Revisiting Event Study Designs: Robust and Efficient Estimation

Kirill Borusyak
UCL and CEPR

Xavier Jaravel
LSE and CEPR

Jann Spiess
Stanford*

May 7, 2021

Abstract

A broad empirical literature uses "event study," or "difference-in-differences with staggered rollout," research designs for treatment effect estimation: settings in which units in the panel receive treatment at different times. We show a series of problems with conventional regression-based two-way fixed effects estimators, both static and dynamic. These problems arise when researchers conflate the identifying assumptions of parallel trends and no anticipatory effects, implicit assumptions that restrict treatment effect heterogeneity, and the specification of the estimand as a weighted average of treatment effects. We then derive the efficient estimator robust to treatment effect heterogeneity for this setting, show that it has a particularly intuitive "imputation" form when treatment-effect heterogeneity is unrestricted, characterize its asymptotic behavior, provide tools for inference, and illustrate its attractive properties in simulations. We further discuss appropriate tests for parallel trends, and show how our estimation approach extends to many settings beyond standard event studies.

*Borusyak: k.borusyak@ucl.ac.uk; Jaravel: X.Jaravel@lse.ac.uk; Spiess: jspiess@stanford.edu. This draft supersedes our 2017 manuscript, "Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume." We thank Alberto Abadie, Isaiah Andrews, Raj Chetty, Itzik Fadlon, Ed Glaeser, Peter Hull, Guido Imbens, Larry Katz, Jack Liebersohn, and Jonathan Roth for thoughtful conversations and comments. Two accompanying Stata commands are available by request: `did_imputation` for treatment effect estimation with our imputation estimator and pre-trend testing, and `event_plot` for making dynamic event study plots.

1 Introduction

Event studies are one the most popular tools in applied economics and policy evaluation. In an event study, a difference-in-differences (DiD) design is implemented using panel data with a set of units that receive treatment at different moments in time. We make a series of points about identification and estimation of causal effects in such settings that revolve around the issues of treatment effect heterogeneity. We first develop a simple econometric framework that delineates the assumptions underlying event studies and the estimation target, as some average of heterogeneous causal effects. We apply this framework in two ways. First, we analyze the conventional practice of implementing event studies via two-way fixed effect (FE) regressions and show how the implicit conflation of different assumptions leads to a host of problems. Second, leveraging event study assumptions in an explicit and principled way allows us to derive the robust and efficient estimator, which takes an intuitive "imputation" form, along with appropriate inference methods and tests.

Event studies are frequently used to estimate treatment effects when treatment is not randomized, but the researcher has panel data allowing to compare outcome trajectories before and after the onset of treatment, as well as across units treated at different times. By analogy to conventional DiD designs, event studies are commonly implemented by two-way fixed effect regressions, such as

$$Y_{it} = \alpha_i + \beta_t + \tau D_{it} + \varepsilon_{it}, \tag{1}$$

where outcome Y_{it} and binary treatment D_{it} are measured in periods t and for units i, α_i are unit fixed effects that allow for different baseline outcomes across units, and β_t are period fixed effects that accommodate overall trends in the outcome. The hope is that specifications like (1) allow one to isolate a treatment effect τ from such trends and different baselines. A commonly-used dynamic version of this regression includes "lags" and "leads" of the indicator for the onset of treatment, to capture the dynamics of treatment effects and test for the parallel trajectories of the outcomes before the onset of treatment.

To understand the problems with conventional two-way fixed effect estimators and build a principled econometric framework for event-study designs, in Section 2 we develop a simple framework that makes estimation targets and underlying assumptions explicit and clearly isolated. We suppose that the researcher chooses a particular weighted average (or weighted sum) of treatment effects they are interested in. Our framework nests a broad class of empirically relevant estimands beyond the standard average treatment-on-the-treated (ATT) and similar averages restricted to a given number of periods after the treatment onset (by "horizon"), including group-wise ATTs and flexibly weighted ATTs that account for compositional changes. We make (and later test) two standard DiD identification assumptions: that potential outcomes without treatment are characterized by parallel trends and that there are no anticipatory effects. Finally, we allow for — but do not require — an auxiliary assumption that the treatment effects themselves follow some model that restricts their heterogeneity for *a priori* specified economic reasons. This explicit approach is in contrast to regression specifications like (1), both static and dynamic, which implicitly embed choices of estimation target and assumptions.

Through the lens of this framework, in Section 3 we uncover a set of problems with conventional regression-based event study estimation methods and trace them back to a lack of clearly stated or sepa-

rated goals and assumptions. First, we note that failing to rule out anticipation effects in "fully-dynamic" specifications (with all leads and lags of treatment included) leads to an underidentification problem, where the dynamic path of anticipation and treatment effects over time is not point-identified. We conclude that it is important to separate out testable assumptions about pre-trends from the estimation of dynamic treatment effects under those assumptions. Second, implicit homogeneity assumptions embedded in static event-study regressions may lead to estimands that put negative weights on some — typically long-run — treatment effects. With staggered rollout, regression-based estimation leverages "forbidden comparisons" between groups that got treated over a period of time and reference groups which had been treated earlier. Such comparisons are valid when the homogeneity assumption is true, but when it is violated, can substantially distort the weights the estimator places on treatment effects or even make them negative. Third, in dynamic specifications, implicit assumptions about treatment effect homogeneity lead to the spurious identification of long-run treatment effects for which no DiD contrasts valid under heterogeneous treatment effects are available. The last two problems highlight the danger of imposing implicit treatment effect homogeneity assumptions instead of embracing the heterogeneity and explicitly specifying the target estimand.

From the above discussion, the reader should not conclude that event study designs are cursed with fundamental problems. On the contrary, these challenges only arise due to a mismatch between the regression estimators and the underlying assumptions. In Section 4 we therefore use our framework to avoid these issues and derive robust and efficient estimators from first principles.

We first establish a simple characterization for the most efficient linear unbiased estimator of any pre-specified weighted average of treatment effects, under homoskedasticity. This estimator explicitly incorporates the researcher's estimation goal and assumptions about parallel trends, anticipation effects, and restrictions on treatment effect heterogeneity. It is constructed by estimating a flexible high-dimensional regression that differs from conventional event study specifications, and aggregating its coefficients appropriately. While homoskedasticity is unrealistic in most applications, the principled construction of this estimator yields attractive efficiency properties more generally, as we confirm in simulations. Specifically, we find that imputation estimator has sizable efficiency advantages, with variances of alternative estimators 15–41% higher in a homoskedastic baseline; these gains are preserved under heteroskedasticity and serial correlation of residuals.

The efficient robust estimator takes a particularly transparent "imputation" form in our leading case where the heterogeneity of treatment effects is not restricted. The imputation estimator is constructed in three steps. First, the unit and period fixed effects $\hat{\alpha}_i$ and $\hat{\beta}_t$ are fitted by regression on untreated observations only. Second, they are used to impute the untreated potential outcomes and therefore obtain an estimated treatment effect $\hat{\tau}_{it} = Y_{it} - \hat{\alpha}_i - \hat{\beta}_t$ for each treated observation. Finally, a weighted average of these treatment effect estimates is taken with weights, corresponding to the estimation target.

While we propose and implement this specific, efficient estimator, we also show that the imputation structure is more general. First, any other linear estimator that is unbiased in our framework with unrestricted causal effects can be represented as *an* imputation estimator, albeit with an inefficient way of imputing untreated potential outcomes. Second, even when assumptions that restrict treatment effect heterogeneity are imposed, any unbiased estimator can still be understood as an imputation estimator for

an adjusted estimand. Together, these two results allow us to characterize estimators of treatment effects in event studies as a combination of how they impute unobserved potential outcomes and which weights they put on treatment effects.

For the efficient estimator in our framework, we analyze the large-sample properties, provide results for valid inference, and propose new tests for the identifying assumptions. We provide conditions under which the estimator is consistent and asymptotically Normal in large panels. We then propose standard error estimates that are valid in large samples. These estimates are challenging to obtain under arbitrary treatment effect heterogeneity, because causal effects cannot be separated from the error terms. We instead show how they can be estimated asymptotically conservatively, by attributing some variation in estimated treatment effects to the error terms. We next propose a principled way of testing the identifying assumptions of parallel trends and no anticipation effects, based on regressions with untreated observations only. Thanks both to the properties of our efficient estimator and to the separation of testing from estimation, our test avoids some of the pre-testing problems pointed out by Roth (2018).

While our baseline setting is for panel data with two-way fixed effects, we show how our formal results extend naturally in a number of ways. We allow for more general panel-type specifications that can include, for instance, unit-specific trends or time-varying covariates. In Section 5 we consider extensions to repeated cross-sections, data defined by two cross-sectional dimensions (e.g. regions and age groups), triple-differences designs, and other data structures. We further discuss the implications of our results when treatment is simultaneous rather than staggered, when it can switch on and off, and when multiple treatment events can happen in the same unit.

It is also useful to point out the limitations of our estimation strategy. First, all event study designs assume a restrictive parametric model for untreated outcomes. We do not evaluate when these assumptions may be applicable, and therefore when the event study design are *ex ante* appropriate, as Roth and Sant'Anna (2020) do. We similarly do not consider estimation that is robust to violations of parallel-trend type assumptions, as Rambachan and Roth (2020) propose. We instead take the standard assumptions of event study designs as given and derive optimal estimators, valid inference, and practical tests. Second, we also do not consider event studies as understood in the finance literature, based on high-frequency panel data (MacKinlay, 1997).

For convenient application of our results, we supply a Stata command, `did_imputation`, which implements the imputation estimator and inference for it in a computationally efficient way. Our command handles a variety of practicalities, such as additional covariates and fixed effects, observation weights, and repeated cross-sections. We also provide a second command, `event_plot`, for producing "event study plots" that visualize the estimates with both our estimator and the alternative ones. An empirical application illustrating our ideas and using these commands will be added in future versions of the draft.

Our paper contributes to a growing methodological literature on event studies. To the best of our knowledge, our paper is the first and only one to characterize the underidentification and spurious identification problems that arise in regression-based implementations of event study designs. The negative weighting problem has received more attention. It was first shown by De Chaisemartin and D'Haultfœuille (2015, Supplement 1). The earlier manuscript of our paper (Borusyak and Jaravel, 2017) independently pointed it out and additionally explained how it arises because of forbidden comparisons and why it affects

long-run effects in particular, which we now discuss in Section 3.2 below. The issue has since been further investigated by Goodman-Bacon (2018), Strezhnev (2018), and De Chaisemartin and D'Haultfœuille (2020), while Sun and Abraham (2021) have shown similar problems with dynamic specifications. Sun and Abraham (2021) and Roth (2018) have further uncovered problems with conventional pre-trend tests, and Schmidheiny and Siegloch (2020) have characterized the problems which arise from binning multiple lags and leads in dynamic specifications. Besides being the first to point out some of these issues, our paper provides a unifying econometric framework which explicitly relates these issues to the conflation of the target estimand and the underlying assumptions.

Several papers have proposed ways to address these problems with conventional regression-based event studies, introducing estimators that remain valid when treatment effects can vary arbitrarily (De Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Cengiz et al., 2019).[1] An important limitation of these robust estimators is that their efficiency properties are not known. A key contribution of our paper is to derive a robust and efficient estimator from first principles. We show that this estimator takes a particularly transparent form under unrestricted treatment effect heterogeneity, while our construction also yields efficiency when some restrictions on treatment effects are imposed. By clearly separating the testing of underlying assumptions from estimation that imposes them, we simultaneously increase estimation efficiency and mitigate pre-testing bias. In particular, our estimator uses all pre-treatment periods for imputation, as appropriate under the DiD assumptions, while alternative estimators use more limited information; see Section 4.5 for a detailed discussion.[2]

Finally, our paper is related to a nascent literature that develops robust estimators similar to the imputation estimator. To the best of our knowledge, this idea has been first proposed for factor models (Gobillon and Magnac, 2016; Xu, 2017). In contemporaneous and independent work, it has been applied to event studies by Liu et al. (2020) and Gardner (2021). Specifically, the counterfactual estimator of Liu et al. (2020) and the two-stage estimator of Gardner (2021) coincide with the imputation estimator in our model for the estimands their papers consider. Relative to these contributions, we *derive* the imputation estimator from first principles, show its efficiency, provide tools for valid asymptotic inference when unit fixed effects are included, and show its robustness to pre-testing.

2 Setting

We consider estimation of causal effects of a binary treatment D_{it} on some outcome Y_{it} in a panel of units i and periods t. We focus on "staggered rollout" designs, in which being treated is an absorbing state. That is, for each unit there is an event date E_i when D_{it} switches from 0 and 1 and stays there forever: $D_{it} = \mathbf{1}[K_{it} \geq 0]$, where $K_{it} = t - E_i$ is the number of periods since the event date ("relative time"). We

[1] Baker et al. (2021) provide evidence that using robust, rather than regression-based, estimation matters in published empirical research.

[2] A separate strand of literature has considered design-based inference with event studies, where randomness arises in the treatment assignment and timing, following a known model (Athey and Imbens, 2018; Roth and Sant'Anna, 2021; Arkhangelsky and Imbens, 2019). Our identification and inference approach differs in that we condition on treatment timing and take the panel as given. With this approach, we follow Abadie et al. (2014) in focussing on an in-sample estimand, which in our case is a weighted sample average treatment effect for those observations that were actually treated.

also allow that some units are never treated, denoted by $E_i = \infty$.[3] Units with the same event date are referred to as a cohort.

We do not make any random sampling assumptions and work with a fixed set of N observations $it \in \Omega$, which may or may not form a balanced panel. We similarly view the event date for each unit, and therefore all treatment indicators, as fixed. We define the set of treated observations by $\Omega_1 = \{it \in \Omega\colon\ D_{it} = 1\}$ of size N_1 and the set of untreated observations by $\Omega_0 = \{it \in \Omega\colon\ D_{it} = 0\}$ of size N_0.

We denote by $Y_{it}(0)$ the period-t potential outcome of unit i if it is never treated. We are then interested in causal effects $\tau_{it} = Y_{it} - Y_{it}(0)$ on the treated observations $it \in \Omega_1$.[4] Without loss of generality, we consider treatment effects τ_{it} non-stochastic, too; the only source of randomness is therefore in $Y_{it}(0)$.[5]

We suppose a researcher is interested in a statistic which sums or averages treatment effects $\tau = (\tau_{it})_{it\in\Omega_1}$ over the set of treated observations with pre-specified non-stochastic weights $w_1 = (w_{it})_{it\in\Omega_1}$ that can depend on treatment assignment and timing, but not on realized outcomes:

Estimation Target. $\tau_w = \sum_{it\in\Omega_1} w_{it}\tau_{it} \equiv w_1'\tau$.

For notation brevity we consider scalar estimands.

Different weights are appropriate for different research questions. The researcher may be interested in the overall ATT, formalized by $w_{it} = 1/N_1$ for all $it \in \Omega_1$. In "event study" analyses a common estimand is the average effect h periods since treatment for a given "horizon" $h \geq 0$: $w_{it} = \mathbf{1}\left[K_{it} = h\right] / \left|\Omega_{1,h}\right|$ for $\Omega_{1,h} = \{it\colon\ K_{it} = h\}$. One may also be interested in its "balanced" version: the average treatment effects at horizon h computed only for the subset of units also observed at horizon h', such that the gap between two or more estimates is not confounded by compositional differences. Finally, we do not require the w_{it} to add up to one; for example, a researcher may be interested in the difference between average treatment effects at different horizons or across some groups of units, corresponding to $\sum_{it\in\Omega_1} w_{it} = 0$.

We note that in practice researchers are not always explicit about their estimand of interest and may be happy with any "reasonable" average of treatment effects, informally understood. In Section 3 we will show why specifying the estimand (or perhaps a class of estimands) is important.

To identify τ_w we consider three assumptions. We start with the parallel trends assumption, which imposes a two-way fixed effect (TWFE) model on the untreated potential outcome, which is standard in difference-in-differences designs.[6]

[3] In principle always-treated units are also allowed for, by $E_i = -\infty$, but in practice they will not be useful for causal identification with flexible treatment effect heterogeneity.

[4] This formulation allows treatment effects to be not only heterogeneous across observations but also dynamic, i.e. to causally depend on the number of periods since treatment. While Sun and Abraham (2021) and Callaway and Sant'Anna (2021) use dynamic potential outcomes as a starting point, we save on notation by defining the realized treatment effect τ_{it} directly. Since we condition on treatment timing, the two formulations are isomorphic in our setting.

[5] All our results go through with stochastic τ_{it} if the estimands are based on $\mathbb{E}\left[\tau_{it}\right]$. Appendix A.5 further shows how this framework can be derived from one in which the set of observations and treatment timing are stochastic, too, by conditioning on them.

[6] An interesting line of work by Athey and Imbens (2018) and Roth and Sant'Anna (2021) develops methods for causal inference in staggered adoption designs which do not require parallel trends, but are instead based on randomness of the event date. Rambachan and Roth (2020) instead consider a relaxed version of Assumption 1 which imposes bounds on non-parallel trends and leads to set identification of causal effects.

Assumption 1 (Parallel trends)**.** *There exist non-stochastic* α_i *and* β_t *such that* $Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}$ *with* $\mathbb{E}[\varepsilon_{it}] = 0$ *for all* $it \in \Omega$.[7]

Our framework extends immediately to richer models of $Y_{it}(0)$, which may include time-varying controls, unit-specific trends or additional fixed effects (see Assumption 1′ below). Similarly, it applies in settings where unit fixed effects are not appropriate, as with repeated cross-sections, or even if the data do not have a panel structure (see Section 5). Assumption 1, however, is inherently asymmetric in that it imposes restrictions on $Y_{it}(0)$ but not on τ_{it}. This asymmetry is natural and reflects the standard practice in staggered rollout DiD designs, hence our focus on this class of models.

