Two-Stage Differences in Differences

John Gardner
University of Mississippi

Neil Thakral
Brown University

Linh T. Tô
Boston University

Luther Yap
Princeton University

May 2024*

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 estimator that compares treated and untreated outcomes after removing group and period effects identified from a regression using untreated observations. Our regression-based approach enables us to conduct inference within a conventional GMM asymptotic framework. It easily facilitates fairly standard extensions such as estimating dynamic treatment effects and triple differences; incorporating time-varying controls, individual unit fixed effects, and different approaches to testing parallel trends; and considering violations of the parallel trends assumption. To understand the finite sample properties of our estimator, we conduct simulations of randomly generated laws in state-level wage data, extending the "placebo law" analysis of Bertrand, Duflo and Mullainathan (2004) to a setting with heterogeneous treatment effects and staggered treatment timing. Our method outperforms alternative approaches for estimation and inference based on precision and rejection rates. Even with homogeneous treatment effects, our approach yields similar standard errors as two-way fixed effects regressions, unlike other proposed heterogeneity-robust estimators. Across seven empirical applications, we compare the relative performance of the different methods by analyzing the rate of extreme t-statistics and outlying standard errors relative to one another. Our two-stage approach thus 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 Michael Briskin, Kyle Butts, Carolina Caetano, Gregorio Caetano, Brantly Callaway, Scott Cunningham, David Drukker, Len Goff, Zhanyuan Tian, Tao Wang, Taylor Wright, and participants at the Society of Labor Economists annual meeting, the Southern Economic Association annual meeting, and the Western Economics Association International conference 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 that is robust to treatment-effect heterogeneity when adoption of the treatment is staggered over time. The first stage regresses outcomes on group and period fixed effects using the subsample of untreated observations. 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. Furthermore, we demonstrate how to extend this approach to recover a variety of treatment effect measures, including event-study analyses of pre-trends and duration-specific average treatment effects.

We derive the asymptotic distribution of the treatment effect estimates by interpreting the two-stage procedure as a joint GMM estimator. The two-stage estimator, along with valid asymptotic standard errors, can thus be implemented easily using standard statistical software, with little programming or computational time beyond that required to estimate a regression.[1] Our approach to estimation and inference can be extended to a variety of settings. This includes estimating individual fixed effects or any linear combination of coefficients in

[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 in Appendix A; also see the `did2s` Stata and R packages (Butts, 2021).

the regression model, as well as accommodating time-varying covariates. We also discuss how to extend our approach to settings with partial violations of the parallel trends assumption, continuous treatments, triple differences, and design-based settings.[2]

To evaluate the performance of our proposed estimator, we conduct simulations of randomly generated laws using state-level wage data. This 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 with randomly drawn treated states, their associated years of passage, and treatment effects. Using a 42-year panel, we compare the rejection rates at the 5 percent significance level from recently proposed heterogeneity-robust estimators (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Wooldridge, 2021; Borusyak, Jaravel and Spiess, 2024; de Chaisemartin and d'Haultfoeuille, 2024).[3] The simulations highlight the value of our approach to estimation and inference by demonstrating its finite-sample performance. Our estimator consistently offers the best performance in terms of rejection rates, computational speed, and efficiency. This holds even in comparison with the imputation approach from Borusyak, Jaravel and Spiess (2024) 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 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 reduce 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 Callaway and Sant'Anna (2021)

[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.

[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.

estimator yields large standard errors with a large number of treatment cohorts (e.g., Bailey and Goodman-Bacon, 2015; Deryugina, 2017), while the Sun and Abraham (2021) and de Chaisemartin and d'Haultfoeuille (2024) estimators perform relatively poorly with a relatively large number of small cohorts (e.g., Bailey and Goodman-Bacon, 2015; Deryugina, 2017).

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 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. Borusyak, Jaravel and Spiess (2024) show that TWFE identifies a regression-weighted mean of the average effect of the treatment in each post-treatment period, and de Chaisemartin and d'Haultfoeuille (2020) 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-specific average treatment effects (de Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021).

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

[5] See, for example, de Chaisemartin and d'Haultfoeuille (2020); Goodman-Bacon (2021); Imai and Kim (2021); Sun and Abraham (2021); Athey and Imbens (2022); Borusyak, Jaravel and Spiess (2024).

