prior to treatment. This untreated subsample is used to estimate predicted outcomes for state s at time t if it were to have remained untreated. With this in hand, observed outcomes can be compared to these counterfactual outcomes to construct a DID estimate.

3.2.1 Implementing the Fixed Effect Imputation Method

While packages[5] are available to implement this method, outlining the steps is instructive for understanding intuition and for potentially customizing the method for different contexts.

5. For implementation, see Stata package `did_imputation` in https://github.com/borusyak/did_imputation or R package `didimputation` in https://github.com/kylebutts/didimputation.

Fixed Effect Imputation Method

1. **Estimate Untreated Outcomes**
   Restrict the sample to untreated observations ($D_{s,t} = 0$), which again includes all observations for never-treated units and pre-treatment observations for treated units. In this subsample, fit the model for untreated outcomes above. In other words, regress outcomes for untreated observations on unit fixed effects (α_s) and time fixed effects (γ_t). Because the sample is only untreated outcomes, these unit and time fixed effects are not influenced at all by the treatment effects. Using these estimates, we can predict untreated outcomes as:

$$\hat{Y}_{s,t}(0) = \hat{\alpha}_s + \hat{\gamma}_t \tag{3}$$

   Notice that this step uses only untreated observations ($D_{st} = 0$), but it produces fitted values for both untreated units (in-sample) and treated units (out-of-sample). For treated units, these fitted values serve as the basis for constructing their "counterfactual" outcomes.
2. **Extrapolate**
   For each treated observation, compute the individual-level estimated treatment effect as the difference between the observed and imputed untreated outcomes:

$$\hat{\beta}_{s,t} = Y_{s,t}^{obs} - \hat{Y}_{s,t}(0) \quad \text{for } D_{s,t} = 1 \text{ units} \tag{4}$$

   Because $\hat{Y}_{s,t}(0)$ represents the counterfactual outcome had unit s not been treated, the difference between the observed outcome and this quantity represents the treatment effect for unit s at time t. Note, this process explicitly relies on the common trends assumption: in the absence of treatment, treated units are assumed to have followed the same trend as the untreated units, which is represented by how $\hat{Y}_{s,t}(0)$ changes over time. And because $\hat{Y}_{s,t}(0)$ also incorporates an estimate of the state fixed effect, α_s, it also adjusts for level differences between treated and control units, analogously to a standard DID model.
3. **Aggregation**
   At this point, $\hat{\beta}_{s,t}$ represents separate treatment effects for each unit s in time period t. One advantage of this approach is that the researcher can then aggregate these estimates in many ways depending on the research question. Given our emphasis on event studies in this paper, one natural aggregation is to average across treated units lined up in event time. In other words, for each period in event time, e, construct a weighted average across adoption units at that event time, where for each adoption unit, event time is calendar time minus adoption time ($t - a$). Using Ψ to indicate the set of units that are treated, construct:

$$\hat{\beta}_e^w = \sum_{s \in \Psi} w_s \hat{\beta}_{s,t-a} \tag{5}$$

   The w_s term can be any set of weights the researcher would like to use. Treated units could simply be equally weighted, or the researcher could weight based on the population within each unit, s, etc.
   As in the Stacked DID context, this method can also be applied to repeated cross section data where the data is at the individual level rather than the group level. In this case, the imputation estimates a treatment effect for each individual in the data set. The aggregation step is then the same, but aggregates individuals rather than groups. This also gives one the ability to conduct subgroup analysis by applying weights that are set to 0 for some individuals and not others.

Borusyak, Jaravel, and Spiess 2024 develop a method for statistical inference in their imputation model grounded in asymptotic theory. The main challenge in Borusyak, Jaravel, and Spiess 2024 statistical inference arises from the nature of their imputation procedure. Their imputation method will fit all treated observations ($D_{st} = 1$) perfectly. That is, for these treated units, the regression residuals are identically zero. This would incorrectly suggest that treated observations contain no idiosyncratic noise at all. Consequently, it is impossible to disentangle how much of the "observed difference" between Y_{st}^{obs} and $\hat{Y}_{st}(0)$ reflects the true treatment effect versus random error.