To be able to identify α_i and β_t, we next rule out anticipation effects, i.e. the causal effects of being treated in the future on current outcomes:

Assumption 2 (No anticipation effects)**.** $Y_{it} = Y_{it}(0)$ *for all* $it \in \Omega_0$.

It is straightforward to weaken this assumption, e.g. by allowing anticipation for some k periods before treatment: this simply requires redefining event dates to earlier ones. However, some form of this assumption is necessary for DiD identification, as there would be no reference periods for treated units otherwise. Assumptions 1 and 2 together imply that the observed outcomes Y_{it} for untreated observations follow the TWFE model.

Finally, researchers sometimes impose restrictions on causal effects, explicitly or implicitly. For instance, τ_{it} may be assumed to be homogeneous for all units and periods, or only depend on the number of periods since treatment (but be otherwise homogeneous across units and periods), or perhaps to be time-invariant for each unit. We will consider such restrictions as a possible auxiliary assumption:

Assumption 3 (Restricted causal effects)**.** $B\tau = 0$ *for a known* $M \times N_1$ *matrix* B *of full row rank.*

If restrictions on the treatment effects are implied by economic theory, imposing them will increase estimation power.[8] Often, however, this assumption is imposed without an *ex ante* justification, but just because it yields a simple model for the outcome. We will show in Section 3.2 how imposing this assumption, when violated, impedes estimation of reasonable averages of treatment effects, let alone a particular estimand τ_w.

It will be more convenient for us to work with an equivalent formulation of Assumption 3, based on $N_1 - M$ free parameters driving treatment effects rather than M restrictions on them:

Assumption 3′ (Model of causal effects)**.** $\tau = \Gamma\theta$*, where* θ *is a* $(N_1 - M)\times 1$ *vector of unknown parameters and* Γ *is a known* $N_1 \times (N_1 - M)$ *matrix of full column rank.*

Assumption 3′ imposes a parametric model of treatment effects. For example, the assumption that treatment effects all be the same, $\tau_{it} \equiv \theta_1$, corresponds to $N_1 - M = 1$ and $\Gamma = (1, \ldots, 1)'$. In contrast, a "null model" $\tau_{it} \equiv \theta_{it}$ that imposes no restrictions is captured by $M = 0$ and $\Gamma = \mathbb{I}_{N_1}$.

[7] In estimation, we will set the fixed effect of either one unit or one period to zero, such as $\beta_1 = 0$. This is without loss of generality, since the TWFE model is otherwise over-parameterized.

[8] In some cases τ_w may even not be identified without such restrictions, while it is identified with them.

3 Conventional Practice and Associated Problems

Causal effects in staggered adoption DiD designs have traditionally been estimated via Ordinary Least Squares (OLS) regressions with two-way fixed effects. While details may vary, the following specification covers many studies:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{\substack{h=-a \\ h\neq -1}}^{b-1} \tau_h \mathbf{1}\left[K_{it} = h\right] + \tau_{b+} \mathbf{1}\left[K_{it} \geq b\right] + \tilde{\varepsilon}_{it}, \tag{2}$$

Here $\tilde{\alpha}_i$ and $\tilde{\beta}_t$ are the unit and period ("two-way") fixed effects, $a \geq 0$ and $b \geq 0$ are the numbers of included "leads" and "lags" of the event indicator, respectively, and $\tilde{\varepsilon}_{it}$ is the error term. The first lead, $\mathbf{1}\left[K_{it} = -1\right]$, is often excluded as a normalization, while the coefficients on the other leads (if present) are interpreted as measures of "pre-trends," and the hypothesis that $\tau_{-a} = \cdots = \tau_{-2} = 0$ is tested visually or statistically. Conditionally on this test passing, the coefficients on the lags are interpreted as a dynamic path of causal effects: at $h = 0, \ldots, b-1$ periods after treatment and, in the case of τ_{b+}, at longer horizons binned together.[9] We will refer to this specification as *"dynamic"* (as long as $b > 0$) or, if it includes all available leads and lags except $h = -1$, *"fully dynamic."*

Viewed through the lens of the Section 2 framework, these specifications make implicit assumptions on potential outcomes, anticipation and treatment effects, and the estimand of interest. First, they make Assumption 1 but, for $a > 0$, do not fully impose Assumption 2, allowing for anticipation effects for a periods before treatment.[10] Typically this is done as a means to *test* Assumption 2 rather than to *relax* it, but the resulting specification is the same.

Second, Equation (2) imposes strong restrictions on causal effect heterogeneity (Assumption 3), with treatment (and anticipation) effects assumed to only vary by horizon h and not across units and periods otherwise. Most often, this is done without an *a priori* justification. If the lags are binned into the term with τ_{b+}, the effects are further assumed time-invariant once b periods have elapsed since the event.

Finally, dynamic specifications do not explicitly define the estimands τ_h as particular averages of heterogeneous causal effects, even though researchers estimating them often admit that the effects may indeed vary across observations. Instead, the OLS coefficients for τ_h are interpreted as averages of causal effects for horizon h with unspecified weights, which are presumed reasonable.

Besides dynamic specifications, Equation (2) also nests a very common specification used when a researcher is interested in a single parameter summarizing all causal effects. With $a = b = 0$, we have the *"static"* specification in which a single treatment indicator is included:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \tau^{\text{static}} D_{it} + \tilde{\varepsilon}_{it}. \tag{3}$$

In line with our Section 2 setting, the static equation imposes the parallel trends and no anticipation Assumptions 1 and 2. However, it also makes a particularly strong version of Assumption 3 — that all

[9] Schmidheiny and Siegloch (2020) discuss the problems that arise when the bin $\mathbf{1}\left[K_{it} \geq b\right]$ is not included.

[10] One can alternatively view this specification as imposing Assumption 2 but making a weaker Assumption 1 which includes some pre-trends into $Y_{it}(0)$. This difference in interpretation is immaterial for our results.

treatment effects are the same. Moreover, the target estimand is again not written out as an explicit average of potentially heterogeneous causal effects.

In the rest of this section we turn to the problems with OLS estimation of Equations (2) or (3). We explain how these issues result from the conflation of the target estimand, Assumption 2 and Assumption 3, providing a new and unified perspective on the problems of static and dynamic OLS-based methods. Specifically, in Sections 3.1, 3.2 and 3.3 we focus on the three problems originally identified in De Chaisemartin and D'Haultfœuille (2015) (for negative weights) and our original manuscript, Borusyak and Jaravel (2017); some of them were further investigated by Strezhnev (2018) and Goodman-Bacon (2018)): underidentification of the fully-dynamic regression, negative weighting in the static regression, and spurious identification of the long-run effects. In Section 3.4 we discuss how our framework also relates to other problems that have since been pointed out by Roth (2018) and Sun and Abraham (2021).

3.1 Under-Identification of the Fully-Dynamic Specification

The first problem pertains to fully-dynamic specifications and arises because a strong enough Assumption 2 is not imposed. We show that those specifications are under-identified if there is no never-treated group:

Lemma 1. *If there are no never-treated units, the path of $\{\tau_h\}_{h\neq -1}$ coefficients is not point identified in the fully-dynamic OLS specification. In particular, adding a linear trend to this path, $\{\tau_h + \kappa\,(h+1)\}$ for any $\kappa \in \mathbb{R}$, fits the data equally well with the fixed effect coefficients appropriately modified.*

Proof. All proofs are given in Appendix B. □

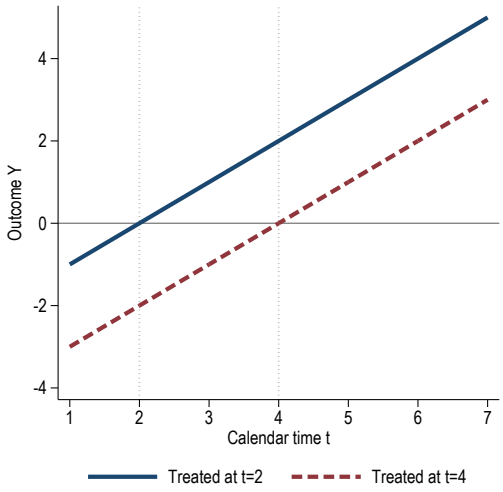
This result can be illustrated with a simple example, which we intentionally make extreme and free from additional effects and noise. Consider Figure 1, which plots the outcomes for a simulated dataset with two units (or cohorts): one treated early at $t = 2$ (solid line) and the other one later at $t = 4$ (dashed line). Both units are observed for periods $t = 1, \ldots, 7$, and the outcomes exhibit linear growth with the same slope of one, but starting from different levels. There are two interpretations of what could cause such dynamics. On one hand, treatment could have no impact on the outcome, in which case the level difference corresponds to the unit FEs, while trends are just a common feature of the environment, formalized by the period FEs. On the other hand, note that the outcome equals the number of periods since the event for both groups and all time periods: it is zero at the moment of treatment, negative before and positive after. So a possible interpretation is that the outcome is entirely driven by causal effects of treatment and anticipation of treatment. One cannot hope to distinguish between unrestricted dynamic causal effects and a combination of unit effects with time trends.[11]

Formally, the problem arises because a linear time trend t and a linear term in the cohort E_i (subsumed by the unit FEs) can perfectly reproduce a linear term in relative time $K_{it} = t - E_i$. Therefore, a complete set of treatment leads and lags, which is equivalent to the FE of relative time, is collinear with the unit and period FEs.[12]

[11] Note that in the former case, the solid line is a vertical shift up (a level shift) from the dashed line, while in the latter case it is a horizontal shift to the left that is due to the differential timing. With straight lines, these are observationally equivalent.

[12] The mechanics of this issue are essentially the same as in the well-known "age-cohort-time" problem, where the set of units treated at a given time $E_i = e$ can be viewed as a birth cohort and the relative time $K_{it} = t - E_i$ serves as age.

Figure 1: Underidentification of Fully-Dynamic Specification



The problem may be important in practice, as statistical packages may resolve this collinearity by dropping an arbitrary unit or period indicator. Some estimates of $\{\tau_h\}$ would then be produced, but because of an arbitrary trend in the coefficients they may suggest a violation of parallel trends even when the specification is in fact correct, i.e. Assumptions 1 and 2 hold and there is no heterogeneity of treatment effects for each horizon (Assumption 3).

To break the collinearity problem, stronger restrictions on anticipation effects, and thus on Y_{it} for untreated observations, have to be introduced. One could consider imposing minimal restrictions on the specification that would make it identified. In typical cases, only a linear trend in $\{\tau_h\}$ is not identified in the fully dynamic specification, while nonlinear paths cannot be reproduced with unit and period fixed effects. Therefore, just one additional normalization, e.g. $\tau_{-a} = 0$ in addition to $\tau_{-1} = 0$, breaks multicollinearity.[13]

However, minimal identified models rely on *ad hoc* identification assumptions which are *a priori* unattractive. For instance, just imposing $\tau_{-a} = \tau_{-1} = 0$ means that anticipation effects are assumed away 1 and a periods before treatment, but not in other pre-periods. This assumption therefore depends on the *ex post* realizations of treatment timing.

Instead, a systematic approach is to impose the assumptions — some forms of no anticipation effects and parallel trends — that the researcher has an *a priori* argument for and which motivated the use of DiD to begin with. Such assumptions on anticipation effects give much stronger identification power relative to the unnecessarily flexible specifications, of which the fully-dynamic specification is an extreme example. Importantly, our suggestion to impose identification assumptions at the estimation stage does not mean that those assumptions should not also be tested; we discuss testing in detail below. Rather, the separation of estimation and testing makes the identification argument explicit, while stronger-than-minimal assumptions allow for more powerful testing.

[13]There are some exceptions in which additional collinearity arises, e.g. when treatment is staggered but happens at periodic intervals.

Table 1: Two-Unit, Three-Period Example

$\mathbb{E}[Y_{it}]$	$i = A$	$i = B$
$t = 1$	α_A	α_B
$t = 2$	$\alpha_A + \beta_2 + \tau_{A2}$	$\alpha_B + \beta_2$
$t = 3$	$\alpha_A + \beta_3 + \tau_{A3}$	$\alpha_B + \beta_3 + \tau_{B3}$
Event date	$E_i = 2$	$E_i = 3$

Notes: without loss of generality, we normalize $\beta_1 = 0$.

3.2 Negative Weighting in the Static Regression

We now show how, by imposing Assumption 3 instead of specifying the estimation target, the static TWFE specification does not identify a reasonably-weighted average of heterogeneous treatment effects. The underlying weights may be negative, particularly for the long-run causal effects.[14]

First, we show that, if the parallel trends and no anticipation assumptions hold, static OLS identifies *some* weighted average of treatment effects:

Lemma 2. *If Assumptions 1 and 2 hold, then the estimand of the static OLS specification in (3) satisfies* $\tau^{static} = \sum_{it \in \Omega_1} w_{it}^{OLS} \tau_{it}$ *for some weights* w_{it}^{OLS} *that do not depend on the outcome realizations and add up to one,* $\sum_{it \in \Omega_1} w_{it}^{OLS} = 1$.

The underlying weights w_{it}^{OLS} can be easily estimated from the data using the Frisch–Waugh–Lovell theorem (see Equation (17) in the proof of Lemma 2) and only depend on the timing of treatment for each unit and the set of observed units and periods.

The static OLS estimand, however, cannot be interpreted as a *proper* weighted average, as some weights can be negative. We illustrate this problem with a simple example:

Lemma 3. *Suppose Assumptions 1 and 2 hold and the data consist of two units, A and B, treated in periods 2 and 3, respectively, both observed in periods* $t = 1, 2, 3$ *(as shown in Table 1). Then the estimand of the static OLS specification (3) can be expressed as* $\tau^{static} = \tau_{A2} + \frac{1}{2}\tau_{B3} - \frac{1}{2}\tau_{A3}$.

This example illustrates the severe short-run bias of TWFE OLS: the long-run causal effect, corresponding to the early-treated unit A and the late period 3, enters with a negative weight $(-1/2)$. Thus, the larger the effects are in the long-run, the smaller the coefficient will be.

This problem results from what we call "forbidden comparisons" performed by OLS. Recall that the original idea of DiD estimation is to compare the evolution of outcomes over some time interval for the units which got treated during that interval relative to a reference group of units which didn't, identifying the period FEs. In the Lemma 3 example, such an "admissible" contrast is between units A and B in periods 2 and 1, $(Y_{A2} - Y_{A1}) - (Y_{B2} - Y_{B1})$. However, panels with staggered treatment timing also lend

[14] Since De Chaisemartin and D'Haultfœuille (2015) and our earlier draft pointed out this issue, it has been extensively studied. Strezhnev (2018) and Goodman-Bacon (2018) provide two decompositions of the static regression's estimand, while De Chaisemartin and D'Haultfœuille (2020) characterize the weights analytically in a special case of complete panels (or, if group FEs are included instead of the unit FEs, a complete set of group-by-period cells). We complement this discussion by attributing the fundamental cause of the problem to the conflation of assumptions delineated in our framework from Section 2. Early insights into this problem in specific applications can be found in Wolfers (2006) and Meer and West (2016).

themselves to a second type of contrasts — which we label forbidden — in which the reference group has been treated throughout the relevant period. For units in this group, the treatment indicator D_{it} does not change over the relevant period, and so OLS uses them to identify period FEs, too. The contrast between units B and A in periods 3 and 2, $(Y_{B3} - Y_{B2}) - (Y_{A3} - Y_{A2})$, in Lemma 3 is a case in point. While a contrast like this is appropriate and increases efficiency when treatment effects are homogeneous (which OLS was designed for), forbidden comparisons are problematic under treatment effect homogeneity. For instance, subtracting $Y_{A3} - Y_{A2}$ not only removes the gap in period FEs, $\beta_3 - \beta_2$, but also deducts the evolution of treatment effects $\tau_{A3} - \tau_{A2}$, placing a negative weight on τ_{A3}. OLS leverages contrasts of both types and estimates the treatment effect by $\hat{\tau}^{static} = (Y_{B2} - Y_{A2}) - \frac{1}{2}(Y_{B1} - Y_{A1}) - \frac{1}{2}(Y_{B3} - Y_{A3})$.[15]

Fundamentally, this problem arises because OLS estimation imposes very strong restrictions on treatment effect homogeneity, i.e. Assumption 3, instead of acknowledging the heterogeneity and specifying a particular target estimand (or perhaps a class of estimands that the researcher is indifferent between). Such conflation is likely a consequence of the common perception that regression estimators generally recover reasonably-weighted averages of treatment effects. Even if one gets some variance-weighted average instead of the policy-relevant ATT, the benefit of convenience dominates. This perception is correct in regressions with saturated controls, as the seminal paper of Angrist (1998) has shown. However, it fails in staggered adoption DiD designs where the set of controls (i.e. unit and period FEs) is non-saturated and complex.

With a large never-treated group, our setting becomes closer to a classical non-staggered DiD design, and therefore negative weights disappear. Yet, the weights may remain highly unequal and diverge from the estimands that the researcher is interested in. Committing to the estimation target and embracing treatment effect heterogeneity, except where some form of Assumption 3 is *ex ante* appropriate, is therefore our preferred strategy.

3.3 Spurious Identification of Long-Run Causal Effects

Another consequence of inappropriately imposing Assumption 3 concerns estimation of long-run causal effects. Dynamic OLS specifications (except those subject to the underidentification problem) yield *some* estimates for all τ_h coefficients, provided sufficiently many restrictions on pre-trends are imposed. Yet, for large enough h, no averages of treatment effects are identified under Assumptions 1 and 2 with treatment effect heterogeneity. Therefore, OLS estimates are fully driven by unwarranted extrapolation of treatment effects across observations and may not be trusted, unless strong *ex ante* reasons for Assumption 3 exist.