[6] Gibbons, Suárez Serrato and Urbancic (2018) suggest an approach like this for fixed effects models; Borusyak, Jaravel and Spiess (2024) 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.

out by Borusyak, Jaravel and Spiess (2024). 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 (Wooldridge, 2021). Several limitations arise when applying these methods. First, the stacked estimator identifies 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.

Given the multitude of alternative approaches for DD estimation, our simulation and empirical exercises constitute a distinct contribution to this literature. 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. 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 two-stage method achieves the closest standard error to that of a TWFE estimator that imposes a null effect in the pre-treatment periods.[8] 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.

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

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

The paper proceeds as follows. In Section 2, we introduce the main idea in a simple setting with group and period fixed effects and without covariates. We provide intuition for why the TWFE approach to DD estimation may not identify the average effect of the treatment on the treated and show how our proposed two-stage regression-based approach is robust to treatment-effect heterogeneity in settings with variation in treatment timing. While Section 3 presents theoretical results in a more general setting that can include covariates and individual fixed effects, applied readers may benefit more from Sections 4 and 5, which demonstrate the performance of the two-stage approach compared to alternative proposals in simulations and empirical applications. We conclude in Section 6.

2 Motivating the two-stage approach in a simplified setting

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. Let Y_{gpit}, $Y_{gpit}(d = 1)$ and $Y_{gpit}(d = 0)$ denote the observed, treated, and untreated potential outcomes for the ith member of group g during time t of period p, let D_{gp} be an indicator for whether members of group g are treated in period p, and let $\beta_{gp} = \mathbb{E}\left[Y_{gpit}(d = 1) - Y_{gpit}(d = 0) \,\middle|\, g, p\right]$ 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 3.3). Under parallel trends, mean outcomes satisfy

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

[9]These expressions implicitly hold treatment times fixed at their observed values. Causal effects for the never-treated group may be normalized to zero.

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.

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:

$$Y_{gpit} = \lambda_g + \alpha_p + \beta D_{gp} + \varepsilon_{gpit}. \tag{2}$$

It follows from Equation (1) that the coefficient on D_{gp} in Equation (2) identifies the average effect of the treatment on the treated, $\mathbb{E}\left[Y_{gpit}(d=1) - Y_{gpit}(d=0) \,\middle|\, D_{gp} = 1\right]$.[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 (de Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Imai and Kim, 2021; Athey and Imbens, 2022; Borusyak, Jaravel and Spiess, 2024), when the average effect of the treatment varies across groups and over periods, the coefficient on D_{gp} 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 (see Section 2.4 below), an obvious candidate is the average $\mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right] = \mathbb{E}\left[Y_{gpit}(d=1) - Y_{gpit}(d=0) \,\middle|\, D_{gp} = 1\right]$ 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

[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_{gp} in Equation (2) can be obtained by regressing Y_{gpit} 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.

can be expressed as

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

The difficulty with the regression approach is that, except in special cases, the "error term" $[\beta_{gp} - \mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right]] D_{gp}$ 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 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 (de Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Borusyak, Jaravel and Spiess, 2024).

2.2 The difference-in-differences regression estimand

In light of the preceding argument, we discuss what the DD regression identifies. To provide additional insight into the difference-in-differences estimand, it can be shown that, under parallel trends, the coefficient on D_{gp} from the difference-in-differences regression specification (2) identifies the following weighted average of β_{gp}:

$$\beta^* = \sum_{g=1}^{G} \sum_{p=g}^{P} \omega_{gp} \beta_{gp},$$

with weights that take the form

$$\omega_{gp} = \frac{\tilde{\omega}_{gp}}{\sum_{g'=1}^{G} \sum_{p'=g'}^{P} \tilde{\omega}_{g'p'}}, \tag{3}$$

where

$$\tilde{\omega}_{gp} = \left[\left(1 - \Pr\left(D_{gp} = 1 \,\middle|\, g\right)\right) - \left(\Pr\left(D_{gp} = 1 \,\middle|\, p\right) - \Pr\left(D_{gp} = 1\right)\right)\right] \Pr(g, p),$$