To address this issue, Borusyak, Jaravel, and Spiess 2024 introduce a conservative inference procedure based on "pseudo-residuals" which is

$$\tilde{\epsilon}_{st} = Y_{st}^{obs} - \hat{a}_s - \hat{b}_t - \hat{\beta}_{st}^{\text{avg}}, \tag{6}$$

where $\hat{\beta}_{st}^{\text{avg}}$ is "some average" of estimated treatment effects over many units. This avoids setting residuals to zero while preserving meaningful variation. The resulting variance

$$\sigma_w^2 = E\left[\sum_s \left(\sum_t v_{st}\tilde{\epsilon}_{st}\right)^2\right] \tag{7}$$

is asymptotically *conservative* – that is, it weakly overstates the true variance in large samples.

Conceptually, this approach parallels heteroskedasticity-robust inference: both use noisy proxies for the true error variance but yield valid coverage once aggregated. Although conservative inference is less efficient, it provides reliable standard errors for average treatment effects in staggered adoption settings with heterogeneous responses. Borusyak, Jaravel, and Spiess 2024 outline a cluster robust version of their standard error estimate as well, and these can be implemented through their posted packages.

Figure 5 applies this framework to the PFL application. Averaging the post-period estimates $\hat{\beta}_0^w$ through $\hat{\beta}_5^w$ yields an average post-period ATT of 0.048 (SE$= 0.009$). The pre-treatment coefficients ($e = -5$ through $e = -2$) are close to zero, with the exception of a negative estimate at $e = -5$ that is not statistically distinguishable from zero, providing evidence consistent with the parallel trends assumption. The post-period estimates are generally positive, rising through $e = 2$–4 before returning toward zero at $e = 5$. The estimate at $e = 1$ is slightly below zero, likely reflecting sampling variability. Compared with the stacked DID estimate of 0.033 (SE$= 0.018$), the imputation-based estimate is somewhat larger in magnitude and substantially more precise. This pattern is consistent with the efficiency gains from using all untreated observations to construct counterfactual outcomes

rather than restricting to clean control groups.