This issue is well illustrated in the example of Lemma 3. To identify the long-run effect τ_{A3} under Assumptions 1 and 2, one needs to form an admissible DiD contrast, comparing the outcome growth over some period between unit A and another unit not yet treated in period 3. But by period 3 both units have been treated. Mechanically, this problem arises because the period fixed effect β_3 is not identified separately from the treatment effects τ_{A3} and τ_{B3} in this example, absent restrictions on treatment effects.

[15] The proof of Lemma 2 shows why long-run effects in particular are subject to the negative weights problem. In general, negative weights arise for the treated observations, for which the residual from an auxiliary regression of D_{it} on the two-way FEs is negative. De Chaisemartin and D'Haultfœuille 2020 show that, in balanced panels, the unit FEs are higher for early-treated units (which are observed treated for a larger shares of periods) and period FEs are higher for later periods (in which a larger shares of units are treated). The early-treated units observed in later periods correspond to the long-run effects.

Indeed, in Table 1, we cannot determine τ_{A3} and τ_{B3} because we cannot distinguish them from the fixed effect β_3 — in contrast to τ_{A2}, an unbiased estimate of which can be obtained from period 1 and 2 outcomes by the usual DiD estimator $\hat{\tau}_{A2} = (Y_{A2} - Y_{A1}) - (Y_{B2} - Y_{B1})$. In general, the gap between the earliest and the latest event times observed in the data puts an upper bound on the number of dynamic coefficients that can identified. This result, which follows by the same logic of non-identification of the later period effects, is formalized by our next lemma:

Lemma 4. *Suppose there are no never-treated units and let $\bar{H} = \max_i E_i - \min_i E_i$. Then, for any non-negative weights w_{it} defined over the set of observations with $K_{it} \geq \bar{H}$ (that are not identically zero), the weighted sum of causal effects $\sum_{it:\ K_{it} \geq \bar{H}} w_{it}\tau_{it}$ is not identified by Assumptions 1 and 2.*[16]

This result applies to all estimators, not just those based on OLS. However, robust estimators, including the one we characterize in Section 4, would not be possible to compute for non-identified estimands, never resulting in spurious estimates.

3.4 Other Limitations of Conventional Practice

We finally discuss how our framework relates to two other limitations of conventional practice pointed out in other work.

First, Roth (2018) uncovers undesirable consequences of conditioning on a pre-trend test passing. He shows that OLS pre-trend estimates are correlated with the estimates of treatment effects obtained from the same dynamic specification. As a consequence, when Assumptions 1 and 2 are in fact satisfied, pre-trend testing makes statistical inference on the coefficients invalid; if the assumptions are instead violated, this can exacerbate estimation bias. In Section 4.4 we will show that this problem can be attributed to the fact that Assumption 2 is not imposed when estimating the treatment effects by dynamic OLS specifications, and testing is not separated from estimation—similarly to the underidentification problem discussed above.

Second, Sun and Abraham (2021) show that the negative weighting problem, that we initially pointed out for static regressions, is also relevant in dynamic OLS specifications. Specifically, causal estimates for one horizon may be confounded by heterogeneous effects at others. In light of our framework, this problem arises because Assumption 3 is imposed instead of specifying a target estimand.

Sun and Abraham (2021) further show that pre-trend tests may reject Assumption 2 when it is in fact satisfied or pass when the assumption fails, all because of heterogeneous treatment effects. That problem is another variation on how the Section 2 assumptions may be conflated: the conventional pre-trend test for Assumptions 1 and 2 is in fact a joint test of those assumptions together with Assumption 3 and can reject because of the latter.

We conclude that OLS estimation of both static and dynamic specifications suffers from a host of problems, which all arise from the conflation of the target estimand, Assumptions 1, 2 and 3. Importantly, these problems pertain only to OLS estimation of event studies, not to the event study design itself, as we show next.

[16] The requirement that the weights are non-negative rules out some estimands on the *gaps* between treatment effects which are in fact identified. For instance, adding period $t = 4$ to the Table 1 example, the difference $\tau_{A4} - \tau_{B4}$ would be identified (by $(Y_{A4} - Y_{B4}) - (Y_{A1} - Y_{B1})$), even though neither τ_{A4} nor τ_{B4} is.

4 Imputation-Based Estimation and Testing

To overcome the challenges of conventional practice we now derive the robust and efficient estimator from first principles, and show that it takes a particularly transparent "imputation" form when no restrictions on treatment-effect heterogeneity are assumed. We then perform asymptotic analysis, establishing the conditions for the estimator to be consistent and asymptotically normal, deriving conservative standard error estimates for it, and discussing appropriate pre-trend tests. We then compare the imputation estimator with other methods recently proposed in response to some problems with OLS estimation. We conclude the section by providing simulation evidence that the efficiency gains from using our estimator are sizable, that its sensitivity to some parallel trend violations is no larger than that of the alternatives, and that our inference tools perform well in finite samples.

Our theoretical results apply in a class of models more general than those that satisfy Assumption 1. We therefore relax it to have:

Assumption 1′ (General model of $Y(0)$)**.** *For all $it \in \Omega$, $Y_{it}(0) = \underbrace{A_{it}'\lambda_i + X_{it}'\delta}_{=Z_{it}'\pi} + \varepsilon_{it}$, where λ_i is a vector of unit-specific nuisance parameters, A_{it} and X_{it} are known non-stochastic vectors, and $\mathbb{E}[\varepsilon_{it}] = 0$.*

The first term in this model of $Y_{it}(0)$ nests unit FEs, but also allows to interact them with some observed covariates, e.g. to include unit-specific trends. The second term nests period FEs but additionally allows any time-varying controls, i.e.[17]

$$Y_{it}(0) = \alpha_i + \beta_t + \tilde{X}_{it}'\tilde{\delta} + \varepsilon_{it}.$$

To streamline notation, we write $A_{it}'\lambda_i + X_{it}'\delta \equiv Z_{it}'\pi$, where all parameters λ_i and δ are collected into a single column vector π, with the corresponding covariates collected in Z_{it}. Below, we will use Z for the matrix with N rows Z_{it}', and Z_1 and Z_0 for its restrictions to observations Ω_1 and Ω_0, respectively.

Like before, we further suppose that a researcher has chosen estimation target τ_w and made Assumption 2. Some model of treatment effects (Assumption 3) is also assumed, although our main focus is on the null model, under which treatment effect heterogeneity is unrestricted.

4.1 Efficient Estimation

For our efficiency result we impose an additional homoskedasticity assumption:[18]

Assumption 4 (Homoskedastic residuals)**.** *Residuals ε_{it} are homoskedastic and mutually uncorrelated across all $it \in \Omega$: $\mathbb{E}[\varepsilon\varepsilon'] = \sigma^2\mathbb{I}_N$.*

Then we have:

Proposition 1 (Efficient estimator)**.** *Suppose Assumptions 1′, 2, 3′ and 4 hold. Then among the linear unbiased estimators of τ_w, the (unique) efficient estimator $\hat{\tau}_w^*$ can be obtained with the following steps:*

[17] Without further restrictions on $X_{it}'\delta$, there is redundancy in this formulation. However, we will impose such restrictions when considering the asymptotic behavior of the estimator below. We note that Assumption 1′ also nests specifications that exclude unit or period FEs.

[18] Our efficiency results are straightforward to relax to any other known form of heteroskedasticity or mutual dependence.

1. *Estimate θ by $\hat{\theta}$ from the linear regression (where we assume that θ is identified)*

$$Y_{it} = A'_{it}\lambda_i + X'_{it}\delta + D_{it}\Gamma'_{it}\theta + \varepsilon_{it}; \tag{4}$$

2. *Estimate the vector of treatment effects τ by $\hat{\tau} = \Gamma\hat{\theta}$;*

3. *Estimate the target τ_w by $\hat{\tau}^*_w = w'_1\hat{\tau}$.*

Further, this estimator $\hat{\tau}^*_w$ is unbiased for τ_w under Assumptions 1′, 2 and 3′ alone, even when residuals are not homoskedastic.

Under assumptions Assumptions 1′, 2 and 3′, regression (4) is correctly specified. Thus, this estimator for θ is unbiased by construction, and efficiency under homoskedasticity of residuals is a direct consequence of the Gauss–Markov theorem. Moreover, OLS yields the most efficient estimator for any linear combination of θ, including $\tau_w = w'_1\Gamma\theta$. While assuming homoskedasticity may be unrealistic in practice, we think of this assumption as a natural benchmark to decide between the many unbiased estimators of τ_w. Proposition 1 assumes that the parameter vector θ is identified in the regression model in (4). This, for instance, rules out estimation of long-run treatment effects (absent strong restrictions on treatment effects), in line with Lemma 4.[19]

In the important special case of unrestricted treatment effect heterogeneity, $\hat{\tau}^*_w$ has a useful "imputation" representation. The idea is to estimate the model of $Y_{it}(0)$ using the untreated observations $it \in \Omega_0$ and extrapolate it to impute $Y_{it}(0)$ for treated observations $it \in \Omega_1$. Then observation-specific causal effect estimates can be averaged appropriately. Perhaps surprisingly, the estimation and imputation steps are identical regardless of the target estimand. Applying any weights to the imputed causal effects yields the efficient estimator for the corresponding estimand. We have:

Corollary 1 (Imputation representation for the efficient estimator)**.** *With a null Assumption 3 (that is, if $\Gamma = \mathbb{I}_{N_1}$), the efficient linear unbiased estimator $\hat{\tau}^*_w$ of τ_w from Proposition 1 can be obtained as an imputation estimator:*

1. *Within the untreated observations only ($it \in \Omega_0$), estimate the λ_i and δ (by $\hat{\lambda}_i, \hat{\delta}$) by OLS in the regression*

$$Y_{it}(0) = A'_{it}\lambda_i + X'_{it}\delta + \varepsilon_{it}; \tag{5}$$

2. *For each treated observation ($it \in \Omega_1$) with $w_{it} \neq 0$, set $\hat{Y}_{it}(0) = A'_{it}\hat{\lambda}_i + X'_{it}\hat{\delta}$ and $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$ to obtain the estimate of τ_{it};*

3. *Estimate the target τ_w by a weighted sum $\hat{\tau}^*_w = \sum_{it\in\Omega_1} w_{it}\hat{\tau}_{it}$.*

The "imputation" structure of the estimator is related to the "direct estimation of the counterfactual" form considered by Gobillon and Magnac (2016) for linear factor models. In the special case of two-way fixed effects as Assumption 1′, the estimator from this corollary yields the "counterfactual" and "two-stage"

[19] We note that for the Proposition 1 result we do not necessarily need to identify all fixed effects in $Y_{it}(0)$ separately; it is sufficient that the fitted value $\mathbb{E}[Y_{it}(0)] = A'_{it}\lambda_i + X'_{it}\delta$ is identified. For example, we could shift all unit FEs up and all period FEs down by the same constant, but this change is immaterial for the fitted values.

estimators proposed in contemporaneous work by Liu et al. (2020) and Gardner (2021) for the specific choices of the weights w_1 they consider.

The imputation representation offers conceptual and computational benefits. First, it is computationally efficient as it only requires estimating a simple TWFE model, for which fast algorithms are available (Guimarães and Portugal, 2010; Correia, 2017). This is in contrast to the OLS estimator from Proposition 1, as regression (4) has regressors $\Gamma_{it} D_{it}$ in addition to the fixed effects, which are high-dimensional unless a low-dimensional model of treatment effect heterogeneity is imposed.

Second, the imputation approach is intuitive and transparently links the parallel trends and no anticipation assumptions to the estimator. Indeed, Imbens and Rubin (2015) write: *"At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others"* (p. 141). We formalize this statement in the next proposition, which shows that *any* estimator unbiased for τ_w can be represented in the imputation way, but other imputation procedures are no longer efficient.[20]

Proposition 2 (Imputation representation for all unbiased estimators)**.** *Under Assumptions 1′ and 2, any unbiased linear estimator $\hat{\tau}_w$ of τ_w that allows for arbitrary treatment-effect heterogeneity (that is, a null Assumption 3) can be written as an imputation estimator:*

1. *For every treated observation, estimate expected potential control outcomes $A'_{it}\lambda_i + X'_{it}\delta$ by some unbiased linear estimator $\hat{Y}_{it}(0)$ using data from the untreated observations only;*
2. *For each treated observation, set $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$;*
3. *Estimate the target by a weighted sum $\hat{\tau}_w = \sum_{it\in\Omega_1} w_{it}\hat{\tau}_{it}$.*

Furthermore, if Assumption 3 is non-trivial, then there exists a set of weights $\tilde{w}_{it}$ such that this representation still holds with $\hat{\tau}_w = \sum_{it\in\Omega_1} \tilde{w}_{it}\hat{\tau}_{it}$ in the third step, where the weights $\tilde{w}_1 = (\tilde{w}_{it})_{it\in\Omega_1}$ satisfy $\Gamma\tilde{w}_1 = \Gamma w_1$.

4.2 Asymptotic Properties

Having derived the linear unbiased estimator $\hat{\tau}_w$ for τ_w in Proposition 1 that is also efficient under homoskedastic errors terms, we now consider its asymptotic properties without imposing homoskedasticity. We study convergence along a sequence of panels indexed by the sample size N, where randomness only stems from the error terms ε_{it}, as in Section 2. Our approach applies to asymptotic sequences where both the number of units and the number of time periods may grow, although our main results are most applicable when the number of time periods remains constant or grows slowly, as in short panels. By viewing the set of observations Ω, treatment timing, and all FEs and controls as non-stochastic, we do not have to impose assumptions on the sampling of the weights w_1 themselves, but can take them as given.

Instead of making Assumption 3, that error terms are homoskedastic and uncorrelated, we now assume that error terms are clustered by units i.

[20] Corollary A2 in the Appendix extends Corollary 1 by showing that the efficient estimator of Proposition 1 has an imputation representation even under a non-trivial Assumption 3. The first two steps are identical but the third step differs. With restrictions on treatment effects, there is a class of weighted sums of $\hat{\tau}_{it}$ which are all unbiased for τ_w. The third step searches for the minimum-variance one.

Assumption 5 (Clustered error terms)**.** *Residuals ε_{it} are uncorrelated across units and have bounded variance:* $\mathrm{Cov}\left[\varepsilon_{it}, \varepsilon_{js}\right] = 0$ *for* $i \neq j$*, and* $\mathrm{Var}\left[\varepsilon_{it}\right] \leq \bar{\sigma}^2$ *uniformly.*

The key role in our results is played by the weights that the Proposition 1 estimator places on each observation. Since the estimator is linear in the observed outcomes Y_{it}, we can write it as $\hat{\tau}_w^* = \sum_{it\in\Omega} v_{it}^* Y_{it}$ with non-stochastic weights v_{it}^*. Corollary A1 in the appendix derives v_{it}^* in the general case where the treatment-effect model Γ may be non-trivial. If treatment effects can vary arbitrarily, the expression simplifies: for all treated observations, $v_{it}^* = w_{it}$, while the weights on the untreated observations are obtained as

$$v_{it}^* = -Z_{it}'(Z_0'Z_0)^{-1}Z_1'w_1, \qquad it \in \Omega_0. \tag{6}$$

These weights will later be helpful to analyze the asymptotic distribution of $\hat{\tau}_w^*$.

We now formulate high-level conditions on the sequence of weight vectors that ensure consistency, asymptotic Normality, and allow us to provide valid inference. These results apply to any unbiased linear estimator $\hat{\tau}_w = \sum_{it\in\Omega} v_{it} Y_{it}$ of τ_w, not just the efficient estimator $\hat{\tau}_w^*$ from Proposition 1 – that is, if the respective conditions are fulfilled for the weights v_{it}, then consistency, asymptotic Normality, and valid inference follow as stated. For the specific estimator $\hat{\tau}_w^*$ introduced above, we then provide sufficient low-level conditions for short panels.

First, we obtain consistency for $\hat{\tau}_w = \sum_{it\in\Omega} v_{it} Y_{it}$ under a Herfindahl condition on the weights v that takes the clustering structure of error terms into account.

Assumption 6 (Herfindahl condition)**.** *Along the asymptotic sequence,*

$$\|v\|_{\mathrm{H}}^2 \equiv \sum_i \left(\sum_t |v_{it}|\right)^2 \to 0.$$

The condition on the clustered Herfindahl index $\|v\|_{\mathrm{H}}^2$ states that the sum of squared weights vanishes, where weights are aggregated by units. One can think of the inverse of the sum of squared weights, $n_H = \|v\|_{\mathrm{H}}^{-2}$, as a measure of effective sample size, which Assumption 6 requires to grow large along the asymptotic sequence. If it is satisfied, and variances are uniformly bounded, then the difference between the estimator $\hat{\tau}_w$ and the estimand τ_w vanishes asymptotically, and we obtain consistency:

Proposition 3 (Consistency of $\hat{\tau}_w$)**.** *Under Assumptions 1′, 2, 3′, 5 and 6,* $\hat{\tau}_w - \tau_w \xrightarrow{\mathcal{L}_2} 0$ *for an unbiased estimator* $\hat{\tau}_w$ *of* τ_w*, such as* $\hat{\tau}_w^*$ *in Proposition 1.*[21]

We next consider the asymptotic distribution of the estimator around the estimand.