$\Pr(D_{gp} = 1 \,|\, p)$ is the fraction of units that are treated in period p, $\Pr(D_{gp} = 1 \,|\, g)$ is the fraction of periods in which members of group g are treated, $\Pr(D_{gp} = 1)$ is the fraction of unit×times that are treated, and $\Pr(p, g)$ is the population share of observations that correspond to group g and period p. This representation can be obtained from Theorem 1 of de Chaisemartin and d'Haultfoeuille (2020), who note that the weights ω_{gp} may also be negative. Our Appendix B presents an alternative derivation based on population regression algebra.[11]

Appearances notwithstanding, this weighting scheme is deeply intuitive. Specification (2) assumes a conditional expectation function that is linear in group, period, and treatment status. When misspecified, it will attribute some of the heterogeneous impacts of the treatment to group and period fixed effects.[12] As a group's observed treatment duration increases (i.e., the greater $\Pr(D_{gp} = 1 \,|\, g)$ is), more of that group's treatment effects will be absorbed by group fixed effects. Similarly, as the probability of being treated in a particular period (i.e., $\Pr(D_{gp} = 1 \,|\, p)$) increases, more of that period's treatment effects will be absorbed by period effects. Larger groups also receive more weight.

2.3 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
$$Y_{gpit} = \lambda_g + \alpha_p + u_{gpit}$$
on the sample of observations for which $D_{gp} = 0$, retaining the estimated group and time effects $\hat{\lambda}_g$ and $\hat{\alpha}_p$.

2. Regress adjusted outcomes $Y_{gpit} - \hat{\lambda}_g - \hat{\alpha}_p$ on D_{gp}.

[11] One immediate implication of Equation (3) is that the weights must sum to one. Another is that $\omega_{11} = 1$ when there is only one treatment group, so the regression DD specification (2) identifies the average effect of the treatment on the treated, as noted above.

[12] This is consistent with the intuition that Equation (2) uses already-treated units as controls for newly treated ones (de Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Borusyak, Jaravel and Spiess, 2024).

Since parallel trends implies that

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

where $\mathbb{E}\left[[\beta_{gp} - \mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right]] D_{gp} \,\middle|\, D_{gp}\right] = 0$, 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.

Unbiasedness of the first-stage (and hence second-stage) estimates follows from standard arguments. If P is fixed as the sample size increases, so does the consistency of the first stage for the group and period effects. The consistency of the second-stage for $\mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right]$ follows from the consistency of the first stage for the group and period effects and the continuous mapping theorem.[13] As we show below in Section 3, a variation on the usual within estimator can be used to apply this procedure using individual, rather than group, fixed effects.

In DD analyses based on two-way fixed-effects regression, it is common to control for observable time-varying covariates by simply including them in the regression. The two-stage approach can readily be adapted to allow for such covariates: simply include them in the first-stage regression and amend the second-stage to

2'. Regress $Y_{gpit} - \widehat{\lambda}_g - \widehat{\alpha}_p - X'_{gpit}\widehat{\gamma}$ on D_{gp} (where X_{gpit} is the vector of covariates and $\widehat{\gamma}$ is the estimated vector of coefficients on X_{gpit} from the first-stage regression).

While this approach allows the effect of the treatment to depend arbitrarily on observable covariates, it does implicitly rule out the possibility of feedback from the treatment to the covariates and, as Sant'Anna and Zhao (2020) note, covariate-specific trends.[14]

2.4 2SDD estimands

Implemented as described, the two-stage difference-in-differences (2SDD) estimator identifies $\mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right]$, where the expectation is implicitly taken with respect to all observed units and periods. This expectation can be expressed as

$$\mathbb{E}\left[\beta_{gp} \,\middle|\, D_{gp} = 1\right] = \sum_{g=1}^{G} \sum_{p=g}^{P} \beta_{gp} \Pr\left(g, p \,\middle|\, D_{gp} = 1\right), \tag{4}$$

[13] Also note that restricting the sample to untreated observations does not introduce sample-selection bias because the selection is with respect to treatment status.