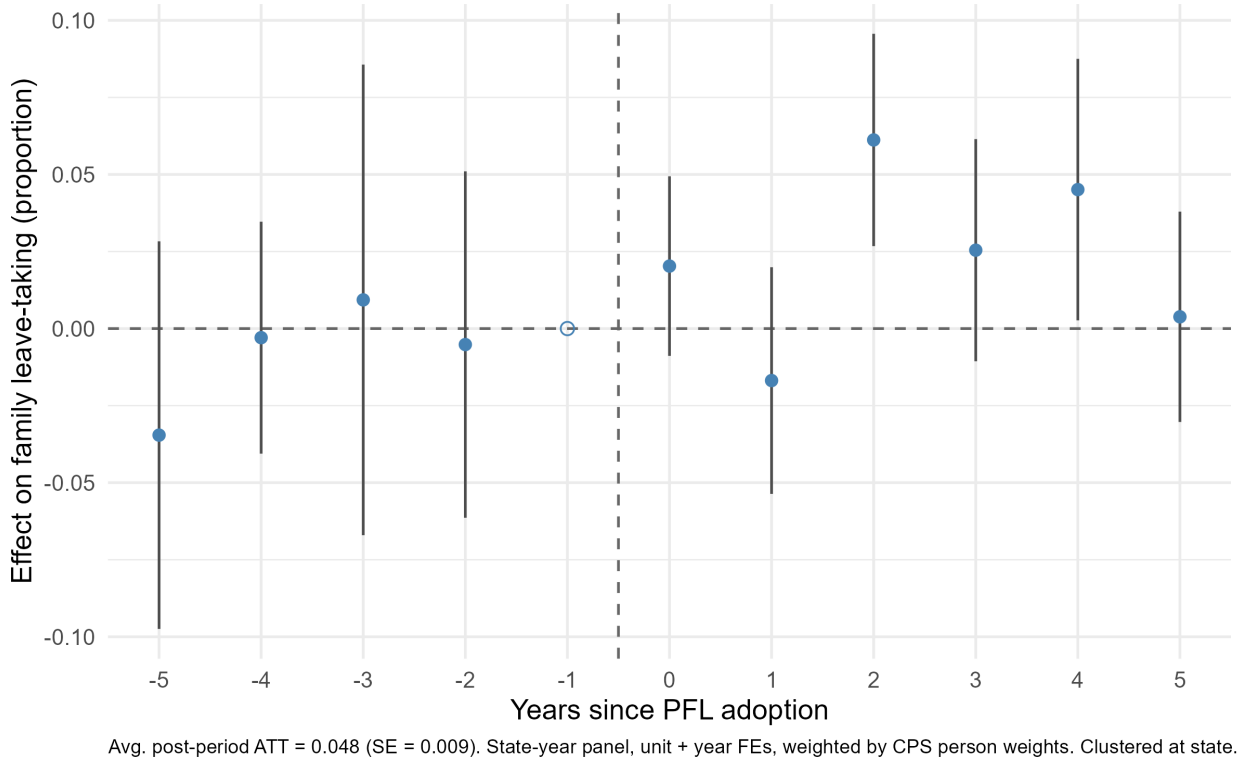


Figure 5 — Fixed effect imputation estimates of the effect of state Paid Family Leave programs on family leave-taking

Note: The sample is mothers of infants (youngest child under age 1) in PFL-adopting states, collapsed to a state-year panel and weighted by CPS person weights. The model includes unit and year fixed effects estimated on untreated observations only; treated counterfactuals are imputed out-of-sample. Six adoption cohorts: California (2004), New Jersey (2009), Rhode Island (2014), New York (2018), and the District of Columbia and Washington (both 2020), with a symmetric event window of $\kappa_{pre} = \kappa_{post} = 5$ years. Open circle: reference period ($e = -1$). Avg. post-period ATT$= 0.048$ (SE$= 0.009$). Standard errors clustered at the state level; 95% confidence intervals shown. Data: IPUMS CPS ASEC, 1999–2025 (Rossin-Slater, Ruhm, and Waldfogel 2013; Borusyak, Jaravel, and Spiess 2024).

3.2.2 Key Advantages of Fixed Effect Imputation

One major advantage of the Borusyak, Jaravel, and Spiess (2024) imputation method is that it makes use of all available observations in the data (i.e., it does not discard potential controls). Other approaches such as Callaway and P. H. Sant'Anna (2020) or the Wing et al. 2023 stacked DID design typically rely on the "clean controls" requirement. Only never-treated or not-yet-treated controls are kept, and other controls are dropped. While this exclusion is an important identification step, it comes at the cost of a smaller effective sample size, leading to wider confidence intervals and less precise inference. The Borusyak,

Jaravel, and Spiess (2024) style of using all untreated observations usually produces smaller standard errors or tighter confidence intervals. This is shown in their simulation exercise [6].

Another key strength is its flexibility and transparency. The estimator is built directly by imputing untreated potential outcomes and averaging treatment effects with user-specified weights. It is straightforward to define and estimate a wide range of causal parameters. For example, researchers can easily target calendar-time ATTs or subgroup-specific averages simply by adjusting the weights placed on each observation. The explicit imputation structure in BJS makes the logic of the estimator clear. Outcomes are simply decomposed into fixed effects, predicted untreated outcomes, and treatment effects. This intuitive and transparent design makes it easier to communicate results to both economists and general audiences.

Additionally, the analyst can easily add a set of control variables in BJS method. In the first-stage regression where untreated observations are used to impute counterfactuals,

$$Y_{s,t}(0) = \alpha_s + \gamma_t + \epsilon_{s,t} \tag{8}$$

the analyst can simply add $X_{s,t}$.

$$Y_{s,t}(0) = \alpha_s + \gamma_t + \gamma X_{s,t} + \epsilon_{s,t} \tag{9}$$

The researcher can augment the specification with covariates $X_{s,t}$ where γ is a set of coefficients for each covariate. This flexibility makes it straightforward to control for time-varying variables that may improve the plausibility of the parallel trends assumption. Incorporating covariates is far less natural in other approaches. For example, in the Wing et al. 2023's stacked DID framework, the estimand weights are derived from a saturated model without controls. Adding additional regressors breaks the weighting structure and invalidates the interpretation.Callaway and P. H. Sant'Anna (2020) (technically) allows for covariates, but in the form of baseline, non-time-varying characteristics. Their scheme uses a set of baseline covariates, rather than time-varying controls, to create propensity score for matched staggered DID.[7] The BJS setup provides a more flexible way to integrate co-

6. See the simulation example comparing the efficiency of different DID estimators at https://github.com/borusyak/did_imputation/blob/main/five_estimators_example.png.

7. Callaway and P. H. Sant'Anna (2020)directly estimates cohort-specific ATTs, $ATT(g,t)$, by comparing each adoption cohort to a not-yet-treated or never-treated comparison group. These group-time effects are then aggregated using researcher-specified weights into summary parameters such as an overall ATT or an event-study plot. A key strength of the CS approach is that it allows the parallel trends assumption to hold

variates into the analysis.

