Two-Stage Differences in Differences

John Gardner | Neil Thakral | Linh T. Tô | Luther Yap

University of Mississippi | Brown University | Boston University | Princeton University

April 2025[*]

Abstract

This paper develops a framework for estimation and inference to analyze the effect of a policy or treatment in settings with treatment-effect heterogeneity and variation in treatment timing. We propose a two-stage difference-in-differences (2SDD) estimator that compares treated and untreated outcomes after removing group and period effects identified using untreated observations. Our regression-based approach enables us to conduct inference within a conventional GMM asymptotic framework. It easily facilitates extensions such as dynamic treatment effects, triple differences, continuous treatments, time-varying controls, and violations of parallel trends. Simulations of randomly generated placebo laws in state-level wage data demonstrate that 2SDD outperforms alternatives in terms of precision and rejection rates. Under homogeneous treatment effects, 2SDD yields similar standard errors as TWFE regressions, unlike other heterogeneity-robust estimators. Analyzing the rate of extreme t-statistics and outlying standard errors for various methods across seven empirical applications, 2SDD stands out as a practical choice for applied researchers.

[*] Gardner: Department of Economics, University of Mississippi (email: jrgardne@olemiss.edu). Thakral: Department of Economics, Brown University (email: neil_thakral@brown.edu). Tô: Department of Economics, Boston University (email: linhto@bu.edu). Yap: Department of Economics, Princeton University (email: lyap@princeton.edu). We thank Stéphane Bonhomme, Michael Briskin, Kyle Butts, Carolina Caetano, Gregorio Caetano, Brantly Callaway, Kenneth Chay, Scott Cunningham, David Drukker, Len Goff, David S. Lee, Zhanyuan Tian, Tao Wang, Taylor Wright, and participants at the ASSA, SOLE, SEA, and WEAI conferences for helpful comments and suggestions. This paper supersedes earlier versions that circulated under the same title by Gardner (2020).

1 Introduction

Difference-in-differences (DD) estimation has emerged as an indispensable tool for empirical researchers seeking to evaluate the impact of a given intervention or policy. Its appeal stems in part from the conceptual simplicity of comparing changes in outcomes for groups affected by an intervention to changes for unaffected groups. A potential reason for the widespread use of two-way fixed-effects (TWFE) in settings with multiple groups and time periods is a presumption that it should identify the average effect of the treatment on the treated. Although this intuition is accurate when the heterogeneous treatment effects are distributed identically across groups and periods (a condition that is automatically satisfied in the classic two-group, two-period setting), it does not hold in general. When these distributions are not identical, conditional mean outcomes are no longer linear in group, period, and treatment status, causing the TWFE regression model to be misspecified for conditional mean outcomes, and thus it is unable to identify the average treatment effect on the treated.

This paper develops a two-stage regression-based approach to identification, along with an accompanying GMM-based variance estimator, which is robust to treatment-effect heterogeneity when adoption of the treatment is staggered over time. The two-stage difference-in-differences (2SDD) estimator, along with valid asymptotic standard errors, can be implemented easily using standard statistical software, with little programming or computational time beyond that required to estimate a regression.[1] Alternative implementations of the two-stage framework rely on variance estimation methods that are asymptotically conservative and more computationally intensive (Borusyak, Jaravel and Spiess, 2024).

The proposed two-stage methodology regresses outcomes on group and period fixed effects using the subsample of untreated observations in the first stage. The second stage subtracts the estimated group and period effects from observed outcomes and regresses the resulting residualized outcomes on treatment status. Under the usual parallel trends assumption, this procedure identifies the overall average effect of the treatment on the treated (i.e., across groups and periods), even when average treatment effects are heterogeneous over groups and periods. This approach preserves the intuition behind identification in the two-group, two-period case: it recovers the average difference in outcomes between treated and untreated units, after removing group and period effects.

Our approach inherits the flexibility for which researchers have come to appreciate

[1] This contrasts with alternative approaches that rely on bootstrapping for inference (e.g., de Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021). We provide example Stata syntax that shows how to implement the two-stage difference-in-differences approach (with valid asymptotic standard errors) via GMM or the `did2s` Stata package (Butts, 2021) in Appendix A; also see the `did2s` R package (Butts and Gardner, 2022).

regression as a tool for applied empirical analysis, making it adaptable to the wide variety of settings where difference-in-differences analyses are used in practice, with standard inference procedures naturally applying. This extensibility is aided by our straightforward GMM approach, which naturally and efficiently maintains valid inference within a familiar regression-based framework without requiring additional computational steps. For instance, our basic approach can easily be extended to estimate the dynamic effects of treatments, implement tests of parallel trends, and handle models with individual fixed effects using a variation of the within estimator. It can also accommodate settings where parallel trends holds only after conditioning on time-varying covariates by simply including those covariates as regressors in the first stage, or where covariate-specific trends are needed by interacting time-invariant covariates with time indicators. In settings with continuous treatments, by replacing binary treatment status with a continuous treatment variable in the second stage, our approach preserves interpretability and computational simplicity while ensuring valid inference. Similarly, our method extends seamlessly to triple-differences analyses with only a simple modification to the first stage, offering clear advantages in ease of implementation over alternative methods. We also discuss how to extend our approach to settings with partial violations of parallel trends, and how it can be adapted to design-based analyses.[2]

In the canonical two-period, two-group DD setup, we show that our proposed two-stage estimator exactly matches the variance of the traditional TWFE estimator. By contrast, other popular heterogeneity-robust estimators (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Wooldridge, 2021; de Chaisemartin and d'Haultfoeuille, 2024) yield strictly wider confidence intervals—even in this most basic setting. Thus, while heterogeneity-robustness typically entails an efficiency penalty, the equivalence between 2SDD and TWFE under the simplest designs demonstrates that such a trade-off is not inevitable. From an applied perspective, 2SDD not only retains the intuitive clarity, interpretability, and extensibility of TWFE, but also avoids much of the precision losses inherent in other heterogeneity-robust estimators.

To further evaluate the performance of the proposed variance estimation procedure and the broader methodology, we conduct simulations of randomly generated laws using state-level wage data. This simulation builds on the seminal work of Bertrand, Duflo and Mullainathan (2004), extending their analysis to a setting with heterogeneous treatment effects and staggered treatment timing. In particular, we analyze the performance of various DD estimators and their associated inference procedures with randomly drawn treated states, their associated

[2] Our approach to inference accommodates design-based sources of uncertainty. As Abadie et al. (2020) emphasize, the design-based perspective provides a coherent interpretation for standard errors, particularly for empirical settings where the source of randomness is known.

years of passage, and treatment effects. Using a 42-year panel, we compare rejection rates at the 5 percent significance level from recently proposed heterogeneity-robust procedures (CS2021; SA2021; W2021; BJS2024; dCDH2024).[3] Simulations highlight the value of our approach to estimation and inference by demonstrating its finite-sample performance. Our procedure consistently offers the best performance in terms of rejection rates, computational speed, and efficiency. This result holds even in comparison with the imputation approach from BJS2024 that provides identical point estimates to ours with different variance estimators.[4] We obtain similar results using independent and identically distributed data. Furthermore, we document these advantages even in cases with homogeneous treatment effects. In such cases, our method yields comparable standard errors to the TWFE estimator, while other heterogeneity-robust estimators tend to yield much larger standard errors.

We also compare the relative performance of the different estimators across seven empirical applications. In these applications, unlike in the simulation environment, the "true" treatment effects remain unknown. This mirrors the challenge empirical researchers face, where the choice of estimator can potentially influence their conclusions. In such situations, a method that produces fewer outliers or inconsistencies relative to the alternatives reduces the potential for skewed results. Our approach consistently provides stable conclusions across the empirical applications, with the lowest rate of extreme t-statistics and the fewest outlying standard errors relative to the other estimators. In contrast, the CS2021 estimator yields large standard errors with a large number of treatment cohorts (e.g., Bailey and Goodman-Bacon, 2015; Deryugina, 2017), while the SA2021 and dCDH2024 estimators perform relatively poorly with a relatively large number of small cohorts (e.g., Lafortune, Rothstein and Schanzenbach, 2018; Tewari, 2014; Ujhelyi, 2014).

Our work adds to an emerging body of research highlighting limitations of the traditional TWFE approach for DD estimation in the presence of staggered treatment timing and when the effects of a treatment vary across groups and time.[5] We motivate our approach by elucidating how misspecified TWFE regression models project heterogeneous treatment effects

[3] Our analysis does not include the local projections approach (Dube et al., 2023) due to its lack of a theoretical framework for inference.

[4] The "imputation" estimator, which first appears in Borusyak, Jaravel and Spiess (2021), is numerically identical to the two-stage estimators initially proposed by Gardner (2020), Thakral and Tô (2020), and Liu, Wang and Xu (2019). However, they develop a different asymptotic theory, resulting in an asymptotically conservative default variance estimator and a leave-one-out modification which they show results in improved finite-sample performance. BJS2024 note that similar approaches have been proposed for factor models by Gobillon and Magnac (2016) and Xu (2017), while Arkhangelsky and Imbens (2024) note that the two-stage approach can be viewed as a particular extension of a broader class of estimation strategies for panel models developed by Chamberlain (1992) and extended by Graham and Powell (2012) and Arellano and Bonhomme (2012).

[5] See, for example, dCDH2020, Goodman-Bacon (2021), Imai and Kim (2021), SA2021, Athey and Imbens (2022), and BJS2024.

onto treatment status, group effects, and period fixed effects. The simple observation that untreated outcomes are linear in group and period effects under parallel trends then naturally leads to our proposed two-stage method. Several papers provide alternative representations of the TWFE estimand. BJS2024 show that TWFE identifies a regression-weighted mean of the average effect of the treatment in each post-treatment period, and dCDH2020 show that all TWFE regression estimates (which include DD regressions as a special case) identify weighted averages of group- and period-specific average treatment effects. Since the weights in both of these representations can be negative, interpreting the TWFE estimand becomes challenging. Goodman-Bacon (2021) further shows that the TWFE estimate represents a weighted average of all two-group, two-period differences in differences, which under parallel trends identifies a combination of weighted averages of group×period-specific average treatment effects and changes over time in those effects. These decomposition results tend to motivate alternative methodologies based on manually averaging cohort×period-specific average treatment effects (dCDH2020; CS2021; SA2021). Although Harmon (2024) shows how these procedures are more efficient than imputation-based procedures such as BJS2024 when residuals are strongly serially correlated, the simplicity of our procedure makes it easy to adapt to be more efficient in the presence of serial correlation.

In the presence of staggered treatment adoption, several alternatives to the TWFE regression approach exhibit robustness to heterogeneity across groups and periods. One alternative, as mentioned earlier, is to estimate separate average treatment effects for each group and period, which can then be aggregated to form measures of the overall effect of the treatment.[6] In comparison to this approach, our regression-based methodology offers simplicity in estimation and inference, significant computational speed advantages, and strong finite-sample performance. In addition, our approach retains the efficiency advantages pointed out by BJS2024. We also discuss how to mitigate bias from violations of parallel trends by using an appropriate subset of untreated observations in the first stage.

Alternative regression-based approaches include the "stacked" difference-in-differences (see, e.g., Gormley and Matsa, 2011; Deshpande and Li, 2019; Cengiz et al., 2019; Dube et al., 2023), which attempts to transform the staggered adoption setting to a two-group, two-period design (in which difference in differences identifies the overall average effect of the treatment on the treated) by stacking separate datasets containing observations on treated and control units for each treatment cohort, and the extended TWFE approach (W2021). Several limitations arise when applying these methods. First, the stacked estimator identifies

[6] Gibbons, Suárez Serrato and Urbancic (2018) suggest an approach like this for fixed effects models; Borusyak, Jaravel and Spiess (2021) suggest such a solution for DD models in which the duration-specific effects of the treatment are identical across groups, as do Callaway and Sant'Anna, 2021 for the case when treatment effects vary by group and duration and Sun and Abraham, 2021 in the event-study context.

a particular weighted average of group-specific average treatment effects that depends on arbitrary features of the data, making the resulting estimate more challenging to interpret.[7] Second, implementing the stacked approach requires defining a fixed event time window and ensuring a balanced panel throughout this period. Third, stacking involves using the same control groups across different stacked datasets but lacks a theoretical framework for inference. Finally, the extended TWFE approach only considers time-invariant covariates and assumes a linear relationship between covariates and treatment effects. Our method overcomes these issues by delivering clear and interpretable estimates, providing a theoretical framework for inference, and allowing for flexible implementation across various contexts, including those with time-varying covariates that interact arbitrarily with treatment effects and the other extensions of our procedure discussed above.

Given the multitude of alternative approaches for DD estimation, our empirical and simulation exercises constitute a distinct contribution to this literature. Our empirical exercises complement recent work by Chiu et al. (2023), which reanalyzes a set of political science publications that estimate TWFE regressions. They emphasize that TWFE estimates correlate strongly with the estimates from alternative methods but find that the latter tend to be less precise. Under homogeneous treatment effects, as our simulation results demonstrate, our proposed approach to inference achieves the closest standard error to that of a TWFE estimator that imposes a null effect in the pre-treatment periods.[8] Our simulation exercises present a novel and systematic comparison of standard errors and rejection rates across different estimators in staggered adoption settings based on typical empirical applications. Subsequent work by Egerod and Hollenbach (2024) and Weiss (2024) also compare heterogeneity-robust estimators using simulations. Weiss (2024) focuses on the CS2021, SA2021, BJS2024, and dCDH2024 approaches (but does not examine 2SDD) and shows that their built-in variance estimators can perform poorly in small samples. Egerod and Hollenbach (2024) reinforce our finding of systematically low coverage among most methods, with CS2021 and 2SDD providing significantly better coverage. They further document power issues that arise in small samples and show that BJS2024 exhibits a notably high probability of reporting the wrong sign for the estimated effect conditional on a significant result. As our method for achieving robustness to treatment-effect heterogeneity entails minimal efficiency loss in settings where TWFE provides an unbiased estimate, it offers arguably the most compelling alternative to the TWFE approach in practice.

The paper proceeds as follows. In Section 2, we provide intuition for why the TWFE

[7] The weights depend on the relative sizes of the group-specific datasets and the variance of treatment status within those datasets, as Appendix G shows.

[8] As a result, our proposed estimator results in greater precision than the fully dynamic event-study specification using TWFE.

approach to DD estimation may not identify the average effect of the treatment on the treated, show how our proposed two-stage regression-based approach is robust to treatment-effect heterogeneity in settings with variation in treatment timing, and discuss extensions. Sections 3 to 5 demonstrate the performance of the two-stage approach compared to alternative proposals in simulations and empirical applications. We conclude in Section 6.

2 Inference with a two-stage approach

2.1 The problem with difference-in-differences regression

Difference-in-differences (DD) research designs attempt to identify the causal effects of treatments under the parallel or common trends assumption. This assumption asserts that, absent the treatment, treated units would experience the same change in outcomes as untreated units. Mathematically, this amounts to the assumption that average untreated potential outcomes decompose into additive group and period effects. Let i index units (e.g., states or, with microdata, individuals) and t index calendar time (often years). Further partition units and time into treatment groups $g \in \{0, 1, \dots, G\}$ and periods $p \in \{0, 1, \dots, P\}$ defined by the adoption of the treatment among successive groups, so that members of group 0 are untreated in all periods, only members of group 1 are treated in period 1, members of groups 1 and 2 are treated in period 2, and so on, and define corresponding group and time variables G_i and P_t. Let Y_{it}, $Y_{it}(d = 1)$ and $Y_{it}(d = 0)$ denote the observed, treated, and untreated potential outcomes for unit i at time t, let D_{it} be an indicator for whether i is treated at t, and let $\beta_{gp} = \mathbb{E}[Y_{it}(d = 1) - Y_{it}(d = 0) \mid g, p]$ denote the average causal effect of the treatment for members of g in p.[9] Assume for simplicity that the treatment is both irreversible and unanticipated (though these assumptions can be at least partially relaxed, as detailed in Section 2.5). Under parallel trends, mean outcomes satisfy

$$\mathbb{E}[Y_{it} \mid g, p, D_{it}] = \lambda_g + \alpha_p + \beta_{gp} D_{it}. \tag{1}$$

The idea behind differences in differences is to eliminate the permanent group effects λ_g and secular period effects α_p in order to identify the average effect of the treatment. In the classic setup, there are only two periods (pre and post) and two groups (treatment and control). In this setting, within-group differences over time eliminate the group effects and within-period differences between groups eliminate the period effects. Hence the between-group difference in post-pre differences (i.e., the difference in differences) identifies the average effect of the treatment for members of the treatment group during the post-treatment period.

[9] Causal effects for the never-treated group may be normalized to zero, since they are not identified.

The two-period, two-group difference-in-differences estimate can be obtained using a regression of outcomes on group and period fixed effects and a treatment-status indicator:

$$Y_{it} = \lambda_{g(i)} + \alpha_{p(t)} + \beta D_{it} + \varepsilon_{it}, \tag{2}$$

where $g(i)$ is the group index for unit i and $p(t)$ is the period index at time t. It follows from Equation (1) that the coefficient on D_{it} in Equation (2) identifies the average effect of the treatment on the treated, $\mathbb{E}[Y_{it}(d=1) - Y_{it}(d=0) \,|\, D_{it} = 1]$.[10]

The regression approach suggests a natural way to extend the DD idea to settings with multiple groups and time periods. Unfortunately, as several authors have noted (dCDH2020; Goodman-Bacon, 2021; Imai and Kim, 2021; Athey and Imbens, 2022; BJS2024), when the average effect of the treatment varies across groups and over periods, the coefficient on D_{it} in specification (2) does not always identify an easily interpretable measure of the "typical" effect of the treatment. Although this result is now well established, because it is also somewhat counterintuitive, it bears further clarification.

While there are multiple ways to think about the typical effect of the treatment when that effect varies across groups and over time, an obvious candidate is the average $\mathbb{E}\left[\beta_{gp} \,\middle|\, D_{it} = 1\right] = \mathbb{E}\left[Y_{it}(d=1) - Y_{it}(d=0) \,\middle|\, D_{gp} = 1\right]$ of group- and period-specific average treatment effects, taken over all units that receive the treatment and all times during which they receive it (i.e., the expectation of β_{gp} over the joint distribution of g and p, conditional on being treated). This is analogous to the average $\mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right]$ identified by difference in differences in the two-period, two-group case. Hence, parallel trends can be expressed as

$$\mathbb{E}\left[Y_{it} \,\middle|\, g, p, D_{gp}\right] = \lambda_g + \alpha_p + \mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right] D_{gp} + \left[\beta_{gp} - \mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right]\right] D_{gp}.$$

The difficulty with the regression approach is that, except in special cases, the "error term" $\left[\beta_{gp} - \mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right]\right] D_{gp}$ in this expression varies at the group×period level, and is not mean zero *conditional on group membership, period, and treatment status.* Consequently, the regression is misspecified in the sense that the conditional expectation $\mathbb{E}\left[Y_{gpit} \,\middle|\, g, p, D_{gp}\right]$ is not a linear function of those variables (at least, not one in which the coefficient on D_{gp} is $\mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right]$.) In contrast to the two-group, two-period case, the coefficient on D_{gp} from

[10] There are several equivalent variations on this regression. Specification (2) is identical to a regression of outcomes on an indicator $Post_{it}$ for whether t occurs in the post-treatment period, an indicator $Treat_{it}$ for whether i belongs to the treatment group, and an interaction between the two. Often, the group and period effects λ_g and α_p in Equation (2) are replaced with individual and time effects λ_i and γ_t. By the Frisch-Waugh-Lovell theorem, the coefficient on D_{it} in Equation (2) can be obtained by regressing Y_{it} on the residuals from a regression of treatment status on group and period effects. Since treatment status only varies by group and period, these residuals are the same as those from a regression of treatment status on individual and time effects, so the coefficients on treatment status from both specifications are identical.

the regression DD specification (2) does not identify $\mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right]$ unless those average effects are independent of group and period (in which case $\beta_{gp} = \mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right] = \beta$). Outside of this special case, when average treatment effects vary across groups and periods, and the adoption of the treatment by different groups is staggered over time, difference-in-differences regression does not recover a simple group×period average treatment effect (dCDH2020; Goodman-Bacon, 2021; BJS2024). In Appendix B, we derive a simple expression for what regression DD does identify in the special case of cohort fixed effects and no covariates.

2.2 A two-stage approach

The observation that the problem arises from misspecification of Equation (2) suggests a simple two-stage average treatment effect estimator for the multiple group and period case. As long there are untreated and treated observations for each group and period, λ_g and α_p are identified from the subpopulation of untreated groups and periods. The overall group×period average effect of the treatment on the treated is then identified from a comparison of mean outcomes between treated and untreated groups, after removing the group and period effects.

This logic suggests the following regression-based two-stage estimation procedure:

1. Estimate the model
$$Y_{it} = \lambda_{g(i)} + \alpha_{p(t)} + u_{it}$$
on the sample of observations for which $D_{it} = 0$, retaining the estimated group and time effects $\hat{\lambda}_g$ and $\hat{\alpha}_p$.

2. Regress adjusted outcomes $Y_{it} - \hat{\lambda}_{g(i)} - \hat{\alpha}_{p(t)}$ on D_{it}.

Since parallel trends implies that

$$\mathbb{E}\left[Y_{it} \,\middle|\, g, p, D_{it}\right] - \lambda_g - \alpha_p = \beta_{gp} D_{it} = \mathbb{E}\left[\beta_{gp} \,\middle|\, D_{it} = 1\right] D_{it} + \left[\beta_{gp} - \mathbb{E}\left[\beta_{gp} \,\middle|\, D_{it} = 1\right]\right] D_{it},$$

where $\mathbb{E}\left[\left[\beta_{gp} - \mathbb{E}\left[\beta_{gp} \,\middle|\, D_{it} = 1\right]\right] D_{it} \,\middle|\, D_{it}\right] = 0$, this procedure identifies $\mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right]$, even when the adoption and average effects of the treatment are heterogeneous with respect to groups and periods. A straightforward application of the continuous mapping theorem shows that this two-stage estimator is consistent for the overall average effect of the treatment.

We now formalize the logic of this argument, and extend it to specifications that include time-varying covariates and unit-level fixed effects. Suppose that we observe (Y_{it}, X_{it}, D_{it}), where Y_{it} denotes outcomes, D_{it} treatment status, and X_{it} a vector of covariates, where $i \in \{1, 2, \cdots, N\}$ indexes individuals and $t \in \{1, 2, \cdots, T\}$ indexes time, so that there are NT observations in a balanced panel.

We assume that outcomes are generated from a potential outcomes model with $Y_{it}(d) = Y_{it}(0) + \beta_{it}d$, with observed outcomes given by $Y_{it} = Y_{it}(D_{it})$. This notation implicitly assumes that there are no anticipation effects, as $Y_{it} = Y_{it}(0)$ for all it such that $D_{it} = 0$. When the treatment D_{it} is binary, $\beta_{it} = Y_{it}(1) - Y_{it}(0)$ can be interpreted as the treatment effect. We further assume that the treatment is irreversible and that the covariates are unaffected by treatment status, so that $X_{it} = X_{it}(0)$, where $X_{it}(0)$ denotes the counterfactual untreated value of the covariate.[11]

As in CS2021, SA2021, and W2021, we assume that $\{Y_{it}(0), \beta_{it}, D_{it}, X_{it}\}_{t=1}^{T}$ are iid (across individuals), and hence that the observed Y_{it} is also iid.[12] We also assume that T is fixed and $N \to \infty$, which the most common case encountered by researchers.

Our object of interest is the overall average effect of the treatment on the treated (ATT), which we define as[13]

$$\beta := E\left[\beta_{it} \mid D_{it} = 1\right].$$

We implicitly assume that this object is well-defined in that $E\left[D_{it}\right] > 0$. This assumption is easily satisfied in event studies. This average treatment effect parameter can be also written as a ratio of moments:

$$E\left[\beta_{it} \mid D_{it} = 1\right] = \frac{E\left[\beta_{it} D_{it}\right]}{E\left[D_{it}\right]}.$$

Our difference-in-differences approach is based on the following parallel trends assumption.

Assumption 1 (Parallel trends). *There exist non-stochastic γ, λ_i such that*

$$E\left[Y_{it}(0) \mid \{X_{it}\}_{t=1}^{T}, G_i = g\right] = \lambda_i + X_{it}'\gamma \tag{3}$$

for all it.

This notation subsumes the familiar time fixed effects into the vector X_{it} as time indicators (i.e., if X_{it} consists of a set of time indicators, this assumption takes form $E[Y_{it}(0) \mid G_i = g] = \lambda_i + \alpha_t$, where α_t are time fixed effects). Also note that conditioning on the treatment group g is equivalent to conditioning on a particular sequence $\{D_{it}\}_{t=1}^{T}$ of treatment paths.

[11] This assumption is actually stronger than necessary. As Callaway, Goodman-Bacon and Sant'Anna (2021) note, it suffices to assume that the treated and untreated covariate distributions are the same.

[12] This iid setup is somewhat stronger than merely having clustered error terms, but our derivation will still go through even if the observations are inid.

[13] This definition implicitly treats the treatment effects β_{it} as a random variable drawn from a hierarchical distribution that varies with time (rather than a series of time-specific random variables). One advantage of this notation is that it clarifies the weights with which our estimator aggregates group- and period-specific treatment effects, since regardless of whether covariates are included in the first stage, we have that $E(\beta_{it}|D_{it} = 1) = \sum_{g,p} \beta_{gp} \Pr(g,p)$.

Under Assumption 1, we can express observed outcomes using the regression specification

$$Y_{it} = \lambda_i + \beta D_{it} + X_{it}'\gamma + \varepsilon_{it}, \tag{4}$$

where X_{it} absorbs a time trend such that time fixed effects α_t are a part of γ, and can be consistently estimated.