Proposition 4 (Asymptotic Normality)**.** *Under the assumptions of Proposition 3, a balance assumption on higher moments of the weights (Assumption A1), and if* $\liminf n_H \sigma_w^2 > 0$ *for* $\sigma_w^2 = \mathrm{Var}\left[\hat{\tau}_w\right]$*, we have*

[21] The Herfindahl condition can be restrictive since it allows for a worst-case correlation of error terms within units. When such correlations are limited, other sufficient conditions may be more appropriate instead, such as $R\left(\sum_{it} v_{it}^2\right) \to 0$ with $R = \max_i \left(\text{largest eigenvalue of } \Sigma_i\right)/\bar{\sigma}^2$, where $\Sigma_i = \left(\mathrm{Cov}\left[\varepsilon_{it}, \varepsilon_{is}\right]\right)_{t,s}$. Here R is a measure of the maximal joint covariation of all observations for one unit. If error terms are uncorrelated, then $R \leq 1$, since the maximal eigenvalue of Σ_i corresponds to the maximal variance of an error term ε_{it} in this case, which is bounded by $\bar{\sigma}^2$. An upper bound for R is the maximal number of periods for which we observe a unit, since the maximal eigenvalue of Σ_i is bounded by the sum of the variances on its diagonal.

that

$$\sigma_w^{-1}(\hat{\tau}_w - \tau_w) \xrightarrow{d} \mathcal{N}(0,1).$$

This result establishes conditions under which the difference between estimator and estimand is asymptotically Normal. Besides regularity, this proposition requires that the estimator variance σ_w^2 does not decline faster than $1/n_H$. It is violated if the clustered Herfindahl formula is too conservative: for instance, if the number of periods is growing along the asymptotic sequence while the within-unit over-time correlation of residuals remains small. Alternative sufficient conditions for asymptotic Normality can be established in such cases, e.g. along the lines of Footnote 21.

So far, we have formulated high-level conditions on the weights v_{it} of any linear unbiased estimator of τ_w. We now present low-level sufficient conditions for consistency and asymptotic Normality of the imputation estimator $\hat{\tau}_w^*$ for the benchmark case of a panel with unit and period fixed effects, a fixed or slowly growing number of periods, and no restrictions on treatment effects. Unlike Propositions 3 and 4, these conditions are imposed directly on the weights w_1 chosen by the researcher, and not on the the implied weights v_{it}^*, and are therefore easy to verify.

Corollary 2 (Consistency and asymptotic Normality in short panels)**.** *Consider the estimator $\hat{\tau}_w^*$ for a panel with I units and T periods in data where no unit is treated in the first period, with two-way fixed effects (Assumption 1), no anticipation effects (Assumption 2), treatment effects that vary arbitrarily (trivial Assumption 3′), and clustered error terms (Assumption 5). Suppose Assumption A4 holds, which requires that all units are observed in the first period, the weights w_1 fulfill a Herfindahl condition, T does not grow too fast, and there is a sufficient number of untreated observations. Then $\hat{\tau}_w^*$ from Proposition 1 is consistent for τ_w. Moreover, under the stronger Assumption A5 that requires that the panel is complete, the number of periods T is fixed, treatment-effect weights w_1 do not vary too much within cohort-period cells, and the size of each cohort increases, $\hat{\tau}_w^*$ is $\sqrt{I}$-consistent and asymptotically Normal.*

The idea behind these sufficient conditions is that sufficiently large groups of untreated observations in each relevant period cell allow for the consistent estimation of period fixed effects. With period FEs estimated well, we can then derive the properties of weights v_{it}^* for all observations from the conditions on w_1, and thus invoke the above high-level conditions.

4.3 Conservative Inference

Our goal next is to estimate the variance of $\hat{\tau}_w = \sum_{it\in\Omega} v_{it} Y_{it}$, which with clustered error terms (Assumption 5) equals to $\sigma_w^2 = \mathbb{E}\left[\sum_i \left(\sum_t v_{it}\varepsilon_{it}\right)^2\right]$. We start with the case where treatment effect heterogeneity is unrestricted (i.e. $\Gamma = \mathbb{I}$). As in Section 4.2, the inference tools we propose apply to a generic linear unbiased estimator but we use them for the efficient estimator $\hat{\tau}_w^*$.

Estimating σ_w^2 presents two challenges. To see them, consider a natural starting point: the plug-in estimator, which replaces error terms ε_{it} with the residuals $\hat{\varepsilon}_{it}$ from the regression model in Equation (4).[22] First, $\hat{\varepsilon}_{it}$ may not well-approximate ε_{it} for the untreated observations. This is because of overfitting, as

[22]Note that these residuals from Equation (4) can be used, with the adjustments described below, even when estimating the variance for estimators other than $\hat{\tau}_w^*$.

the unit-specific coefficient estimates $\hat{\lambda}_i$ may not be consistent for λ_i. For example, in short panels there is only a small number of observations to identify each unit's fixed effect. This can be seen by writing:

$$\hat{\varepsilon}_{it} = \varepsilon_{it} - A'_{it}\left(\hat{\lambda}_i - \lambda_i\right) - X'_{it}\left(\hat{\delta} - \delta\right) \qquad \text{for } it \in \Omega_0,$$

and assuming that the coefficients δ on other covariates (such as period FEs) are estimated well in large samples.

Second, when treatment effects are allowed to vary arbitrarily, they are estimated by fitting the corresponding outcomes Y_{it} perfectly (see Corollary 1). Thus, the residuals from (4) are identically zero for all treated observations.

It turns out that the first challenge does not generate problems when clustered standard errors are used, as Stock and Watson (2008) have shown in a related context. Since the estimator $\hat{\tau}_w$ is unbiased under Assumptions $1'$ and 2, and therefore invariant to the values of the fixed effects or, more generally, the $A'_{it}\lambda_i$ terms, unit-clustered standard errors are invariant, too.

We address the second challenge by showing that σ_w^2 can still be estimated conservatively. This issue is more serious than the first, as it is generally impossible to distinguish between τ_{it} and ε_{it} for the treated observations and thus isolate the variance of ε_{it} only. We make progress by attributing some of the variation in estimated treatment effects across treated observations to the residuals. Specifically, for each $it \in \Omega_1$ we choose some statistic $\hat{\bar{\tau}}_{it}$ — some weighted average of estimated treatment effects over a large group of observations — which in large samples converges to a non-stochastic limit $\bar{\tau}_{it}$. We then replace ε_{it} with $\tilde{\varepsilon}_{it} = \hat{\tau}_{it} - \hat{\bar{\tau}}_{it}$ for the treated observations in estimating σ_w^2. To show how this solves the second challenge, we use $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$ and write:

$$\tilde{\varepsilon}_{it} \approx \varepsilon_{it} + (\tau_{it} - \bar{\tau}_{it}) - X'_{it}\left(\hat{\delta} - \delta\right) - A'_{it}\left(\hat{\lambda}_i - \lambda_i\right), \qquad it \in \Omega_1. \tag{7}$$

In large samples the third term in (7) is small under appropriate regularity conditions. Moreover, the last term, which reflects the first challenge, cancels out when using clustered standard errors. Thus, we approximately recover the sum of the true residual ε_{it} and the nuisance term $\tau_{it} - \bar{\tau}_{it}$, which reflects the heterogeneity of treatment effects in ways not captured by the researcher's choice of $\bar{\tau}_{it}$. Unless $\tau_{it} = \bar{\tau}_{it}$, this makes inference conservative. We have:

Proposition 5 (Conservative clustered standard error estimates)**.** *Assume that the assumptions of Proposition 3 hold, that the model of treatment effects is trivial ($\Gamma = \mathbb{I}$), that the estimates $\hat{\bar{\tau}}_{it}$ converge to some non-random $\bar{\tau}_{it}$ in the sense that $\|v\|_{\mathrm{H}}^{-2}\sum_i\left(\sum_t v_{it}(\hat{\bar{\tau}}_{it} - \bar{\tau}_{it})\right)^2 \xrightarrow{p} 0$, that $\hat{\delta}$ from Proposition 1 is sufficiently close to δ in the sense of an appropriately weighted mean-squared error (Assumption A2), and that additional regularity conditions on the moments of the model parameters and weights hold (Assumption A3). Then the variance estimate*

$$\hat{\sigma}_w^2 = \sum_i \left(\sum_{t;D_{it}=0} v_{it}\hat{\varepsilon}_{it} + \sum_{t;D_{it}=1} v_{it}\tilde{\varepsilon}_{it}\right)^2 \tag{8}$$

is asymptotically conservative: $\|v\|_{\mathrm{H}}^{-2}(\hat{\sigma}_w^2 - \sigma_w^2 - \sigma_\tau^2) \xrightarrow{p} 0$ where $\sigma_\tau^2 = \sum_i\left(\sum_{t;D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it})\right)^2 \geq 0$.

It remains to discuss how to choose the averages $\hat{\bar{\tau}}_{it}$. This choice aims to maintain a balance between

avoiding overly conservative variance estimates and ensuring consistency. A conservative choice is to choose $\hat{\bar{\tau}}_{it}$ to be a broad average of treatment effects, perhaps the same $\hat{\bar{\tau}}$ for all observations. It would then converge to some limit $\bar{\tau}$ in large samples only under mild conditions. However, any deviations of treatment effects τ_{it} from that $\bar{\tau}$ would inflate the variance estimates, by $\sigma_\tau^2 = \sum_i \left(\sum_{t;D=1} v_{it}(\tau_{it} - \bar{\tau})\right)^2$ according to Proposition 5.

A smart choice of $\bar{\tau}$ and $\hat{\bar{\tau}}$ can to some extent mitigate this excessive conservativeness. Specifically, σ_τ^2 is minimized by

$$\bar{\tau} = \frac{\sum_i \left(\sum_{t;D_{it}=1} v_{it}\right)\left(\sum_{t;D_{jt}=1} v_{it}\tau_{it}\right)}{\sum_i \left(\sum_{t;D_{it}=1} v_{it}\right)^2}. \tag{9}$$

We therefore recommend using its feasible analog, which replaces τ_{it} with $\hat{\tau}_{it}$, to construct the standard errors based on just one $\hat{\bar{\tau}}$.[23]

However, if the sample is large enough, a more natural approach is to have $\hat{\bar{\tau}}_{it}$ vary across more narrow — but still large (in the asymptotic sense) — groups of observations. In particular, if the treated cohorts are large enough, we propose to use the analog of (9) restricted to cohort-by-period groups Ω_{et} that partition Ω_1 and within which treatment effect variation may be smaller:

$$\bar{\tau}_{et} = \frac{\sum_{i;E_i=e} v_{it}^2 \tau_{it}}{\sum_{i;E_i=e} v_{it}^2}, \tag{10}$$

with estimated analogs $\hat{\bar{\tau}}_{et}$.[24] As an intermediate solution, analogous averages by horizon (but pooled across cohorts) could also be used, depending on the sample size and the most likely dimension of treatment effect heterogeneity.

We make several final remarks on Proposition 5. First, it extends directly to conservative estimation of variance-covariance matrices for vector-valued estimands, e.g. for average treatment effects at multiple horizons h. Second, the proof of Corollary 2 shows that the low-level conditions which establish asymptotic Normality of the imputation estimator $\hat{\tau}_w^*$ with unrestricted treatment effect heterogeneity also suffice for the conditions of Proposition 5. Finally, Proposition 5 can be extended to the case with a non-trivial treatment-effect model imposed in Assumption 3. By Corollary A2, the general efficient estimator can be represented as an imputation estimator for a modified estimand, i.e. by changing w_1 to some w_1^*. Proposition 5 then yields a conservative variance estimate for it. However, with strong restrictions on treatment effects, asymptotically exact inference may be possible, as the residuals $\hat{\varepsilon}_{it}$ for treated observations in (4) may be estimated consistently (except for the inconsequential noise in $\hat{\lambda}_i$).

The computation of $\hat{\sigma}_w^2$ for the estimator $\hat{\tau}_w^*$ from Proposition 1 involves the implied weights v_{it}^*, which, according to (6) and the more general formula in Corollary A1, require inverting the $Z_0'Z_0$ matrix and is therefore computationally challenging with multiple sets of high-dimensional FEs. In Appendix A.4

[23] Alternatively, one could use $\hat{\bar{\tau}} \equiv \hat{\tau}_w$, which is simpler, as no additional calculation is required, but more conservative. Also note that the denominator of (9) is zero if the estimand makes only within-unit comparisons of treatment effects over time; in that case the choice of $\bar{\tau}$ is inconsequential, as it cancels out in (8).

[24] Unlike $\bar{\tau}$ in (9), $\bar{\tau}_{et}$ does not exactly minimize $\sigma_\tau^2 = \sum_i \left(\sum_{t;D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{E_i t})\right)^2$. However, the solution to the minimization problem for σ_τ^2 in this case may not have a simple representation if non-zero weights are placed on multiple observations of the same unit.

we develop a computationally efficient algorithm for computing v_{it}^* based on the iterative least squares algorithm for conventional regression coefficients (Guimarães and Portugal, 2010).

4.4 Testing for Parallel Trends

We now discuss a principled way to falsify parallel trends with no anticipatory effects or, more generally, Assumptions 1′ and 2. Our proposal departs from traditional tests in that we perform pre-trend testing using the untreated sample only. This separation is preferable conceptually, as it prevents the conflation of using an identifying assumption and validating it — a practice which is common in DiD but not in other designs. Additionally, we show that under homoskedasticity our approach solves the problem explained by Roth (2018): while conditioning on a conventional pre-trend test passing makes inference invalid, this does not happen asymptotically with our test and a class of estimators that impose parallel trends, including the imputation estimator.

We start by reviewing two existing strategies to test parallel trends. Traditionally, researchers estimated a dynamic specification which includes both lags and leads of treatment, and test — visually or statistically — that the coefficients on leads are equal to zero. More recent papers (e.g. De Chaisemartin and D'Haultfœuille, 2020) replace it with a placebo strategy: pretend that treatment happened k periods earlier for all eventually treated units, and estimate the average effects $h = 0, \dots, k-1$ periods after the placebo treatment using the same estimator as for actual estimation. The imputation estimator could be used with that approach.

Both of these strategies have drawbacks. Because the traditional regression-based test uses the full sample, including treated observations, it is *not* a test for Assumptions 1′ and 2 only. Rather, it is a joint test that is also sensitive to violations of the implicit Assumption 3: that treatment effects τ_{it} for $it \in \Omega_1$ are homogeneous within each horizon (Sun and Abraham, 2021). Even if the researcher has reasons to impose a non-trivial Assumption 3, a test for parallel trends specifically should not use those restrictions on the treatment effect heterogeneity. With a null Assumption 3, treated observations are not useful for testing, and our test only uses the untreated ones, for which $Y_{it}(0)$ is directly observed under Assumption 2.

Tests based on placebo estimates appropriately use untreated observations only, but they suffer from a conceptual problem that may further result in a power loss. The problem arises because the placebo strategy does not make a conceptual distinction between estimation and testing, and specifies a precise placebo estimation target — typically the placebo ATT. However, with the standard logic of statistical assumption testing, the choice of the test should be based on a guess about the most plausible alternative, while we see little intrinsic value in having the test mimic the estimator. Insisting on a particular estimand makes the estimator more noisy when the placebo effect heterogeneity is not in fact present, as in the neighborhood around the null.

We therefore propose the following test for Assumption 1:

Test 1.

1. *Choose an alternative model for $Y_{it}(0)$ that is richer than Assumption 1:*

$$Y_{it}(0) = A_{it}'\lambda_i + X_{it}'\delta + W_{it}'\gamma + \tilde{\varepsilon}_{it}; \tag{11}$$

2. *Estimate γ by $\hat{\gamma}$ using OLS on untreated observations $it \in \Omega_0$ only;*

3. *Test $\gamma = 0$ using the* F*-test (or visually).*

A natural choice for W_{it}, which inherits from prior conventional pre-trend tests, is a set of indicators for observations $1, \ldots, k$ periods before treatment, with periods before $E_i - k$ serving as the reference group. This choice is appropriate, for instance, if the researcher's main worry is the possible effects of treatment anticipation, i.e. violations of Assumption 2.[25] When Equation (11) is correctly specified and $\tilde{\varepsilon}_{it}$ are homoskedastic, this test is asymptotically most powerful among unbiased tests.[26]

We now show an additional advantage of the proposed test: if the researcher conditions on the test passing (i.e. does not report the results otherwise), inference on $\hat{\tau}^*_w$ is still asymptotically valid under the null of no violations of Assumptions 1′ and 2 and under homoskedasticity. This is in contrast to the results by Roth (2018) who find distorted inference in the context of dynamic event study regressions.

Proposition 6. *Suppose Assumptions 1, 2 and 4 hold. Consider $\hat{\tau}^*_w$ constructed as in Proposition 1 for some estimation target and with some Γ. Then $\hat{\tau}^*_w$ is uncorrelated with any vector $\hat{\gamma}$ constructed as in Test 1. Consequently, if the vector $(\hat{\tau}^*_w, \hat{\gamma})$ is asymptotically Normal, $\hat{\tau}^*_w$ and $\hat{\gamma}$ are asymptotically independent, and conditioning on $\hat{\gamma} \notin R_\gamma$ for a non-stochastic rejection region R_γ does not asymptotically affect the distribution of $\hat{\tau}^*_w$.*[27]

The key ingredients to this proposition are that estimation imposes Assumptions 1′ and 2, while testing uses untreated observations only. The formal logic behind it is similar to that of Hausman tests: under Assumptions 1, 2 and 4 and additionally imposing appropriate Assumption 3, $\hat{\tau}^*_w$ is efficient for τ_w. Under the same assumptions, $\hat{\gamma}$ is unbiased for zero and thus should be uncorrelated with $\hat{\tau}^*_w$, or else $\hat{\tau}^*_w + \zeta'\hat{\gamma}$ would be more efficient for some ζ.