[14] In principle, the two-stage approach can be modified to accommodate the more stringent notion of conditional parallel trends introduced by Callaway and Sant'Anna (2021) by interacting the covariates with time indicators. Caetano et al. (2022) discuss how a two-stage approach (in addition to methods based on inverse-probability weighting) can be used when the treatment also affects the covariates.

where $\Pr(g, p \,|\, D_{gp} = 1)$ is the population share of treated unit-times that correspond to group g in period p. While this is a natural summary measure of group×period-specific average treatment effects, and can be interpreted as an average effect on the treated (ATT), it may not be especially informative for program evaluation and policy analysis. For example, even if the effects of the treatment are identical across groups, this measure will put more weight on groups that are in early stages of the treatment.[15] Callaway and Sant'Anna (2021) provide a much richer discussion of how heterogeneous average treatment effects can be summarized.

If there is some treatment duration $\bar{P}$ such that a subset of groups has been treated for $\bar{P}$ periods, then an alternative summary measure is the $\bar{P}$-period average

$$\sum_{g=1}^{G} \sum_{p=g}^{g+\bar{P}-1} \beta_{gp} \Pr(g \,|\, D_g = 1)/\bar{P}, \tag{5}$$

where $\Pr(g \,|\, D_g = 1)$ is the fraction of treated units that belong to group g. Because this measure averages the group-specific average effects of the treatment for a common set of completed durations, it may provide a more balanced picture of the typical effect of the treatment, although it ignores the effects of the treatment for durations longer than $\bar{P}$ periods. The two-stage procedure can be modified to identify this measure by restricting the sample used in the second step to untreated observations and treated observations with durations no greater than $\bar{P}$.

It is worth noting that the two-stage procedure is equivalent to estimating the two-way fixed effects model

$$Y_{gpit} = \lambda_g + \alpha_p + B'_{gpit}\theta + e_{it},$$

where B_{gpit} is a saturated set of interactions between group and period indicators for all treated observations, then aggregating the group×period-specific treatment effects estimates in θ as the sample analog of $\mathbb{E}\left[\beta_{gp} \,|\, D_{gp} = 1\right]$. One way to see this is to note that, by the Frisch-Waugh-Lovell (FWL) theorem, estimates of the λ_g and α_p can be obtained by regressing Y_{gpit} on the residuals from auxiliary regressions of group and period indicators on B_{gpit}. Since B_{gpit} perfectly predicts group and time for all treated observations, the residuals from these auxiliary regressions will be zero for all treated units. Consequently, λ_g and α_p are identified from variation in untreated outcomes, as they are in the two-stage procedure. In either case, the overall ATT is identified as $\mathbb{E}\left[\beta_{gp} \,|\, D_{gp} = 1\right] = \mathbb{E}\left[Y_{gpit} - \lambda_g - \alpha_p \,|\, D_{gp} = 1\right]$.[16]

[15] When treatment effects vary by group, it is unclear whether any summary measure will be informative about how the treatment might affect future groups. External validity with this type of heterogeneity is inherently challenging.

[16] The same equivalence applies when covariates are included in the first stage of the two-stage procedure,

2.5 Event studies

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

$$Y_{gpit} = \lambda_g + \alpha_p + \sum_{r=-\underline{R}}^{\overline{R}} \eta_r W_{rgp} + u_{gpit}, \tag{6}$$

where for $r \leq 0$ the $W_{rgp} \in \{W_{-\underline{R}gp}, \dots, W_{0gp}\}$ are $(r+1)$-period leads of treatment adoption, and for $r > 0$ the $W_{rgp} \in \{W_{1gp}, \dots, W_{\overline{R}gp}\}$ are r-period lags of adoption (i.e., indicators for being r periods since treatment).[17] In principle, such regressions serve a dual purpose. First, they can be used to show how the effect of the treatment evolves over the course of the treatment. Second, the coefficients on the treatment adoption leads can be used as placebo tests for the plausibility of parallel trends.