To remove the individual fixed effects, we define the transformation $\tilde{Y}_{it} := Y_{it} - \frac{1}{T_i^0}\sum_{t=1}^{T_i^0} Y_{it}$, where $T_i^0 := \sum_{t=1}^{T}(1 - D_{it})$ is the last time period before individual i gets treated (or equivalently, the number of periods in which i is observed in an untreated state). For $\tilde{Y}_{it}$ to be well-defined, we assume that, for all i, there is some t where $D_{it} = 0$. The transformed $\tilde{X}_{it}$ and $\tilde{\varepsilon}_{it}$ are defined in a similar manner.

Using this transformation, (4) can also be written

$$\tilde{Y}_{it} = \beta D_{it} + \tilde{X}_{it}'\gamma + \tilde{\varepsilon}_{it}.$$

Lemma 1. *If Assumption 1 holds, then* $E\left[\sum_t D_{it}\tilde{\varepsilon}_{it}\right] = 0$ *and* $E\left[\sum_t \tilde{X}_{it}\tilde{\varepsilon}_{it} \mid D_{it} = 0\right] = 0$.

Assumption 1 is hence the key identifying assumption for this procedure, as it implies moment conditions for regressions. Due to Lemma 1, running a standard regression of $\tilde{Y}_{it}$ on $\tilde{X}_{it}$ for observations with $D_{it} = 0$ yields a consistent estimator for γ. We refer to this regression that obtains γ as the first-stage regression. Then, using the estimated $\hat{\gamma}$, we can regress $\tilde{Y}_{it} - \tilde{X}_{it}\hat{\gamma}$ on D_{it} to obtain an estimate of β; we refer to this regression as the second-stage regression.

To be precise about these regressions, we define a few objects. Using X_{kit} to denote regressor k for individual i at time t, we define the $T \times K$ matrix $\tilde{X}_{0i}$ as

$$\tilde{X}_{0i} = \begin{bmatrix} \left(X_{1i1} - \frac{1}{T_i^0}\sum_{t=1}^{T} X_{1it}(1 - D_{it})\right)(1 - D_{i1}) & \cdots & \left(X_{Ki1} - \frac{1}{T_i^0}\sum_{t=1}^{T} X_{Kit}(1 - D_{it})\right)(1 - D_{i1}) \\ \vdots & \ddots & \vdots \\ \left(X_{1iT} - \frac{1}{T_i^0}\sum_{t=1}^{T} X_{1it}(1 - D_{it})\right)(1 - D_{iT}) & \cdots & \left(X_{KiT} - \frac{1}{T_i^0}\sum_{t=1}^{T} X_{Kit}(1 - D_{it})\right)(1 - D_{iT}) \end{bmatrix}.$$

Similarly, $\tilde{Y}_{0i}$ is a $T \times 1$ vector of the form

$$\tilde{Y}_{0i} = \left[\tilde{Y}_{i1}(1 - D_{i1}) \quad \cdots \quad \tilde{Y}_{iT}(1 - D_{iT})\right]'.$$

The coefficient estimator from the first stage regression is then

$$\hat{\gamma} = \left(\sum_{i=1}^{N} \tilde{X}_{0i}'\tilde{X}_{0i}\right)^{-1}\left(\sum_{i=1}^{N} \tilde{X}_{0i}'\tilde{Y}_{0i}\right).$$

Observe that both sums are over independent individuals i (the sum over time is already implicit in the matrix multiplication) and that the second stage-regression is done without a constant, because the data are already demeaned. Hence, the two-stage difference in difference estimator is

$$\hat{\beta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\tilde{Y}_{it} - \tilde{X}_{it}\hat{\gamma}\right)\right).$$

To summarize, the two-stage procedure now becomes:

1. Regress $\tilde{Y}_{0i}$ on $\tilde{X}_{0i}$ to obtain $\hat{\gamma}$.

2. Regress adjusted outcomes $\tilde{Y}_{it} - \tilde{X}'_{it}\hat{\gamma}$ on D_{it} to obtain $\hat{\beta}$.

For these regressions to be feasible, we need to assume a rank condition such that $\left(\sum_{i=1}^{N} \tilde{X}'_{0i}\tilde{X}_{0i}\right)$ is invertible. This invertibility condition rules out identification of unit and time fixed effects separately in environments where treatment cohorts are too small and we are using too few periods.

Assumption 2 (Rank condition). *$E[\tilde{X}'_{0i}\tilde{X}_{0i}]$ and $E[D_{it}]$ are invertible.*

Inference can then proceed by standard GMM arguments. To apply limit theorems, it suffices to assume that the moments of stochastic objects are bounded.

Assumption 3 (Bounded moments). *There exists $C < \infty$ such that $E[Y_{it}(0)^4] \leq C$, $E[X_{kit}^8] \leq C$, and $E[\beta_{it}^4] \leq C$ for all t.*

The first- and second-stage regressions can be written as the solution to the sample analog of the following moment conditions:

$$E\left[\begin{array}{c} \tilde{X}'_{0i}\left(\tilde{Y}_{0i} - \tilde{X}_{0i}\gamma\right) \\ \sum_{t=1}^{T} D_{it}\left(\tilde{Y}_{it} - \tilde{X}'_{it}\gamma - \beta D_{it}\right) \end{array}\right] = 0,$$

where we have $K+1$ moment conditions, with K in the first stage and one in the second stage.[14]

Theorem 1. *If Assumptions 1 to 3 hold, then $\hat{\gamma} \xrightarrow{p} \gamma$, $\hat{\beta} \xrightarrow{p} \beta$ and $\sqrt{N}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N(0, V)$, where $V = G_\beta^{-1} E\left[\left(g_i + G_\gamma \psi_i\right)\left(g_i + G_\gamma \psi_i\right)'\right] G_\beta^{-1\prime}$, with*

$$G_\beta = -E\left[\sum_{t=1}^{T} D_{it}\right],$$

[14] Note that the joint solution to this GMM problem is numerically identical to the "manual" two-stage procedure described above.

$$
\begin{aligned}
G_\gamma &= -E\left[\sum_{t=1}^{T} D_{it}\tilde{X}_{it}\right], \\
\psi_i &= E\left[\tilde{X}_{0i}'\tilde{X}_{0i}\right]^{-1}\tilde{X}_{0i}'\left(\tilde{Y}_{0i} - \tilde{X}_{0i}\gamma\right),
\end{aligned}
$$

and

$$
g_i = \sum_{t=1}^{T} D_{it}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\gamma - \beta D_{it}\right)
$$

The sample analogs are:

$$
\begin{aligned}
\hat{G}_\beta &= -\frac{1}{N}\sum_i\sum_t D_{it}, \\
\hat{G}_\gamma &= -\frac{1}{N}\sum_i\sum_t D_{it}\tilde{X}_{it}, \\
\hat{\psi}_i &= \left[\frac{1}{N}\sum_i \tilde{X}_{0i}'\tilde{X}_{0i}\right]^{-1}\left(\tilde{X}_{0i}'\left(\tilde{Y}_{0i} - \tilde{X}_{0i}\hat{\gamma}\right)\right),
\end{aligned}
$$

and

$$
\hat{g}_i = \sum_t D_{it}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\hat{\gamma} - \hat{\beta}D_{it}\right).
$$

With $\hat{V} = \hat{G}_\beta^{-1}\frac{1}{N}\sum_i\left[\left(\hat{g}_i + \hat{G}_\gamma\hat{\psi}_i\right)\left(\hat{g}_i + \hat{G}_\gamma\hat{\psi}_i\right)'\right]\hat{G}_\beta^{-1}$*, we have* $\hat{V} \xrightarrow{p} V$.

Theorem 1 states that the 2SDD estimator is consistent for β, and is asymptotically normal. Further, the plug-in variance estimator is consistent, which suffices for feasible inference. The proof in Appendix C proceeds by standard arguments (see Newey and McFadden 1994).

2.3 Event studies

DD analyses are often accompanied by event-study regressions of the form

$$
Y_{it} = \lambda_{g(i)} + \alpha_t + \sum_{r=-\underline{R}}^{\overline{R}} \eta_r W_{rit} + u_{it}, \tag{5}
$$

where for $r \leq 0$ the $W_{rit} \in \{W_{-\underline{R}it}, \dots, W_{0it}\}$ are $(r+1)$-period leads of treatment adoption, and for $r > 0$ the $W_{rit} \in \{W_{1it}, \dots, W_{\overline{R}it}\}$ are r-period lags of adoption (i.e., indicators for being r periods since treatment). SA2021 show that, when duration-specific average treatment effects vary across groups, event-study regressions suffer from the same problem as

DD regressions.[15]

The two-stage procedure developed above can be extended to the event-study setting by amending the second stage of the procedure to:

2'. Regress $Y_{it} - \hat{\lambda}_{g(i)} - \hat{\alpha}_t$ on $W_{-\underline{R}it}, \cdots, W_{0it}, \cdots, W_{\overline{R}it}$.

Following the logic of the previous section, because $\mathbb{E}[Y_{it} \,|\, g, t, (W_{rit})] - \lambda_g - \alpha_t$ is linear in the W_{rit}, the coefficients on the W_{rit}, $r > 0$, identify the average effects $\mathbb{E}[\eta_{rit} \,|\, W_{rit} = 1]$.[16] For $r \leq 0$, the coefficients on the W_{rit} can be used to test the hypothesis that $\mathbb{E}[Y_{it} \,|\, g, t, W_{rit} = 1] = \lambda_g + \alpha_t$ (i.e., that the mean first-stage population residual is zero for all units who are $r + 1$ periods away from adopting the treatment), as implied by parallel trends. Note that, by the same logic, the treatment-duration indicators in step 2' can be replaced with group- or period-specific treatment-status indicators in order to identify group- or period-specific ATTs.

To formalize and extend this argument, let $t^*(i)$ denote the time at which individual i becomes treated, and let $W_{rit} = 1\left[t - t^*(i) = r\right]$ denote whether individual i is r periods away from treatment at time t, with $r \in \{-\underline{R}, \cdots, \overline{R}\}$, where we assume that there is an $\underline{R}^*$ such that parallel trends holds for all observations $t - t^*(i) < -\underline{R}^*$. Researchers typically assume that $\underline{R}^* = 0$, so that parallel trends holds in all pre-treatment periods. Allowing $\underline{R}^* > 0$ allows us to relax the assumption that the treatment is unanticipated, and as we discuss below, provides us with one method of testing whether parallel trends is satisfied.

The elements of W_{rit} can be stacked into an $\underline{R} + \overline{R} + 1$-dimensional vector W_{it}, where w is a potential value of W_{it}. To adapt the reasoning from the DD setup, we have potential outcomes $Y_{it}(w) = w'\eta_{it} + Y_{it}(0)$, where w is a vector of treatment-duration indicators.[17]

The object of interest is now $\eta = (\eta_{-\underline{R}}, \cdots, \eta_{\overline{R}})$, where $\eta_r := E[\eta_{rit} \mid t - t^*(i) = r]$ is the average causal effect across individuals who are r periods away from receiving the treatment. For $r > 0$, η_r can be interpreted as the treatment effect r periods after treatment. Since our parallel trends assumption implies that $\eta_r = 0$ for all $r < -\underline{R}^*$, the estimated η_r, $r < -\underline{R}^*$ provide a test of whether the data are consistent with this assumption, even in the presence of $\underline{R}^*$ periods of treatment-anticipation effects. Below, we discuss several alternative approaches to testing parallel trends.

[15] An argument similar to the one presented for DD regressions in Section 2.1 can also be used to show that the coefficients on the W_{rit} do not identify the average effect of being treated for r periods.

[16] This expectation is taken over all groups with treatment durations of at least r. Since under staggered adoption the completed treatment duration varies by group, the groups over which these duration-specific effects are averaged will vary across durations. These averages are also what the interaction-weighted estimator proposed by SA2021 identifies. If all groups are treated for at least $\bar{P}$ periods, an alternative is to exclude observations corresponding to treatment durations longer than $\bar{P}$ periods from the second-stage sample, in which case the two-stage approach identifies duration-specific treatment effects, averaged over all groups.

[17] Here, $Y_{it}(0)$ should be understood as the potential outcome for $t - t^*(i) < -\underline{R}^*$. Then, the parallel trends assumption as stated in Assumption 1 can still be used.

Define $Q_{it} := 1\left[t - t^*(i) < -\underline{R}^*\right]$. Then, by analogy to the case for the overall ATT, define $T_i^Q := \sum_{t=1}^T Q_{it}$. Now,

$$\tilde{Y}_{it} = Y_{it} - \frac{1}{T_i^Q} \sum_{t=1}^{T_i^Q} Y_{it} Q_{it},$$

with a similar definition applying to $\tilde{X}_{it}$ and $\tilde{\varepsilon}_{it}$. Analogously,

$$\tilde{X}_{Qi} = \begin{bmatrix} \left(X_{1i1} - \frac{1}{T_i^Q}\sum_{t=1}^T X_{1it}Q_{it}\right) Q_{i1} & \cdots & \left(X_{Ki1} - \frac{1}{T_i^Q}\sum_{t=1}^T X_{Kit}Q_{it}\right) Q_{i1} \\ \vdots & & \\ \left(X_{1iT} - \frac{1}{T_i^Q}\sum_{t=1}^T X_{1it}Q_{it}\right) Q_{iT} & \cdots & \left(X_{KiT} - \frac{1}{T_i^Q}\sum_{t=1}^T X_{Kit}Q_{it}\right) Q_{iT} \end{bmatrix},$$

and

$$\tilde{Y}_{Qi} = \left[\tilde{Y}_{i1}Q_{i1} \quad \cdots \quad \tilde{Y}_{iT}Q_{iT}\right].$$

In this environment, our analogous two-stage procedure becomes:

1. Regress $\tilde{Y}_{Qi}$ on $\tilde{X}_{Qi}$ to obtain $\hat{\gamma}$.
2. Regress adjusted outcomes $\tilde{Y}_{it} - \tilde{X}_{it}'\hat{\gamma}$ on W_{it} to obtain $\hat{\eta}$.

Hence, the first- and second-stage estimators are

$$\hat{\gamma} = \left(\frac{1}{N}\sum_i \tilde{X}_{Qi}'\tilde{X}_{Qi}\right)^{-1}\left(\frac{1}{N}\sum_i \tilde{X}_{Qi}'\tilde{Y}_{Qi}\right),$$

and

$$\hat{\eta} = \left(\sum_{i=1}^N\sum_{t=1}^T W_{it}W_{it}'\right)^{-1}\left(\sum_{i=1}^N\sum_{t=1}^T W_{it}\left(\tilde{Y}_{it} - \tilde{X}_{it}\hat{\gamma}\right)\right).$$

As before, the estimators can be written as the solution to a GMM problem:

$$E\begin{bmatrix} \tilde{X}_{Qi}'\left(\tilde{Y}_{Qi} - \tilde{X}_{Qi}\gamma\right) \\ \sum_{t=1}^T W_{it}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\gamma - W_{it}'\eta_{it}\right) \end{bmatrix} = 0$$

The normality and consistency results are analogous.

2.4 Alternative approaches to testing parallel trends

There are alternative approaches to testing the validity of parallel trends within the two-stage framework. BJS2024 recommend testing for parallel trends by including leads of treatment status in the first stage of the estimator, noting that their approach can, under some conditions, circumvent concerns about conditioning difference-in-differences estimates on passing tests for parallel trends (note that inference in this approach is based on standard OLS asymptotics). Another approach is to assume that parallel trends holds up to $\underline{R}^* + 1$ periods before the adoption of the treatment, then use the two-stage procedure to estimate the $\underline{R}^*$ pre-treatment placebo ATTs (i.e., the coefficients on D_{rit} for $r \in \{-\underline{R}^*, \dots, -1\}$). This approach is also suggested by Liu, Wang and Xu (2022), who develop an equivalence test to increase the power of tests based on this idea.[18]

The two-stage framework suggests another approach still, this one motivated by the fact that it is not necessary to use all pre-treatment periods to identify the group (or individual) and time effects used by the second stage of the estimator. For example, instead of using all untreated observations, the first stage can be estimated from the sample of all observations for never-treated units (from which the period effects and group effects for never-treated units are identified) as well as all observations for eventually-treated units in the period immediately before they adopt the treatment (from which the group effects for treated units are identified).[19] Under the normalization that parallel trends holds in the last pre-treatment period (i.e., that eventually-treated units experience the same time effects in that period as never-treated units), the coefficients on the D_{rit} for $r \in \{-\underline{R}, \dots, -1\}$ for this variant of the two-stage procedure identify average pre-treatment deviations among eventually-treated units from never-treated units' trends.[20] Although this restriction of the first-stage sample may reduce the efficiency of the second-stage estimates, it addresses some of the challenges associated with interpreting coefficients that represent tests of parallel trends from within the two-stage framework (cf. footnote 18 and Roth 2024). In Figure A1, we show that two-stage estimates obtained using this modified procedure correctly identify both pre- and post-treatment trends in the setting where Roth (2024) shows that the default dCDH2020, CS2021, and BJS2024 estimators do not. While the coefficients on leads of treatment status from this modified procedure are more readily interpretable, the analogous coefficients from

[18] While all of the methods discussed above are capable of identifying violations of parallel trends, none of them reliably identify parameters that can be interpreted as average deviations from trends in pre-treatment periods. Second-stage coefficients on leads of treatment status test whether average first-stage residuals are close to zero in pre-treatment periods, first-stage coefficients on such leads presumably identify a (potentially non-convex) weighted average of deviations from trend for all groups and periods, and placebo ATTs only represent such deviations under the assumption that parallel trends holds prior to the adoption of the placebo treatment. This contrasts with traditional event-studies based on two-way fixed-effects regressions with homogeneous duration-specific average treatment effects, in which the coefficients on leads can be interpreted as average deviations from trends, subject to a normalization.

[19] The last treated cohort can be used as the never-treated cohort in the absence of a pure control group.

[20] The normalization required here is the same as that required for traditional two-way fixed-effects event studies.

the "standard" two-stage approach (i.e., using the full untreated sample in the first stage) still represent valid tests of parallel trends, even if they cannot be interpreted as average deviations from never-treated trends.

A further advantage of this modified procedure is that it may offer superior performance in cases when the divergence between untreated outcomes between eventually- and never-treated units increases over time (an advantage that dCDH2024 argue is shared by other estimators that do not compare treated observations to all untreated observations).

2.5 Discussion of assumptions and extensions

The simplicity of our regression-based estimation and inference procedure allows us to flexibly incorporate several extensions and relaxations of the assumptions. We provide sample `Stata` code in Appendix A both using the in-built `gmm` command and the `did2s` package.

Other average treatment effects. While the discussion so far has focused on identification of the overall and duration-specific ATTs, as these are the most common objects of interest in difference-in-difference analyses, the logic behind our two-stage approach also applies to other average treatment effect measures. In particular, it can be used to identify group×time-specific ATTs by amending the second stage to include a full set of group×time indicators, which can be examined individually or aggregated to form further summary measures of the treatment. Similarly, modifying the second stage to include full sets of group or time indicators identifies group- or time-specific ATTs.

Parallel trends assumption. The theoretical results presented above assume parallel trends for every group and between every pair of consecutive time periods, as in dCDH2020 and SA2021. This assumption is stronger than the one used in CS2021, who only require parallel trends between treated groups and never-treated groups after their treatment time. While this accommodates cases where parallel trend fails prior to treatment time, the distinction may not matter in practice, as testing for parallel trends prior to treatment time is often used as a proxy for the infeasible test for parallel trends post treatment. If the stronger version of parallel trends fails, researchers tend to have little confidence in the weaker version.

Nevertheless, our procedure can be modified to accommodate the weaker version. If we only believe in the weak version of parallel trends, our procedure can be modified to estimate the first stage using only never-treated observations and the first pre-treatment period for eventually-treated units (which is necessary in order to estimate the unit FEs for treated units). In contrast, even if the stronger version of parallel trend holds, the CS2021 approach does not yield a more precise estimate because it does not make use of data from treated observations before relative time -1.

Our approach can be adapted to reduce bias when the parallel trends assumption is violated. The potential for a larger bias arises if the parallel trends assumption does not hold exactly and the difference in trends between groups increases over time. To reduce the bias, we can simply estimate the first stage using untreated data within a few periods of being treated. Group-specific linear trends may also be included in the regression-based approach to remove the group trends directly. In particular, X_{it} in Equation (3) may include $1\{g(i) = g\}t$, where $g(i)$ denotes the group to which observation i belongs.

Triple differences. Triple differences-in-differences use the evolution of observed outcomes for auxiliary untreated units to control for potential violations of parallel trends, and can easily be accommodated from within our two-stage framework. For example, if only members of states that belong to a particular subgroup are treated, then the first-stage regression can be modified to include state, time, and state×time, state×subgroup and time×subgroup fixed effects, which can be collected into the vector X_{it}. Under the hypothesis that state-level deviations from parallel trends are identical across subgroups within states, two-stage estimates will recover the overall ATT.

Serial correlation. It is possible to adapt the procedure to obtain an efficient estimator even in the presence of serial correlation. Since the estimator is identical to the imputation estimator of BJS2024, it is known that the estimator is efficient in the canonical normal homoskedastic model. If there is serial correlation in the error term following an AR(1) process for each i, we can make a simple adjustment to the regression. If ε_{it} in Equation (3) is AR(1) with correlation parameter ρ, then Equation (3) can be written as $Y_{it} = \rho Y_{it-1} + \delta_{it} D_{it} + \tilde{\lambda}_i + \tilde{\gamma}_t + \nu_{it}$, with $\nu_{it} | D_{it}, Y_i^{t-1} \sim \mathcal{N}(0, \sigma^2)$ for an appropriately defined $\tilde{\lambda}_i, \tilde{\gamma}_t$ and $Y_i^{t-1} := \{Y_{i1}, Y_{i2}, \cdots, Y_{it-1}\}$. Our procedure can be analogously implemented by: (1) regressing Y_{it} on Y_{it-1}, X_{it}, and fixed effects for observations with $D_{it} = 0$, (2) regressing $Y_{it} - \hat{\rho} Y_{it-1} - (1 - \hat{\rho}) \widehat{\lambda_{g(i)}} - \widehat{\alpha_t} + \hat{\rho} \hat{\alpha}_{t-1}$ on D_{it}. Since the OLS estimator coincides with the maximum likelihood estimator, the resulting estimator is efficient under the normal homoskedastic benchmark. While the estimated coefficients are not the coefficients of interest due to the Nickell bias, they can be combined to recover the coefficients of interest in a manner detailed in the appendix.[21]

Continuous and multivalued treatments. Our approach to estimation and inference extends to continuous treatments and discrete (non-binary) treatments. In this setting, observations have $D_{it} = 0$ prior to treatment, but the treatment value may be continuous post-treatment. The two-stage procedure still applies, with the first-stage consisting of a regression of the outcome on X for $D_{it} = 0$ to obtain $\hat{\gamma}$, and the second stage consisting of

[21] While Harmon (2024) emphasizes the efficiency advantages of alternative approaches under strongly correlated errors, the procedure outlined here highlights how our estimator can be adapted to such settings.

a regression of $Y_{it} - X_{it}'\hat{\gamma}$ on D_{it} to obtain $\hat{\beta}$. With a continuous treatment, the two-stage procedure identifies a (positive) Yitzhaki-weighted average of the derivatives of the causal response function (Yitzhaki, 1996; Angrist and Krueger, 1999; Angrist and Pischke, 2009). Inference proceeds through GMM as before.[22]

Anticipation effects. The procedure can be extended to accommodate anticipation effects. If the treatment is anticipated for r periods before adoption, we can redefine treated to mean having adopted the treatment for at least r periods.

Reversible treatment and several treatments. The procedure can be extended to accommodate reversible treatment and having several treatments. If the treatment is reversible, one way to apply our results is to use the (potentially strong) assumption that there are no within-unit spillovers of the treatment to future periods. Alternatively, if there are within-unit spillovers of the treatment to future periods, we can define W_{it} in Section 2.3 as a vector of indicators for the treatment path, which is defined as the sequence of treatment indicators since first treatment.[23] Observations that have been treated prior to t are excluded from the first stage. The asymptotics hold when there are many observations with the same treatment path. The estimands remain interpretable as the corresponding coefficients are effects relative to the untreated group. If there are several treatments, say D_1 and D_2, then we can similarly define each treatment path as a tuple of the treatment duration of (D_1, D_2), so W_{it} is a vector of indicators for every combination of these tuples. The rest of the procedure and interpretation are identical to that of having reversible treatment with within-unit spillovers.

Design-based analysis. The 2SDD estimand is also interpretable in a design-based world. Let A_{1i} denote the number of periods that individual i has been treated, with $A_{1i} = 1$ in the period that i was first treated. Further, assume that $\beta_{it} = \beta_i$ for all t. Define the estimand as $\beta := \mathbb{E}[\beta_{it} \mid D_{it} = 1]$. Using an argument similar to the proof of Lemma C.1 in Appendix C, $\beta = \text{plim}\left(\sum_{i=1}^N \sum_{t=1}^T D_{it}\right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T D_{it}\beta_{it}\right)$. Hence, in the setting with staggered treatment adoption, $\beta = \text{plim}\left(\sum_{i=1}^N A_i\right)^{-1} \left(\sum_{i=1}^N A_i\beta_i\right)$. Under the Athey and Imbens (2022) setup where the adoption time of treatment is as good as random, A_i is randomly assigned across individuals in our setting, so $\frac{1}{N}\sum_{i=1}^N A_i \xrightarrow{p} a$, and $\mathbb{E}[A_i] = a$ for all i.[24] Thus, the estimand becomes $\beta = \frac{1}{aN}\mathbb{E}\left[\sum_{i=1}^N A_i\beta_i\right] = \frac{1}{aN}\sum_{i=1}^N \mathbb{E}[A_i]\beta_i = \frac{1}{N}\sum_{i=1}^N \beta_i$,

[22] When using this approach to approximate the causal response to a multivalued treatment, the second-stage regression should include a constant term.