While we use Proposition 6 for the efficient estimators from Proposition 1, we note that it applies well beyond. This is because it does not require that Assumption 3 used to construct $\hat{\tau}^*_w$ (via Γ) actually holds. Thus, the result applies, for instance, to the static TWFE estimator which is a special case of Proposition 1 with all treatment effects assumed (perhaps incorrectly) to be the same. Similarly, it applies to dynamic OLS specifications that include only lags of treatment and no leads.[28]

In contrast, Proposition 6 fails when estimation and testing are done simultaneously. This point is most simply illustrated in conventional DiD designs where treatment happens at the same time in the treatment group and never in the control group. Then the period right before treatment is used as the sole

[25] Alternatively, the researcher may focus on possible violations of the parametric structure in Assumption 1′. For instance, with data spanning many years one could test for the presence of a structural break, e.g. that α_i are the same before and after the Great Recession (up to a constant shift captured by the period FEs).

[26] The optimal choice of k is a challenging question that we leave to future research. As usual with F-tests, choosing a k that is too large can lead to low power against many alternatives, in particular those that generate large biases in treatment effect estimates that impose invalid Assumption 1.

[27] For estimators that do not satisfy this property, Roth (2018) shows how to construct an adjustment that removes the dependence, provided the covariance matrix between $\hat{\tau}^*$ and $\hat{\gamma}$ can be estimated. By Proposition 6, this adjustment is not needed for the Proposition 1 estimator under homoskedasticity.

[28] Furthermore, Proposition 6 also holds when testing is done with the placebo imputation estimator, rather than by OLS as we suggested above. This is because placebo treatment effects γ_{it} can be recovered from (11), by stacking the dummies for all individual observations up to k periods before treatment into W_{it}. Thus, the placebo test is a special case of Test 1, except with $\hat{\gamma}_{it}$ averaged for each horizon before applying the F-test.

reference period for both estimation and testing with a fully-dynamic regression, creating a correlation between them.

4.5 Comparison to Other Estimators

Since the problems of OLS event study regressions have been investigated by Borusyak and Jaravel (2017), Goodman-Bacon (2018), Sun and Abraham (2021) and others, several alternative approaches robust to treatment effect heterogeneity have been developed. We now analyze these proposals through the lens of our framework and contrast our efficient imputation-based estimator with strategies based on manual aggregation of admissible DiD contrasts and "stacked" regressions, focusing on the case where treatment-effect heterogeneity is not constrained ($\Gamma = \mathbb{I}$). We also connect our imputation strategy to a related idea proposed for factor models.

We can employ our framework to compare proposed estimators in terms of their estimands and the (implicit) ways they impute counterfactual outcomes. In Proposition 2 above, we showed that any linear estimator that is unbiased for some weighted estimand without making assumptions about treatment-effect heterogeneity can be represented in an imputation form. This includes estimators considered in De Chaisemartin and D'Haultfœuille (2020), Sun and Abraham (2021) and Callaway and Sant'Anna (2021), who all propose explicit formulas that aggregate admissible DiD contrasts in varying ways. All of them start by estimating "cohort-average treatment effects" $CATT_{e,t}$: the average treatment effects for all units first treated in some period e and observed in period $t \geq e$; those estimates are then aggregated across cohorts and periods. Hence, these estimators put equal weight on treatment effects within cohort–period cells, while we accommodate arbitrary weights within these cells. To estimate $CATT_{e,t}$, they use DiD contrasts for the cohort e in period t against some reference group (e.g. never-treated or not-yet-treated by t) and period $e-1$, which directly precedes treatment.[29]

We can now compare these different strategies by how efficiently they impute counterfactual outcomes. The key advantage of the imputation estimator we characterize above is that it imputes the potential untreated outcomes $Y_{it}(0)$ for treated observations $it \in \Omega_1$ from the full set of untreated observations, yielding the efficiency properties formalized by Corollary 1. Alternative estimators rely on *some* estimates of $Y_{it}(0)$, but those are not the efficient choices when parallel trends are assumed and there are no anticipation effects. In particular, these specific estimators only use data from period $e-1$ for $CATT_{e,t}$, while the imputation estimator leverages all pre-periods. Although the imputation estimation is only most efficient under homoskedasticity, the use of all available data should make it more efficient with other heteroskedasticity and mutual correlation structures, as we find in the simulation in Section 4.6.

Some imputation strategies that differ from the efficient choice in Corollary 1 may have the advantage of imposing weaker assumptions. However, the use of limited information by estimators that obtain $CATT_{e,t}$ with $e-1$ as the only reference period cannot generally by justified by weaker parallel trends or no anticipation effect assumptions. We therefore do not see a clear robustness-efficiency tradeoff between estimators. To illustrate this point, we extend the Table 1 example by adding unit C treated in period 4 (see Table 2). The alternative estimators would use $Y_{B2} - Y_{B1}$ and $Y_{C2} - Y_{C1}$ as contrasts for $Y_{A2} - Y_{A1}$ (when estimating τ_{A2}) and similarly $Y_{C3} - Y_{C2}$ for $Y_{B3} - Y_{B2}$ (when estimating τ_{B3}), ruling out any

[29] Fadlon and Nielsen (2015) propose a similar approach based on matching treated and not-yet-treated units.

Table 2: A Three-Unit, Three-Period Example

$\mathbb{E}[Y_{it}]$	$i = A$	$i = B$	$i = C$
$t = 1$	α_A	α_B	α_C
$t = 2$	$\alpha_A + \beta_2 + \tau_{A2}$	$\alpha_B + \beta_2$	$\alpha_C + \beta_2$
$t = 3$	$\alpha_A + \beta_3 + \tau_{A3}$	$\alpha_B + \beta_3 + \tau_{B3}$	$\alpha_C + \beta_3$
Event date	$E_i = 2$	$E_i = 3$	$E_i = 4$

anticipation effects for units B and C. However, they would stop short of also using $Y_{C3} - Y_{C1}$ as a contrast to $Y_{B3} - Y_{B1}$ to estimate τ_{B3}.

The imputation estimator offers four additional benefits. First, by transparently mapping the model of $Y_{it}(0)$ into the estimator, it is immediately extended to models richer than TWFE, such as with unit-specific trends or in triple-difference strategies, with the same efficiency properties (see Section 5). Second, by separating testing of parallel trends from estimation, it solves the pre-testing problem of Roth (2018), as we have shown in Section 4.4. This is in contrast to estimators which use the same reference period, from $E_i - 1$ to E_i, for both estimation and testing, generating a correlation between the resulting coefficients. Third, since our framework does not require random sampling, we allow for a more general class of estimands. Specifically, any weighted sum of treated observations is allowed for, while other frameworks only work with cohort-average treatment effects. Finally, we provide analytical formulas for valid standard errors with a computationally efficient procedure to implement them. This complements De Chaisemartin and D'Haultfœuille (2020) and Callaway and Sant'Anna (2021) who use the bootstrap, while Sun and Abraham (2021) obtain analytical standard errors from an "interaction-weighted" regression which may be computationally challenging with many cohorts and periods.[30]

Another approach to event study estimation is the "stacked regression" method developed in a more applied work by Cengiz et al. (2019, Appendix D). For each "event" in which some unit j has been treated, they construct a separate dataset that includes a window around the event time and contains unit j along with all units not yet treated by the end of the window. Then all datasets are stacked to estimate the OLS regression on pooled data:

$$Y_{it}^{(j)} = \alpha_i^{(j)} + \beta_t^{(j)} + \sum_{h \neq -1} \tau_h \mathbf{1}\left[K_{it} = h\right] + \varepsilon_{it}^{(j)}. \tag{12}$$

Here $Y_{it}^{(j)}$ is the outcome of observation it in the dataset for event j, and the τ_h are supposed to measure the average effects (or pre-trend coefficients) h periods after treatment. This approach has two limitations in addition to those arising with the alternative estimators we discussed above. First, its estimand is not obvious, i.e. we cannot easily read off the weights used by OLS to pool the different events.[31] Second,

[30] Our standard errors are only conservative, while the analytical standard errors of Sun and Abraham (2021) are asymptotically exact. However, this difference arises because we consider a fixed sample. In the case where the estimand is based on cohort average treatment effects (with $\bar{\tau}_{it}$ similarly specified) and deviations from these averages are seen as random (as would be the case for random sampling), then our standard errors are exact, too.

[31] One can hope that, in the absence of forbidden comparisons in each individual dataset, these weights will be positive. However, we are not aware of a proof, and the standard intuition of Angrist (1998) need not apply as Equation (12) is not saturated.

this estimator uses limited information, for two reasons: the data are restricted to a window around E_j, and units treated later than j but before the end of the window are excluded.

Our imputation approach is instead closely related to the "direct estimation of the counterfactual" approach proposed by Gobillon and Magnac (2016) in a different context, for linear factor models. Like in Corollary 1, they estimate the model of $Y_{it}(0)$ — albeit a different model — on untreated observations only and extrapolate it to the treated observations. Xu (2017) notes the applicability of this approach in DiD settings too, which is developed in a follow-up paper (Liu et al., 2020) that is independent from our work; Gardner (2021) similarly proposes a "two-stage DiD estimator" based on a specification with cohort and period FEs. Both estimators coincide with the imputation estimator for the estimands they consider: the overall average treatment effect and the average effect a given number of periods after treatment . Compared with these studies, we *derive* the imputation estimator as the most efficient one for a class of problems, allow for restrictions on treatment effects, for a general class of target estimands, and consider a series of extensions. Moreover, we develop asymptotic theory for the estimator and provide analytical standard errors for the standard case when fixed effects are on the individual level.[32]

4.6 Monte-Carlo Simulations

We now quantify the efficiency properties of the imputation estimator in a simulated datasets, both under homoskedastic and serially uncorrelated residuals and without those assumptions, in comparison to the alternative robust estimators of De Chaisemartin and D'Haultfœuille (2020) and Sun and Abraham (2021). We also verify correct coverage of our inference procedure and check sensitivity of different estimators to anticipation effects.

In our baseline simulation we consider a complete panel of $I = 250$ units observed for $T = 6$ periods. The event happens for each unit in periods 2–7 with equal probabilities; units with $E_i = 7$ are therefore never treated in the observed sample. Treatment effects depend on the horizon, as $\tau_{it} = K_{it}+1$ for $it \in \Omega_1$, but are otherwise homogeneous. We impose Assumptions 1 and 2 and set the FEs to $\alpha_i = -E_i$ and $\beta_t = 3t$. Finally, in each of the 500 simulations we generate homoskedastic and mutually independent residuals, $\varepsilon_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. In line with Section 2, we generate the E_i realizations only once, viewing them (along with α_i and β_t) as non-stochastic. Our target estimands are the ATTs τ_h for each horizon $h = 0, \dots, 4$ observed in the data; while τ_0 is an average of the short-run effects on 205 units, τ_4 corresponds to 41 unit only (those treated at $E_i = 2$ and observed at $t = 6$).

Besides the imputation estimator of Corollary 1, we consider the De Chaisemartin and D'Haultfœuille (2020) estimator (denoted DCDH) which uses all non-yet-treated units as the reference group but only $t = E_i - 1$ as the reference period, as well the Sun and Abraham (2021) estimator (denoted SA) which further restricts the reference group to the latest-treated cohort $E_i = 7$. The Callaway and Sant'Anna (2021) estimator is equivalent to SA in our setting with no additional covariates. Importantly, the estimands are exactly the same for all three estimators we consider. We implement the imputation estimator via our Stata command `did_imputation`, and the alternative estimators by the commands provided by the authors:

[32] Gardner (2021) derives large-sample theory by interpreting the estimator as a Generalized Method of Moments estimator with cohort fixed effects, while we consider unit fixed effects, which unlike cohort fixed effects cannot be estimated consistently in the case of short panels.

Table 3: Efficiency and Bias of Alternative Estimators

Horizon	Estimator	Baseline simulation		More pre-periods	Heterosk. residuals	AR(1) residuals	Anticipation effects
		Variance (1)	Coverage (2)	Variance (3)	Variance (4)	Variance (5)	Bias (6)
$h = 0$	Imputation	0.0099	0.942	0.0080	0.0347	0.0072	-0.0569
	DCDH	0.0140	0.938	0.0140	0.0526	0.0070	-0.0915
	SA	0.0115	0.938	0.0115	0.0404	0.0066	-0.0753
$h = 1$	Imputation	0.0145	0.936	0.0111	0.0532	0.0143	-0.0719
	DCDH	0.0185	0.948	0.0185	0.0703	0.0151	-0.0972
	SA	0.0177	0.948	0.0177	0.0643	0.0165	-0.0812
$h = 2$	Imputation	0.0222	0.956	0.0161	0.0813	0.0240	-0.0886
	DCDH	0.0262	0.958	0.0262	0.0952	0.0257	-0.1020
	SA	0.0317	0.950	0.0317	0.1108	0.0341	-0.0850
$h = 3$	Imputation	0.0366	0.928	0.0255	0.1379	0.0394	-0.1101
	DCDH	0.0422	0.930	0.0422	0.1488	0.0446	-0.1087
	SA	0.0479	0.952	0.0479	0.1659	0.0543	-0.0932
$h = 4$	Imputation	0.0800	0.942	0.0546	0.3197	0.0773	-0.1487
	DCDH	0.0932	0.950	0.0932	0.3263	0.0903	-0.1265
	SA	0.0932	0.954	0.0932	0.3263	0.0903	-0.1265

Notes: See Section 4.6 for a detailed description of the data-generating processes and reported statistics.

`did_multiplegt` and `eventstudyinteract`, respectively. We also compute the weights v_{it} underlying these estimators to calculate exact properties of the estimators, such as their finite-sample variance. For inference on the imputation estimator, we use the results from Section 4.3 with $\bar{\tau}_{it}$ defined as simple averages of treatment effects by cohort-period cells, as in (10). In the absence of treatment effect heterogeneity within these cells, inference is exact rather than conservative. Standard errors for the DCDH estimator are based on bootstrap with 100 replications, while SA standard errors are based on large-cohort asymptotic results, as described in the respective papers.

Column 1 of Table 3 reports the exact variance of each estimator, for each horizon-specific estimand. In line with Proposition 1, the imputation estimator is most efficient in all cases, but the simulation is useful in quantifying the magnitude of the efficiency gain. Under homoskedasticity, the variances of the DCDH and SA estimators are 15–41% higher than the variance of the imputation estimator, i.e. a 15–41% larger sample would be needed to obtain confidence intervals of a similar length if these estimators are used. Relative to the DCDH estimator, the efficiency gains tend to be stronger at shorter horizons, while they are non-monotonic relative to SA.

Column 2 shows that the inference procedures accompanying each of the three estimators perform well. Specifically, we report the simulated coverage: the fraction of the 500 simulations in which a t-test

does not reject the null of $\tau_h = h + 1$ at the 5% significance level. Coverage close to 95% is found in all simulations of Table 3, and therefore is not reported later.

Since the imputation estimator uses all pre-periods as reference periods, while the alternative estimators only use the period directly preceding treatment, it is not surprising that the efficiency gains are even higher when we add four more pre-periods, $t = -3, \ldots, 0$, in Column 3. In this simulation variances of the DCDH and SA estimators are 44–97% higher.

In Columns 4 and 5 of Table 3, we report estimator variances under deviations from Assumption 4, such that the relative efficiency of the imputation estimator is no longer guaranteed by Proposition 1. In Column 4 we make the residuals heteroskedastic (while still mutually independent): $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_{it}^2)$ for $\sigma_{it}^2 = t$, such that the variance is higher in later periods. In Column 5 we instead suppose that ε_{it} follow a stationary AR(1) process with $\text{Var}[\varepsilon_{it}] = 1$ and $\text{Cov}[\varepsilon_{it}, \varepsilon_{it'}] = 0.5^{|t-t'|}$, with ε_{it} still Normally distributed and independent across units. The imputation estimator remains the most efficient of the three, with variances of DCDH and SA higher by 2–51% in Column 4 and 5–42% in Column 5. The only exception is the estimator for $h = 0$ in Column 5, where the alternative estimators are 3–8% more efficient.

Finally, in Column 6 we consider the sensitivity of the three estimators to violations of Assumptions 1 and 2. Specifically, we add an anticipation effect of $1/\sqrt{I} = 0.0632$ to the outcomes of each unit corresponding to the period right before treatment, $t = E_i - 1$, and report the exact bias of each estimator.[33] While the DCDH estimator is always more sensitive to anticipation effects than SA (except $h = 4$ where the two estimators coincide), we find no clear relationship between the imputation estimator and its alternatives. The imputation estimator is more sensitive for longer horizons $h = 3, 4$ but more robust for $h = 0, 1, 2$.

Taken together, these results suggest that the imputation estimator has sizable efficiency advantages that extend even to heteroskedasticity and serial correlation of residuals, which do not come at a cost of higher sensitivity to some parallel trend violations. Moreover, our analytical inference tools perform well in finite samples. Naturally, these results may be specific to the data-generating processes we considered, and we recommend that researchers perform similar simulations based on their data.

5 Extensions

Our benchmark setting assumed panel data and treatments that happen at different times but stay on forever. We now show that our results extend naturally to a range of related settings used in applied economics, deviating from these benchmark conditions in various ways. We consider three alternative data structures: repeated cross-sections, datasets with two-dimensional cross-sectional variation in one period, and triple-difference designs. We also comment on how our assumptions and results translate to generic datasets which need not have a panel structure. On treatment timing, we discuss plain vanilla DiD designs without staggered timing, scenarios where treatment switches on and off, and settings where the same unit goes through multiple events. We focus on robust and efficient estimation throughout the section but also point out how OLS-based procedures continue to be problematic.

[33] The $\sqrt{I}$ normalization makes the bias comparable in magnitude to standard errors.