3.3. Summary

In sum, each alternative to the TWFE model for staggered adoption has its own strengths and weaknesses summarized below:

Stacked DID

- Best when one wants to avoid using already-treated units as controls by construction, while keeping the analysis close to a standard regression with familiar clustered SEs.
- Requires re-shaping data into event-time stacks and assumes treatment effects are best organized in event-time.

Fixed Effect Imputation (Borusyak–Jaravel–Spiess)

- Best when not-yet-treated comparisons become scarce late in the rollout, or when efficiency matters and you are willing to assume a credible model for untreated potential outcomes.
- Relies on correct specification of the untreated outcome model and assumptions may be harder to justify in applied settings.

With estimators in hand, Section 4 turns to a challenge that applies across all of these approaches: statistical inference, particularly when the number of treated clusters is small.

4 — Statistical Inference

4.1. Statistical Inference for Stacked DID

The stacked dataset structure raises specific inferential challenges beyond those of a standard DID. As discussed in Section 3, clean control units may appear in multiple subexperiments, so the same unit can contribute observations more than once. This creates dependence across subexperiments for a given unit. Standard errors must therefore be clustered at the unit level (e.g., state level), rather than at the unit-by-subexperiment level, which would treat these repeated observations as independent. In many policy applications, this implies clustering over a relatively small number of units, making the choice of inference method especially consequential.

---

only after conditioning on baseline (pre-treatment) covariates X, using outcome regression, inverse probability weighting, or doubly-robust estimands. Like BJS, the CS estimator avoids contaminated late-versus-early comparisons and recovers a well-defined causal aggregate. In practice, the choice between them often matters less than the underlying identifying assumptions and the research design. R package for CS estimator is available here: https://cran.r-project.org/web/packages/did/index.html

A central issue in DID designs is that observations within the same group (e.g. states, hospitals, or schools) tend to be correlated over time. This can be due to either similarities across members of the same group (intra-cluster correlation) and correlated errors over time, known as serial correlation or autocorrelation. Ignoring this dependence leads to underestimating the uncertainty in the results because the data contain fewer truly independent sources of variation than they appear to be. To correct this, researchers use cluster-robust standard errors as developed in Liang and Zeger (1986), which allow for arbitrary correlation of residuals within groups while assuming independence across them.

In DID settings, the key decision is *at what level to cluster*. Generally, researchers should cluster at the level where treatment varies, because that level defines the true source of independent variation. For example, if policy adoption occurs at the state level, standard errors should be clustered at the state level, even if the regression includes data for many subgroups or time periods within states. Simulations in Wing et al. (2023) show that clustering at the state-level—rather than the state-by-subexperiment—produces valid inference and avoids overcounting independent variation.

Often, however, DID research applications include a small number of clusters. For example, a study could use fewer than 50 states due to data limitations. As a rough rule of thumb, cluster-robust standard errors may be unreliable with fewer than about 50 clusters, and inference problems become more severe when the number of clusters falls below 20. In these settings, the wild cluster bootstrap becomes especially important. Standard cluster-robust standard errors rely on asymptotic approximations that may perform poorly. With few clusters, the estimated variance of the treatment effect can fluctuate a lot depending on the sample, leading to downward-biased standard errors and overconfident significance levels. In these cases, alternative inference methods are needed.

One flexible and widely recommended solution is the wild cluster bootstrap. The bootstrap is a resampling method where you repeatedly re-draw the data with replacement to approximate the sampling distribution of an estimator. The wild cluster bootstrap is a version tailored for clustered data. Instead of resampling individuals, it resamples entire clusters while allowing the residuals within clusters to keep their correlation structure. This works well even when the number of clusters is small, because it avoids relying on large-sample approximations. This provides an estimate that is much less likely to have an overly small p-value, but if clusters are too unbalanced in size, the estimate may be unstable. Among available inference methods, the wild cluster bootstrap is generally considered the most reliable and straightforward option for DID studies with few clusters, as it performs well across a range of simulation settings. For implementation details, see MacKinnon and

Webb (2018).[8]

Algorithm: Wild Cluster Bootstrap

1. Estimate the stacked DID regression model and obtain residuals $\hat{m}_{st}$.
2. For each cluster s, draw a random multiplier w_s from the distribution

$$w_s = \begin{cases} +1 & \text{with probability } 0.5, \\ -1 & \text{with probability } 0.5. \end{cases}$$