[23] For instance, $(0, 1, 0, 1)$ and $(0, 1, 1, 1)$ are two different treatment paths. While both groups begin being treated in the second period, the only first group becomes untreated in the third period. The coefficient on the third and fourth periods are then allowed to be different for the two groups to accommodate the different treatment paths.

[24] The difference in identifying assumptions not only affects the interpretation of the estimand, but also affects inference. The design-based environment models the assignment process A_i instead of $Y_i(0)$, so

which is exactly the average treatment effect (ATE). Inference is straightforward due to the GMM framework: it suffices to cluster on i since A_i is randomly assigned.

Linear combination of treatment effects. The procedure can also be extended to estimating any linear combination of coefficients. Recall that we have the model $Y_{it} = D_{it}\beta_{it} + X_{it}'\gamma + \varepsilon_{it}$ with $\mathbb{E}\left[\varepsilon_{it} \,\middle|\, \{D_{it}, X_{it}\}_{t=1}^{T}\right] = 0$. This model implies that $\mathbb{E}[Y_{it} - D_{it}\beta_{it} - X_{it}'\gamma] = 0$. We are interested in $\tau := w_{it}\beta_{it}$, where w_{it} is a nonstochastic weight. Due to the moment condition, and w_{it} being nonstochastic,

$$\mathbb{E}[w_{it}Y_{it} - w_{it}X_{it}'\gamma] - \mathbb{E}[D_{it}]w_{it}\beta_{it} = 0.$$

Assume that heterogeneity in $\mathbb{E}[D_{it}]$ occurs at some level h, and ζ is the vector of values it can take, so that $\mathbb{E}[D_{it}] = 1(h)_{it}'\zeta$. Assume that ζ is either known or can be consistently estimated, and all elements of ζ are nonzero. Then, by summing $w_{it}\beta_{it} = \mathbb{E}[w_{it}Y_{it} - w_{it}X_{it}'\gamma]/\mathbb{E}[D_{it}]$ over i, t:

$$\tau = \sum_{i,t} w_{it}\beta_{it} = \sum_{i,t} w_{it}\mathbb{E}\left[\frac{Y_{it} - X_{it}'\gamma}{1(h)_{it}'\zeta}\right]$$

Hence, writing everything as a system of moment conditions,

$$\mathbb{E}\begin{bmatrix} 1(h)_{it}\left(D_{it} - 1(h)_{it}'\zeta\right) \\ X_{it}\left(1 - D_{it}\right)\left(Y_{it} - X_{it}'\gamma\right) \\ \tau - w_{it}\left(\frac{Y_{it} - X_{it}'\gamma}{1(h)_{it}'\zeta}\right) \end{bmatrix} = 0$$

The just-identified system of equations enables the application of GMM in the same way as before.

Test for treatment effect heterogeneity. The GMM approach also allows us to test for treatment effect heterogeneity. One approach is to use the fact that, under the null of constant treatment effects, the two-way fixed effects estimator has the same probability limit as the 2SDD estimator that is robust to heterogeneity. Then, we can test if the two estimators are equal. An alternative approach is to obtain coefficients in the second stage that correspond to group (or covariate-specific) treatment effects. We can then use a standard F test to jointly test if the coefficients are equal.

parallel trends is no longer required to interpret β as the ATE. For inference in the design-based environment, if there is clustered assignment for A_i on dimension C where several individuals i can belong to some cluster c, then researchers should cluster at the c level instead of i. However, if we instead assume parallel trends in our original environment, even if there is clustered assignment, it suffices to cluster on unit i if we are targeting a conditional estimand (i.e., the ATT): by conditioning on D_{it}, the dependence structure is irrelevant for inference.

3 Comparison with other approaches to inference

The remaining sections of the paper compare 2SDD against several leading estimators frequently used in empirical research. We begin with the simplest possible setting that underlies standard difference-in-differences logic—a 2 × 2 design—and show that 2SDD produces exactly the same variance as the classical TWFE estimator. While 2SDD does not sacrifice any precision in the most basic environment, we demonstrate that other heterogeneity-robust procedures perform strictly worse under this canonical benchmark. We also clarify the relationship between 2SDD and the BJS2024 imputation estimator.

3.1 Equivalence with TWFE and improved precision over alternatives in the 2×2 case

We provide a simple example to demonstrate that none of the other commonly used DD methods produce the same confidence interval as TWFE in canonical 2 × 2 designs. While TWFE and our proposed two-stage approach yield identical estimates and confidence intervals, other popular methods produce confidence intervals that are strictly wider.

We construct a minimal 2 × 2 example with four units observed for two periods, with two units receiving treatment in the second period, and we choose the outcome data so that the treatment effect is 1 and the variance estimate using TWFE (clustering standard errors at the unit level) is 1.[25] To examine the behavior of confidence intervals when the sample grows, we replicate the dataset either 1, 10, or 100 times and add a random normal error with standard deviation 0.05 to the outcome variable for each observation.

We apply several recently proposed heterogeneity-robust DD approaches (CS2021; SA2021; W2021; dCDH2024) and present the 95% confidence intervals in Table 1.[26] The TWFE and 2SDD confidence intervals are identical for all three different expansions of the original data. By contrast, the other heterogeneity-robust methods entail efficiency losses relative to TWFE even in the basic 2 × 2 case.

We show theoretically that the TWFE and 2SDD variances are asymptotically identical in this setting, with full details provided in Appendix D.[27]

[25] Units 1 and 2 are never treated, while units 3 and 4 eventually receive treatment. Letting y_{it} denote the outcome for unit i in period t, we specify that $y_{11} = 2$, $y_{12} = 0$, $y_{21} = y_{22} = 3$, $y_{31} = 1$, $y_{32} = 0$, $y_{41} = 2$, and $y_{42} = 3$.

[26] We do not include the imputation estimator from BJS2024 in the table because it coincides with 2SDD in this case. Section 3.2 compares the two estimators more generally, highlighting how the default BJS2024 variance estimator is anti-conservative if are treatment cohorts that are not very large.

[27] Proofs of numerical equivalence of the coefficient and variance estimators under homoskedasticity are also available upon request.

3.2 Relationship with imputation estimator

We conclude this section with a brief discussion of how our approach to inference differs from that developed in BJS2024, whose imputation estimator produces numerical point estimates to our two-stage estimator.

First, BJS2024 primarily consider a setting in which treatment status is non-stochastic (fixed in repeated samples) and the treatment effects are also non-stochastic. Consequently, the variance of their target parameter is smaller than the variance of ours, and, accordingly, they use a different variance estimator. When treatment status is in fact random, their default variance estimator is no longer appropriate, while our approach already accounts for the potential stochasticity of treatment status.

Second, even in settings where treatment status is fixed, our approach is more robust to small-sample issues—particularly those arising from having very few observations per cohort. While both variance estimators achieve consistency as the sizes of the treatment cohorts grow without bound, the small-sample performance of the variance estimators can differ regardless of whether treatment status is fixed or random. The standard errors for the imputation estimator developed in BJS2024 are constructed based on the residuals $\tilde{\varepsilon}_{it} = \hat{\tau}_{it} - \hat{\hat{\tau}}_{it}$, where $\hat{\tau}_{it}$ is the estimated treatment effect for unit i at time t and $\hat{\hat{\tau}}_{it}$ is some average of these estimated individual treatment effects. While they consider different values for $\hat{\hat{\tau}}_{it}$, their recommended default is to use cohort-period averages.[28] When the groups are small, their variance estimator can therefore become anti-conservative (which results in size distortions); in the extreme case with a group consisting of only one observation, $\hat{\hat{\tau}}_{it} = \hat{\tau}_{it}$, so $\tilde{\varepsilon}_{it} = 0$. This problem is not alleviated even with the leave-one-out modification to their variance estimator: with one observation in a group, there are no observations to calculate $\hat{\hat{\tau}}_{it}$. In contrast, our approach uses $\hat{\hat{\tau}}_{it} = \hat{\beta}$, which does not share the same small-sample problem.

4 Rejection rates for randomly generated interventions

This section conducts Monte Carlo simulation exercises inspired by Bertrand, Duflo and Mullainathan (2004) to evaluate the two-stage approach and provide insight into how various difference-in-differences (DD) methods perform under realistic conditions. First, we aim to assess finite-sample performance in environments that resemble common empirical applications. Second, acknowledging that theoretical frameworks often rely on the assumption of i.i.d. data, we simulate scenarios that incorporate autocorrelation and reflect real-world datasets more accurately. Third, the proliferation of recently proposed alternatives for DD estimation necessitates a comparative analysis to discern their relative strengths and

[28] To be precise, $\hat{\hat{\tau}}_{it} = \left(\sum_{i:g(i)=g} \sum_{t:p(t)=p} \hat{\tau}_{it}\right) / \left(\sum_{i:g(i)=g} \sum_{t:p(t)=p} 1\right)$.

weaknesses. Lastly, since the BJS2024 method shares point estimates with ours, it becomes essential to assess the distinct approaches to inference.

4.1 Data and methodology

Our primary dataset consists of wage data for women between the ages of 25 and 50 from the Current Population Survey (CPS). We define wage as the natural logarithm of weekly earnings, which are recorded in the fourth interview month in the Merged Outgoing Rotation Group of the CPS.[29] The data span a 42-year period from 1979 to 2020 and contain over one million women reporting strictly positive weekly earnings. Using data from 50 states, we construct a state-by-year panel dataset comprising average wages in 2,100 state-year cells for our Monte Carlo exercises. In such environments, the theoretical results of BJS2024 regarding efficiency, which also hold for our estimator, may not apply (though see our discussion in Section 2.5). In addition, we generate an i.i.d. dataset by drawing the outcome variable from a normal distribution with the same mean and variance as wages in our CPS sample.

Our simulation study adopts a "random design" strategy. This approach introduces stochasticity by randomly drawing treated states, treatment effects, and treatment timing in each iteration. By doing so, we create a more realistic representation of real-world scenarios where the assignment of treatments may not follow a fixed pattern (Athey and Imbens, 2022). Importantly, we also document some inherent limitations of considering treatment and treatment timing as non-stochastic as in the "fixed design" approach of BJS2024.[30]

To simulate a staggered treatment setting, we randomly assign states to the treatment group and generate treatments that occur randomly over a specified period. This contrasts with the original exercise by Bertrand, Duflo and Mullainathan (2004), in which the placebo treatment timing is homogeneous across treated states and drawn uniformly at random. In all cases, we restrict the earliest treatment year to 1982 and the latest treatment year to 2014. Since treatment is an absorbing state, this ensures that we observe outcome data in all treated states for at least 5 years after the treatment event.

We estimate the effects of the randomly generated interventions using the two-stage approach (with our analytical standard errors) as well as a number of alternative methods for comparison. In particular, we consider the imputation approach from BJS2024, using both their "default" asymptotically conservative standard errors and "leave-out" version with improved finite-sample performance, as well as various alternative estimators (CS2021;

[29] Using the logarithmic transformation excludes women with zero weekly earnings. While many recent papers use quasi-logarithmic transformations to incorporate zero-valued observations, Thakral and Tô (2023) document substantial biases arising from the use of such transformations, and thus we focus on women with strictly positive earnings following Bertrand, Duflo and Mullainathan (2004).

[30] See Appendix E for further discussion, though we note that our conclusions do not require random designs.

SA2021; dCDH2024; W2021).[31] Standard errors are adjusted for clustering at the state level, following Bertrand, Duflo and Mullainathan (2004).

4.2 Simulation results

We conduct an event-study analysis to estimate the effect of the randomly generated interventions in each of the five years starting from the time of treatment. The primary measure we use to evaluate the performance of each method is the relative frequency of rejecting the null hypothesis of the true generated effect size at the 5 percent significance level over 500 simulations. We also report the mean bias, root-mean-square error (RMSE), and average per-simulation computational speed.

The baseline environment consists of states being treated over a 20-year period, which corresponds to an empirical example highlighted in the recent Miller (2023) guide to event-study models (the impact of state-level school finance reforms in 26 states from 1990–2011 from Lafortune, Rothstein and Schanzenbach, 2018). However, we consider 40 treated states in our baseline environment and ensure at least 2 treated states per year, with the goal of providing the BJS2024 approach to inference with a more balanced assessment since computing their leave-out variance estimator requires that no treatment cohort consists only of a single state. Treatment effects are heterogeneous and drawn from a normal distribution, with an average value randomly drawn between 2 percent and 5 percent of the average wage and a standard deviation equal to 10 percent of the average wage.

Table A1 reports results from the baseline environment, in which the average true effect is approximately 0.2. Our proposed two-stage method with the GMM approach to inference leads to rejection rates near 5 percent, with standard errors around 0.10. Despite having the same point estimates, the default BJS2024 variance estimator leads to the most substantial levels of over-rejection, ranging from 13 percent to 16.8 percent, with standard errors around 0.08. Their leave-out variance estimator, on the other hand, leads to overly conservative estimates, with rejection rates around 1 percent and standard errors around 0.14. Compared to the leave-out variance estimator, the SA2021 method leads to similar rejection rates with a larger standard error (around 0.17) and the CS2021, dCDH2024, and W2021 methods result in similar standard errors (around 0.14) but achieve rejection rates closer to 5 percent.

The two-stage approach and the imputation approach share a speed advantage, outper-

[31] We conduct these analyses in Stata using the packages `did2s` (Butts, 2021), `did_imputation` (Borusyak, 2021), `csdid` (Rios-Avila, Sant'Anna and Callaway, 2023), `eventstudyinteract` (Sun, 2021), `did_multiplegt_dyn` (de Chaisemartin et al., 2023), and `jwdid` (Rios-Avila, Nagengast and Yotov, 2022). For CS2021, we use the asymptotic standard errors (the default in the Stata package, obtained via influence functions), though we note that the wild bootstrap generally yields wider confidence intervals. With the exception of `jwdid`, the authors of the respective methodological papers were directly involved in the development of the packages.

forming most alternatives by a factor of 100 or more. This highlights the simplicity of the two-stage estimator, which can be computed straightforwardly using OLS regressions, and the advantage of having analytical standard errors based on the familiar GMM approach to inference.

Next we alter the baseline environment to allow for a correlation between treatment assignment and untreated potential outcomes. In particular, we use the treatment assignment probability model from Arkhangelsky et al. (2021) to construct an environment that "reflect[s] actual differences across states with respect to important economic variables." The 2SDD and CS2021 approaches continue to yield rejection rates around 5 percent, though the latter results in larger standard errors, as Table A2 shows. To emphasize the nontrivial nature of this robustness exercise, note that the change in treatment assignment mechanism can meaningfully impact rejection rates: the default BJS2024, dCDH2024, and W2021 methods all result in higher rejection rates than before. The BJS2024 leave-out variance estimator and the SA2021 method continue to lead to under-rejection.

To further evaluate these methods, we proceed to vary the minimum number of treated states in each year, the number of years during which the treatment can occur, and the total number of treated states. We then extend our analysis to environments with homogeneous treatment effects and i.i.d. data.

4.2.1 Size of treatment cohorts

Many datasets, such as the setting from Lafortune, Rothstein and Schanzenbach (2018), have the feature that treatment cohorts may consist of only a single treated unit. To accommodate such instances, we remove the restriction that at least two states must be treated in each period. In this case, the leave-out variance estimator from BJS2024 can no longer be computed. Aside from that, removing the restriction leads to similar results for all methods (Table S1). The same holds when using the alternative non-uniform treatment assignment probability model from Arkhangelsky et al. (2021) (Table S2). With the (overly) conservative leave-out option no longer available, over-rejection becomes a significant concern with the imputation approach.

4.2.2 Number of treatment cohorts

Table 2 shows how the results change after increasing the number of treatment cohorts to 30 from the baseline of 20. This change has little effect on two-stage approach and the CS2021 estimator, with both leading to similar rejection rates (near 5 percent) and standard errors (around 0.10 for 2SDD and around 0.14 for CS) as before. The SA2021 standard error also changes little and leads to similar rates of under-rejection as before. In contrast, the default BJS2024 variance estimator leads to even more severe over-rejection rates than before, ranging from 25 percent to 30 percent, with much smaller standard errors of around

0.06. In this case, the leave-out variance estimator cannot be computed. Additionally, the dCDH2024 and W2021 estimators lead to smaller standard errors than before (0.10 and 0.12, respectively), leading to over-rejection (rates around 20 percent and 10 percent respectively).

Decreasing the number of treatment cohorts to 15 similarly has little effect on the performance of the two-stage approach, the CS2021 estimator, and the SA2021 estimator, as Table S3 show. The default BJS2024 variance estimator continues to lead to over-rejection, though with a rejection rate of only around 10–12 percent, while the dCDH2024 and W2021 estimators lead to slightly higher rejection rates than before.

Overall, these results highlight the anti-conservativeness of the default imputation approach to inference. This can be attributed to over-fitting in finite samples. This observation also explains why the imputation default performs poorly when the number of groups increases relative to N.[32] Due to over-fitting when the group size is small, the extent of over-rejection using that approach becomes more severe if treatment timing is staggered over a longer period. In practice, we find evidence of over-rejection using the imputation default variance estimator even if the treatment is staggered over fewer periods (see Tables S4 to S6).

4.2.3 Number of treatment units

The baseline environment consists of 40 treated states. However, many empirical examples such as the Lafortune, Rothstein and Schanzenbach (2018) setting consist of fewer treated units (26 states in that case). Before proceeding, we note that BJS2024 suggest a minimum effective number of treated observations of 30 because, as their documentation states, "inference on coefficients which are based on a small number of observations is unreliable" (see the Herfindahl condition in their paper). Given the prevalence of empirical examples with smaller numbers of treated units, we evaluate the performance of the various methods in such settings to shed light on their relative strengths and weaknesses.

To hold fixed the number of treatment cohorts while ensuring that the BJS2024 leave-out variance estimator can be computed even when the number of treated states is only 30, we consider settings with 15 treatment cohorts. In all cases except for the default BJS2024 variance estimator and the W2021 estimator, when decreasing the number of treated states from 40 (Table S3) to 30 (Table S7), the standard error appreciably increases and the resulting rejection rates remain stable. These simulation results suggest that most difference-in-difference methods may still apply reliably in empirical settings with smaller numbers of treated units and, furthermore, highlight an important advantage of the GMM approach to inference.

[32] Since their default is to use $\hat{\bar{\tau}}_{it} = \hat{\bar{\tau}}_{gp}$, as G increases, the groups become finer, so $\tilde{\varepsilon}_{it} \to 0$, which underestimates the variance. This problem is avoided if the imputation method were to use the largest group available, where $\hat{\bar{\tau}}_{it} = \hat{\bar{\tau}} = \hat{\beta}$, as GMM does.

4.2.4 Homogeneous treatment effects

While the possibility of misspecification under the TWFE regression model in situations with heterogeneous treatment effects motivates the development of alternative methods for DD estimation, the case of homogeneous treatment effects provides a useful benchmark for comparing different methods. The various alternative approaches eliminate bias that arises when estimating average treatment effects in the presence of treatment effect heterogeneity with staggered treatment timing. A natural question, however, is whether the reduction in bias comes at the cost of considerably increasing the variance even when the TWFE model is correctly specified.

We therefore conduct a set of simulations in which treatment effects are homogeneous across units and time periods. In these simulations, the normal distribution from which treatment effects are drawn has an average value equal to 5 percent of the average wage, the maximum value of the range from before.

When treatment effects are homogeneous, we find that the two-stage approach performs almost as well as a TWFE estimator that imposes a null effect in the pre-treatment periods, as Table A3 shows. Both methods achieve rejection rates around 5 percent, though TWFE gives slightly smaller standard errors (an average of 0.102 instead of 0.103).[33] Since homogeneous treatment effects is a special case of our setup, the 2SDD estimand converges to the true treatment effect β, which is the same limit as TWFE.[34]

The other methods, however, are markedly outperformed by TWFE. The BJS2024 default variance estimator gives much smaller standard errors of about 25 percent smaller than under TWFE, leading to rejections of the null hypothesis of the true effect about three times as often as under TWFE. The BJS2024 leave-out variance estimator (standard error 0.14) and the SA2021 approach (standard error 0.17) reject only 20–40 percent as often as TWFE. The CS2021, dCDH2024, and W2021 estimators yield similar rejection rates as our approach and TWFE, but with a relatively large standard errors (around 0.13–0.14). The two-stage approach, in comparison, provides the most natural way to extend DD estimation to achieve robustness to treatment effect heterogeneity without much efficiency loss.

[33] In comparison, the fully dynamic event-study specification using TWFE yields an average standard error of 0.137.

[34] Due to the FWL theorem, the estimator $\hat{\beta}_{\text{TWFE}}$ is numerically identical to the result we would obtain by first regressing Y_{it} and D_{it} on $1(g)'_{it}, 1(p)'_{it}$ for all observations, and then regressing the residual of Y_{it} on the residual of D_{it}. In the first stage, $\hat{\lambda}_{\text{TWFE}} = \lambda(1 + o_P(1))$ and $\hat{\alpha}_{\text{TWFE}} = \alpha(1 + o_P(1))$, when there are homogeneous treatment effects. The 2SDD approach is similar, except that the first stage regression uses only the untreated observations, so $\hat{\lambda}_{\text{2SDD}} = \lambda(1 + o_P(1)), \hat{\alpha}_{\text{2SDD}} = \alpha(1 + o_P(1))$. Then, asymptotically, the residual generated in both procedures will be $\tilde{Y}_{it} = Y_{it} - \hat{\lambda}'1(g)_{it} - \hat{\alpha}'1(p)_{it} = Y_{it} - \lambda'1(g)_{it} - \alpha'1(p)_{it} + o_P(1)$. 2SDD and TWFE hence only differ in the second stage: TWFE regresses $\tilde{Y}_{it}$ on the residual of D_{it} while 2SDD regresses $\tilde{Y}_{it}$ on D_{it}. Since both estimators converge to the same limit, the only difference in inference is the variance.

4.2.5 I.i.d. data

The data that we use for our primary simulation exercises exhibit realistic features such as higher-order serial correlation. However, we note that the advantages of the two-stage approach do not rely on this particular feature of the data. We show this by conducting the much simpler exercise of generating i.i.d. data and comparing the performance of the different estimators.

All of our conclusions persist in the i.i.d. environment. The baseline environment (Table A4) continues to show rejection rates close to 5 percent for the two-stage approach. The same also holds for the CS2021, dCDH2024, and W2021 estimators, though with standard errors around 30 percent larger. Also as before, the default BJS2024 variance estimator leads to over-rejection (rejection rates ranging from 14.4 percent to 17.6 percent), while their leave-out variance estimator is overly conservative (rejection rates ranging from 0.2 percent to 1.8 percent), as is the SA2021 estimator. The same patterns hold in the simple case of homogeneous treatment effects (Table S8). The comparison between Tables S9 to S12 shows, as before, that a larger number of treatment cohorts leads to smaller standard errors for all methods but keeps rejection rates stable for all except the default BJS2024 variance estimator, the dCDH2024 estimator, and the W2021 estimator, for which rejection rates reach as high as 31.7 percent, 21.2 percent, and 13.2 percent, respectively. Analogously, the comparison between Tables S13 and S14 shows, as before, that decreasing the number of treated states leads to notably larger standard errors and correspondingly stable rejection rates for all methods except the default BJS2024 variance estimator (for which rejection rates increase from 10.6–12.6 percent to 14.2–19.0 percent as the number of treated states decreases from 40 to 30) and the W2021 estimator (for which rejection rates increase to 8.6–10.6 percent as the number of treated states decreases to 30).

5 Empirical applications

This section illustrates the performance of our two-stage estimator through a variety of empirical applications. In particular, we replicate all papers with variation in treatment timing and a single treatment event listed in Table 1 of SA2021 using the existing heterogeneity-robust estimators that have been published as well as 2SDD.[35] The seven papers that we reanalyze appear in Table 3, with the number of treatment cohorts ranging from 5 to 21. The

[35] As the different packages mentioned in Section 4 have different data processing requirements, we provide a Stata package `didio` to harmonize the input and output format for all of them, which can be used by other researchers interested in implementing any subset of methods. Compared to Section 4, the larger datasets and extensive set of covariates commonly used in empirical applications magnify the differences in runtime. Some methods can take many hours (SA2021) or even days (CS2021) to run for a single outcome, while the equivalent TWFE regressions run within minutes.

applications cover a range of fields including development, education, environmental, finance, health, political economy, and public economics. This allows us to assess the estimator's performance in real-world scenarios that deviate from the stylized setting in the simulations.

Going beyond our Monte Carlo analysis of wage data poses an inherent challenge because the true effects are no longer known. However, except for the two-way fixed effects (TWFE) estimator, the different methods tend to produce fairly comparable point estimates with one another (Figure 1), and all of the methods that we compare have at least some theoretical justification for their approach to inference, so we do not strictly need to observe a "true" baseline to learn about their reliability. Our discussion therefore largely focuses on the extent of agreement in terms of confidence intervals and t-statistics across the different methods, assessing which methods tend to yield outliers in repeated applications. While we provide the full set of results in the online appendix, we emphasize in the main text notable instances in which conclusions may differ depending on the choice of method.