5.1 Deviations from the Panel Data Structure

Repeated Cross-Sections. Suppose in each period data are available for different random samples of units i (e.g., individuals) belonging to the same set of groups $g(i)$ (e.g., regions), and treatment timing varies at the group level. In that case no estimation with units FEs is possible, as only one observation is available for each unit. However, DiD strategies are still applicable to repeated cross sections, with group FEs replacing unit FEs. In our framework this means replacing Assumption 1 with $Y_{it}(0) = \alpha_{g(i)} + \beta_t + \varepsilon_{it}$. Proposition 1 then directly extends to this setting.

Two-Dimensional Cross-Sections. DiD designs are also used when the outcome is measured in a single period but across two cross-sectional dimensions, such as regions i and age groups g. To fix ideas, suppose some policy is implemented in a set of regions always for the older groups, such that the treatment indicator is defined as $D_{ig} = \mathbf{1}[g \geq E_i]$, and the age cutoff E_i varies across regions. With the untreated potential outcomes modeled in a TWFE way, as $Y_{ig} = \alpha_i + \beta_g + \varepsilon_{ig}$, the setting is isomorphic to our benchmark. Thus, TWFE OLS may suffer from negative weighting, in particular for the oldest groups in the regions with lower cutoffs. Moreover, the imputation estimator is robust to heterogeneous effects and efficient under homoskedasticity of ε_{it}.

Triple-Differences. In triple-difference designs, the data have three dimensions, such as regions i, demographic groups g, and periods t. Conventional static OLS estimation is based on the following regression:

$$Y_{igt} = \alpha_{ig} + \alpha_{it} + \alpha_{gt} + \tau D_{igt} + \varepsilon_{igt}.$$

For the same reasons as in Section 3.2, the OLS estimand for τ may not properly average heterogeneous effects τ_{igt} when different regions and groups are treated at different times. The imputation estimator based on the model $Y_{igt}(0) = \alpha_{ig} + \alpha_{it} + \alpha_{gt} + \varepsilon_{igt}$ is robust and efficient under homoskedasticity.

Generic Data. Ultimately, Proposition 1 and Corollary 1 apply to generic datasets with observations indexed by j (which may or may include a time dimension), if one assumes the analogs of Assumptions $1'$, 2 and $3'$:

- a model of untreated potential outcomes: $Y_j(0) = Z_j'\pi + \varepsilon_j$ for some covariates Z_i (perhaps including one or more sets of group FEs) and $\mathbb{E}[\varepsilon_j] = 0$;
- that the observed outcome $Y_j = Y_j(0)$ if the treatment indicator D_j is zero;
- and a (possibly trivial) model of treatment effects $\tau_j \equiv Y_j - Y_j(0) = \Gamma_j\theta$.

The asymptotic results generalize similarly.

5.2 Deviations from Staggered Treatment Timing

Simultaneous Treatment. Consider plain vanilla DiD designs in which treatment happens at a single date (in the treatment group) or never (in the control group). In this case there are no forbidden

DiD contrasts, in which a unit switches its treatment status in a period when another unit has already been treated. Thus, only admissible contrasts are available, and OLS estimation does not suffer from negative weights. The presence of a never-treatment group also prevents the underidentification problem.

These nice properties of conventional DiD designs breaks if more covariates are included, in particular in presence of unit-specific trends or at least a treatment group-specific time trend. Static OLS estimates unit-specific trends by using the data both pre- and post-treatment, and therefore contaminates them with the dynamics of treatment effects. Similarly, the underidentification problem reappears, as the fully-dynamic specification cannot distinguish between a linear path in treatment effects and a combination of a time trend and a treatment group-specific effect.

In contrast, the imputation estimator continues to apply, providing robust estimates which are efficient under homoskedasticity.

Treatment Switching On and Off. In some applications, the treatment of interest may switch on and off for the same unit over time. Our results extend directly if it is appropriate to write $Y_{it} = \alpha_i + \beta_t + \tau_{it} D_{it} + \varepsilon_{it}$ and therefore extrapolate the TWFE outcome structure from the untreated to the treated periods, regardless of how they are ordered relative to each other.

This characterization of Y_{it}, however, here requires an additional assumption that there are no within-unit spillovers from the periods of treatment to the future untreated periods. If the lags of treatment affect current outcomes, the observed outcomes may not equal the never-treated potential outcome $Y_{it}(0)$ even when $D_{it} = 0$ and there are no anticipation effects. All periods since the first treatment date may thus be contaminated by treatment. The imputation approach is therefore applicable when treatment effects are heterogeneous across units and periods but not dynamic, i.e. the potential outcome today depends only on the treatment status today.[34]

Multiple Events per Unit. A related scenario arises when units experience more than one event of interest. For instance, Adda (2016) considers the effects of school holidays on the epidemics. He leverages school holidays that happen in a staggered manner across regions, but his data span several years and therefore several holidays. If each holiday is viewed as a separate event that may have its own effects which potentially last forever and may change over time, causal identification is clearly impossible: one cannot distinguish between the effects of all past holidays.

However, with the events sufficiently spaced out in time, natural restrictions may be introduced via Assumption 3. For instance, one may be willing to assume that holidays can have no effects more than a few weeks after, which is still well before the next holiday. Alternatively, one may assume that the effects stabilize for each unit after that period of time, even if not at zero. Such assumptions imply different "true" models of outcomes Y_{it} which have a lot of flexibility, yet enough structure for identification. Analogs of Proposition 1 then apply.

[34] Another assumption that becomes less natural when the treatment may switch on and off is the asymmetry between $Y_{it}(0)$ and $Y_{it}(1) = Y_{it}(0) + \tau_{it}$ — the asymmetry that motivates us to impose parametric assumptions on $Y_{it}(0)$ while keeping treatment effects and therefore $Y_{it}(1)$ unrestricted.

6 Application

[Will be added in later drafts of the paper.]

7 Conclusion

In this paper we revisited a popular class of empirical designs: event studies, or difference-in-differences with staggered rollout. We provided a unified framework that can serve as acomprehensive guide to these designs for applied economists. Our simple framework aims to make explicit the goals and assumptions that previously were often driven by implicit specification choices and to derive an efficient estimator. Focussing on robustness to treatment effect heterogeneity, we showed that conventional OLS methods suffer from identification and negative weighting issues that we related to the conflation of estimation goals, fundamental assumptions on no-treatment potential outcomes and anticipation effects, and auxiliary assumptions on treatment effects. We then solved for the efficient estimator within our framework. In a baseline case where treatment-effect heterogeneity remains unrestricted, this robust and efficient estimator takes a particularly simple "imputation" form that estimates fixed effects among the untreated observations only, imputes untreated outcomes for treated observations, and then forms treatment-effect estimates as weighted averages over the differences between actual and imputed outcomes. We then developed results for asymptotic inference and testing, discussed how our approach naturally extends to a number of related settings, and compared our approach to other estimators. We also highlighted the importance of separating testing of identification assumptions from estimation based on them in order to both increase estimation efficiency and mitigate issues with pre-testing.

References

Abadie, Alberto, Guido W. Imbens, and Fanyin Zheng, "Inference for Misspecified Models With Fixed Regressors," *Journal of the American Statistical Association*, 2014, *109* (508), 1601–1614.

Adda, Jerome, "Economic Activity and the Spread of Viral Diseases: Evidence from High Frequency Data," *Quarterly Journal of Economics*, 2016, *131* (2), 891–941.

Angrist, Joshua, "Estimating the labor market impact of voluntary military service using social security data on military applicants," *Econometrica*, 1998, *66* (2), 249–288.

Arkhangelsky, Dmitry and Guido W. Imbens, "Double-Robust Identification for Causal Panel Data Models," 2019.

Athey, Susan and Guido W. Imbens, "Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption," *Working Paper*, 2018.

Baker, Andrew C., David F. Larcker, and Charles C. Y. Wang, "How Much Should We Trust Staggered Difference-In-Differences Estimates?," *Working Paper*, 2021.

Borusyak, Kirill and Xavier Jaravel, "Revisiting Event Study Designs," *Working Paper*, 2017.

Callaway, Brantly and Pedro H.C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment," *Journal of Econometrics*, 2021.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zippere, "The Effect of Minimum Wages on Low-Wage Jobs," *The Quarterly Journal of Economics*, 2019, pp. 1405–1454.

Correia, Sergio, "Linear Models with High-dimensional Fixed Effects: An Efficient and Feasible Estimator," *Working Paper*, 2017, (March).

De Chaisemartin, Clément and Xavier D'Haultfœuille, "Fuzzy Differences-in-Differences," *arXiv preprint*, 2015.

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

Fadlon, Itzik and Torben Heien Nielsen, "Household Responses to Severe Health Shocks and the Design of Social Insurance," *Working Paper*, 2015.

Gardner, John, "Two-stage differences in differences," *Working Paper*, 2021.

Gobillon, Laurent and Thierry Magnac, "Regional policy evaluation: Interactive fixed effects and synthetic controls," *Review of Economics and Statistics*, 2016, *98* (3), 535–551.

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," *Working Paper*, 2018.

Guimarães, Paulo and Pedro Portugal, "A simple feasible procedure to fit models with high-dimensional fixed effects," *Stata Journal*, 2010, *10* (4), 628–649.

Imbens, Guido W. and Donald B Rubin, *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press, 2015.

Liu, Licheng, Ye Wang, and Yiqing Xu, "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data," *SSRN Electronic Journal*, 2020.

MacKinlay, A. Craig, "Even Studies in Economics and Finance," *Journal of Economic Literature*, 1997, *XXXV*, 13–39.

Meer, Jonathan and Jeremy West, "FEffects of the Minimum Wage on Employment Dynamics," *Journal of Human Resources*, 2016, *51* (2), 500–522.

Rambachan, Ashesh and Jonathan Roth, "An Honest Approach to Parallel Trends," *Working Paper*, 2020.

Roth, Jonathan, "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," *Working Paper*, 2018.

_ **and Pedro H. C. Sant'Anna**, "Efficient Estimation for Staggered Rollout Designs," *Working Paper*, 2021.

_ **and Pedro H.C. Sant'Anna**, "When Is Parallel Trends Sensitive to Functional Form?," *Working Paper*, 2020.

Schmidheiny, Kurt and Sebastian Siegloch, "On Event Study Designs and Distributed-Lag Models: Equivalence , Generalization and Practical Implications," *Working Paper*, 2020.

Stock, James H. and Mark W. Watson, "Heteroskedasticity-robust standard errors for fixed effects panel data regression," *Econometrica*, 2008, *76* (1), 155–174.

Strezhnev, Anton, "Semiparametric weighting estimators for multi-period difference-in-differences designs," *Working Paper*, 2018.

Sun, Liyang and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2021.

Wolfers, Justin, "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results," *American Economic Review*, 2006, pp. 1802–1820.

Xu, Yiqing, "Generalized synthetic control method: Causal inference with interactive fixed effects models," *Political Analysis*, 2017, *25* (1), 57–76.

A Details and Additional Results

A.1 General Weight and Imputation Representation for the Efficient Estimator

In this section, we provide additional details on the efficient estimator from Proposition 1. We first provide an explicit expression for the weights implied by the estimator:

Corollary A1 (Weight representation of OLS)**.** *The OLS estimator from Proposition 1 can be represented as* $\hat{\tau}_w^* = v'^* Y = v_1'^* Y_1 + v_0'^* Y_0$ *with the weight vector*

$$v^* = \begin{pmatrix} \mathbb{I} - Z_1(Z'Z)^{-1}Z_1' \\ -Z_0(Z'Z)^{-1}Z_1' \end{pmatrix} \Gamma(\Gamma'(\mathbb{I} - Z_1(Z'Z)^{-1}Z_1')\Gamma)^{-1}\Gamma' w_1$$

that does not depend on the realization of the Y_{it}*. Specifically, when* $\Gamma = \mathbb{I}_M$ *, then* $v_1^* = w_1$ *and*

$$v_0^* = -Z_0(Z_0'Z_0)^{-1}Z_1' w_1.$$

We next extend the imputation representation from Corollary 1 to non-trivial treatment-effect models:

Corollary A2 (Imputation representation, treatment-effect model)**.** *With a non-trivial model* $\tau = \Gamma\theta$ *for the treatment effects, the estimator from Proposition 1 can be written as an imputation estimator with alternative weights in three steps:*

1. *Within the untreated observations only ($it \in \Omega_0$), estimate the λ, β (by $\hat{\lambda}, \hat{\beta}$) by OLS in the regression*

$$Y_{it} = A'_{it}\lambda_i + X'_{it}\beta + \varepsilon_{it}.$$

2. *For each observation in the treatment group ($it \in \Omega_1$), set $\hat{Y}_{it}(0) = A'_{it}\hat{\lambda}_i + X'_{it}\hat{\beta}$ and $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$ to obtain the estimate $\hat{\tau}$ of τ.*

3. *Compute the weighted sum $\hat{\tau}_w = w_1^{*\prime}\hat{\tau}$,*

where the weights w_1^* are obtained in one of the following equivalent ways:

a. *If $W_1^* = \frac{\partial}{\partial Y_1'}\hat{\theta}$ is the $(N_1 - M) \times N_1$ matrix of OLS weights on the treated units $Y_{it}, it \in \Omega_1$ in the OLS estimator $\hat{\theta}$ from Proposition 1, then $w_1^* = (w_1'\Gamma W_1^*)'$.*

b. *With $B_1 = \mathbb{I} - Z_1(Z_1'Z_1 + Z_0Z_0)^{-1}Z_1'$, w_1^* minimizes $w_1^{*\prime}B_1^{-1}w_1^*$ subject to $\Gamma' w_1^* = \Gamma' w_1$.*

The characterization of efficient weights in terms of the minimization problem clarifies that weights w_1^* can differ from weights w_1 if treatment-effect estimates $\hat{\tau}_{it}$ can be combined in a way that reduces variance according to it variance-covariance matrix B_1^{-1}, while ensuring that the expectation of the resulting estimator is still τ_w. Equivalently, we can obtain a more efficient estimator $w_1^{*\prime}\hat{\tau}$ than $w_1'\hat{\tau}$ when there is an unbiased estimator $(w_1^* - w_1)'\hat{\tau}$ of zero that is negatively correlated with $w_1'\hat{\tau}$.

A.2 Asymptotic Conditions on Weights

In this section, we spell out the high-level conditions on weights used in Section 4.2. In order to extend consistency to asymptotic Normality, we impose an additional assumption on the concentration of weights:[35]

Assumption A1 (Higher moments of weights)**.** *There exists $\delta > 0$ such that $\mathbb{E}\left[|\varepsilon_{it}|^{2+\delta}\right]$ uniformly bounded and*

$$\sum_i \left(\frac{\sum_t |v_{it}|}{\|v\|_{\mathrm{H}}}\right)^{2+\delta} \to 0.$$

Second, we assume that $\hat{\delta}$ is close to δ in the sense of a weighted mean-squared error:

[35] An example for a sequence of weights that sums to one, fulfills the Herfindahl condition $\|v\|_{\mathrm{H}}^2 \to 0$ from Assumption 6 and thus yields consistency (assuming bounded variance), but does not satisfy Assumption A1 is

$$v_{11} = 1/\sqrt{I}, \qquad v_{i1} = \frac{1 - 1/\sqrt{I}}{I - 1}\ (i > 1)$$

where we only observe one observation per cluster. For this choice, $\|v\|_{\mathrm{H}}^2 = 1/I + \frac{(1-1/\sqrt{I})^2}{I-1} \to 0$, but since weights are concentrated on the first unit,

$$\sum_i \left(\frac{\sum_{t\in i} |v_{it}|}{\|v\|_{\mathrm{H}}}\right)^{2+\delta} = \left(\frac{1/\sqrt{I}}{\sqrt{1/I + \frac{(1-1/\sqrt{I})^2}{I-1}}}\right)^{2+\delta} + (I-1)\left(\frac{\frac{1-1/\sqrt{I}}{I-1}}{\sqrt{1/I + \frac{(1-1/\sqrt{I})^2}{I-1}}}\right)^{2+\delta} \to \left(\frac{1}{\sqrt{2}}\right)^{2+\delta}.$$

Assumption A2 (Consistent estimation of $\hat{\delta}$)**.** *For the estimate $\hat{\delta}$ of δ we assume that*

$$\|v\|_{\mathrm{H}}^{-2} \sum_i \left(\sum_t v_{it} X_{it}' (\hat{\delta} - \delta) \right)^2 \xrightarrow{p} 0.$$

This condition expresses that the fitted values $X_{it}'\hat{\delta}$ are close to $X_{it}'\delta$ according to a norm given by the weights on related units. We develop sufficient conditions below for the case of short panels when δ express time fixed effects.

Finally, we impose additional moment restrictions for estimating standard errors:

Assumption A3 (Moment conditions for inference)**.** *$|\tau_{it}|$, $|\bar{\tau}_{it}|$ and $\mathbb{E}[\varepsilon_{it}^4]$ are uniformly bounded and* $\sum_i \left(\frac{\sum_{t \in i} |v_i|}{\|v\|_{\mathrm{H}}} \right)^4 \to 0$.

A.3 Sufficient Asymptotic Conditions for Inference in Short Panels

In this section we develop low-level sufficient conditions on the weights w_1 on treated observations and cohort sizes in the case of a complete panel with I units and T time periods with respective fixed effects. We first state sufficient condition for consistency in a panel where the number of periods T is allowed to grow slowly.