3. Form pseudo-residuals for each observation:

$$\hat{m}^*_{s,t} = w_s \, \hat{m}_{s,t}.$$

4. Construct pseudo-outcomes using the fitted values from the original regression:

$$Y^*_{s,t} = \hat{Y}_{s,t} + \hat{m}^*_{s,t}.$$

5. Re-estimate the regression using $Y^*_{s,t}$ to obtain the bootstrapped estimate β^*.
6. Repeat steps 2–5 many times (e.g. 999 iterations) to form the empirical distribution of β^*.
7. Compute the bootstrap p-value as the fraction of β^* estimates that are as extreme as or more extreme than the observed $\hat{\beta}$.

Other methods exist to correct for a small number of clusters.[9] Though these other methods can be helpful, the wild cluster bootstrap is the most flexible and accurate approach when the number of clusters is small (MacKinnon and Webb 2018).

4.2. Statistical Inference for Fixed Effect Imputation

Borusyak, Jaravel, and Spiess 2024 develop a method for statistical inference in their imputation model grounded in asymptotic theory. The main challenge in Borusyak, Jaravel, and Spiess 2024 statistical inference arises from the nature of their imputation procedure. Their imputation method will fit all treated observations ($D_{s,t} = 1$) perfectly. That is, for these treated units, the regression residuals are identically zero. This would incorrectly suggest that treated observations contain no idiosyncratic noise at all. Consequently, it is impossible to disentangle how much of the "observed difference" between $Y^{\text{obs}}_{s,t}$ and $\hat{Y}_{s,t}(0)$ reflects the true treatment effect versus random error.

To address this issue, Borusyak, Jaravel, and Spiess 2024 introduce a conservative infer-

8. Software implementations include the `fwildclusterboot` package in R (Fischer and Roodman 2021) and the `boottest` command in Stata (Roodman et al. 2019).

9. These include the CR2/CR3 bias-corrected estimators with Satterthwaite degrees-of-freedom adjustments and cluster-level aggregation, where inference is conducted on the average outcome per cluster.

ence procedure based on "pseudo-residuals" which is

$$\tilde{\epsilon}_{s,t} = Y_{s,t}^{obs} - \hat{a}_s - \hat{b}_t - \hat{\beta}_{s,t}^{\text{avg}}, \tag{10}$$

where $\hat{\beta}_{st}^{\text{avg}}$ is "some average" of estimated treatment effects over many units. This avoids setting residuals to zero while preserving meaningful variation. The resulting variance

$$\sigma_w^2 = E\left[\sum_s \left(\sum_t v_{s,t}\tilde{\epsilon}_{s,t}\right)^2\right] \tag{11}$$

is asymptotically *conservative* – that is, it weakly overstates the true variance in large samples.

Conceptually, this approach parallels heteroskedasticity-robust inference: both use noisy proxies for the true error variance but yield valid coverage once aggregated. Although conservative inference is less efficient, it provides reliable standard errors for average treatment effects in staggered adoption settings with heterogeneous responses. Borusyak, Jaravel, and Spiess 2024 outline a cluster robust version of their standard error estimate as well, and these can be implemented through their posted packages.

This chapter focuses on settings where treatment is binary and absorbing—once a unit adopts a policy, it remains treated for the duration of the sample. Extensions to continuous treatments (such as minimum wage levels or subsidy amounts) and to non-absorbing policies that can be reversed or interrupted are important in many applied settings (Chaisemartin, D'Haultfœuille, and Vazquez-Bare 2024). A growing body of work addresses these settings; readers interested in these extensions should consult the evolving literature.

5 — Conclusion

This chapter has introduced the core tools of modern difference-in-differences analysis. The canonical two-group, two-period DID provides an intuitive starting point: under parallel trends and no-anticipation, a simple double difference identifies the average treatment effect on the treated. But real policy settings rarely fit the two-by-two mold. Treatments are adopted at different times by different units, treatment effects may vary across cohorts and over event time, and the number of treated clusters is often small. The stacked DID and fixed effect imputation estimators address the first challenge; the wild cluster bootstrap addresses the last.