Several patterns emerge from our analysis of the differences in results across estimators. We first investigate differences in coverage of confidence intervals, including a discussion of differences between our method and that of BJS2024. Next, we examine consistency between the t-statistics and standard error estimates across methods, highlighting instances where they disagree most. We then address differences in the pre-event coefficient estimates. Finally, we discuss how the various estimators compare with TWFE.

5.1 Event-study estimates

For the first event-study analysis in each empirical paper, we provide the corresponding estimates using the different methods in Figure 1.[36] Figure 2 reports the average standard error for each method applied to each of these papers (also see regression results in Table S15), which we discuss below. We only consider the default BJS2024 variance estimator because five out of seven of our empirical settings contain treated cohorts with only one unit.

The Bailey and Goodman-Bacon (2015) analysis of the effects of increasing access to primary care on mortality rates consists of data on more than 3,000 U.S. counties over a 40-year period, with 15 treatment cohorts and a never-treated group. The large standard errors using the CS2021 estimator create difficulty in visually discerning the differences between the other methods in Figure 1a, but Figure 2 and Table S15 make the comparison clearer. In this case, the SA2021 estimator seems to perform best, yielding the most precise estimates, followed by 2SDD and then the dCDH2024 estimator. Notably, the BJS2024 estimator in some cases produces a standard error over twice as large as that of 2SDD. This

[36] The exception is Kuziemko, Meckel and Rossin-Slater (2018), discussed in Appendix F, with estimates presented in Figure A2. The remaining estimates appear in Figures S1 to S5.

possibility can arise in finite samples when the covariance between the residual and the estimated treatment effect for unit i at time t contributes a sufficiently negative component to the BJS2024 variance estimator.[37]

Deryugina (2017) studies the effect of hurricanes on government transfers using data from over 1,000 U.S. counties over a 44-year period, with 15 treatment cohorts and a never-treated group. In this case, we find the most precise estimates using the SA2021 and dCDH2024 estimators and the least precise estimates again using the CS2021 estimator. Unlike in the previous example, for all 11 post-treatment periods and all 15 outcome variables, the BJS2024 estimator results in smaller standard errors compared to 2SDD.

He and Wang (2017) examine the impact of increased bureaucrat quality on the effectiveness of social assistance programs in rural China. The authors present a case study, including field interviews with local officials and bureaucrats, administrative records, and online surveys, as well as analyze a panel dataset consisting of a representative sample of 255 villages over a 12-year period, with between 1 and 30 villages being treated in each of 8 treatment cohorts. Only 2SDD and the BJS2024 event-study estimates provide evidence that supports the case study results by showing evidence of a significant improvement in the delivery of public services to poor households for all four outcomes. Among the other methods, the estimates from dCDH2024 support the finding of a significant effect on one outcome (increase in subsidized population), shown in Figure 1c.

Next, consider the Lafortune, Rothstein and Schanzenbach (2018) analysis of school finance reforms that largely aim for "higher spending in low-income than in high-income districts, to compensate for the out-of-school disadvantages that low-income students face." Their data consist of 49 states over a 25-year period, with 11 treatment cohorts consisting of only a single state, 6 treatment cohorts consisting of only two states, and the remaining treatment cohort consisting of only three states. While most methods agree about the resulting sustained increase in state transfers per pupil in the lowest-income districts (Figure 1d), only the BJS2024 variance estimator indicates a significant increase for the highest-income districts, and it does so for five out of the first nine years following the reform (Figure S5b). On average, the BJS2024 approach generates standard errors that are half the size of those produced by the other methods, and their conservative leave-out variance estimator remains

[37] To be precise, let v_{it} denote the weights on residual ε_{it} when constructing the variance of the coefficient, hats denote the estimators for the various coefficients, and $\hat{\tau}_{it}$ denote the treatment effect for unit i at time t estimated by the BJS2024 shrinkage method. The variance estimator of BJS2024 uses $\hat{\sigma}^2_{\text{BJS}} = \sum_i \big(\sum_t v_{it}(Y_{it} - X'_{it}\hat{\gamma} - D_{it}\hat{\tau}_{it})\big)^2$, while we use $\hat{\sigma}^2_{\text{2SDD}} = \sum_i \big(\sum_t v_{it}(Y_{it} - X'_{it}\hat{\gamma} - D_{it}\hat{\tau})\big)^2 = \hat{\sigma}^2_{\text{BJS}} + \sum_i \big(\sum_t v_{it} D_{it}(\hat{\tau}_{it} - \hat{\tau})\big)^2 + \sum_i \big(\sum_t v_{it} D_{it}(\hat{\tau}_{it} - \hat{\tau})\big)\big(\sum_t v_{it}(Y_{it} - X'_{it}\hat{\gamma} - D_{it}\hat{\tau})\big)$. Hence, the BJS2024 variance estimator can be larger when $\sum_i \big(\sum_t v_{it} D_{it}(\hat{\tau}_{it} - \hat{\tau})\big)^2 + \sum_i \big(\sum_t v_{it} D_{it}(\hat{\tau}_{it} - \hat{\tau})\big)\big(\sum_t v_{it}(Y_{it} - X'_{it}\hat{\gamma} - D_{it}\hat{\tau})\big) < 0$.

infeasible. Excluding their method, the 2SDD approach results in the smallest standard errors, followed by the CS2021 approach. We note that the estimates in Figure 2 understate the advantage of 2SDD because the table conditions on post-treatment periods when all five methods produce estimates, and the dCDH2024 approach only produces treatment-effect estimates up to 10 years after the event (11 estimates instead of 20) for all outcomes. In the periods when dCDH2024 does not produce an estimate, the difference in standard errors between 2SDD and SA2021 nearly triples, and the difference with CS2021 grows fivefold.

Tewari (2014) studies how mortgage access changed following the removal of geographic restrictions on banks using a dataset of 39 states over a 32-year period. The data consist of 20 treatment cohorts, including 13 cohorts each consisting of only a single state and 3 cohorts each consisting of only two states. The SA2021 approach provides extremely precise estimates and implies a treatment effect that fluctuates between a significant positive and significant negative effect, while the other methods always generate positive point estimates. The dCDH2024 and CS2021 approaches show some evidence supporting a significant positive effect of deregulation on homeownership. However, we note that only 2SDD is able to estimate the full set of dynamic treatment effects. In particular, the CS2021, SA2021, and dCDH2024 methods do not yield point estimates for the effect of the treatment 9 years after the event. Additionally, the CS2021 and BJS2024 methods do not yield point estimates for the effect of the treatment 9–11 years before the event, the dCDH2024 method does not yield point estimates for the effect of the treatment 6–11 years before the event. The most apparent feature of Figure 1e is the difference between the 2SDD and BJS2024 approaches in the periods preceding the event, which we discuss in Section 5.3.

Finally, Ujhelyi (2014) investigates the impact of the state-level adoption of merit-based recruitment systems for civil service on government expenditure patterns. The data consist of 48 states over a 25-year period, with 10 treatment cohorts each consisting of only a single state, 6 treatment cohorts each consisting of only two states, and the 5 remaining treatment cohorts each consisting of only three states. While the SA2021 and dCDH2024 methods give the widest confidence intervals on average, we tend to find the narrowest confidence intervals using the CS2021 method. However, the precision of the CS2021 estimator varies substantially across periods.[38] As Figure 1f shows, for the year of the introduction of the merit system, their standard error is about three times larger than that of the other methods. In comparison with CS2021, the 2SDD and BJS2024 approaches give slightly higher, but less variable, standard errors.

[38] In addition, the CS2021 method does not provide estimates for any of the periods before the event in this application.

5.2 Comparison of performance across methods

5.2.1 Synthesizing results on confidence interval coverage

In the simulations from Section 4, the 2SDD and CS2021 estimators both deliver rejection rates closest to 5 percent, albeit with a larger standard error for the latter. In Figure 2, we obtain comparable standard error estimates using the 2SDD and CS2021 estimators in four settings (He and Wang, 2017; Lafortune, Rothstein and Schanzenbach, 2018; Tewari, 2014; Ujhelyi, 2014). However, we find substantially larger standard errors using the CS2021 estimator in the remaining settings (Bailey and Goodman-Bacon, 2015; Deryugina, 2017), highlighting the merits of our approach.

We present a concise summary of the points discussed in the preceding section in Table 4, which primarily focuses on comparing standard errors as a measure of performance. While these results derive from a limited sample of empirical papers, the 2SDD estimator stands out as a practical choice for applied researchers.

The SA2021 and dCDH2024 approaches offer notable advantages when the number of groups is large (Bailey and Goodman-Bacon, 2015; Deryugina, 2017) but are outperformed by the CS2021 and 2SDD estimators with a relatively large number of small cohorts (Lafortune, Rothstein and Schanzenbach, 2018; Tewari, 2014; Ujhelyi, 2014). On the other hand, the CS2021 estimator seems to perform particularly poorly, yielding larger standard errors, when the number of groups is large. With a medium-sized number of groups (He and Wang, 2017), all of the methods seem to perform adequately. In five empirical applications (with Bailey and Goodman-Bacon, 2015 as the exception), the BJS2024 estimator produces smaller standard errors than 2SDD does.[39]

5.2.2 Consistency between standard errors

To further examine the level of consistency between the various dynamic treatment effect estimators, we compare the standard error of each event-study coefficient with the average of the standard errors across the other four methods for the same coefficient, normalized by the average standard error for that coefficient. We similarly compute, for each event-study coefficient, the difference between each method's t-statistic and its associated leave-out mean, normalized by the average of the absolute value of the t-statistics for that coefficient. Both sets of normalized differences roughly follow a normal distribution. To highlight discrepancies between the different estimators, we focus on outliers in these distributions. Outliers in the right tail of the distribution represent imprecise estimates, while outliers in the left tail suggest overly precise estimates.

[39] The differences are significant except in two settings where the sample size of estimates is small (Table S15).

Estimates for which a given method's standard error diverges from its counterparts' average standard error appear in Figure A3.[40] Each row in the figure corresponds to a single event-study coefficient for which the normalized difference falls in the top or bottom 5 percent of the distribution, along with the normalized differences for all five methods. A negative normalized difference indicates that a method produces a more precise estimate than its counterparts, while a positive normalized difference indicates the opposite. This representation shows several striking patterns. First, we find the greatest number of outliers for the CS2021 estimator, despite excluding estimates for the Bailey and Goodman-Bacon (2015) paper. Nearly all of the outliers using this method fall in the imprecise end of the distribution. 2SDD results in the fewest outliers, mostly in cases where it produces more conservative standard error estimates than other methods, rather than for producing overly precise estimates. On the other hand, for the BJS2024, SA2021, and dCDH2024 methods, most outliers arise because the estimates are unusually precise. For the SA2021 estimator, as previously noted, this occurs in part due to the overly precise standard errors for the Tewari (2014) paper. For the dCDH2024 estimator, the issue relates to its high precision in estimating short-term treatment effects, and relatively low precision in estimating longer-term treatment effects, which we highlight next.

5.2.3 Standard errors across periods

While the preceding discussions focus on the average standard error estimates across methods, we now consider how the estimates within the same method vary across time since treatment. In settings with staggered treatment timing, the presence of later-treated cohorts increases the effective sample size for estimating shorter-run treatment effects but not longer-run treatment effects. Thus, all methods exhibit less precision for treatment effect estimates over longer time horizons. To compare performance along this dimension, for each paper, we take the sample of all dynamic treatment effect estimates produced by all five methods and regress the standard errors on indicators for time since treatment, indicators for each method, and method-specific linear period trends. We report the difference between each method's linear period trend and that of 2SDD in Table A5. A positive value for a given method indicates that it produces relatively less precise estimates of longer-term treatment effects. In four of the empirical applications (Deryugina, 2017; Lafortune, Rothstein and Schanzenbach, 2018; Tewari, 2014; Ujhelyi, 2014), the dCDH2024 estimator results in significantly lower precision for longer-term effects compared to 2SDD. In three of these cases Deryugina (2017); Lafortune, Rothstein and Schanzenbach (2018); Ujhelyi (2014), the

[40] We exclude estimates from the Bailey and Goodman-Bacon (2015) paper; otherwise, that paper would account for all the outliers due to the large standard errors that arise when applying the CS2021 estimator in this setting.

SA2021 estimator also leads to significantly larger standard errors for longer-term treatment effects, and the same holds for the CS2021 estimator in the first two cases. Other than cases in which other methods yield overly precise standard errors (i.e., Lafortune, Rothstein and Schanzenbach, 2018 for Borusyak, Jaravel and Spiess, 2024, and Tewari, 2014 for Sun and Abraham, 2021), we find only two instances in which another method yields relatively greater precision for long-run treatment effects compared to 2SDD (He and Wang, 2017 and Ujhelyi, 2014 for Callaway and Sant'Anna, 2021) at the 10 percent significance level.

5.2.4 Consistency between t-statistics

To build on our discussion of the consistency between standard error estimates, we present a complementary analysis of t-statistics in Figure 3. The normalized difference between t-statistics falls in the top or bottom 5 percent of the distribution most often for the CS2021 and dCDH2024 estimators and least often for the 2SDD estimator. Using the top or bottom 1 percent of the distribution as the cutoff, the dCDH2024 and SA2021 estimators result in the most outliers. Analyzing absolute t-statistics rather than normalized differences reveals additional insights (Table A6 Panel A). Compared to the 2SDD approach, the BJS2024, SA2021, and dCDH2024 estimators produce larger absolute t-statistics on average (column 1) and a higher share of statistically significant event-study coefficients (column 2). Moreover, those estimators produce a higher share of estimates with extreme levels of statistical significance, defined using as thresholds the 90th percentile of the distribution (approximately 4.3, $p < 10^{-5}$) and the 99th percentile of the distribution (approximately 7.4, $p < 10^{-13}$) of t-statistics in our sample. The CS2021 estimator leads to significantly smaller absolute t-statistics and a significantly smaller share of significant event-study coefficients, but no significant reduction in extreme levels of statistical significance. These conclusions continue to hold if we use weights to adjust for differences in the number of periods for each outcome variable. In fact, when weighting by the inverse of the number of outcomes for each paper (Table A6 Panel C), the CS2021 estimator produces higher absolute t-statistics, more significant event-study coefficients, and a greater proportion of extremely statistically significant estimates. We also find similar results when restricting the sample to the subset of estimates that all five estimators agree are statistically significant (Table S16), as well as when expanding the sample to include estimates that only a subset of methods produce and adding paper-outcome-period fixed effects (Table S17). Overall, the 2SDD estimator appears to demonstrate more moderate performance compared to the alternatives, particularly given the low frequency of normalized t-statistic differences in the tails (Figure 3); this moderation places the 2SDD estimator toward the conservative end of the spectrum, evident from its low rate of extreme t-statistics (Table A6).

5.3 Differences in pre-event coefficients

One of the most noticeable features of the event-study graphs is the difference in estimates in the periods leading up to the event. While the dCDH2024 and SA2021 estimators tend to produce greater statistical significance in the post-treatment period (Table A6), we do not find the same pattern in the pre-treatment periods (Table S18), where greater statistical significance would indicate violations of parallel trends.[41] In the case of the CS2021 estimator, we see significantly fewer significant coefficients in the pre-treatment periods.[42] This suggests that the 2SDD and BJS2024 methods may offer a more conservative approach.

The discrepancy in pre-event coefficient estimates between 2SDD and BJS2024 requires further discussion. These differences do not stem from a fundamental distinction in the methodologies. Instead, they reflect different choices about what pre-event coefficients to estimate, with both methods accommodating either choice. One approach, which BJS2024 advocate, is to estimate the pre-event coefficients in the first stage of estimation, which uses only untreated observations. This approach results in more outlier standard errors (Figure A4). Another option is to estimate them in the second stage alongside the dynamic treatment effects.

The first- and second-stage approaches would both lead to appropriate rejection rates in our simulations. We note, however, that they estimate distinct quantities. Under the first-stage approach, pre-event coefficients are estimated in a separate regression from and are thus not directly comparable to the post-event coefficients. Estimates using both approaches can still serve a useful role in testing the validity of the parallel trends assumption. When parallel trends fails, the first- and second-stage pre-treatment coefficients identify different parameters, although they should both approach zero when parallel trends holds. While our event-study figures follow the convention of displaying the pre-treatment and post-treatment period estimates on the same figure, this representation may not be as suitable for the first-stage approach.

5.4 Comparison with TWFE

To take stock of our results, we address the concluding remarks of the recent survey by de Chaisemartin and d'Haultfoeuille (2023), which states, "It is also important to stress that at this stage, it is still unclear whether researchers should systematically abandon TWFE estimators." Our analysis provide some clarity on this issue, suggesting that 2SDD should

[41] These comparisons exclude the period immediately preceding the event because some of the methods (SA2021; dCDH2024) normalize the effect in this period to zero.

[42] CS2021 impose a weaker parallel trends assumption than the other methods, though applied researchers may question whether treatment cohorts could be expected to follow the same trend as the never-treated group once they are treated if they were on different trends beforehand.

replace TWFE as the default approach for estimating dynamic treatment effects in settings with staggered treatment timing.

First, for nearly one-sixth of the dynamic treatment effect estimates in our sample, the conclusions based on the TWFE estimator—regarding whether an effect is significantly positive, significantly negative, or insignificant—do not align with those based on any of the heterogeneity-robust estimators.[43] This is not a problem of the heterogeneity-robust estimators simply being imprecise: In about 40 percent of these instances, the discrepancy arises because all of the heterogeneity-robust estimates are statistically significant with the same sign while the TWFE estimate is not significantly different from zero. While de Chaisemartin and d'Haultfoeuille (2023) conjecture that such issues are less likely to arise for "simple designs (e.g.: a single binary and staggered treatment)," our findings suggest that they are not uncommon even in such settings.

Second, while other heterogeneity-robust estimators show pronounced reductions in precision in environments with treatment effect homogeneity, the 2SDD estimator does not share this limitation (recall Table A3). Considering the prevalence of discrepancies between the conclusions of TWFE and heterogeneity-robust estimators highlighted above, defaulting to an assumption of homogeneity seems unjustified. Using 2SDD with inference via GMM yields similar results as using TWFE in settings with homogeneous treatment effects while safeguarding against potential bias due to heterogeneity.

6 Conclusion

When adoption of a treatment is staggered across time, and the average effects of the treatment vary by group and period, the usual difference-in-differences regression specification does not identify an easily interpretable measure of the typical effect of the treatment. When the duration-specific effects are also heterogeneous, neither do the coefficients from the usual event-study specification. The ultimate source of these identification failures is that outcomes are not necessarily linear in group, period, and treatment status, as difference-in-differences and event-study regression specifications assume.

The two-stage approach developed in this paper is motivated by the observation that, under parallel trends, untreated outcomes are linear in group and period effects. Those effects are therefore identified from a first-stage regression estimated using the sample of untreated observations. The average effect of the treatment on the treated is then identified from a regression of outcomes on treatment status, after removing group and period effects.

[43] This issue occurs in four out of the six papers for which we can estimate dynamic treatment effects using all the methods (Bailey and Goodman-Bacon, 2015; Deryugina, 2017; Lafortune, Rothstein and Schanzenbach, 2018; Tewari, 2014), and for nearly half of the outcomes in our data.

This procedure transparently handles the complexities of staggered treatment adoption with familiar and straightforward tools, analogous to traditional regression methods. Estimation and inference are simple and intuitive, and can be easily extended to a variety of different treatment effect measures, including event studies, group-specific treatment effects, design-based analyses, continuous treatments, and triple-difference analyses.

Monte Carlo simulations demonstrate that the two-stage estimator correctly identifies informative average treatment effect measures, outperforming the more complex and computationally demanding alternative methods. Examining these methods across a series of empirical exercises also supports our two-stage approach to estimation and inference as a viable and effective option for applied research. More broadly, the close relationship between our two-stage approach and the traditional TWFE estimator suggests that the two-stage approach provides the most natural extension of the difference-in-differences method to settings with heterogeneous treatment effects. This facilitates adaptations to a variety of problems, and indeed, the general approach proposed in this paper has already been developed by other authors to address settings where time-varying covariates are affected by the treatment (Caetano et al., 2022), to interactive fixed effects models (Brown and Butts, 2023), and to local-projections estimation (Dube et al., 2023).

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2020. “Sampling-Based versus Design-Based Uncertainty in Regression Analysis.” *Econometrica*, 88(1): 265–296.

Angrist, Joshua D., and Alan B. Krueger. 1999. “Chapter 23 - Empirical Strategies in Labor Economics.” In . Vol. 3 of *Handbook of Labor Economics*, , ed. Orley C. Ashenfelter and David Card, 1277–1366. Elsevier.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist’s companion.* Princeton university press.

Arellano, Manuel, and Stéphane Bonhomme. 2012. “Identifying Distributional Characteristics in Random Coefficients Panel Data Models.” *The Review of Economic Studies*, 79(3): 987–1020.

Arkhangelsky, Dmitry, and Guido Imbens. 2024. “Causal models for longitudinal and panel data: A survey.” *The Econometrics Journal*, utae014.

Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. 2021. “Synthetic difference-in-differences.” *American Economic Review*, 111(12): 4088–4118.

Athey, Susan, and Guido W Imbens. 2022. “Design-based analysis in difference-in-differences settings with staggered adoption.” *Journal of Econometrics*, 226(1): 62–79.

Bailey, Martha J, and Andrew Goodman-Bacon. 2015. “The War on Poverty’s experiment in public medicine: Community health centers and the mortality of older Americans.” *American Economic Review*, 105(3): 1067–1104.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How much should we trust differences-in-differences estimates?" *The Quarterly journal of economics*, 119(1): 249–275.

Borusyak, Kirill. 2021. "DID_IMPUTATION: Stata module to perform treatment effect estimation and pre-trend testing in event studies." *Statistical Software Components, Boston College Department of Economics.*

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2021. "Revisiting event study designs: Robust and efficient estimation." *arXiv preprint arXiv:2108.12419.*

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting event study designs: Robust and efficient estimation." *Review of Economic Studies.*

Brown, Nicholas, and Kyle Butts. 2023. "Dynamic Treatment Effect Estimation with Interactive Fixed Effects and Short Panels." *Mimeo.*

Butts, Kyle. 2021. "DID2S: Stata module to estimate a TWFE model using the two-stage difference-in-differences approach." *Statistical Software Components S458951*, Revised: Apr 28, 2023.

Butts, Kyle, and John Gardner. 2022. "did2s: Two-Stage Difference-in-Differences." *R Journal*, 14(3): 162–173.

Caetano, Carolina, Brantly Callaway, Stroud Payne, and Hugo Sant'Anna Rodrigues. 2022. "Difference in differences with time-varying covariates." *arXiv preprint arXiv:2202.02903.*

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with a continuous treatment." *arXiv preprint arXiv:2107.02637.*

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The effect of minimum wages on low-wage jobs." *The Quarterly Journal of Economics*, 134(3): 1405–1454.

Chamberlain, Gary. 1992. "Efficiency Bounds for Semiparametric Regression." *Econometrica*, 60(3): 567–596.

Chiu, Albert, Xingchen Lan, Ziyi Liu, and Yiqing Xu. 2023. "What to do (and not to do) with causal panel analysis under parallel trends: Lessons from a large reanalysis study." *arXiv preprint arXiv:2309.15983.*

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

de Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2023. "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey." *The Econometrics Journal*, 26(3): C1–C30.

de Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2024. "Difference-in-differences estimators of intertemporal treatment effects." *Review of Economics and Statistics*, 1–45.

de Chaisemartin, Clément, Xavier D'Haultfoeuille, Mélitine Malézieux, and Doulo Sow. 2023. "DID_MULTIPLEGT_DYN: Stata module to estimate event-study Difference-in-Difference (DID) estimators in designs with multiple groups and periods, with a potentially non-binary treatment that may increase or decrease multiple times." *Statistical Software Components, Boston College Department of Economics.*

Deryugina, Tatyana. 2017. "The fiscal cost of hurricanes: Disaster aid versus social insurance." *American Economic Journal: Economic Policy*, 9(3): 168–198.

Deshpande, Manasi, and Yue Li. 2019. "Who is screened out? Application costs and the targeting of disability programs." *American Economic Journal: Economic Policy*, 11(4): 213–248.

Dube, Arindrajit, Daniele Girardi, Oscar Jorda, and Alan M Taylor. 2023. "A local projections approach to difference-in-differences event studies." National Bureau of Economic Research.

Egerod, Benjamin CK, and Florian M Hollenbach. 2024. "How many is enough? Sample Size in Staggered Difference-in-Differences Designs." *OSF Preprint*.

Gallagher, Justin. 2014. "Learning about an infrequent event: Evidence from flood insurance take-up in the United States." *American Economic Journal: Applied Economics*, 206–233.

Gardner, John. 2020. "Two-stage differences in differences." *Mimeo*.

Gibbons, Charles E, Juan Carlos Suárez Serrato, and Michael B Urbancic. 2018. "Broken or fixed effects?" *Journal of Econometric Methods*, 8(1): 20170002.

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

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2): 254–277.

Gormley, Todd A, and David A Matsa. 2011. "Growing out of trouble? Corporate responses to liability risk." *The Review of Financial Studies*, 24(8): 2781–2821.

Graham, Bryan S., and James L. Powell. 2012. "Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models." *Econometrica*, 80(5): 2105–2152.

Harmon, Nikolaj A. 2024. "Difference-in-differences and efficient estimation of treatment effects." *Working paper*.

He, Guojun, and Shaoda Wang. 2017. "Do college graduates serving as village officials help rural China?" *American Economic Journal: Applied Economics*, 9(4): 186–215.

Imai, Kosuke, and In Song Kim. 2021. "On the use of two-way fixed effects regression models for causal inference with panel data." *Political Analysis*, 29(3): 405–415.