Assumption A4 (Sufficient conditions for consistency)**.** *Assume that every unit is present and not treated in the first period, and that*

1. $\sum_{i=1}^I \left(\sum_{t; D_{it}=1} |w_{it}| \right)^2 \to 0$*, that is, the weights on treatment effects fulfill a (clustered) Herfindahl condition (this bounds the variance from the treated observations themselves);*
2. $T \sum_{i=1}^I \left(\sum_{t; D_{it}=1} w_{it} \right)^2 \to 0$*, that is, the Herfindahl concentration of unit net weights decreases fast enough (this bounds the variance from estimating unit-fixed effects from the untreated observations);*
3. $T^2 \sum_{t=2}^T \frac{(\sum_{i; D_{it}=1} w_{it})^2}{\sum_{i; D_{it}=0} 1} \to 0$*, that is, the sum of squared total weight on observations treated at t relative to the number of untreated observations in the same period vanishes sufficiently quickly (this bounds the variance from estimating time-fixed effects from the untreated observations).*

The first two conditions express that the weights do not concentrate on too few units. They are similar, but not redundant; when some weights within a unit are negative and some positive, the weights may cancel out within units, yielding the second condition even when the first one is not fulfilled.

Next, we develop sufficient conditions, which will imply asymptotic Normality and valid inference in the special case of a complete panel with fixed number T of periods and staggered adoption.

Assumption A5 (Sufficient conditions for asymptotic Normality and inference)**.** *Assume that we have a complete panel with staggered adoption where no unit is treated in the first of a fixed number T of periods,and that*

1. *There is some uniform constant C such that for all t and i, j with $E_i = E_j$, $D_{it} = 1 = D_{jt}$ and $w_{it} > 0$, we have that $\frac{|w_{it}|}{|w_{jt}|} \leq C$, that is, weights within cohort and period do not vary too much;*
2. $\sum_{i; E_i = t} 1 \to \infty$ *for all t, that is, the size of the cohort of units first treated at t grows.*

A.4 Calculation of Efficient Weights

We first establish a general result about regressions: that any linear combination of estimated coefficients can be represented as a weighted sum of outcomes, where the weights are themselves a linear combination of the regressors.

Lemma A1. *Consider some scalar estimator $\hat{\psi}_w = w'\hat{\psi}$ obtained from an arbitrary point-identified regression $y_j = \psi' z_j + \varepsilon_j$. Like every linear estimator, it can be uniquely represented as $\hat{\psi}_w = v'y$, with y collecting y_j and with implied weights $v = (v_j)_j$ that do not depend on the outcome realizations. Then weights v_j can be represented as a linear combination of z_j in the sample, i.e. $v_j = z_j'\check{\psi}$ for some vector $\check{\psi}$, the same for all j.*

Proof. By standard OLS results, $\hat{\psi}_w = w'(z'z)^{-1} z'y$. Thus, $\hat{\psi}_w = v'y$ for weights $v = z(z'z)^{-1} w$ which do not depend on y. Moreover, these weights can be rewritten as $v_j = z_j'\check{\psi}$ for $\check{\psi} = (z'z)^{-1} w$. □

We now apply this lemma to Proposition 1, with the general model of $Y_{it}(0) = Z_{it}'\pi + \varepsilon$. Then $\hat{\tau}_w^* = v'Y$, where weights v can be represented as

$$v = Z_{it}'\check{\pi} + D_{it}\Gamma_{it}'\check{\theta}.$$

It remains to find the unknown $\check{\pi}$ and $\check{\theta}$ to obtain the v weights. To do so, we use the properties of $\hat{\tau}_w^*$. First, it equals to zero if Y_{it} is linear in Z_{it}. Second, letting $\mu = \Gamma' w_1$, $\hat{\tau}_w^* = \mu_j$ if $Y_{it} = \Gamma_{it,j} D_{it}$, as in that case $\hat{\theta}_j = 0$ and $\hat{\theta}_{-j} = 0$. Thus we have a system of equations which determine $\check{\pi}$ and $\check{\theta}$:

$$\sum_{it \in \Omega} Z_{it}\left(Z_{it}'\check{\pi} + D_{it}\Gamma_{it}'\check{\theta}\right) = 0; \tag{13}$$

$$\sum_{it \in \Omega_1} \Gamma_{it}\left(Z_{it}'\check{\pi} + D_{it}\Gamma_{it}'\check{\theta}\right) = \Gamma' w_1. \tag{14}$$

When Z_{it} has a block structure in which some covariates are FEs, solving this system iteratively is most convenient and computationally efficient. For instance, suppose $Z_{it}'\pi \equiv \alpha_i + X_{it}'\delta$ and $\Gamma = \mathbb{I}_{N_1}$ (i.e. Assumption 3 is trivial). Then Lemma A1 implies $v_{it} = \check{\alpha}_i + X_{it}'\check{\delta} + D_{it}'\check{\theta}_{it}$ for all $it \in \Omega$, and (14) simplifies to $v_{it} = w_{it}$ for all $it \in \Omega_1$. Using this and the structure of Z_{it}, we rewrite (13) as a system

$$\sum_{t,\ it \in \Omega_0} \left(\check{\alpha}_i + X_{it}'\check{\delta}\right) = -\sum_{t,\ it \in \Omega_1} w_{it}; \tag{15}$$

$$\sum_{it \in \Omega_0} X_{it}\left(\check{\alpha}_i + X_{it}'\check{\delta}\right) = -\sum_{it \in \Omega_1} X_{it} w_{it}. \tag{16}$$

This system suggests an iterative algorithm, similar to iterative OLS (e.g. Guimarães and Portugal, 2010):

1. Given a guess of $\check{\delta}$, set $\check{\alpha}_i$ for each unit to satisfy (15);
2. Given $\check{\alpha}_i$, set $\check{\delta}$ to satisfy (16);
3. Repeat until convergence.

A.5 Stochastic Event Timing

TBA.

B Proofs

In this section, we collect proofs for the results in the main text and in the appendix. Here, we extend the matrix notation from Section 4 to simplify notation. Specifically, we stack the vectors λ_i into a single vector $\lambda = (\lambda_i)_i$. We set $Z_{it} = \begin{pmatrix} (1_{i=j} A_{jt})_j \\ X_{it} \end{pmatrix}, \pi = \begin{pmatrix} \lambda \\ \beta \end{pmatrix}$ to summarize the nuisance component of the model. In matrix-vector notation, we write Y for the vector of outcomes, $Z = (A, X)$ for the covariate matrix, D for the matrix of indicators for treated units, ε for the vector of error terms, and $\Sigma = \mathrm{Var}[\varepsilon]$ for their variance. We write $Y_1, Z_1 = (A_1, X_1), D_1, \varepsilon_1$ for the rows corresponding to treated observations ($it \in \Omega_1$); in particular, $D_1 = \mathbb{I}$. Analogously, we write $Y_0, Z_0 = (A_0, X_0), D_0, \varepsilon_0$ for the rows corresponding to untreated observations ($it \in \Omega_0$) ; in particular, $D_0 = \mathbb{O}$. We write $\tau = (\tau_{it})_{it \in \Omega_1}$ for the vector of treatment effects of the treated units, $\theta = (\theta_m)_{m=1}^{N_1 - M}$ for the vector of underlying parameters, $\Gamma = (\Gamma_{it,m})_{it \in \Omega_1, m \in \{1,\dots,N_1 - M\}}$ for the matrix linking the two, and $w_1 = (w_{it})_{it \in \Omega_1}$ for the weight vector. Then we can write model and estimand as

$$Y = Z\pi + D\tau + \varepsilon, \qquad \tau = \Gamma\theta, \qquad \tau_w = w_1'\tau$$

with $\mathbb{E}[\varepsilon] = \mathbf{0}, \mathrm{Var}[\varepsilon] = \Sigma$, where Σ has block structure according to units i. For unit i, we write

$$A_i = (A_{it})_t, X_i = (X_{it})_t, Y_i = (Y_{it})_t, \varepsilon_i = (\varepsilon_{it})_t, v_i = (v_{it})_t$$

and denote by $\Sigma_i = \mathrm{Var}[\varepsilon_i]$ the within-unit variance–covariance matrix of error terms.

B.1 Proofs from Main Text

Proof of Lemma 1. In the absence of never-treated units and defining $\tau_{-1} = 0$, we can write $\sum_{h \neq -1} \tau_h \mathbf{1}[K_{it} = h] = \tau_{K_{it}}$.

Now consider some collection of τ_h (with $\tau_{-1} = 0$) and FEs $\tilde{\alpha}_i$ and $\tilde{\beta}_t$. For any $\kappa \in \mathbb{R}$, let $\tau_h^\star = \tau_h + \kappa(h+1)$, $\tilde{\alpha}_i^\star = \tilde{\alpha}_i + \kappa(E_i - 1)$, and $\tilde{\beta}_t^\star = \beta_t - \kappa t$. Then for any observation it,

$$\begin{aligned} \tilde{\alpha}_i^\star + \tilde{\beta}_t^\star + \tau_{K_{it}}^\star &= \tilde{\alpha}_i + \beta_t + \tau_h + \kappa(E_i - 1) - \kappa t + \kappa(t - E_i + 1) \\ &= \tilde{\alpha}_i + \beta_t + \tau_h, \end{aligned}$$

and Equation (2) has exactly the same fit under the original and modified FEs and τ_h coefficients, indicating perfect collinearity. □

Proof of Lemma 2. By the Frisch–Waugh–Lovell theorem, τ^{static}can be obtained by a regression of $\mathbb{E}[Y_{it}] = \alpha_i + \beta_t + \tau_{it} D_{it}$ on $\tilde{D}_{it}$ (without a constant), where $\tilde{D}_{it} = D_{it} - \check{\alpha}_i - \check{\beta}_t$ are the residuals from the auxiliary

regression of D_{it} on the unit and period FEs. Thus,

$$\tau^{\text{static}} = \frac{\sum_{it\in\Omega} \tilde{D}_{it} \left(\alpha_i + \beta_t + \tau_{it} D_{it}\right)}{\sum_{it\in\Omega} \tilde{D}_{it}^2}.$$

We have $\sum_{it\in\Omega} \tilde{D}_{it}\alpha_i = \sum_i \alpha_i \sum_{t:\ it\in\Omega} \tilde{D}_{it} = 0$ because the residuals in the auxiliary regression are orthogonal to all unit indicators. Analogously, $\sum_{it\in\Omega} \tilde{D}_{it}\beta_t = 0$. Defining

$$w_{it}^{\text{OLS}} = \frac{\tilde{D}_{it}}{\sum_{it\in\Omega} \tilde{D}_{it}^2}, \tag{17}$$

we have that $\tau^{\text{static}} = \sum_{it\in\Omega} w_{it}^{\text{OLS}} \tau_{it} D_{it} = \sum_{it\in\Omega_1} w_{it}^{\text{OLS}} \tau_{it}$, as required.

Clearly, w_{it}^{OLS} do not depend on the outcome realizations. Moreover, these weights add up to one:

$$\begin{aligned}
\sum_{it\in\Omega_1} \tilde{D}_{it} &= \sum_{it\in\Omega} \tilde{D}_{it} D_{it} \\
&= \sum_{it\in\Omega} \tilde{D}_{it} \left(\tilde{D}_{it} + \check{\alpha}_i + \check{\beta}_t\right) \\
&= \sum_{it\in\Omega} \tilde{D}_{it}^2,
\end{aligned}$$

where the last equality holds because $\tilde{D}_{it}$ are orthogonal to the unit and period FEs. □

Proof of Lemma 3. We use the characterization of OLS weights in Equation (17). Given the balanced panel, the regression of D_{it} on TWFE produces residuals

$$\tilde{D}_{it} = D_{it} - \bar{D}_{i\cdot} - \bar{D}_{\cdot t} + \bar{D}_{\cdot\cdot},$$

where $\bar{D}_{i\cdot} = \frac{1}{3}\sum_{t=1}^{3} D_{it}$, $\bar{D}_{\cdot t} = \frac{1}{2}\sum_{i=A,B} D_{it}$, and $\bar{D}_{\cdot\cdot} = \frac{1}{6}\sum_{i=A,B}\sum_{t=1}^{3} D_{it}$ (De Chaisemartin and D'Haultfœuille, 2020). Plugging in $\bar{D}_{\cdot A} = 1/3$, $\bar{D}_{\cdot B} = 2/3$, $\bar{D}_{\cdot 1} = 0$, $\bar{D}_{\cdot 2} = 1/2$, $\bar{D}_{\cdot 3} = 1$, and $\bar{D}_{\cdot\cdot} = 1/2$, and computing $\sum_{i,t\in\Omega_1} \tilde{D}_{it} = 1/3$, we have

$$\begin{aligned}
\hat{\tau}^{\text{static}} &= \frac{\sum_{i,t\in\Omega} \tilde{D}_{it} Y_{it}}{\sum_{i,t\in\Omega_1} \tilde{D}_{it}} \\
&= (Y_{B2} - Y_{A2}) - \frac{1}{2}\left(Y_{B1} - Y_{A1}\right) - \frac{1}{2}\left(Y_{B3} - Y_{A3}\right).
\end{aligned}$$

Similarly, the OLS estimand equals

$$\begin{aligned}
\tau^{\text{static}} &= \frac{\sum_{i,t\in\Omega_1} \tilde{D}_{it} \tau_{it}}{\sum_{i,t\in\Omega_1} \tilde{D}_{it}} \\
&= \tau_{B2} - \frac{1}{2}\left(\tau_{B3} - \tau_{A3}\right).
\end{aligned}$$

□

Proof of Lemma 4. For any observation it, $K_{it} = t - E_i \geq \bar{H}$ implies $t \geq E_i + \bar{H} \geq \min_i E_i + \bar{H} \geq \max_i E_i$. Thus, all observations considered by the estimand correspond to the periods in which all units are already treated. Consider one such period $t^\star$ for which the total weights are non-zero, $\sum_{i:\ K_{it^\star} \geq \bar{H}} w_{it} \neq 0$. (It exists because all weights w_{it} are assumed non-negative and are not identically zero.) Then consider a data-generating process (DGP) in which $\beta_{t^\star}$ is replaced with $\beta_{t^\star} - \kappa$ for some $\kappa \neq 0$ and $\tau_{it^\star}$ is replaced with $\tau_{it^\star} + \kappa$ for all i. This DGP is observationally equivalent in terms of the observed Y_{it} and continues to satisfy , Assumptions 1 and 2. Yet, the estimand differs by any arbitrary $\kappa \sum_{i:\ K_{it^\star} \geq \bar{H}} w_{it} \neq 0$, and it therefore not identified. □

Proof of Proposition 1. By construction, $\hat{\tau}_w = w_1' \Gamma \hat{\theta}$ for the OLS estimator $\hat{\theta}$ of θ. We now relate efficiency of $\hat{\tau}_w$ to efficiency of $\hat{\theta}$. Any linear estimator $\tilde{\tau}_w$ that is unbiased for τ_w for all θ can be related to a linear unbiased estimator $\tilde{\theta}$ of θ (with variance $\Sigma_{\tilde{\theta}}$) for which $\tilde{\tau}_w = w_1' \Gamma \tilde{\theta}$, for example by setting $\tilde{\theta} = \hat{\theta} + \Gamma' w (w' \Gamma \Gamma' w)^{-1} (\tilde{\tau}_w - w_1' \Gamma \hat{\theta})$. Indeed, for that choice,

$$\begin{aligned}
\mathbb{E}[\tilde{\theta}] &= \mathbb{E}[\hat{\theta}] + \Gamma' w_1 (w_1' \Gamma \Gamma' w_1)^{-1} \mathbb{E}[\tilde{\tau}_w] - \mathbb{E}[w_1' \Gamma \hat{\theta}] \\
&= \theta + \Gamma' w_1 (w_1' \Gamma \Gamma' w_1)^{-1} (w_1' \Gamma \theta - w_1' \Gamma \theta) = \theta, \\
w_1' \Gamma \tilde{\theta} &= w_1' \Gamma \hat{\theta} + w_1' \Gamma \Gamma' w_1 (w_1' \Gamma \Gamma' w_1)^{-1} (\tilde{\tau}_w - w_1' \Gamma \hat{\theta}) = \tilde{\tau}_w.
\end{aligned}$$

Under homoskedasticity, the OLS estimator $\hat{\theta}$ is the BLUE for θ in the regression $Y = Z\pi + D\Gamma\theta + \varepsilon$, with variance $\Sigma_{\hat{\theta}}$ that is minimal (in the partial ordering implied by positive definiteness) among the variance of linear unbiased estimators of θ by Gauss–Markov. Hence, $\mathrm{Var}(w_1' \Gamma \hat{\theta}) - \mathrm{Var}(w_1' \Gamma \tilde{\theta}) = w_1' (\Sigma_{\tilde{\theta}} - \Sigma_{\hat{\theta}}) w_1 \leq 0$, establishing efficiency. The efficient linear estimator of τ_w is also unique; indeed, if there was some unbiased linear estimator $\tilde{\tau}_w$ with $\mathrm{Var}(\tilde{\tau}_w) = \mathrm{Var}(\hat{\tau}_w)$ but $\mathbb{E}[(\tilde{\tau}_w - \hat{\tau}_w)^2] > 0$ (and thus $\mathrm{Cov}\,[\tilde{\tau}_w, \hat{\tau}_w] < \mathrm{Var}(\hat{\tau}_w)$), then $\frac{\hat{\tau}_w + \tilde{\tau}_w}{2}$ would be an unbiased linear estimator with lower variance. □

Proof of Corollary 1. By Frisch–Waugh–Lovell applied to residualization of Y and Z with respect to D, the estimate $\hat{\pi}$ of π in the linear regression $Y = Z\pi + D\tau + \varepsilon$ is the same as the estimate of π in the restricted linear regression $Y_0 = Z_0 \pi + \varepsilon$. The OLS estimator $\hat{\tau}$ of τ in $Y = Z\pi + D\tau + \varepsilon$ is then $\hat{\tau} = Y_1 - Z_1 \hat{\pi}$ since $D_1 = \mathbb{I}, D_0 = \mathbb{O}$, yielding the estimator in the proposition. □