The rise of econometric theory in the DID literature can seem to have taken what was once a transparent and simple design and buried it in jargon. This chapter argues that this

is the wrong way to think about it. At its core, DID is about forming clean comparisons and aggregating them sensibly. The problem with conventional TWFE in staggered adoption settings is not a failure of the underlying logic—it is that TWFE silently mixes clean and contaminated comparisons. The new estimators do not replace DID; they discipline it, making explicit which comparisons are valid and ensuring that only those enter the aggregate. Applied researchers who understand these tools are better positioned to defend their designs, communicate their results, and avoid drawing causal conclusions that the data cannot support.

References

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting event-study designs: robust and efficient estimation." *Review of Economic Studies* 91 (6): 3253–3285. (Cited on pages 4, 5, 19, 27, 28, 30, 31, 35, 36).

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods" [in en]. *Journal of Econometrics*, Themed Issue: Treatment Effect 1, 225, no. 2 (December): 200–230. ISSN: 0304-4076, accessed January 3, 2023. https://doi.org/10.1016/j.jeconom.2020.12.001. https://www.sciencedirect.com/science/article/pii/S0304407620303948. (Cited on page 5).

———. 2020. "Difference-in-differences with multiple time periods." *Journal of Econometrics*, (cited on pages 19, 31, 32).

Chaisemartin, Clément de, Xavier D'Haultfœuille, and Gonzalo Vazquez-Bare. 2024. "Difference-in-Difference Estimators with Continuous Treatments and No Stayers." *AEA Papers and Proceedings* 114:610–613. https://doi.org/10.1257/pandp.20241049. (Cited on page 36).

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review* 110 (9): 2964–2996. (Cited on pages 3, 15).

Fischer, David, and David Roodman. 2021. *fwildclusterboot: Fast Wild Cluster Bootstrap Inference for Linear Regression Models*. R package. Available from https://cran.r-project.org/package=fwildclusterboot. (Cited on page 35).

Gardner, John. 2022. *Two-stage differences in differences*. ArXiv:2207.05943 [econ], July. Accessed November 12, 2022. https://doi.org/10.48550/arXiv.2207.05943. http://arxiv.org/abs/2207.05943. (Cited on page 20).

Goodman-Bacon, Andrew. 2018. *Difference-in-differences with variation in treatment timing*. Technical report. National Bureau of Economic Research. (Cited on pages 3, 15).

Liang, Kung-Yee, and Scott L Zeger. 1986. "Longitudinal data analysis using generalized linear models." *Biometrika* 73 (1): 13–22. (Cited on page 34).

MacKinnon, James G, and Matthew D Webb. 2018. "The wild bootstrap for few (treated) clusters." *The Econometrics Journal* 21 (2): 114–135. (Cited on pages 34, 35).

Meyer, Bruce D., W. Kip Viscusi, and David Durbin. 1995. "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment." *American Economic Review* 85 (3): 322–340. (Cited on page 3).

Roodman, David, Morten Ørregaard Nielsen, James G. MacKinnon, and Matthew D. Webb. 2019. "Fast and Wild: Bootstrap Inference in Stata Using `boottest`." *The Stata Journal* 19 (1): 4–60. https://doi.org/10.1177/1536867X19830877. (Cited on page 35).

Rossin-Slater, Maya, Christopher J. Ruhm, and Jane Waldfogel. 2013. "The Effects of California's Paid Family Leave Program on Mothers' Leave-Taking and Subsequent Labor Market Outcomes." *Journal of Policy Analysis and Management* 32 (2): 224–245. https://doi.org/10.1002/pam.21676. (Cited on pages 3, 4, 13, 14, 26, 31).

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701. https://doi.org/10.1037/h0037350. (Cited on page 6).

Sun, Liyang, and Sarah Abraham. 2020. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, (cited on page 19).

Wing, Coady, Seth Freedman, Alex Hollingsworth, and Koaali Simon. 2023. *Stacked Difference in Differences*. Working Paper. (Cited on pages 3, 5, 19, 21, 24, 25, 26, 31, 32, 34).

Wing, Coady, Madeline Yozwiak, Alex Hollingsworth, Seth Freedman, and Kosali Simon. 2024. "Designing Difference-in-Difference Studies with Staggered Treatment Adoption: Key Concepts and Practical Guidelines." *Annual Review of Public Health* 45 (Volume 45, 2024): 485–505. ISSN: 1545-2093. https://doi.org/https://doi.org/10.1146/annurev-publhealth-061022-050825. https://www.annualreviews.org/content/journals/10.1146/annurev-publhealth-061022-050825. (Cited on page 4).