Kuziemko, Ilyana, Katherine Meckel, and Maya Rossin-Slater. 2018. "Does managed care widen infant health disparities? Evidence from Texas Medicaid." *American Economic Journal: Economic Policy*, 10(3): 255–283.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018. "School finance reform and the distribution of student achievement." *American Economic Journal: Applied Economics*, 10(2): 1–26.

Liu, Licheng, Ye Wang, and Yiqing Xu. 2019. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data."

Liu, Licheng, Ye Wang, and Yiqing Xu. 2022. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." *American Journal of Political Science*.

Miller, Douglas L. 2023. "An Introductory Guide to Event Study Models." *Journal of Economic Perspectives*, 37(2): 203–230.

Newey, Whitney K, and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." *Handbook of econometrics*, 4: 2111–2245.

Rios-Avila, Fernando, Arne J. Nagengast, and Yoto V. Yotov. 2022. "JWDID: Stata module to estimate Difference-in-Difference models using Mundlak approach."

Rios-Avila, Fernando, Pedro Sant'Anna, and Brantly Callaway. 2023. "CSDID: Stata module for the estimation of Difference-in-Difference models with multiple time periods."

Roth, Jonathan. 2024. "Interpreting event-studies from recent difference-in-differences methods." *Mimeo.*

Sun, Liyang. 2021. "EVENTSTUDYINTERACT: Stata module to implement the interaction weighted estimator for an event study." *Statistical Software Components, Boston College Department of Economics.*

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, 225(2): 175–199.

Tewari, Ishani. 2014. "The distributive impacts of financial development: Evidence from mortgage markets during us bank branch deregulation." *American Economic Journal: Applied Economics*, 6(4): 175–196.

Thakral, Neil, and Linh Tô. 2020. "Anticipation and consumption." *Available at SSRN 3756188.*

Thakral, Neil, and Linh T Tô. 2023. "When Are Estimates Independent of Measurement Units?" *Mimeo.*

Ujhelyi, Gergely. 2014. "Civil service rules and policy choices: evidence from US state governments." *American Economic Journal: Economic Policy*, 6(2): 338–380.

Weiss, Amanda. 2024. "How Much Should We Trust Modern Difference-in-Differences Estimates?" Center for Open Science.

Wooldridge, Jeffrey M. 2021. "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators." *Available at SSRN 3906345.*

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

Yitzhaki, Shlomo. 1996. "On using linear regressions in welfare economics." *Journal of Business & Economic Statistics*, 14(4): 478–486.

Table 1: Estimated 95% confidence intervals for the treatment effect in 2 × 2 simulations

	(1)		(2)		(3)	
Spec.	Lower	Upper	Lower	Upper	Lower	Upper
TWFE	-1.20211	2.87373	0.37141	1.61166	0.80560	1.19847
GTTY	-1.20211	2.87373	0.37141	1.61166	0.80560	1.19847
CS	-1.48911	3.16073	0.36838	1.61469	0.80552	1.19855
SA	-3.68518	5.35680	0.33505	1.64802	0.80450	1.19956
dCDH (dyn)	-2.04625	3.71786	0.35530	1.62777	0.80510	1.19896
dCDH (old)	-1.39136	3.06298	0.35969	1.62338	0.80451	1.19956
W	-1.94852	3.62014	0.35541	1.62766	0.80510	1.19896

Note: The table reports the estimated 95% confidence intervals for the treatment effect in simulations of a 2 × 2 design. The first pair of columns report the lower and upper bounds of the confidence interval for the 8-data-point example described in Section 3.1. The second pair of columns report the same information for a tenfold expansion of the dataset, and the third pair of columns report the same information for a 100-fold expansion of the dataset. TWFE refers to the two-way fixed effects estimator and GTTY refers to the method proposed in the current paper, while CS, SA, dCDH (dyn), dCDH (old), and W refer to the methods proposed by Callaway and Sant'Anna (2021), Sun and Abraham (2021), de Chaisemartin and d'Haultfoeuille (2024), de Chaisemartin and d'Haultfoeuille (2020), and Wooldridge (2021), respectively.

Table 2: Simulations (CPS wage data, heterogeneous treatment effects): 40 states treated over 30 years (at least 1 per year)

Method	Period	Rejection rate	S.E.	Bias	RMSE	Speed (secs)
GTTY	0	4.39	0.1014	-0.0006	0.0962	0.07
	1	5.59	0.1024	-0.0006	0.1016	
	2	3.79	0.1030	-0.0005	0.1011	
	3	4.59	0.1030	0.0006	0.1029	
	4	6.99	0.1040	0.0074	0.1076	
BJS	0	27.94	0.0550	-0.0029	0.1041	0.21
	1	25.95	0.0554	0.0077	0.0961	
	2	29.74	0.0551	-0.0078	0.1041	
	3	28.54	0.0566	0.0002	0.1028	
	4	25.35	0.0573	0.0012	0.0984	
CS	0	3.79	0.1404	0.0015	0.1327	47.32
	1	6.39	0.1398	0.0092	0.1459	
	2	5.99	0.1407	0.0040	0.1398	
	3	4.19	0.1412	-0.0046	0.1385	
	4	4.79	0.1418	0.0049	0.1438	
SA	0	1.40	0.1654	-0.0016	0.1370	143.55
	1	1.20	0.1663	-0.0036	0.1371	
	2	1.40	0.1677	-0.0011	0.1350	
	3	1.60	0.1666	-0.0017	0.1347	
	4	2.00	0.1693	0.0068	0.1372	
dCDH	0	22.95	0.0924	-0.0024	0.1403	3.64
	1	20.36	0.0938	-0.0024	0.1399	
	2	17.96	0.0949	-0.0023	0.1370	
	3	19.36	0.0938	-0.0011	0.1354	
	4	18.36	0.0965	0.0060	0.1394	
W	0	12.38	0.1155	-0.0017	0.1378	180.58
	1	11.18	0.1173	-0.0037	0.1373	
	2	10.78	0.1187	-0.0012	0.1354	
	3	10.38	0.1180	-0.0018	0.1349	
	4	9.98	0.1210	0.0067	0.1376	

Note: The table reports results from 501 simulations of 40 treated states over 30 years, with at least one treated state in each of those years. The data consist of log wages for women between the ages of 25 and 50 from the CPS. Treatment effects are heterogeneous and drawn from a normal distribution, with an average value drawn uniformly at random between 2 and 5 percent of the average wage and a standard deviation equal to 10 percent of the average wage. Rejection rate denotes the percentage of simulations in which the specified parameter estimate significantly differs from the true value at the 5 percent significance level. S.E. denotes the standard error averaged across all simulations. Bias denotes the average difference between the point estimate and the true value. RMSE denotes the root-mean-square error. GTTY refers to the method proposed in the current paper. BJS and BJS (leave out) refer to the default asymptotic standard errors and leave-out versions from BJS2024. CS, SA, dCDH, and W refer to the methods proposed by CS2021, SA2021, dCDH2024, and W2021, respectively. Average speed per simulation using the corresponding Stata package for each method is reported in seconds.

Table 3: Empirical applications: List of references

Paper	Groups	Periods	Treatment cohorts	Always treated	Never treated
Tewari (2014)	39 states	1976–2007	20	✓	
Ujhelyi (2014)	48 states	1960–1984	21	✓	✓
Bailey and Goodman-Bacon (2015)	3062 counties	1959–1998	9		✓
Deryugina (2017)	1183 counties	1969–2012	15		✓
He and Wang (2017)	255 villages	2000–2011	8	✓	✓
Kuziemko, Meckel and Rossin-Slater (2018)	250 counties	1993–2001	5		✓
Lafortune, Rothstein and Schanzenbach (2018)	49 states	1990–2014	18		✓

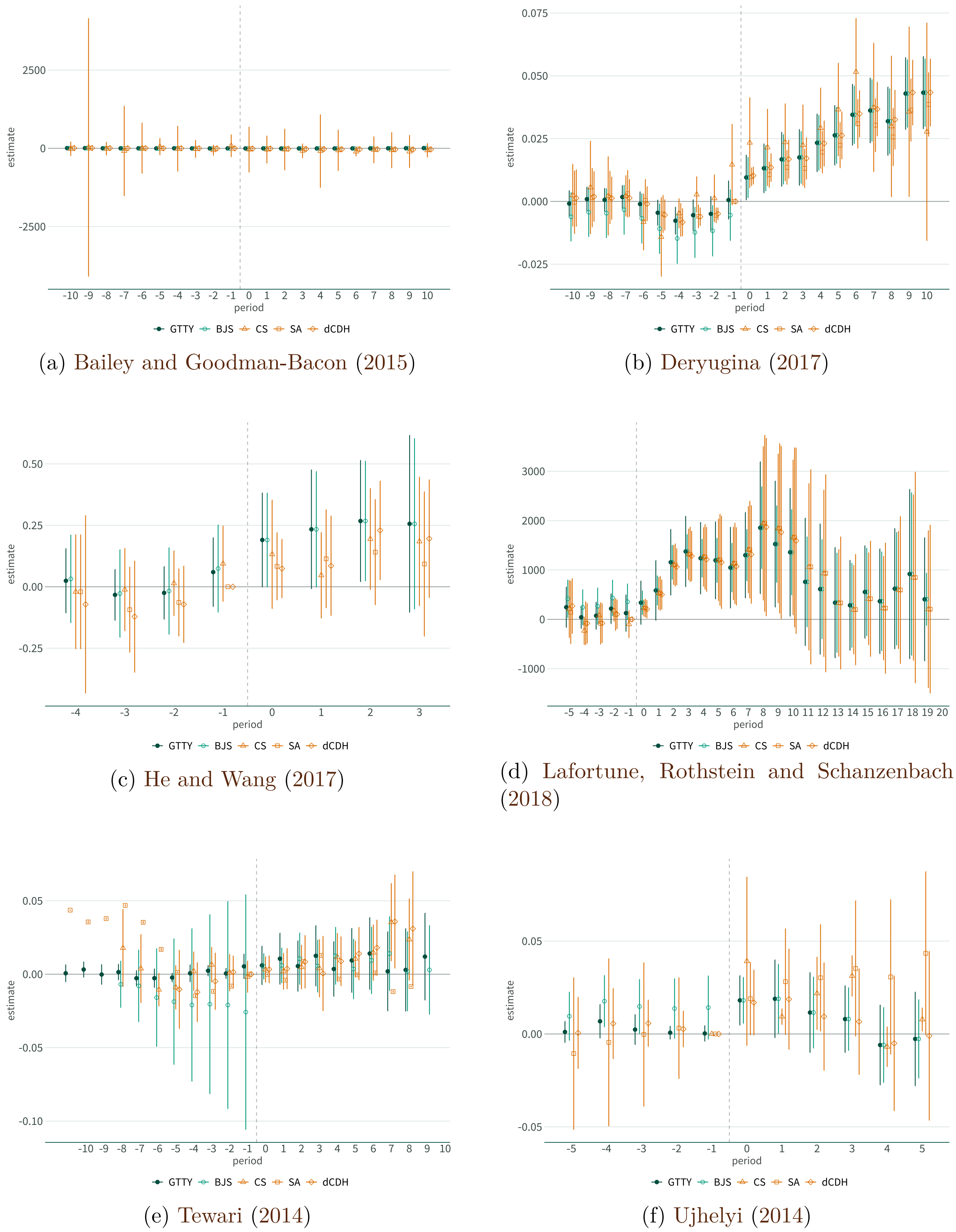
Note: This table describes the set of empirical papers that we reexamine using publicly available data and code. We exclude one from the set of main empirical settings because the paper reports treatment effect estimates at the yearly level while the timing of treatment is at the monthly level (Kuziemko, Meckel and Rossin-Slater, 2018); see Appendix F for further discussion. The list derives from Table 1 of Sun and Abraham (2021), which reports eight papers with variation in treatment timing. We exclude one paper (Gallagher, 2014) due to the presence of multiple treatments.

Table 4: Empirical applications: Comparison of standard errors

	BGB2015	D2017	HW2017	LRS2018	T2014	U2014
GTTY	(+) Small s.e.		(+) Significant	(+) Small s.e.	(+) Full set of estimates	
BJS	(-) Large s.e.		(+) Significant	(-) Overly small s.e.		
CS	(-) Largest s.e.	(-) Largest s.e.			(+) Small s.e., significant	(+) Smallest s.e.
SA	(+) Smallest s.e.	(+) Smallest s.e.		(-) Largest s.e.	(-) Overly small s.e.	(-) Largest s.e.
dCDH	(+) Small s.e.	(+) Small s.e.	(+) Significant	(-) Missing estimates	(+) Significant	(-) Large s.e.

Note: This table summarizes the findings discussed in Section 5.1. The full set of event-study estimates appear in Figure 1 and figures S1 to S5.

Figure 1: Empirical applications: Event-study estimates



(a) Bailey and Goodman-Bacon (2015)

(b) Deryugina (2017)

(c) He and Wang (2017)

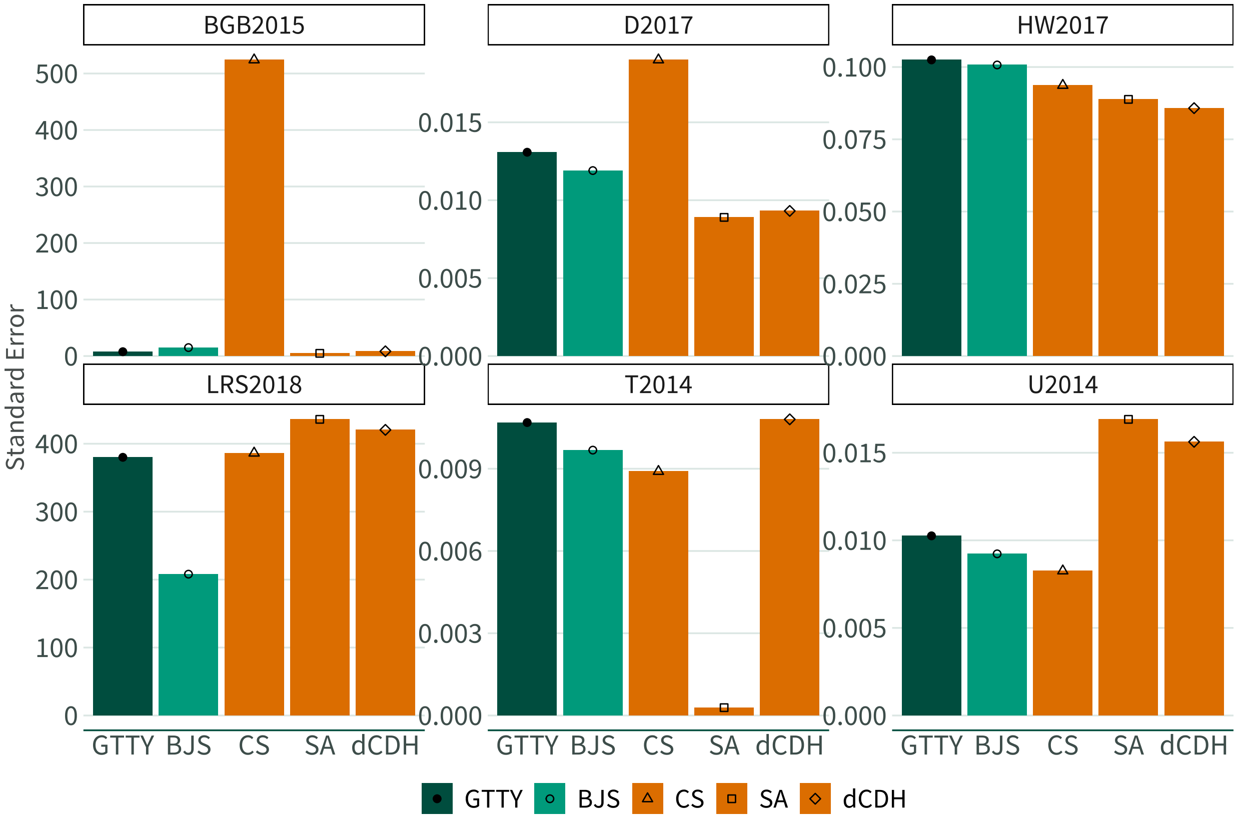
(d) Lafortune, Rothstein and Schanzenbach (2018)

(e) Tewari (2014)

(f) Ujhelyi (2014)

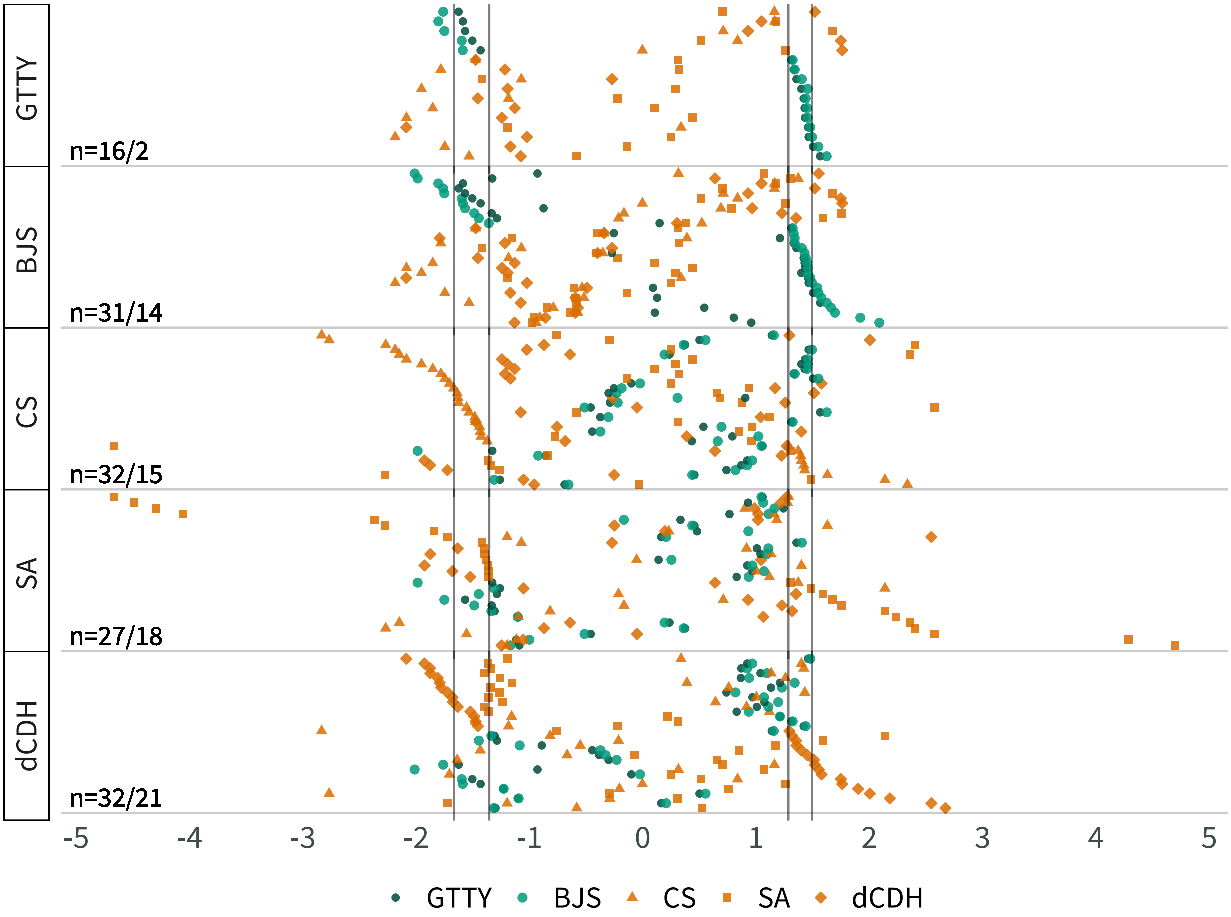
Note: This table reports event-study estimates from applying each estimator to the first event-study specification for each of the main empirical settings in Table 3.

Figure 2: Empirical applications: Comparison of standard errors



Note: This figure reports the average standard error across all dynamic treatment effect estimates for each replicated paper and each estimation method. The set of papers corresponds to the main empirical settings from Table 3.

Figure 3: Empirical applications: Outlier post-treatment normalized t-statistic differences



Note: Each panel of this figure corresponds to one of the five estimators we investigate. Each entry for a given estimator corresponds to an estimate (associated with a particular post-treatment period, outcome variable, and empirical setting) for which that estimator's t-statistic significantly deviates from the average of the other methods' t-statistics. Each entry displays the difference between each method's t-statistic and its associated leave-out mean, normalized by the average of the absolute value of the t-statistics for that coefficient. The criterion for determining that an estimator's t-statistic significantly deviates from that of the other estimators is that the normalized difference falls in the top 2.5 percent or bottom 2.5 percent of the distribution (vertical bars closer to zero as thresholds), excluding estimates from the Bailey and Goodman-Bacon (2015) paper. The numbers in the bottom left of each panel indicate the number of such outlier estimates at the 5 percent level and 1 percent level, respectively.

Online Appendix

A Stata syntax

Suppose that `y` refers to the outcome, `year` the year, `id` the group, and `d` treatment status. In the simplest case, the two-stage difference-in-differences estimator can be obtained, along with valid cluster-robust asymptotic standard errors, via GMM using the single Stata command:

```
gmm (eq1: (y - {xb: i.year} - {xg: ibn.id})*(1-d)) ///
(eq2: y - {xb:} - {xg:} - {delta}*d), ///
instruments(eq1: i.year ibn.id) ///
instruments(eq2: d, noconstant) winitial(identity) ///
onestep quickderivatives vce(cluster id)
```

The `did2s` package (Butts, 2021) implements the same procedure more efficiently and scales more easily with individual fixed effects. The analogous implementation is as follows.

```
did2s y, first_stage(i.year ibn.id) second_stage(d) treatment(d) cluster(id)
```

Variations on the two-stage estimator (such as the the two-stage event-study estimator) can be obtained using similar syntax, which we briefly mention below. We present both the `gmm` and `did2s` implementations.

Event Studies. We let `wr` be an indicator for having r periods since treatment. Let `d0` denote the untreated observations.

```
gmm (eq1: (y - {xb: i.year} - {xg: ibn.id})*d0) ///
(eq2: y - {xb:} - {xg:} - {b1}*w1 ... - {bR}*wR) ///
instruments(eq1: i.year ibn.id) ///
instruments(eq2: w1 ... wR, noconstant) winitial(identity) ///
onestep quickderivatives vce(cluster id)
```

```
did2s y, first_stage(i.year ibn.id) second_stage(w1... w5) treatment(d)
cluster(id)
```

The coefficients `b1` to `bR` are then our event study coefficients. 2SDD is a special case of the above where we only use `b1`. Design-based analysis does not change the code—it only changes the interpretation.

Parallel trends failure. If we want to estimate the first stage only using untreated data within a R periods of being treated, then we would define `d0` accordingly.

```
gen d0 = (year - treatment_year > -R)*(1-d)
```

For did2s, use the following.

```
gen md0 = 1 - d0
did2s y, first_stage(i.year ibn.id) second_stage(d) treatment(md0) cluster(id)
```

Triple differences. With state, time, and subgroup as described in our main text, we can use the same gmm command as before, just that eq1 is modified to be the following.

```
(eq1: (y - {xb: i.state##i.time i.state##i.subgroup i.time##i.subgroup})*d0)
```

Continuous and multivalued treatments. We can allow d to be continuous or multivalued without changing the syntax. For instance, the original 2SDD implementation would use the following. With a multivalued d, we need a constant when approximating the causal response function for treated units.

```
gen d0 = d==0
gmm (eq1: (y - {xb: i.year} - {xg: ibn.id})*d0) ///
(eq2: y - {xb:} - {xg:} - {delta}*d), ///
instruments(eq1: i.year ibn.id) ///
instruments(eq2: d) winitial(identity) ///
onestep quickderivatives vce(cluster id)
```

```
gen md0 = 1 - d0
did2s y, first_stage(i.year ibn.id) second_stage(contX) treatment(md0)
cluster(id)
```

Reversible treatment and several treatments. We write our approach for several treatments as coding for the treatment path with reversible treatments is analogous. Suppose we have two binary treatments d1 and d2.

```
gen d0 = d1==0 & d2==0
gen p1 = d1*(1-d2)
gen p2 = d2*(1-d1)
gen p3 = d1*d2
gmm (eq1: (y - {xb: i.year} - {xg: ibn.id})*d0) ///
(eq2: y - {xb:} - {xg:} - {b1}*p1- {b2}*p2 - {b3}*p3), ///
instruments(eq1: i.year ibn.id) ///
instruments(eq2: p1 p2 p3, noconstant) winitial(identity) ///
onestep quickderivatives vce(cluster id)
```

```
did2s y, first_stage(i.year ibn.id) second_stage(p1 p2 p3) treatment(d)
cluster(id)
```

Test for treatment effect heterogeneity. We can conduct the joint test of whether interactions of treatment with covariates are significant. Suppose we have a `region` covariate.

```
gmm (eq1: (y - {xb: i.year} - {xg: ibn.id})*(1-d)) ///
(eq2: y - {xb:} - {xg:} - {xf: region#d d}) ///
instruments(eq1: i.year ibn.id) ///
instruments(eq2: region#d d, noconstant) winitial(identity) ///
onestep quickderivatives vce(cluster id)
test _b[xf:0b.region#1.d] == _b[xf:2.region#1.d] == 0
```

```
did2s y, first_stage(i.year ibn.id) second_stage(region#d d) treatment(d)
cluster(id)
test _b[0b.region#1.d] == _b[2.region#1.d] == 0
```

B The TWFE estimand

From Equation (1), we can write

$$Y_{it} = \lambda_{g(i)} + \alpha_{p(t)} + \sum_{h=1}^{G} \sum_{q=h}^{P} \beta_{hq} 1(h,q)_{it} + e_{it}, \tag{6}$$

where $1(h,q)_{it}$ is an indicator for whether observation (i,t) corresponds to group h and period q, and $\mathbb{E}\left[e_{gpit} \,\middle|\, g, p, (1(h,q)_{it})\right] = 0$.