Sun and Abraham (2021) show that, when duration-specific average treatment effects vary across groups, event-study regressions suffer from the same problem as DD regressions. This can be seen using an argument similar to the one presented for DD regressions in Section 2.1. Let $Y_{gpit}(r)$ denote potential outcomes after r periods of treatment, and $\eta_{rgp} = \mathbb{E}\left[Y_{gpit}(r) - Y_{gpit}(0) \,\middle|\, g, p, W_{rgp} = 1\right]$ be the average effect of being treated for r periods for members of group g in time period p.[18] Under parallel trends,

$$\mathbb{E}\left[Y_{gpit} \,\middle|\, g, p, \{W_{rgp}\}_{r=-\underline{R}}^{\overline{R}}\right] = \lambda_g + \alpha_p + \sum_{r=1}^{\overline{R}} \mathbb{E}\left[\eta_{rgp} \,\middle|\, W_{rgp} = 1\right] D_{rpg} + \sum_{r=1}^{\overline{R}} [\eta_{rgp} - \mathbb{E}\left[\eta_{rgp} \,\middle|\, W_{rgp} = 1\right]] W_{rgp},$$

where, in general, $\mathbb{E}\left[\sum_{r=1}^{P^*} [\eta_{rgp} - \mathbb{E}\left[\eta_{rgp} \,\middle|\, W_{rgp} = 1\right]] W_{rgp} \,\middle|\, g, p, (W_{rgp})\right] \neq 0$. Hence, mean outcomes are not necessarily linear in group, period, and treatment-duration indicators, so the coefficients on the W_{rgp} from Equation (6) do not identify the average effects of being treated for r periods. Sun and Abraham (2021) further show that the coefficients on the adoption leads and duration indicators identify weighted averages of all of the group×period-specific average treatment effects. An important consequence of this is that the coefficients on the treatment-adoption leads W_{rgp}, $r \leq 0$, may be nonzero even if trends are, in fact, parallel.

---

with the caveat that, in this case, the two-way fixed-effects regression should include unit and time (rather than only group and period) indicators, and B_{gpit} should contain a saturated set of unit and time indicators for treated observations.

[17] In event-study regressions, it is common practice to use calendar times t in place of more coarse treatment periods p. When researchers do not wish to include leads, $\underline{R}$ can be set to zero.

[18] There is a one-to-one correspondence between duration- and period-specific treatment effects. In terms of the group×period average treatment effects β_{gp}, the duration-specific effects satisfy $\eta_{rgp} = \beta_{g,p-g+1}$. While in principle the duration-specific average treatment effects for each group might vary over time, in practice we only ever observe each treatment duration at most once for each group.

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 $Y_{gpit} - \hat{\lambda}_g - \hat{\alpha}_p$ on $W_{-\underline{R}gp}, \dots, W_{0gt}, \dots, W_{\overline{R}gp}$.

Following the logic of the previous section, because $\mathbb{E}\left[Y_{gpit} \,\middle|\, g, p, (W_{rgp})\right] - \lambda_g - \alpha_p$ is linear in the W_{rgp}, the coefficients on the W_{rgp}, $r > 0$, identify the average effects $\mathbb{E}\left[\eta_{rgp} \,\middle|\, W_{rgp} = 1\right]$.[19] For $r \leq 0$, the coefficients on the W_{rgp} can be used to test the hypothesis that $\mathbb{E}\left[Y_{gpit} \,\middle|\, g, p, W_{rgp} = 1\right] = \lambda_g + \alpha_p$ (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.

2.6 Alternative approaches to testing parallel trends

There are alternative approaches to testing the validity of parallel trends within the two-stage framework. Borusyak, Jaravel and Spiess (2024) 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 $K + 1$ periods before the adoption of the treatment, then use the two-stage procedure to estimate the K pre-treatment placebo ATTs (i.e., the coefficients on D_{rgp} for $r \in \{-K, \dots, -1\}$). This approach is also suggested by Liu, Wang and Xu (2022), who also develop an equivalence test to increase the power of tests based on this idea.[20]

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)

[19] 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 Sun and Abraham (2021) 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.

[20] 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.

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).[21] 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_{rgp} for $r \in \{-K, \dots, -1\}$ for this variant of the two-stage procedure identify average pre-treatment deviations among eventually-treated units from never-treated units' trends.[22] 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 20 and Roth 2024). In Appendix Figure 1, 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 de Chaisemartin and d'Haultfoeuille (2020); Callaway and Sant'Anna (2021); Borusyak, Jaravel and Spiess (2024) estimators do not. While the coefficients on leads of treatment status from this modified procedure are more readily interpretable, the analogous coefficients from 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 de Chaisemartin and d'Haultfoeuille (2024) argue is shared by other estimators that do not compare treated observations to all untreated observations).