Proof of Proposition 2. Write $w_1^* = \frac{\partial}{\partial Y^{1\prime}} \hat{\tau}_w$ for the weights $\hat{\tau}_w$ puts on the treated observations Y_1. As in the proof of Proposition 1, there exists an unbiased linear estimator $\hat{\theta}$ of θ such that $\hat{\tau}_w = w_1^{*\prime} \Gamma \hat{\theta}$. Let $\hat{C} = Y_1 - \Gamma \hat{\theta}$. Then $\hat{C}$ is a linear estimator with $\mathbb{E}[\hat{C}] = \mathbb{E}[Y_1] - \mathbb{E}[\hat{\tau}] = Z^{1\prime} \pi$. Since $\hat{C}$ is linear, we can write $\hat{C} = UY_1 + VY_0$ for matrices U, V. Since

$$\begin{aligned}
Z^{1\prime} \pi &= \mathbb{E}[\hat{C}] = U \mathbb{E}[Y_1] + V \mathbb{E}[Y_0] \\
&= U\tau + (UZ_1 + VZ_0)\pi
\end{aligned}$$

for all τ, π, we must have $U = \mathbb{O}$. Therefore, $\hat{C}$ satisfies the requirement of the proposition: it is a linear estimator that is unbiased for $Z^{1\prime}\pi$ and does not depend on Y_1, and $\hat{\tau}_w = w'(Y_1 - \hat{C})$. □

Proof of Proposition 3. Writing $v_i = (v_{it})_t$, consistency follows from

$$\begin{aligned}\mathbb{E}[\hat{\tau}_w] &= 0,\\ \mathrm{Var}(\hat{\tau}_w) = \sigma_w^2 &= \sum_{i=1}^{I} v_i'\Sigma_i v_i \leq \min\left\{\sum_i\left(\sum_t |v_{it}|\right)^2, \rho\left(\sum_{it} v_{it}^2\right)\right\}\bar{\sigma}^2 \to 0.\end{aligned}$$

□

Proof of Proposition 4. Write

$$\hat{\tau}_w - \tau_w = \sum_{it} v_{it}\varepsilon_{it} = \sum_{i=1}^{I}\eta_i$$

with

$$\eta_i = v_i'\varepsilon_i, \qquad \mathbb{E}[\eta_i] = 0, \qquad \mathrm{Var}(\eta_i) = v_i'\Sigma_i v_i.$$

Write $p = 2+\delta$ and let q be the solution to $\frac{1}{p}+\frac{1}{q} = 1$ (so in particular $1 < q < 2 < p$). Using Hölder's inequality to establish that

$$\sum_t |v_{it}|^{\frac{1}{q}}\left(|v_{it}|^{\frac{1}{p}}|\varepsilon_{it}|\right) \leq \left(\sum_t |v_{it}|^{\frac{q}{q}}\right)^{\frac{1}{q}}\left(\sum_t |v_{it}|^{\frac{p}{p}}|\varepsilon_{it}|^p\right)^{\frac{1}{p}},$$

and using $\mathbb{E}\left[|\varepsilon_{it}|^p\right] \leq C$, we have that

$$\begin{aligned}\mathbb{E}[|\eta_i|^{2+\delta}] &= \mathbb{E}\left[\left|\sum_t v_{it}\varepsilon_{it}\right|^p\right] \leq \mathbb{E}\left[\left(\sum_t |v_{it}\varepsilon_{it}|\right)^p\right] = \mathbb{E}\left[\left(\sum_t |v_{it}|^{\frac{1}{q}}\,|v_{it}|^{\frac{1}{p}}|\varepsilon_{it}|\right)^p\right]\\ &\leq \mathbb{E}\left[\left(\sum_t |v_{it}|^{\frac{q}{q}}\right)^{\frac{p}{q}}\left(\sum_t |v|^{\frac{p}{p}}|\varepsilon_{it}|^p\right)^{\frac{p}{p}}\right] = \left(\sum_t |v_{it}|\right)^{\frac{p}{q}}\sum_t |v_{it}|\,\underbrace{\mathbb{E}[|\varepsilon_{it}|^p}_{\leq C}\\ &\leq \left(\sum_t |v_{it}|\right)^{\overbrace{\frac{p}{q}+1}^{=p}} C = \left(\sum_t |v_{it}|\right)^p C.\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\sum_i \mathbb{E}[|\eta_i|^{2+\delta}]}{\left(\sum_i \mathrm{Var}(\eta_i)\right)^{\frac{2+\delta}{2}}} &= \frac{\sum_i \mathbb{E}[|\eta_i|^{2+\delta}]}{\sigma_w^{2+\delta}}\\ &\leq \frac{\|v\|_{\mathrm{H}}^{2+\delta}}{\sigma_w^{2+\delta}}\frac{\sum_i\left(\sum_t |v_{it}|\right)^{2+\delta}\bar{c}}{\|v\|_{\mathrm{H}}^{2+\delta}} = \frac{\|v\|_{\mathrm{H}}^{2+\delta}}{\sigma_w^{2+\delta}}\sum_i\left(\frac{\sum_t |v_{it}|}{\|v\|_{\mathrm{H}}}\right)^{2+\delta} C \to 0,\end{aligned}$$

where we have used that $\limsup \|v\|_{\mathrm{H}}^{2+\delta}\|v\|_{\mathrm{H}}^2/\sigma_w^2 < \infty$ and that $\sum_{i=1}^{I}\left(\frac{\sum_t |v_{it}|}{\|v\|_{\mathrm{H}}}\right)^{2+\delta} \to 0$, and so by the

Lyapunov central limit theorem we have that

$$\sigma_w^{-1}(\hat{\tau}_w - \tau_w) \xrightarrow{d} \mathcal{N}(0,1).$$

□

Proof of Proposition 5. We have that

$$\begin{aligned}
&\text{for } D_{it}=0, \qquad \hat{\varepsilon}_{it} = Y_{it} - A_{it}'\hat{\lambda}_i - X_{it}'\hat{\delta} = \varepsilon_{it} - A_{it}'(\hat{\lambda}_i - \lambda_i) - X_{it}'(\hat{\delta} - \delta),\\
&\text{for } D_{it}=1, \qquad \hat{\tau}_{it} - \hat{\bar{\tau}}_w = Y_{it} - A_{it}'\hat{\lambda}_i - X_{it}'\hat{\delta} - \hat{\bar{\tau}}_w = \varepsilon_{it} + \tau_{it} - \hat{\bar{\tau}}_w - A_{it}'(\hat{\lambda}_i - \lambda_i) - X_{it}'(\hat{\delta} - \delta)
\end{aligned}$$

so

$$\begin{aligned}
&\sum_{t;D_{it}=0} v_{it}\hat{\varepsilon}_{it} + \sum_{t;D_{it}=1} v_{it}(\hat{\tau}_{it} - \hat{\bar{\tau}}_w)\\
&= v_i'\varepsilon_i - v_i'A_i(\hat{\lambda}_i - \lambda_i) - v_i'X_i(\hat{\delta} - \delta) + \sum_{t;D_{it}=1} v_{it}(\tau_{it} - \hat{\bar{\tau}}_w).
\end{aligned}$$

Since the estimator $\hat{\tau}_w$ is invariant with respect to a change in λ_i, and λ_i only appears within unit g with covariates A_i, we must have that

$$0 = \frac{\partial}{\partial \lambda_i}\hat{\tau}_w = \frac{\partial}{\partial \lambda_i} v_i'Y_i = v_i'\left(\frac{\partial}{\partial \lambda_i} Y_i\right) = v_i'A_i,$$

so $v_i'A_i = 0$. Hence,

$$\hat{\sigma}_w^2 = \sum_{i=1}^{I}\left(v_i'\varepsilon_i + \sum_{t,D_{it}=1} v_{it}(\tau_{it} - \hat{\bar{\tau}}_{it}) - v_i'X_i(\hat{\delta} - \delta)\right)^2,$$

so

$$\begin{aligned}
&\left|\hat{\sigma}_w^2 - \sum_{i=1}^{I}\left(v_i'\varepsilon_i + \sum_{t,D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it})\right)^2\right|\\
&\leq \sum_{i=1}^{I}\left(\sum_{t,D_{it}=1} v_{it}(\hat{\bar{\tau}}_{it} - \bar{\tau}_{it}) + v_i'X_i(\hat{\delta} - \delta)\right)^2\\
&\quad + 2\left|\sum_{i=1}^{I}\left(\sum_{t,D_{it}=1} v_{it}(\hat{\bar{\tau}}_{it} - \bar{\tau}_{it}) + v_i'X_i(\hat{\delta} - \delta)\right)\left(w_i'\varepsilon_i + \sum_{t,D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it})\right)\right|\\
&\leq \sum_{i=1}^{I}\left(\sum_{t,D_{it}=1} v_{it}(\hat{\bar{\tau}}_{it} - \bar{\tau}_{it}) + v_i'X_i(\hat{\delta} - \delta)\right)^2\\
&\quad + 2\sqrt{\sum_{i=1}^{I}\left(\sum_{t,D_{it}=1} v_{it}(\hat{\bar{\tau}}_{it} - \bar{\tau}_{it}) + v_i'X_i(\hat{\delta} - \delta)\right)^2}\sqrt{\sum_{i=1}^{I}\left(v_i'\varepsilon_i + \sum_{t,D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it})\right)^2}.
\end{aligned}$$

We have that

$$\mathbb{E}\left[\left(v_i'\varepsilon_i + \sum_{t,D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it})\right)^2\right] = w_i'\Sigma_i w_i + \left(\sum_{t,D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it})\right)^2$$

and, for $\mathbb{E}[\varepsilon_{it}^4], |\tau_{it}|^4, |\bar{\tau}_{it}|^4 \leq C$,

$$\begin{aligned}
&\text{Var}\left(\left(v_i'\varepsilon_i + \sum_{t,D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it})\right)^2\right) \leq \mathbb{E}\left[\left(v_i'\varepsilon_i + \sum_{t,D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it})\right)^4\right] \\
&\leq 16\,\mathbb{E}\left[\left(v_i'\varepsilon_i\right)^4 + \left(\sum_{t,D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it})\right)^4\right] \leq 16\,(1+2^4)\,\left(\sum_t |v_i|\right)^4 C.
\end{aligned}$$

Since $\sum_i \left(\frac{\sum_{t\in i}|v_i|}{\|v\|_{\text{H}}}\right)^4 \to 0$ also

$$\text{Var}\left(\|v\|_{\text{H}}^{-2}\sum_i\left(v_i'\varepsilon_i + \sum_{t,D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it})\right)^2\right) \leq 16\,(1+2^4)\,\sum_i\left(\frac{\sum_t |v_i|}{\|v\|_{\text{H}}}\right)^4 C \to 0$$

and thus $\|v\|_{\text{H}}^{-2}\left(\sum_i\left(v_i'\varepsilon_i + \sum_{t,D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it})\right)^2 - \sigma_w^2 - \sigma_\tau^2\right) \xrightarrow{p} 0$. Finally,

$$\begin{aligned}
&\|v\|_{\text{H}}^{-2}\sum_{i=1}^{I}\left(\sum_{t,D_{it}=1} v_{it}(\hat{\bar{\tau}}_{it} - \bar{\tau}_{it}) + v_i'X_i(\hat{\delta} - \delta)\right)^2 \\
&\leq 2\|v\|_{\text{H}}^{-2}\sum_{i=1}^{I}\left(\sum_{t,D_{it}=1} v_{it}(\hat{\bar{\tau}}_{it} - \bar{\tau}_{it})\right)^2 + 2\|v\|_{\text{H}}^{-2}\sum_i\left(\sum_t v_i'X_i(\hat{\delta} - \delta)\right)^2 \\
&\xrightarrow{p} 0
\end{aligned}$$

since $\|v\|_{\text{H}}^{-2}\sum_i\left(\sum_{t,D_{it}=1} v_{it}(\hat{\bar{\tau}}_{it} - \bar{\tau}_{it})\right)^2 \xrightarrow{p} 0$ and Assumption A2 holds. Hence, $\|v\|_{\text{H}}^{-2}(\hat{\sigma}^2 - \sigma_w^2 - \sigma_\tau^2) \xrightarrow{p} 0$. □

B.2 Proofs of Appendix Results

Proof of Corollary A1. Write $B_Z = \mathbb{I} - Z(Z'Z)^{-1}Z'$ for the annihilator matrix with respect to the control variables, then

$$\begin{aligned}
B_1 = B_Z|_{\Omega_1\times\Omega_1} &= \mathbb{I} - Z_1(Z'Z)^{-1}Z_1' \\
&= \mathbb{I} - Z_1(Z_1'Z_1 + Z_0'Z_0)^{-1}Z_1' \\
B_0 = B_Z|_{\Omega_0\times\Omega_1} &= -Z_0(Z'Z)^{-1}Z_1' \\
&= -Z_0(Z_1'Z_1 + Z_0'Z_0)^{-1}Z_1'.
\end{aligned}$$

By Frisch–Waugh–Lovell for the OLS estimator $\hat{\theta}$ of θ,

$$\begin{aligned} v &= B_Z D\Gamma((D\Gamma)' B_Z (D\Gamma))^{-1}\Gamma' w_1 \\ &= \begin{pmatrix} B_1 \\ B_0 \end{pmatrix} \Gamma(\Gamma' B_1 \Gamma)^{-1}\Gamma' w_1 \\ &= \left(\begin{pmatrix} \mathbb{I} \\ \mathbb{O} \end{pmatrix} - Z(Z'Z)^{-1}Z_1' \right) \Gamma(\Gamma'(\mathbb{I} - Z_1(Z'Z)^{-1}Z_1')\Gamma)^{-1}\Gamma' w_1 \end{aligned}$$

When Γ is invertible then $\Gamma(\Gamma' B_1 \Gamma)^{-1}\Gamma' = B_1^{-1}$, simplifying the expression to

$$v = \begin{pmatrix} B_1 \\ B_0 \end{pmatrix} \Gamma(\Gamma' B_1 \Gamma)^{-1}\Gamma' w_1 = \begin{pmatrix} w_1 \\ B_0 B_1^{-1} w_1 \end{pmatrix}$$

where

$$B_0 B_1^{-1} = -Z_0(Z'Z)^{-1}Z_1'(\mathbb{I} - Z_1(Z'Z)^{-1}Z_1')^{-1} = -Z_0(Z_0'Z_0)^{-1}Z_1'$$

since

$$\begin{aligned} &(Z_0'Z_0)^{-1}Z_1'(\mathbb{I} - Z_1(Z'Z)^{-1}Z_1') \\ =&(Z_0'Z_0)^{-1}(Z'Z(Z'Z)^{-1} - Z_1'Z_1(Z'Z)^{-1})Z_1' \\ =&(Z_0'Z_0)^{-1}(Z_0'Z_0)(Z'Z)^{-1}Z_1' = (Z'Z)^{-1}Z_1'. \end{aligned}$$

□

Proof of Corollary A2. Write $B_Z = (\mathbb{I} - Z(Z'Z)^{-1}Z')$ for the annihilator matrix with respect to fixed effects, and $B_1 = B_Z|_{\Omega_1 \times \Omega_1}$ (a $N_1 \times N_1$ matrix) and $B_0 = B_Z|_{\Omega_1 \times \Omega_0}$ (a $N_1 \times N_0$ matrix) for its relevant components, where components are ordered such that $D' B_Z = (B_1, B_0)$ and $N_1 = |\Omega_1|, N_0 = |\Omega_0|$. The OLS estimator in Proposition 1 is

$$\begin{aligned} \hat{\theta} &= (\Gamma' D' B_Z D\Gamma)^{-1}\Gamma' D' B_Z Y \\ &= (\Gamma' B_1 \Gamma)^{-1}\Gamma'(B_1 Y_1 + B_0 Y_0), \end{aligned}$$

which also implies that $W_1^* = (\Gamma' B_1 \Gamma)^{-1}\Gamma' B_1$ and

$$\hat{\tau}_w = w_1'\Gamma(\Gamma' B_1 \Gamma)^{-1}\Gamma'(B_1 Y_1 + B_0 Y_0).$$

By the imputation argument from the proof of Corollary 1, the estimator $\hat{\tau}$ (of τ) obtained by imputation in Step 2 is

$$\hat{\tau} = (D' B_Z D)^{-1} D' B_Z Y = Y_1 - B_1^{-1} B_0 Y_0.$$

Hence, to show Part a,

$$\begin{aligned} w_1'\Gamma W_1^*\hat{\tau} &= w_1'\Gamma(\Gamma' B_1\Gamma)^{-1}\Gamma' B_1(Y_1 - B_1^{-1}B_0Y_0) \\ &= w_1'\Gamma(\Gamma' B_1\Gamma)^{-1}\Gamma'(B_1Y_1 + B_0Y_0) = \hat{\tau}_w. \end{aligned}$$

For Part b, note that $\mathrm{Var}(\hat{\tau})/\sigma^2 = (D'B_ZD)^{-1} = B_1^{-1}$. We can solve the optimization problem e.g. from the Lagrangian relaxation

$$\min_v v'B_1^{-1}v - 2\lambda'\Gamma'(v - w_1)$$

with FOC $\Gamma\lambda = B_1^{-1}v$, which is solved for $v^* = B_1\Gamma(\Gamma' B_1\Gamma)^{-1}\Gamma' w_1$ with $\lambda = (\Gamma' B_1\Gamma)^{-1}\Gamma' w_1$. This is the same weight as in Part a, since $w_1^* = (w_1'\Gamma W_1^*)' = B_1\Gamma(\Gamma' A_1\Gamma)^{-1}\Gamma' w_1 = v^*$. □