Let $\tilde{D}_{it}$ denote the residual from a population regression of D_{it} on group and period fixed effects. By the Frisch-Waugh-Lovell theorem, the coefficient on D_{it} from a population regression of Y_{it} on D_{it} and group and period effects is

$$\begin{aligned}
\beta^* &= \frac{\mathbb{E}\left[\tilde{D}_{it} Y_{it})\right]}{\mathbb{E}\left[\tilde{D}_{it}^2\right]} \\
&= \frac{\mathbb{E}\left[\tilde{D}_{it} \sum_{h=1}^{G} \sum_{q=h}^{P} \beta_{hq} 1(h,q)_{it}\right]}{\mathbb{E}\left[\tilde{D}_{it}^2\right]} \\
&= \sum_{h=1}^{G} \sum_{q=h}^{P} \frac{\mathbb{E}\left[\tilde{D}_{it} 1(h,q)_{it}\right] \beta_{hq}}{\mathbb{E}\left[\tilde{D}_{it}^2\right]} \\
&= \sum_{g=1}^{G} \sum_{p=g}^{P} \omega_{gp} \beta_{gp}.
\end{aligned}$$

where ω_{gp} is the coefficient from a regression of $1(h,q)_{it}$ on D_{it} and group and period fixed effects. The second equality uses the facts that e_{it} is mean-independent of the regressors and that $\tilde{D}_{it}$ is uncorrelated with group and period effects by construction.[44]

The weight ω_{gp} that difference in differences places on β_{gp} is the coefficient on D_{it} from a regression of $1(g,p)_{it}$ on D_{it} and group and period fixed effects. By the Frisch-Waugh-Lovell theorem, this is equivalent to the slope coefficient from a population regression of $1(g,p)_{it}$ on the residual from an auxiliary regression of D_{it} on group and period effects. Using the two-way within or double-demeaned transformation, this residual can be expressed as

$$\tilde{D}_{it} = [D_{it} - \Pr(D_{it} = 1 \,|\, g)] - [\Pr(D_{it} = 1 \,|\, p) - \Pr(D_{it} = 1)]. \tag{7}$$

Since $\mathbb{E}\left[\tilde{D}^2_{it}\right] = \mathbb{E}\left[\tilde{D}_{it} D_{it}\right]$, ω_{gp} can also be expressed as

$$\begin{aligned}
\omega_{gp} &= \frac{\mathbb{E}\left[1(g,p)_{it}\tilde{D}_{it}\right]}{\mathrm{Var}\left[\tilde{D}_{it}\right]} \\
&= \frac{\mathbb{E}\left[\tilde{D}_{it} \,\middle|\, 1(g,p)_{it} = 1\right] \Pr\big(1(g,p)_{gpit} = 1\big)}{\mathbb{E}\left[\tilde{D}_{it} \,\middle|\, D_{it} = 1\right] \Pr(D_{it} = 1)} \\
&= \frac{[1 - \Pr(D_{it} = 1 \,|\, g) - (\Pr(D_{it} = 1 \,|\, p) - \Pr(D_{it} = 1))] \Pr(g,p)}{\sum_{g'=1}^{G} \sum_{p'=g'}^{P} [1 - \Pr(D_{it} = 1 \,|\, g') - (\Pr(D_{it} = 1 \,|\, p') - \Pr(D_{it} = 1))] \Pr(g',p')},
\end{aligned} \tag{8}$$

where the final equality uses Equation (7).

C Proofs

Proof of Lemma 1.

$$\begin{aligned}
E\left[\sum_t D_{it}\tilde{\varepsilon}_{it}\right] &= E\left[\sum_t D_{it}\left(\varepsilon_{it} - \frac{1}{T_i^0}\sum_{t=1}^{T_i^0}\varepsilon_{it}\right)\right] \\
&= E\left[\sum_t D_{it}\left(Y_{it} - D_{it}\beta - X_{it}'\gamma - \frac{1}{T_i^0}\sum_{t=1}^{T_i^0}(Y_{it} - X_{it}'\gamma)\right)\right] \\
&= E\left[\sum_t D_{it}\left(Y_{it} - X_{it}'\gamma - \frac{1}{T_i^0}\sum_{t=1}^{T_i^0}(Y_{it} - X_{it}'\gamma)\right) - \sum_t D_{it}\beta\right] \\
&= E\left[\sum_t D_{it}\left(Y_{it}\left(D_{it}\right) - X_{it}'\gamma - \frac{1}{T_i^0}\sum_{t=1}^{T_i^0}(Y_{it}(0) - X_{it}'\gamma)\right) - \sum_t D_{it}\beta\right]
\end{aligned}$$

[44] This, and the related result in Sun and Abraham (2021), can also be established by thinking of the term $\sum_{h=1}^{G}\sum_{q=h}^{P}\beta_{hq}1(h,q)_{it}$ in Equation (6) as an omitted variable, and taking its projection onto the included regressors.

$$
\begin{aligned}
&= E\left[\sum_t D_{it} E\left[\left(Y_{it}\left(D_{it}\right) - X'_{it}\gamma - \frac{1}{T_i^0}\sum_{t=1}^{T_i^0}\left(Y_{it}(0) - X'_{it}\gamma\right)\right) \mid \{D_{it}\}_{t=0}^{T}\right]\right] - \sum_t E\left[D_{it}\beta\right] \\
&= E\left[\sum_t D_{it} E\left[\left(Y_{it}(0) + \beta_{it}D_{it} - X'_{it}\gamma - \frac{1}{T_i^0}\sum_{t=1}^{T_i^0}\left(Y_{it}(0) - X'_{it}\gamma\right)\right) \mid \{D_{it}\}_{t=0}^{T}\right]\right] - \sum_t E\left[D_{it}\beta\right] \\
&\overset{(A1)}{=} E\left[\sum_t D_{it} E\left[\left(\lambda_i + \beta_{it}D_{it} - \lambda_i\right) \mid \{D_{it}\}_{t=0}^{T}\right]\right] - \sum_t E\left[D_{it}\beta\right] \\
&= E\left[\sum_t D_{it}\left(\beta_{it}D_{it}\right) - \sum_t D_{it}\beta\right] \\
&= E\left[\sum_t \beta_{it}D_{it}\right] - \sum_t E\left[D_{it}\right]\frac{\sum_t E\left[\beta_{it}D_{it}\right]}{\sum_t E\left[D_{it}\right]} = 0
\end{aligned}
$$

$$
\begin{aligned}
&E\left[\sum_t \tilde{X}_{it}\tilde{\varepsilon}_{it} \mid D_{it} = 0\right] \\
&= E\left[\sum_t \tilde{X}_{it}\left(\varepsilon_{it} - \frac{1}{T_i^0}\sum_{t=1}^{T_i^0}\varepsilon_{it}\right) \mid D_{it} = 0\right] \\
&= E\left[\sum_{t=1}^{T_i^0}\left(X_{it} - \frac{1}{T_i^0}\sum_{t=1}^{T_i^0}X_{it}\right)\left(Y_{it} - X'_{it}\gamma - \frac{1}{T_i^0}\sum_{t=1}^{T_i^0}\left(Y_{it} - X'_{it}\gamma\right)\right) \mid D_{it} = 0\right] \\
&= E\left[\sum_{t=1}^{T_i^0}\left(X_{it}\left(Y_{it}(0) - X'_{it}\gamma\right) - \left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}X_{it}\right)\left(Y_{it}(0) - X'_{it}\gamma\right)\right) \mid D_{it} = 0\right] \\
&\quad - E\left[\sum_{t=1}^{T_i^0}X_{it}\left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}\left(Y_{it}(0) - X'_{it}\gamma\right)\right) - \sum_{t=1}^{T_i^0}\left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}X_{it}\right)\left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}\left(Y_{it}(0) - X'_{it}\gamma\right)\right) \mid D_{it} = 0\right]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[E\left[\sum_{t=1}^{T_i^0}\left(X_{it}\left(Y_{it}(0) - X'_{it}\gamma\right) - \left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}X_{it}\right)\left(Y_{it}(0) - X'_{it}\gamma\right)\right) \mid \{X_{it}\}_{t=1}^{T}, \{D_{it}\}_{t=1}^{T}\right] \mid D_{it} = 0\right] \\
&\quad - E\left[E\left[\sum_{t=1}^{T_i^0}X_{it}\left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}\left(Y_{it}(0) - X'_{it}\gamma\right)\right) - \sum_{t=1}^{T_i^0}\left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}X_{it}\right)\left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}\left(Y_{it}(0) - X'_{it}\gamma\right)\right) \mid \{X_{it}\}_{t=1}^{T}, \{D_{it}\}_{t=1}^{T}\right.\right. \\
&\overset{(A1)}{=} E\left[\sum_{t=1}^{T_i^0}\left(X_{it}\lambda_i - \left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}X_{it}\right)\left(\lambda_i\right)\right) - \sum_{t=1}^{T_i^0}X_{it}\left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}\left(\lambda_i\right)\right) + \sum_{t=1}^{T_i^0}\left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}X_{it}\right)\left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}\left(\lambda_i\right)\right) \mid D_{it} = 0\right] \\
&= E\left[\sum_{t=1}^{T_i^0}\left(X_{it}\lambda_i - \left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}X_{it}\right)\lambda_i\right) - \sum_{t=1}^{T_i^0}X_{it}\lambda_i + \sum_{t=1}^{T_i^0}\left(\frac{1}{T_i^0}\sum_{t=1}^{T_i^0}X_{it}\right)\lambda_i \mid D_{it} = 0\right] = 0
\end{aligned}
$$

Lemma C.1. *Under Assumptions 1-3, $\hat{\gamma} \xrightarrow{p} \gamma$ and $\hat{\beta} \xrightarrow{p} \beta$.*

Proof of Lemma C.1. We have

$$\tilde{Y}_{0i} - \tilde{X}_{0i}\gamma = \begin{bmatrix} \left(\tilde{\varepsilon}_{i1}\right)\left(1 - D_{i1}\right) \\ \vdots \\ \left(\tilde{\varepsilon}_{iT}\right)\left(1 - D_{iT}\right) \end{bmatrix} =: \tilde{\varepsilon}_{0i}.$$

Hence,

$$\hat{\gamma} = \gamma + \left(\frac{1}{N}\sum_i \tilde{X}_{0i}'\tilde{X}_{0i}\right)^{-1}\left(\frac{1}{N}\sum_i \tilde{X}_{0i}'\tilde{\varepsilon}_{0i}\right).$$

The rank conditions in Assumption 2 ensure that the limit of the denominator is invertible. Lemma 1 implies that the first moment of the numerator is indeed zero due to Assumption 1. To apply the weak law of large numbers (WLLN) for iid observations and the continuous mapping theorem (CMT) on the respective objects, we just need bounded second moments. In particular, with $\varepsilon_{it} = Y_{it}(0)$, $E[\varepsilon_{it}^4] \leq C, E[X_{kit}^8] \leq C$ from Assumption 3 imply $E[\tilde{\varepsilon}_{it}^4] < \infty$ and $E\left\|\tilde{X}_{0i}'\tilde{X}_{0i}\right\|^4 < \infty$ by repeated applications of the Cauchy-Schwarz inequality. Hence, $E\left\|\tilde{X}_{0i}'\tilde{X}_{0i}\right\|^2 \leq C$ and $E\left\|\tilde{X}_{0i}'\tilde{\varepsilon}_{0i}\right\|^2 \leq E\left\|\tilde{X}_{0i}'\tilde{X}_{0i}\right\| E[\tilde{\varepsilon}_{0i}^2] \leq C$, so WLLN and CMT imply that $\hat{\gamma} \xrightarrow{p} \gamma$.

The OLS estimator is:

$$\begin{aligned}
\hat{\beta} &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\tilde{Y}_{it} - \tilde{X}_{it}\hat{\gamma}\right)\right) \\
&= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\beta_{it}D_{it} + \tilde{\varepsilon}_{it} - \tilde{X}_{it}'\left(\hat{\gamma} - \gamma\right)\right)\right) \\
&= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\beta_{it}\right) + \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\tilde{\varepsilon}_{it} - \tilde{X}_{it}'\left(\hat{\gamma} - \gamma\right)\right)\right).
\end{aligned}$$

Using WLLN, $\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it} \xrightarrow{p} E[\sum_t D_{it}]$ and $\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\beta_{it} \xrightarrow{p} E[\sum_t D_{it}\beta_{it}]$, so $\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\beta_{it}\right) \xrightarrow{p} \beta$. Applying a similar argument for the second term, since both $D_{it}\tilde{\varepsilon}_{it}$ and $D_{it}\tilde{X}_{it}$ have bounded moments $\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\tilde{\varepsilon}_{it}\right) \xrightarrow{p} 0$ and $\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\tilde{X}_{it}'\left(\hat{\gamma} - \gamma\right)\right)\right) = O_P(1)o_P(1) = o_P(1)$. Hence, $\hat{\beta} \xrightarrow{p} \beta$.

□

Proof of Theorem 1. If the conditions of Theorem 6.1 of Newey and McFadden (1994) are satisfied, the result automatically follows. Hence, the proof verifies its conditions. Due to Lemma C.1, we already have $\hat{\gamma} \xrightarrow{p} \gamma$ and $\hat{\beta} \xrightarrow{p} \beta$, fulfilling the probability limit requirement. Next, we want to show the following:

1. β is in the interior of the parameter space.
2. $g(Z;\gamma,\beta)$ is continuously differentiable around β.
3. $\mathbb{E}[g(Z;\gamma,\beta)]=0$ and $\mathbb{E}\left[\|g(Z;\gamma,\beta)\|^2\right]$ is finite.
4. $\mathbb{E}\left[\sup_{(\gamma,\beta)}\|\nabla g(Z;\gamma,\beta)\|\right]<\infty$, where $\nabla g(Z;\gamma,\beta)$ is the derivative of g with respect to (γ',β).
5. $\mathbb{E}[\nabla g(Z;\gamma,\beta)]'\mathbb{E}[\nabla g(Z;\gamma,\beta)]$ is nonsingular.
6. $\frac{1}{N}\sum_{i=1}^N g\left(Z_i;\hat{\gamma},\beta\right)\xrightarrow{p}0$ and $\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T\left(\tilde{X}_{0i}'\left(\tilde{Y}_{0i}-\tilde{X}_{0i}\gamma\right)\right)\xrightarrow{p}0$.

Condition 1 is straightforward as long as no further constraints are imposed on β, which is true in the setting. For condition 2, observe that $\nabla_\beta g(Z;\gamma,\beta)=-\sum_t D_t$, which is continuously differentiable. For condition 3, $\mathbb{E}[g(Z;\gamma,\beta)]=0$ is immediate by assumption, and we have

$$\begin{aligned}\mathbb{E}\left[\|g(Z;\gamma,\beta)\|^2\right]&=\mathbb{E}\left[\left(\sum_{t=1}^T D_{it}\left(\tilde{Y}_{it}-\tilde{X}_{it}'\gamma-D_{it}\beta\right)\right)^2\right]\\&=\mathbb{E}\left[\left(\sum_{t=1}^T\left[\tilde{\varepsilon}_{it}+\left(\beta_{it}-\beta\right)D_{it}\right]D_{it}\right)^2\right]<\infty\end{aligned}$$

due to those objects finite moments and T being finite. Condition 4 is immediate from finite moments, and condition 5 is immediate from the rank condition. For condition 6,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N g(Z_i;\hat{\gamma},\beta)&=\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T D_{it}\left(\tilde{Y}_{it}-\tilde{X}_{it}\hat{\gamma}-D_{it}\beta\right)\\&=\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T\left[\beta_{it}D_{it}+\tilde{\varepsilon}_{it}-\tilde{X}_{it}'\left(\hat{\gamma}-\gamma\right)-D_{it}\beta\right]D_{it}\\&=\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T\left[\tilde{\varepsilon}_{it}-\tilde{X}_{it}'\left(\hat{\gamma}-\gamma\right)+\left(\beta_{it}-\beta\right)D_{it}\right]D_{it}=o_P(1)\end{aligned}$$

due to previous arguments. Finally, the second part of condition 6 is immediate from the WLLN. Hence, we obtain the normality result and it remains to show that the variance estimator is consistent.

By CMT, we can use γ,β in place of $\hat{\gamma},\hat{\beta}$ in the variance since moments of random variables are bounded

$$\hat{V}=G_\beta^{-1}\left(\frac{1}{N}\sum_i E\left[\left(g_i+G_\gamma\psi_i\right)\left(g_i+G_\gamma\psi_i\right)'\right]\right)G_\beta^{-1}+o_P(1)$$

It suffices to consider the meat, since $\hat{G}_\beta$ converges to G_β:

$$g_i+G_\gamma\psi_i=\sum_t D_{it}\left(\tilde{Y}_{it}-\tilde{X}_{it}'\gamma-\beta D_{it}\right)+G_\gamma E\left[\tilde{X}_{0i}'\tilde{X}_{0i}\right]^{-1}\tilde{X}_{0i}'\left(\tilde{Y}_{0i}-\tilde{X}_{0i}\gamma\right).$$

Since observations are independent across i applying WLLN results in the consistency result. To apply the WLLN, the sufficient conditions are $E\left\|\tilde{X}_{0i}'\tilde{\varepsilon}_{0i}\right\|^4 < \infty$, $E\tilde{\varepsilon}_{it}^4 < \infty$, $E(\beta_{it}-\beta)^4 < \infty$, and $E\left\|\tilde{X}_{0i}'\tilde{X}_{0i}\right\|^4 < \infty$. Note that $E\tilde{\varepsilon}_{it}^4 < \infty$, and $E(\beta_{it}-\beta)^4 < \infty$ result from how $\tilde{Y}_{it}$ is also squared. Since $E[\beta_{it}^4] \leq C$ implies $E(\beta_{it}-\beta)^4 < \infty$, $E[X_{kit}^8] \leq C$ implies $E\left\|\tilde{X}_{0i}'\tilde{X}_{0i}\right\|^4 < \infty$, and $E\left\|\tilde{X}_{0i}'\tilde{\varepsilon}_{0i}\right\|^4 \leq E\left\|\tilde{X}_{0i}'\tilde{X}_{0i}\right\|^4 E[\tilde{\varepsilon}_{it}^4]$ by Cauchy-Schwarz inequality, the sufficient conditions are satisfied and the variance estimator is consistent. □

D Comparison with TWFE in 2×2 case

We have two treatment groups indexed by $B \in \{0,1\}$ for the treated and untreated group respectively, and two time periods indexed $T \in \{0,1\}$ for pre and post periods respectively. The treatment variable is $D_{it} = B_{it}T_{it}$, so a unit needs to be both in the treatment group $B_{it} = 1$ and in the post period to be treated. Then, we can write the regression equation as:

$$Y_{it} = \beta_0 + \beta_1 B_{it} + \beta_2 T_{it} + \beta_3 D_{it} + \varepsilon_{it}$$

We focus on the estimand and hence use the actual residual instead of the estimated residual i.e., ε_{it} instead of $\hat{\varepsilon}_{it}$. Let $\tilde{D}$ denote the treatment D with group and time fixed effects (B,T) partialled out. The estimands are:

$$V_{TWFE} = E\left[\sum_t \tilde{D}_{it}^2\right]^{-1} E\left[\left(\sum_t \tilde{D}_{it}\varepsilon_{it}\right)^2\right] E\left[\sum_t \tilde{D}_{it}^2\right]^{-1}$$

$$V_{2SDD} = E\left[\sum_t D_{it}\right]^{-1} E\left[\left(g_i + G_\gamma \psi_i\right)^2\right] E\left[\sum_t D_{it}\right]^{-1}$$

Lemma D.1. $V_{TWFE} = V_{2SDD}$.

Proof. We characterize the estimands entirely in terms of primitive parameters, where b indexes the treatment group and t indexes time. Define $\mu := E\left[B_{it}\right]$ and $v_{bt} := Var\left(Y_{it} \mid b,t\right) = E\left[\varepsilon_{it}^2 \mid b,t\right]$. We show that: $V_{TWFE} = \frac{1}{\mu}\left(v_{11}+v_{10}\right) + \frac{1}{(1-\mu)}\left(v_{01}+v_{00}\right) = V_{2SDD}$.

We start with V_{TWFE}. To construct $\tilde{D}_{it}$, we decompose D_{it} as $D_{it} = \alpha_0 + \alpha_1 B_{it} + \alpha_2 T_{it} + e_{it}$, and express $\alpha_0, \alpha_1, \alpha_2$ in terms of primitives of the model by using moment conditions $E\left[e_{it}\right] = E\left[B_{it}e_{it}\right] = E\left[T_{it}e_{it}\right] = 0$. We can use the results that $E\left[D_{it}\right] = E\left[D_{it}^2\right] = \frac{1}{2}\mu$, $E\left[B_{it}\right] = \mu, E\left[T_{it}\right] = 1/2$ and $E\left[B_{it}D_{it}\right] = E\left[T_{it}D_{it}\right] = \frac{1}{2}\mu$. The moment conditions yield:

$$E\left[D_{it}\right] = \alpha_0 + \alpha_1\mu + \alpha_2\frac{1}{2} = \frac{1}{2}\mu$$

$$\begin{aligned}
E\left[B_{it}e_{it}\right] &= E\left[B_{it}\left(D_{it}-\alpha_0-\alpha_1 B_{it}-\alpha_2 T_{it}\right)\right]\\
&= \frac{1}{2}\mu-\alpha_0\mu-\alpha_1\mu-\alpha_2\frac{1}{2}\mu=0\\
E\left[T_{it}e_{it}\right] &= E\left[T_{it}\left(D_{it}-\alpha_0-\alpha_1 B_{it}-\alpha_2 T_{it}\right)\right]\\
&= \frac{1}{2}\mu-\alpha_0\frac{1}{2}-\alpha_1\frac{1}{2}\mu-\alpha_2\frac{1}{2}=0
\end{aligned}$$

Solving for the α's, we obtain $\alpha_0=-\frac{1}{2}\mu$, $\alpha_1=\frac{1}{2}$ and $\alpha_2=\mu$, so

$$\tilde{D}_{it}=D_{it}+\frac{1}{2}\mu-\frac{1}{2}B_{it}-\mu T_{it}.$$

By splitting the denominator into a component with $t=0$ and another with $t=1$,

$$\begin{aligned}
E\left[\tilde{D}_{i0}^2\right] &= E\left[\frac{1}{2}\mu\left(D_{i0}+\frac{1}{2}\mu-\frac{1}{2}B_{i0}-\mu T_{i0}\right)\right]-E\left[\frac{1}{2}B_{i0}\left(D_{i0}+\frac{1}{2}\mu-\frac{1}{2}B_{i0}-\mu T_{i0}\right)\right]\\
&= \frac{1}{4}\mu\left(1-\mu\right)
\end{aligned}$$

and $E\left[\tilde{D}_{i1}^2\right]=\frac{1}{4}\mu\left(1-\mu\right)$, so $E\left[\sum_t \tilde{D}_{it}^2\right]=\frac{1}{2}\mu\left(1-\mu\right)$. Proceeding with the numerator, we can expand the expression:

$$E\left[\left(\sum_t \tilde{D}_{it}\varepsilon_{it}\right)^2\right]=E\left[\tilde{D}_{i0}^2\varepsilon_{i0}^2+2\tilde{D}_{i0}\varepsilon_{i0}\tilde{D}_{i1}\varepsilon_{i1}+\tilde{D}_{i1}^2\varepsilon_{i1}^2\right].$$

We have: $E\left[\tilde{D}_{i0}\varepsilon_{i0}\tilde{D}_{i1}\varepsilon_{i1}\right]=0$ using the assumption that $E\left[\varepsilon_{it}\mid g,t\right]=0$. Using the moment results that $E\left[D_{i0}\varepsilon_{i0}^2\right]=0$, $E\left[B_{i0}\varepsilon_{i0}^2\right]=\mu v_{10}$, $E\left[T_{i0}\varepsilon_{i0}^2\right]=0$, $E\left[\varepsilon_{i0}^2\right]=\mu v_{10}+\left(1-\mu\right)v_{00}$, $E\left[D_{i1}\varepsilon_{i1}^2\right]=\mu v_{11}$, $E\left[B_{i1}\varepsilon_{i1}^2\right]=\mu v_{11}$, $E\left[T_{i1}\varepsilon_{i1}^2\right]=\mu v_{11}+\left(1-\mu\right)v_{01}$, and $E\left[\varepsilon_{i1}^2\right]=\mu v_{11}+\left(1-\mu\right)v_{01}$, we obtain

$$\begin{aligned}
E\left[\tilde{D}_{i0}^2\varepsilon_{i0}^2\right] &= E\left[\left(D_{i0}+\frac{1}{2}\mu-\frac{1}{2}B_{i0}-\mu T_{i0}\right)^2\varepsilon_{i0}^2\right]\\
&= \frac{1}{4}\mu\left(1-\mu\right)\left(\mu v_{00}+\left(1-\mu\right)v_{10}\right)
\end{aligned}$$

and

$$\begin{aligned}
E\left[\tilde{D}_{i1}^2\varepsilon_{i1}^2\right] = {}& E\left[D_{i1}\left(D_{i1}+\frac{1}{2}\mu-\frac{1}{2}B_{i1}-\mu T_{i1}\right)\varepsilon_{i1}^2\right]+\frac{1}{2}\mu E\left[\left(D_{i1}+\frac{1}{2}\mu-\frac{1}{2}B_{i1}-\mu T_{i1}\right)\varepsilon_{i1}^2\right]\\
&-\frac{1}{2}E\left[B_{i1}\left(D_{i1}+\frac{1}{2}\mu-\frac{1}{2}B_{i1}-\mu T_{i1}\right)\varepsilon_{i1}^2\right]-\mu E\left[T_{i1}\left(D_{i1}+\frac{1}{2}\mu-\frac{1}{2}B_{i1}-\mu T_{i1}\right)\varepsilon_{i1}^2\right]
\end{aligned}$$

$$= \frac{1}{4}\mu(1-\mu)(\mu v_{01} + (1-\mu)v_{11}).$$

Substituting these expressions back into the variance estimand yields the result.