2.7 Inference

The standard errors for the two-stage estimators need to be adjusted to account for the fact that the dependent variable $Y_{gpit} - \widehat{\lambda}_g - \widehat{\alpha}_p$ in the second-stage is generated using estimates obtained from the first stage of the procedure (Dumont et al., 2005). Perhaps the simplest way to obtain valid standard errors is using a bootstrap procedure in which both stages of the estimator are estimated in each bootstrap replication (this is the approach used in Liu,

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

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

Wang and Xu, 2022). The asymptotic distribution of the second-stage estimates can also be obtained by interpreting the two-stage procedure as a joint GMM estimator (Hansen, 1982).[23]

Let $Z_{gpit} = [Y_{gpit}, (1(g)_{gpit}), (1(p)_{gpit}), D_{gp}]$ denote the data for observation (g, p, i, t), consisting of the outcome Y_{gpit}, the G-vector of group-membership indicators $(1(g)_{gpit})$, a P-vector $(1(p)_{gpit})$ of period indicators for periods $p \in \{1, \dots, P\}$, and the treatment-status indicator D_{gp}. Let λ be the G-vector of group fixed effects, and α the P-vector of period fixed effects. The two-stage difference-in-differences estimator solves the sample analog of the moment condition

$$\begin{aligned}\mathbb{E}\left[f(\lambda, \alpha, \beta; W_{gpit})\right] &= \mathbb{E}\begin{bmatrix}[Y_{gpit} - (1(g)_{gpit})'\lambda - (1(p)_{gpit})'\alpha][(1(g)_{gpit}), (1(p)_{gpit})]'(1 - D_{gp}) \\ [Y_{gpit} - (1(g)_{gpit})'\lambda - (1(p)_{gpit})'\alpha - \beta D_{gp}]D_{gp}\end{bmatrix} \\ &= 0.\end{aligned}$$

By Theorem 6.1 of Newey and McFadden (1994, cf. Newey, 1984), and under standard regularity conditions, $\sqrt{N}(\hat{\beta} - \beta) \overset{a}{\sim} \mathcal{N}(0, v)$, where v is the last element of

$$\mathbb{E}\left[\frac{\partial f(\lambda, \alpha, \beta; Z_{gpit})}{\partial(\lambda, \alpha, \beta)}\right]^{-1} \mathbb{E}\left[f(\lambda, \alpha, \beta; Z_{gpit}) f(\lambda, \alpha, \beta; Z_{gpit})'\right] \mathbb{E}\left[\frac{\partial f(\lambda, \alpha, \beta; Z_{gpit})}{\partial(\lambda, \alpha, \beta)}\right]^{-1'}.$$

The preceding expression can be used to manually correct the estimated second-stage variances for the use of a generated dependent variable. With modern statistical software, a simpler approach is to estimate both stages of the procedure simultaneously using a GMM routine.[24] This GMM approach does not require doing inference on group-specific treatment effects, which differs from existing papers—further differences are summarized in Table 1.

3 General theory for 2SDD

This section provides the theoretical results behind the main ideas presented in Section 2, and considers a more general setting with covariates and individual fixed effects, nesting Section 2 as a special case. We observe $\{(Y_{it}, X_{it}, D_{it})\}$, where $i \in \{1, 2, \cdots, N\}$ indexes individuals and $t \in \{1, 2, \cdots, T\}$ indexes time, so there are NT observations in a balanced panel. Since indices i and t are sufficient to determine the group g and period p, the indices g, p are

[23] Borusyak, Jaravel and Spiess (2024) provide an alternative derivation of the asymptotic distribution of the two-stage difference-in-differences and related imputation estimators.