Turning to the 2SDD variance, its components are:

$$\begin{aligned} G_\gamma &= -E\left[\sum_t D_{it}\tilde{X}_{it}\right] \\ \psi_i &= E\left[\tilde{X}_{0i}'\tilde{X}_{0i}\right]^{-1}\left(\tilde{X}_{0i}'\left(\tilde{Y}_{0i} - \tilde{X}_{0i}\gamma\right)\right) \end{aligned}$$

Here, $X = T$, $\tilde{Y}_{it} = Y_{it} - Y_{i0}$ and $\tilde{X}_{it} = X_{it} - X_{i0}$. We also have

$$\begin{aligned} Y_{i1} - Y_{i0} &= \beta_0 + \beta_1 B_{i1} + \beta_2 T_{i1} + \beta_3 D_{i1} + \varepsilon_{i1} - (\beta_0 + \beta_1 B_{i0} + \varepsilon_{i0}) \\ &= \beta_2 + \beta_3 D_{i1} + \varepsilon_{i1} - \varepsilon_{i0}. \end{aligned}$$

In this context, $\tilde{X}_{0i}$ only contains T because the constant and group FE have already been partialled out, so

$$\tilde{X}_{0i} = \begin{bmatrix} (T_{i0} - T_{i0}(1-D_{i0}))(1-D_{i0}) \\ (T_{i1} - T_{i0}(1-D_{i0}))(1-D_{i1}) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 - D_{i1} \end{bmatrix}$$

$$\tilde{Y}_{0i} = \begin{bmatrix} (Y_{i0} - Y_{i0})(1-D_{i0}) \\ (Y_{i1} - Y_{i0})(1-D_{i1}) \end{bmatrix} = \begin{bmatrix} 0 \\ (\beta_2 + \varepsilon_{i1} - \varepsilon_{i0})(1-D_{i1}) \end{bmatrix}.$$

With these expressions, $\gamma = E\left[\tilde{X}_{0i}'\tilde{X}_{0i}\right]^{-1} E\left[\tilde{X}_{0i}'\tilde{Y}_{0i}\right] = \beta_2$, so

$$\begin{aligned} \tilde{X}_{0i}'\left(\tilde{Y}_{0i} - \tilde{X}_{0i}\gamma\right) &= (1-D_{i1})((\beta_2 + \varepsilon_{i1} - \varepsilon_{i0})(1-D_{i1}) - (1-D_{i1})\beta_2) \\ &= (\varepsilon_{i1} - \varepsilon_{i0})(1-D_{i1}) \end{aligned}$$

Hence,

$$\begin{aligned} \psi_i &= E[1-D_{i1}]^{-1}((\varepsilon_{i1} - \varepsilon_{i0})(1-D_{i1})) \\ &= \frac{1}{1-\mu}((\varepsilon_{i1} - \varepsilon_{i0})(1-D_{i1})), \text{ and} \end{aligned}$$

$$E\left[\sum_t D_{it}\tilde{X}_{it}\right] = E\left[D_{i1}\tilde{X}_{i1}\right] = \mu,$$

so $G_\gamma = -E\left[\sum_t D_{it}\tilde{X}_{it}\right] = -\mu$. We finally have:

$$\begin{aligned} g_i &= \sum_t D_{it}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\gamma - \beta_3 D_{it}\right) \\ &= D_{i1}\left(\tilde{Y}_{i1} - \tilde{X}_{i1}'\beta_2 - D_{i1}\beta_3\right) = D_{i1}\left(\varepsilon_{i1} - \varepsilon_{i0}\right) \end{aligned}$$

and

$$\begin{aligned} E\left[\left(g_i + G_\gamma\psi_i\right)^2\right] &= E\left[\left(D_{i1}\left(\varepsilon_{i1} - \varepsilon_{i0}\right) - \mu\frac{1}{1-\mu}\left(\left(\varepsilon_{i1} - \varepsilon_{i0}\right)\left(1 - D_{i1}\right)\right)\right)^2\right] \\ &= \mu\left(v_{11} + v_{10}\right) + \frac{\mu^2}{(1-\mu)}\left(v_{01} + v_{00}\right) \end{aligned}$$

Substituting these components into the variance expression yields the result. □

E Simulations: Random design vs. fixed design

We discuss how our simulation environment compares to that of Borusyak, Jaravel and Spiess (2024), who propose a numerically equivalent estimator with a different asymptotic theory. They propose an asymptotically conservative approach to inference and document that it performs well in finite samples using a series of Monte Carlo simulations. As our discussion of overfitting in Section 4.2.2 highlights, we observe substantial rates of over-rejection using their variance estimator, particularly when treatment timing varies over longer periods.

Furthermore, their simulation environment, and their theory more generally, interprets treatment assignment and event times as non-stochastic. The design—treated states, treatment effects, and treatment timing—is therefore held fixed, and the source of randomness across simulations is the randomly drawn error term for generating outcomes. With the error term in outcomes as the sole source of randomness, the variance of that error term plays a crucial role.

We explore the conditions under which rejection rates can reach 100% in such a setup. A simple example with two periods and two states suffices to illustrate the problems that can arise for a small error term variance. Consider a placebo law, for which the true effect is zero, that applies to a random sample of treated states. In the absence of any true treatment effect, we would expect changes in outcomes for both treated and control states to be similar, and any observed discrepancy between the changes for the two groups would be solely attributed to the random error term. However, a finite difference in outcomes arises because the assignment of states to treatment or control groups is fixed after being drawn only once. This finite difference is not fully absorbed by state and year fixed effects, leading to misspecification. As a result, when the variance of the error term is sufficiently small compared to that finite difference, we observe consistent rejection of the null hypothesis. Assuming a larger error variance, using a large sample of treated states, or using random designs mitigates this issue (and we verify that our conclusions regarding the performance of the various estimators continue to hold under fixed designs with large error variance). Our discussion highlights the conceptual appeal of adopting a "random design" approach, in which stochasticity is incorporated into the simulation by randomly drawing treated states, treatment effects, and treatment timing in each iteration. Under random designs, even with a small error variance, rejection rates remain accurate and avoid spurious over-rejection.

F Empirical applications

F.1 Selection of papers and outcomes

Below is the list of papers included in our empirical analysis, which appear in Table 1 of Sun and Abraham (2021), and the outcomes they study. We omit outcomes that are unavailable in the replication data, or are too slow to run (more than 5 days of runtime) for at least one of the methods.

- Bailey and Goodman-Bacon (2015)
  - Age-adjusted mortality rate (Figure 5)
  - Infant mortality rate (Figure 7.A)
  - Age-adjusted mortality rate: children (1–14) (Figure 7.B)
  - Age-adjusted mortality rate: adults (15–49) (Figure 7.C)
  - Age-adjusted mortality rate: older adults (50+) (Figure 7.D)
- Deryugina (2017)
  - Effect of a hurricane on earnings and transfers (Figure 2)
  - Effect of a hurricane on demographics (Figure 3)
  - Effect of a hurricane on transfer components (Figures 4 and 5)
- He and Wang (2017)
  - Subsidized population (Figure 2.A)
  - Poor-quality housing (Figure 2.B)
  - Registered poor households (Figure 2.C)
  - People with disabilities (Figure 2.D)
- Kuziemko, Meckel and Rossin-Slater (2018)
  - Mortality rates of children born to US-born Black mothers (Figure 2.A)

  - Mortality rates of children born to US-born Hispanic mothers (Figure 2.B)
- Lafortune, Rothstein and Schanzenbach (2018)
  - Mean state revenues in lowest income districts (Figure 3)
  - Mean state revenues in highest income districts (Figure 4)
  - Progressivity of state revenues (Figure 5)
  - Mean total revenues per pupil (Figure A3(a))
  - Mean total revenues per pupil in the lowest income quintile of districts (Figure A3(b))
  - Mean total revenues per pupil in the highest income quintile of districts (Figure A3(c))
  - Difference in mean total revenues per pupil between top and bottom quintile districts (Figure A3(d))
- Tewari (2014)
  - Home ownership (Figure 1)
- Ujhelyi (2014)
  - Share of intergovernmental expenditures in total expenditures (Figure 1)

F.2 Replication of Kuziemko, Meckel and Rossin-Slater (2018)

The Kuziemko, Meckel and Rossin-Slater (2018) paper studies the effect of the transition from Medicaid's public fee-for-service (FFS) plan to private Medicaid Managed Care (MMC) plans on infant mortality rates for US-born Black and Hispanic mothers in Texas. Their analysis uses 250 counties in Texas, with 9 years of data from 1993 to 2001. Of the 250 counties, 3 are treated in 1995, 36 are treated in 1996, 1 is treated in 1997, 8 are treated in 1998, and 9 are treated in 1999.

The dataset contains the month and year in which each treated county switched from FFS to MMC. However, the authors estimate the effect of the transition on infant mortality rates using a two-way fixed effects specification with year-since-treatment event dummies, where years are defined as 12-month periods relative to the event time. We attempt to replicate the analysis of Kuziemko, Meckel and Rossin-Slater (2018) using this kind of specification with the heterogeneity-robust estimators.

The 2SDD approach is easily implemented by using month fixed effects in the first stage and year-since-treatment event dummies in the second stage. To obtain estimates using `csdid` (Rios-Avila, Sant'Anna and Callaway, 2023), `eventstudyinteract` (Sun, 2021), `did_multiplegt_dyn` (de Chaisemartin et al., 2023), and `jwdid` (Rios-Avila, Nagengast and Yotov, 2022), we must define the cohort as the treatment year (not the exact month) to obtain dynamic effects by year since treatment. We present these results in Figure A2. However, we note that the conceptually correct way to do this exercise using those estimators would be to estimate separate effects for each month and then aggregate them into 12-month bins. This process would be somewhat cumbersome, and if undertaken, would require either assuming the distribution of units in each bin is known, or using a bootstrap, or devising a potentially complicated analytical asymptotic adjustment to account for that uncertainty.

This highlights the flexibility and simplicity advantages of 2SDD. With 2SDD, implementing the conceptually correct approach is straightforward: Simply include month fixed effects in the first stage and years-since-treatment indicators in the second stage.[45]

G Extension to stacked differences in differences

In the stacked approach, a new dataset is created for each treated group, containing observations on that group $\overline{R}$ periods before, and $\bar{P}$ periods after, the treatment is adopted, as well as on units that are not yet treated during these periods. These group-specific datasets are stacked, and outcomes are regressed on treatment status and dataset-specific group and period fixed effects:

$$Y_{cit} = \lambda_{cg(i)} + \alpha_{cp(t)} + \beta D_{cit} + \varepsilon_{it},$$

where cit indexes the (i, t)th observation of dataset c.

Let D_{cit} be an indicator for whether i is treated at time t in dataset c, and D_{rcit} be an indicator for whether i has been treated for $r \in \{1, \dots, \bar{P}\}$ periods as of time t in dataset c. Let $\tau = \bar{P}/(\bar{P} + \overline{R} + 1)$ denote the fraction of periods during which treated units in any group-specific dataset are treated, π_c denote the fraction of units in dataset c that belong to the treatment group, and ρ_c denote size of dataset c relative to the stacked dataset.

Extending the logic of the previous section, the weight ω_{rg} that stacked differences in differences places on the r-period average treatment effect β_{rg} for group g is given by the slope coefficient from a population regression of D_{rcit} on the residual $\tilde{D}_{cit}$ from a regression of D_{cit} on dataset×period and dataset×group effects.[46] This residual is

$$\tilde{D}_{cit} = D_{cit} - P(D_{cit} = 1|g, c) - [P(D_{cit} = 1|p, c) - P(D_{cit} = 1|c)],$$

where statements conditional on c are true in the population corresponding to dataset c.

[45] However, we were unable to obtain estimates using the imputation approach (Borusyak, 2021) when adding month fixed effects in the first stage.

[46] This is because the correct model can be expressed as

$$Y_{cit} = \lambda_{cg(i)} + \alpha_{cp(t)} + \sum_{c=1}^{C} \sum_{r=1}^{\bar{R}} \beta_{rc} D_{crit} + e_{cit},$$

where $\beta_{rc} = \beta_{rg}$ when $D_{crit} = 1$.

Using this expression and adapting (8) to the stacked setting,

$$\begin{aligned}\omega_{rg} &= \frac{[1-\tau-(\pi_c-\tau\pi_c)]P(D_{rcit}=1)}{\sum_{c=1}^{G}\sum_{p=1}^{\bar{P}}[1-\tau-(\pi_c-\tau\pi_c)]P(D_{rcit}=1)}\\ &= \frac{(1-\tau)(1-\pi_c)\tau\pi_c\rho_c}{\sum_{c=1}^{G}\sum_{p=1}^{\bar{P}}(1-\tau)(1-\pi_c)\tau\pi_c\rho_c}\\ &= \frac{(1-\pi_c)\pi_c\rho_c}{\bar{P}\sum_{c=1}^{G}(1-\pi_c)\pi_c\rho_c}.\end{aligned}$$

H Supplementary material

H.1 Serial correlation

We consider rewriting an event-studies model in the following manner. Instead of $Y_{it} = \lambda_{g(i)} + \alpha_t + \sum_{r=1}^{\overline{R}} \eta_r W_{rit} + u_{it}$, we write:

$$Y_{it} = \lambda_{g(i)} + \alpha_t + \sum_{t=t^*(i)+1}^{t^*(i)+\overline{R}} \beta_{t-t^*(i)} D_{it} + u_{it}$$

where D_{it} is an indicator for treatment. Since treatments are irreversible, the effect after r periods of treatment is $\eta_r = \sum_{j=1}^{r} \beta_j$, so the β_j's can be interpreted as the effect of an additional period of treatment after $j-1$ periods of treatment. Where required, we set $\beta_0 = 0$ to avoid ambiguity. We assume $\epsilon_{it} = \rho\epsilon_{it-1} + \nu_{it}$ and ν_{it} is white noise.

To recover the β_j, we can run the following first-stage regression for untreated units:

$$Y_{it} = \rho Y_{it-1} + \lambda_{g(i)} + \alpha_t + \epsilon^1_{it},$$

and the following second-stage regression:

$$Y_{it} - \hat{\rho}Y_{it-1} - (1-\hat{\rho})\widehat{\lambda_{g(i)}} - \widehat{\alpha_t} + \hat{\rho}\hat{\alpha}_{t-1} = \sum_{t=t^*(i)+1}^{t^*(i)+\overline{R}} \delta_{t-t^*(i)} D_{it} + \epsilon^2_{it},$$

which is implementable by regressing on W_{rit} and backing out δ_r as described above.

To motivate the procedure, rearranging $Y_{it-1} = \sum_{t=t^*(i)}^{t^*(i)+\overline{R}-1} \beta_{t-t^*(i)} D_{it} + \lambda_{g(i)} + \alpha_{t-1} + \epsilon_{it-1}$ gives

$$\epsilon_{it-1} = Y_{it-1} - \sum_{t=t^*(i)+1}^{t^*(i)+\overline{R}} \beta_{t-t^*(i)} D_{it} - \lambda_g - \alpha_{t-1}.$$

Combining $Y_{it} = \sum_{t=t^*(i)+1}^{t^*(i)+\overline{R}} \beta_{t-t^*(i)} D_{it} + \lambda_{g(i)} + \alpha_t + \epsilon_{it}$ and $\epsilon_{it} = \rho\epsilon_{it-1} + \nu_{it}$, and then using the equation above, we obtain

$$\begin{aligned}
Y_{it} &= \sum_{t=t^*(i)+1}^{t^*(i)+\overline{R}} \beta_{t-t^*(i)} D_{it} + \lambda_{g(i)} + \alpha_t + \rho\epsilon_{it-1} + \nu_{it} \\
Y_{it} &= \sum_{t=t^*(i)+1}^{t^*(i)+\overline{R}} \beta_{t-t^*(i)} D_{it} + \lambda_{g(i)} + \alpha_t + \rho \left(\sum_{t=t^*(i)}^{t^*(i)+\overline{R}-1} \beta_{t-t^*(i)} D_{it} - \lambda_{g(i)} - \alpha_{t-1} \right) + \nu_{it} \\
&= \rho Y_{it-1} + \left(\sum_{t=t^*(i)+1}^{t^*(i)+\overline{R}} \beta_{t-t^*(i)} D_{it} - \rho \sum_{t=t^*(i)}^{t^*(i)+\overline{R}-1} \beta_{t-t^*(i)} D_{it} \right) + (1-\rho)\,\lambda_{g(i)} + \alpha_t - \rho\alpha_{t-1} + \nu_{it} \\
&= \rho Y_{it-1} + \left(\sum_{t=t^*(i)+1}^{t^*(i)+\overline{R}} \beta_{t-t^*(i)} - \rho \sum_{t=t^*(i)}^{t^*(i)+\overline{R}-1} \beta_{t-t^*(i)} \right) D_{it} + (1-\rho)\,\lambda_{g(i)} + \alpha_t - \rho\alpha_{t-1} + \nu_{it}.
\end{aligned}$$

This implies

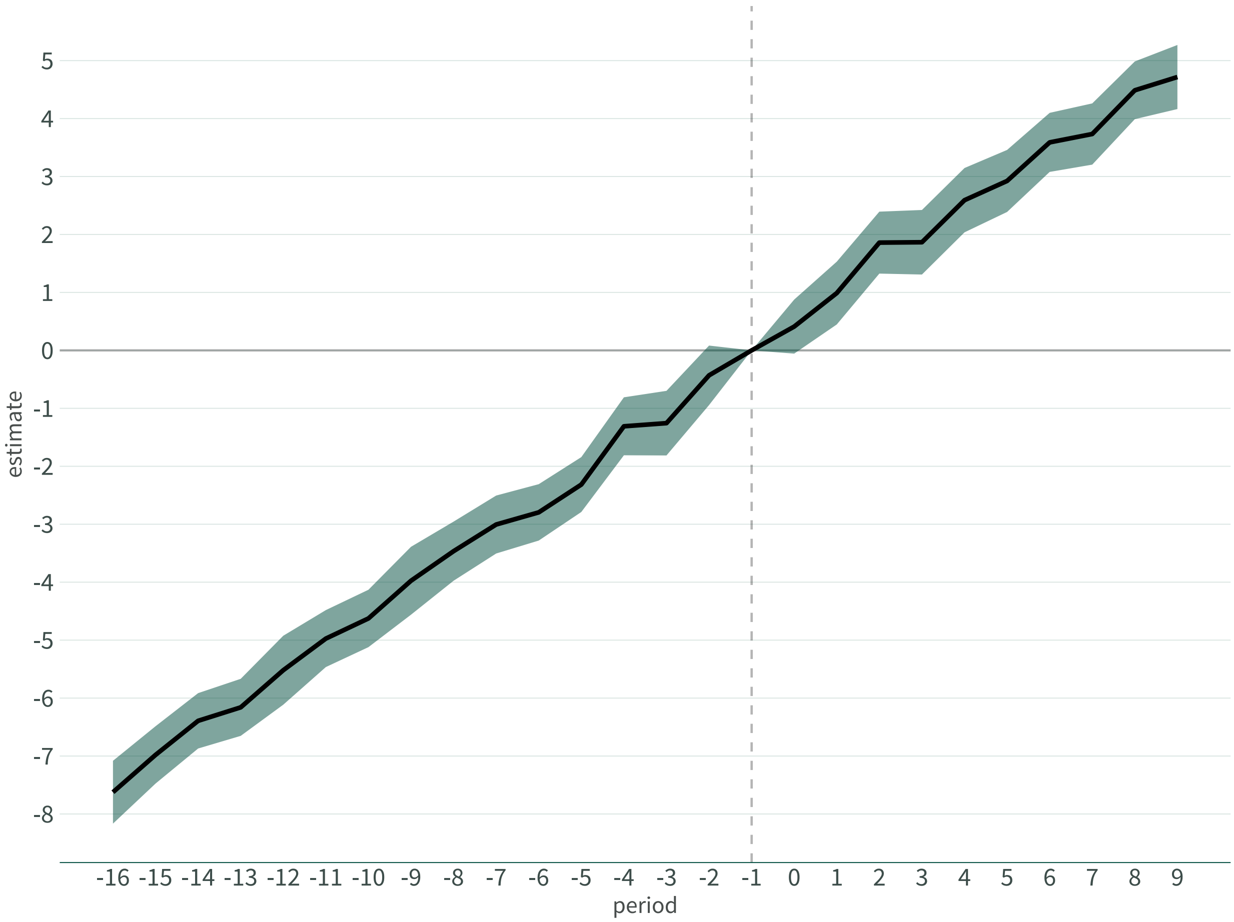
$$\begin{aligned}
\delta_1 &= \beta_1 \\
\delta_2 &= \beta_2 - \rho\beta_1 \\
&\vdots \\
\delta_r &= \beta_r - \rho\beta_{r-1}
\end{aligned}$$

or equivalently

$$\begin{aligned}
\beta_1 &= \delta_1 \\
\beta_2 &= \delta_2 + \rho\beta_1 \\
&= \delta_2 + \rho\delta_1 \\
\beta_3 &= \delta_3 + \rho\beta_2 \\
&= \delta_3 + \rho(\delta_2 + \rho\delta_1) \\
&= \delta_3 + \rho\delta_2 + \rho^2\delta_1 \\
&\vdots \\
\beta_r &= \delta_r + \rho^1\delta_{r-1} + \rho^2\delta_{r-2} + \cdots + \rho^{r-1}\delta_1 \\
&= \sum_{i=0}^{r-1} \rho^i \delta_{r-i}
\end{aligned}$$

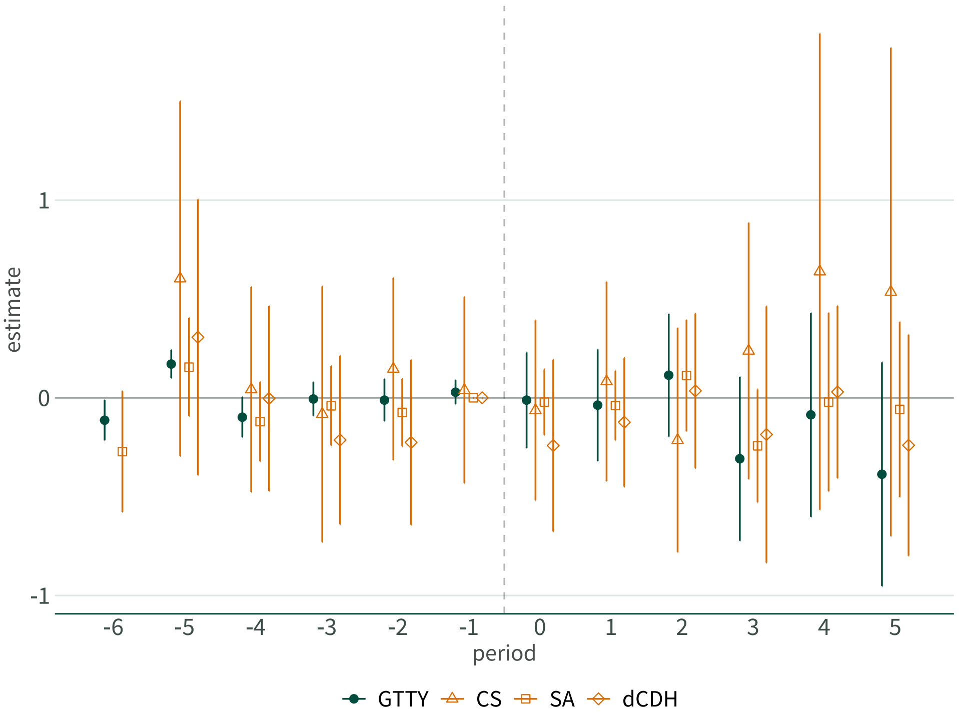
which allows us to back out the parameters of interest.

Figure A1: Event-study in non-staggered setting with pre-trend

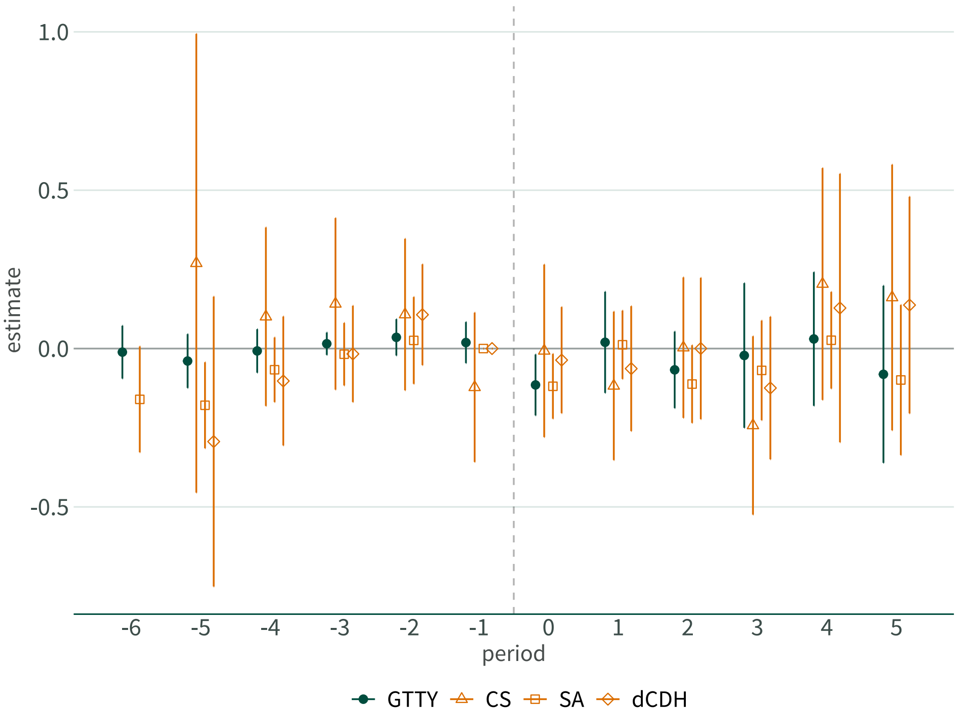


Note: This figure displays event-study estimates for a simulated dataset exhibiting a pre-trend from Roth (2024) by applying 2SDD with the first stage estimated using observations for eventually-treated units in the period immediately before they adopt the treatment as well as all observations for never-treated units. Under this data-generating process, the outcome for treated units follows a linear trend: $Y_{it} = 0.5 \cdot t \cdot D_i + \varepsilon_{it}$, where D_i is an indicator for treatment and ε_{it} are i.i.d. standard normal.

Figure A2: Empirical applications: Kuziemko, Meckel and Rossin-Slater (2018) event study estimates



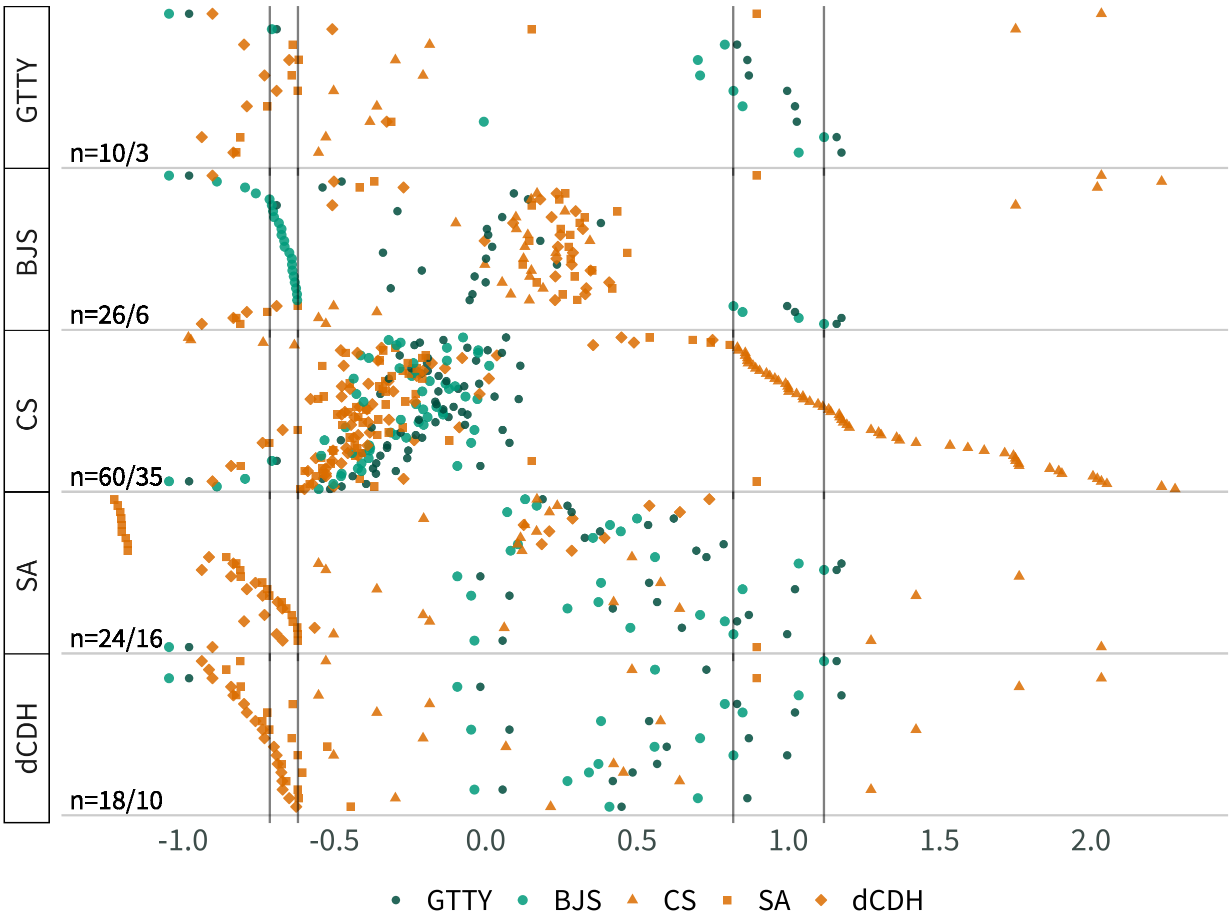
(a) Figure 2(a)



(b) Figure 2(b)

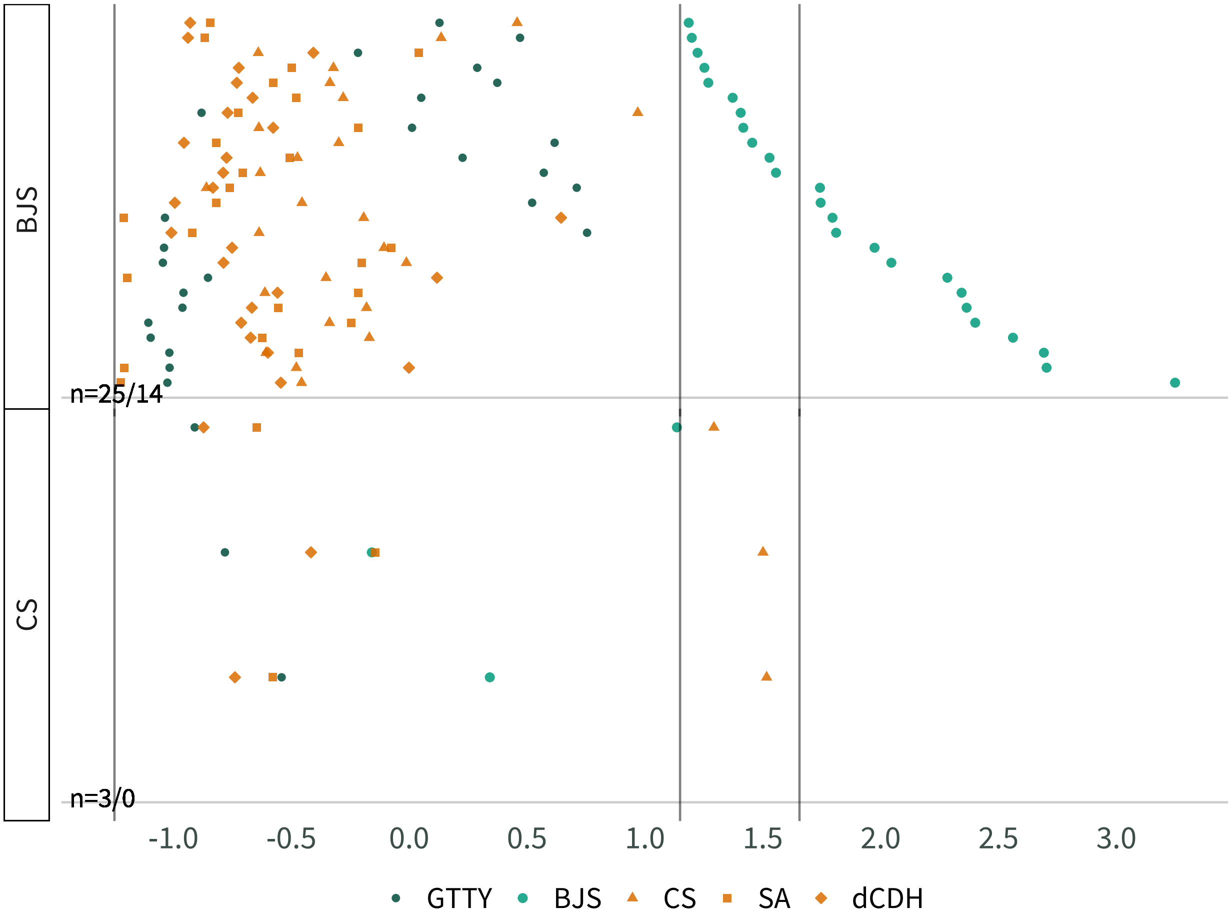
Note: This table reports event-study estimates from applying each estimator to the event-study specifications in Kuziemko, Meckel and Rossin-Slater (2018).

Figure A3: Empirical applications: Outlier post-treatment normalized standard error differences



Note: Each panel of this figure corresponds to one of the five estimators we investigate. Each entry for a given estimator corresponds to an estimate (associated with a particular post-treatment period, outcome variable, and empirical setting) for which that estimator's standard error significantly deviates from the average of the other methods' standard errors. Each entry displays the difference between each method's standard error and its associated leave-out mean, normalized by the average of the absolute value of the standard errors for that coefficient. The criterion for determining that an estimator's standard error significantly deviates from that of the other estimators is that the normalized difference falls in the top 2.5 percent or bottom 2.5 percent of the distribution (vertical bars closer to zero as thresholds), excluding estimates from the Bailey and Goodman-Bacon (2015) paper. The numbers in the bottom left of each panel indicate the number of such outlier estimates at the 5 percent level and 1 percent level, respectively.

Figure A4: Empirical applications: Outlier pre-treatment normalized standard error differences



Note: Among the five estimators we investigate, each panel of this figure corresponds to an estimator for which a standard error estimate (associated with a particular pre-treatment period before −1, outcome variable, and empirical setting) significantly deviates from the average of the other methods' standard errors. Each entry displays the difference between each method's standard error and its associated leave-out mean, normalized by the average of the absolute value of the standard errors for that coefficient. The criterion for determining that an estimator's standard error significantly deviates from that of the other estimators is that the normalized difference falls in the top 2.5 percent or bottom 2.5 percent of the distribution (vertical bars closer to zero as thresholds), excluding estimates from the Bailey and Goodman-Bacon (2015) paper. The numbers in the bottom left of each panel indicate the number of such outlier estimates at the 5 percent level and 1 percent level, respectively.

Table A1: Simulations (CPS wage data, heterogeneous treatment effects): 40 states treated over 20 years (2 per year)

Method	Period	Rejection rate	S.E.	Bias	RMSE	Speed (secs)
GTTY	0	4.79	0.1029	0.0063	0.1001	0.10
	1	5.39	0.1029	0.0062	0.1047	
	2	4.59	0.1041	-0.0075	0.1020	
	3	5.39	0.1043	0.0041	0.1029	
	4	5.79	0.1049	0.0038	0.1049	
BJS	0	12.97	0.0743	-0.0016	0.0971	0.20
	1	16.77	0.0751	-0.0036	0.1046	
	2	15.37	0.0760	-0.0010	0.1074	
	3	12.97	0.0759	-0.0006	0.1003	
	4	16.37	0.0773	-0.0126	0.1049	
BJS (leave out)	0	0.40	0.1405	-0.0016	0.0971	0.19
	1	1.20	0.1408	-0.0036	0.1046	
	2	1.80	0.1423	-0.0010	0.1074	
	3	1.00	0.1412	-0.0006	0.1003	
	4	0.80	0.1432	-0.0126	0.1049	
CS	0	2.99	0.1436	-0.0004	0.1340	31.07
	1	5.39	0.1420	0.0065	0.1381	
	2	4.39	0.1419	-0.0001	0.1302	
	3	3.59	0.1433	0.0055	0.1351	
	4	4.99	0.1433	-0.0026	0.1360	
SA	0	0.60	0.1666	0.0022	0.1274	46.95
	1	1.60	0.1667	0.0013	0.1415	
	2	1.80	0.1672	-0.0138	0.1373	
	3	1.80	0.1676	-0.0023	0.1358	
	4	1.60	0.1681	-0.0012	0.1394	
dCDH	0	4.99	0.1378	0.0013	0.1280	3.64
	1	6.39	0.1374	0.0016	0.1421	
	2	5.19	0.1390	-0.0126	0.1371	
	3	6.59	0.1380	-0.0010	0.1374	
	4	5.19	0.1389	-0.0011	0.1399	
W	0	4.19	0.1345	0.0019	0.1286	87.60
	1	6.79	0.1343	0.0009	0.1433	
	2	5.99	0.1358	-0.0141	0.1384	
	3	5.99	0.1347	-0.0027	0.1363	
	4	6.79	0.1358	-0.0016	0.1398	

Note: The table reports results from 501 simulations of 40 treated states over 20 years, with two treated states in each of those years. See the note accompanying Table 2 for further information.

Table A2: Simulations with non-uniform random treatment assignment (CPS wage data, heterogeneous treatment effects): 40 states treated over 20 years (2 per year)

Method	Period	Rejection rate	S.E.	Bias	RMSE	Speed (secs)
GTTY	0	5.99	0.0186	-0.0006	0.0192	0.06
	1	4.99	0.0185	-0.0004	0.0188	
	2	3.79	0.0186	-0.0003	0.0183	
	3	5.99	0.0186	0.0005	0.0191	
	4	4.79	0.0187	0.0009	0.0184	
BJS	0	15.37	0.0134	-0.0007	0.0184	0.29
	1	18.96	0.0134	-0.0006	0.0201	
	2	19.56	0.0134	-0.0002	0.0193	
	3	16.77	0.0134	0.0003	0.0187	
	4	17.56	0.0137	-0.0010	0.0184	
BJS (leave out)	0	0.40	0.0256	-0.0007	0.0184	0.20
	1	2.00	0.0255	-0.0006	0.0201	
	2	1.20	0.0254	-0.0002	0.0193	
	3	1.00	0.0253	0.0003	0.0187	
	4	0.60	0.0258	-0.0010	0.0184	
CS	0	3.19	0.0242	0.0005	0.0227	24.90
	1	5.39	0.0252	0.0007	0.0242	
	2	6.39	0.0246	-0.0001	0.0243	
	3	5.39	0.0245	0.0024	0.0242	
	4	4.39	0.0249	0.0001	0.0241	
SA	0	2.59	0.0284	-0.0006	0.0249	35.55
	1	2.99	0.0292	-0.0007	0.0239	
	2	1.80	0.0286	-0.0006	0.0243	
	3	2.40	0.0286	-0.0000	0.0252	
	4	1.60	0.0287	0.0006	0.0238	
dCDH	0	7.78	0.0234	-0.0008	0.0247	5.74
	1	6.19	0.0243	-0.0006	0.0242	
	2	7.39	0.0235	-0.0005	0.0243	
	3	8.18	0.0238	0.0002	0.0251	
	4	5.39	0.0239	0.0006	0.0238	
W	0	7.98	0.0228	-0.0006	0.0248	101.97
	1	6.59	0.0237	-0.0007	0.0238	
	2	6.39	0.0229	-0.0005	0.0242	
	3	8.78	0.0232	-0.0000	0.0251	
	4	5.59	0.0232	0.0006	0.0238	

Note: The table reports results from 501 simulations of 40 treated states over 20 years, with two treated states in each of those years. In these simulations, following Arkhangelsky et al. (2021), treatment assignment is correlated with systematic effects. See the note accompanying Table 2 for further information.

Table A3: Simulations (CPS wage data, homogeneous treatment effects): 40 states treated over 20 years (2 per year)

Method	Period	Rejection rate	S.E.	Bias	RMSE	Speed (secs)
GTTY	0	4.99	0.1025	0.0059	0.0997	0.11
	1	5.79	0.1026	0.0064	0.1046	
	2	4.99	0.1038	-0.0075	0.1020	
	3	4.99	0.1039	0.0041	0.1027	
	4	6.39	0.1045	0.0039	0.1047	
BJS	0	12.57	0.0741	-0.0016	0.0967	0.22
	1	16.97	0.0748	-0.0034	0.1034	
	2	16.17	0.0758	-0.0019	0.1073	
	3	12.57	0.0757	-0.0006	0.1000	
	4	15.77	0.0771	-0.0125	0.1045	
BJS (leave out)	0	0.40	0.1401	-0.0016	0.0967	0.59
	1	1.20	0.1402	-0.0034	0.1034	
	2	1.80	0.1419	-0.0019	0.1073	
	3	1.00	0.1407	-0.0006	0.1000	
	4	0.80	0.1428	-0.0125	0.1045	
CS	0	2.99	0.1434	-0.0002	0.1347	42.21
	1	5.59	0.1419	0.0062	0.1378	
	2	4.39	0.1416	0.0006	0.1305	
	3	3.59	0.1429	0.0059	0.1343	
	4	4.79	0.1430	-0.0029	0.1355	
SA	0	0.60	0.1663	0.0018	0.1272	38.92
	1	2.00	0.1664	0.0015	0.1417	
	2	2.00	0.1670	-0.0138	0.1371	
	3	2.00	0.1673	-0.0023	0.1357	
	4	1.80	0.1678	-0.0011	0.1396	
dCDH	0	4.79	0.1375	0.0009	0.1277	5.15
	1	6.59	0.1371	0.0017	0.1422	
	2	4.99	0.1388	-0.0126	0.1370	
	3	5.79	0.1377	-0.0010	0.1373	
	4	5.19	0.1386	-0.0010	0.1400	
W	0	4.59	0.1341	0.0015	0.1284	96.38
	1	6.59	0.1341	0.0011	0.1434	
	2	6.39	0.1357	-0.0141	0.1382	
	3	5.79	0.1344	-0.0027	0.1363	
	4	5.99	0.1355	-0.0015	0.1400	
TWFE	0	2.79	0.1367	0.0005	0.1230	0.16
	1	3.79	0.1365	0.0010	0.1378	
	2	4.39	0.1360	-0.0136	0.1327	
	3	4.99	0.1369	-0.0014	0.1342	
	4	4.59	0.1372	-0.0015	0.1365	
TWFE (no pre)	0	5.39	0.1011	0.0053	0.0969	0.11
	1	5.59	0.1009	0.0059	0.1036	
	2	4.39	0.1020	-0.0087	0.1002	
	3	4.99	0.1022	0.0036	0.1012	
	4	5.19	0.1028	0.0035	0.1042	

Note: The table reports results from 501 simulations of 40 treated states over 20 years, with two treated states in each of those years. Treatment effects are homogeneous and drawn from a normal distribution. TWFE denotes the two-way fixed effects estimator for a fully dynamic specification, estimating both pre-event and post-event coefficients. TWFE (no pre) denotes a two-way fixed effects specification that estimates only post-event coefficients. See the note accompanying Table 2 for further information.

Table A4: Simulations (i.i.d. data, heterogeneous treatment effects): 40 states treated over 20 years (2 per year)

Method	Period	Rejection rate	S.E.	Bias	RMSE	Speed (secs)
GTTY	0	5.19	0.1289	0.0009	0.1269	0.11
	1	4.59	0.1297	0.0073	0.1302	
	2	4.39	0.1300	-0.0105	0.1255	
	3	4.79	0.1307	0.0035	0.1262	
	4	4.99	0.1307	0.0012	0.1272	
BJS	0	16.17	0.0938	-0.0000	0.1301	0.22
	1	14.57	0.0940	-0.0017	0.1231	
	2	16.77	0.0945	0.0023	0.1334	
	3	14.77	0.0944	-0.0041	0.1266	
	4	16.37	0.0970	0.0021	0.1313	
BJS (leave out)	0	1.40	0.1774	-0.0000	0.1301	0.24
	1	0.60	0.1765	-0.0017	0.1231	
	2	1.20	0.1767	0.0023	0.1334	
	3	1.40	0.1758	-0.0041	0.1266	
	4	0.60	0.1797	0.0021	0.1313	
CS	0	3.79	0.1796	-0.0104	0.1702	32.34
	1	4.19	0.1799	-0.0008	0.1771	
	2	4.79	0.1785	0.0010	0.1668	
	3	4.59	0.1791	0.0044	0.1703	
	4	2.79	0.1800	-0.0033	0.1671	
SA	0	1.80	0.2087	-0.0036	0.1685	55.77
	1	1.80	0.2101	0.0013	0.1799	
	2	1.80	0.2098	-0.0170	0.1745	
	3	1.40	0.2090	-0.0031	0.1681	
	4	1.60	0.2092	-0.0034	0.1762	
dCDH	0	5.39	0.1721	-0.0050	0.1682	3.75
	1	6.19	0.1731	0.0020	0.1815	
	2	5.79	0.1731	-0.0163	0.1753	
	3	5.19	0.1724	-0.0024	0.1689	
	4	5.79	0.1728	-0.0046	0.1774	
W	0	5.79	0.1683	-0.0043	0.1696	77.76
	1	6.99	0.1692	0.0007	0.1815	
	2	5.39	0.1695	-0.0175	0.1766	
	3	5.99	0.1685	-0.0039	0.1697	
	4	6.19	0.1694	-0.0042	0.1773	

Note: The table reports results from 501 simulations of 40 treated states over 20 years, with two treated states in each of those years. The outcome data are drawn i.i.d. from a normal distribution with the same mean and variance as that of the wage data used in Table A1. See the note accompanying Table 2 for further information.

Table A5: Empirical applications: Change in standard errors across treatment periods

	BGB2015	D2017	HW2017	LRS2018	T2014	U2014
BJS ×period	1.1382	-0.0000	-0.0003	-12.6865	0.0002	-0.0002
	(0.7426)	(0.0000)	(0.0006)	(4.0584)	(0.0002)	(0.0007)
CS ×period	-12.6385	0.0011	-0.0115	15.6287	0.0001	-0.0035
	(9.2480)	(0.0003)	(0.0032)	(4.6711)	(0.0001)	(0.0019)
SA ×period	-0.2929	0.0003	-0.0008	22.8739	-0.0009	0.0016
	(0.1681)	(0.0001)	(0.0031)	(6.7402)	(0.0002)	(0.0007)
dCDH ×period	0.1798	0.0004	-0.0020	24.0197	0.0005	0.0016
	(0.1225)	(0.0001)	(0.0054)	(7.3067)	(0.0003)	(0.0007)

Note: Each column reports estimates of method-specific linear period trends from a regression of standard error estimates on period fixed effects, method fixed effects, and method-specific linear period trends, corresponding to each of the papers in Table 3. The regression omits the linear period trend for the 2SDD estimator. See Table S15 for additional details.

Table A6: Empirical applications: Comparison of t-statistics (post-treatment periods)

	$\|t\|$	$\mathbb{1}_{\{\|t\|>1.96\}}$	$\mathbb{1}_{\{\|t\|>p_{90}\}}$	$\mathbb{1}_{\{\|t\|>p_{99}\}}$
Panel A: Unweighted				
BJS	0.5056	0.0762	0.0976	0.0244
	(0.1280)	(0.0381)	(0.0238)	(0.0085)
CS	-0.2912	-0.0823	-0.0122	-0.0000
	(0.0983)	(0.0361)	(0.0173)	(0.0000)
SA	0.9228	0.0884	0.1159	0.0213
	(0.2440)	(0.0381)	(0.0246)	(0.0080)
dCDH	0.5061	0.1067	0.0945	0.0122
	(0.1175)	(0.0382)	(0.0237)	(0.0061)
Panel B: Weighted (outcomes)				
BJS	0.4607	0.0760	0.0882	0.0220
	(0.1217)	(0.0393)	(0.0220)	(0.0077)
CS	-0.2664	-0.0765	-0.0064	0.0000
	(0.0966)	(0.0368)	(0.0168)	(0.0000)
SA	0.9153	0.0745	0.1084	0.0230
	(0.2641)	(0.0391)	(0.0232)	(0.0086)
dCDH	0.4339	0.0851	0.0854	0.0110
	(0.1141)	(0.0391)	(0.0219)	(0.0055)
Panel C: Weighted (papers)				
BJS	0.3373	0.0808	0.0635	0.0162
	(0.1409)	(0.0568)	(0.0154)	(0.0060)
CS	0.1307	0.0399	0.0553	-0.0000
	(0.1783)	(0.0618)	(0.0379)	(0.0000)
SA	3.4792	0.1622	0.1833	0.1121
	(1.2457)	(0.0612)	(0.0426)	(0.0416)
dCDH	0.2260	0.0359	0.0380	0.0040
	(0.1325)	(0.0499)	(0.0118)	(0.0021)

Note: This table describes the relationship between each estimator and the absolute t-statistics of the dynamic treatment effect estimates from applying each estimator to the empirical settings in Table 3. Each observation is an estimate of a treatment effect in each post-treatment period associated with each outcome in each paper using each of the five methods. The first column uses the absolute value of the t-statistic as the dependent variable. The second column uses an indicator for significant t-statistics using a conventional threshold (the absolute value of the t-statistic exceeding 1.96) as the dependent variable. The last two columns use an indicator for more extreme levels of statistical significance (the absolute value of the t-statistic exceeding the 90^{th} and 99^{th} percentiles, respectively, of the distribution of estimates in our sample) as the dependent variable; the 90^{th} percentile is approximately 4.3 and the 99^{th} percentile is approximately 7.4. All specifications use a balanced sample of coefficients that all methods can estimate. The estimates in panels B and C use the inverse of the number of periods for each outcome variable and the inverse of the number of outcomes for each paper, respectively, as weights. We report heteroskedasticity-robust standard errors in parentheses.