[24] It is also possible to use the result of Theorem 6.1 of Newey and McFadden (1994, cf. Newey, 1984) to isolate the component of the variance matrix corresponding to the treatment effect estimate(s) (see Butts and Gardner, 2022, for a discussion of this approach).

dropped to avoid notational clutter. The parallel trends assumption now takes the form

$$Y_{it} = \lambda_i + \alpha_t + D_{it}\beta_{it} + X_{it}'\gamma + \varepsilon_{it}, \tag{7}$$

where

$$\mathbb{E}\left[\varepsilon_{it} \,\middle|\, \{D_{it}, X_{it}\}_{t=1}^{T}\right] = 0, \tag{8}$$

i.e., ε is mean independent of X and D. Equation (7) with an assumption on the errors is the parallel trends assumption.

Here, D_{it} is an indicator for treatment status, β_{it} is the heterogeneous treatment effect, and $X_{it} \in \mathbb{R}^K$ is a vector of covariates. Since we consider a setting where N is large and T is fixed, in a slight abuse of notation we now redefine X_{it} to include time indicators, so that the vector γ of coefficients on X_{it} also includes time fixed effects. Note that, compared to the simplified setting in Section 2, Equation (7) now includes individual fixed effects, the coefficient on D_{it} is now β_{it} and, accordingly, the error term is denoted by ε_{it}.

We also make the following substantive assumptions.

Assumption 1. *Assume that:*

1. *(Parallel trends) Outcomes satisfy Equations (7) and (8).*
2. *For all i, there exists some t where $D_{it} = 0$, and $\mathbb{E}\left[\sum_{t=1}^{T} D_{it}\right] > 0$.*
3. *Observations $\{(Y_{it}, D_{it}, X_{it})\}_{t=1}^{T}$ are independent and identically distributed over individuals i.*
4. *T is fixed as $N \to \infty$.*

Assumption 1.1 tells us that our model is correctly specified when heterogeneous treatment effects are accounted for. In the special case without covariates, and heterogeneity of β_{gp} only at the group and period level, Assumption 1.1 reduces exactly to the parallel trends assumption stated in Section 2.1.[25] Assumption 1.2 requires everyone to be untreated for at least one period. As is standard in this environment, Assumption 1.3 does not require independence across time for a given individual, but requires independence over individuals.

[25] To see this, using the notation in Section 2.1, the assumption becomes: $\mathbb{E}\left[Y_{it} - D_{gp}\beta_{gp} - \lambda_g - \alpha_p \,\middle|\, D_{gp}, g, p\right] = 0$, which implies:

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

The assumption of i.i.d. data is unnecessary, but it is assumed for exposition and to avoid notational clutter.[26]

3.1 2SDD

Our general proposed procedure is:

1. Regress Y_{it} on X_{it} and individual fixed effects to obtain $\hat{\gamma}$ and $\hat{\lambda}_i$ for observations with $D_{it} = 0$.
2. Regress adjusted outcomes $Y_{it} - \hat{\lambda}_i - X_{it}'\hat{\gamma}$ on D_{it}.

Due to the FWL theorem, the first step of this procedure is equivalent to running the regression using data that have been transformed into deviations from individual means in the untreated sample. Consequently, we can recover the same $\hat{\gamma}$, and consequently $\hat{\beta}$, even though the λ_i are not consistently estimated.

To be precise, let $T_i^0 := \sum_{t=1}^{T} (1 - D_{it})$ denote the number of periods that individual i is untreated, and define $\tilde{Y}_{it}$ as

$$\tilde{Y}_{it} = Y_{it} - \frac{1}{T_i^0} \sum_{t=1}^{T_i^0} Y_{it} (1 - D_{it}) .$$

Define $\tilde{\varepsilon}_{it}$ similarly, and let $\tilde{X}_{it}$ denote the matrix of deviations of the elements of X_{it} from their individual untreated means. The procedure given above is equivalent to the following:

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 $\hat{\beta}$.

Using X_{kit} to denote regressor k for individual i at time t, $\tilde{X}_{0i}$ is a $T \times K$ matrix of the form

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

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

$$\tilde{Y}_{0i} = \begin{bmatrix} \tilde{Y}_{i1} (1 - D_{i1}) & \cdots & \tilde{Y}_{iT} (1 - D_{iT}) \end{bmatrix}' .$$

[26] Identical distribution of the triple does not contradict heterogeneous treatment effects, because the identical distribution of Y_{it} can be attributed to the identical distribution of ε_{it} instead of β_{it}. Heterogeneous treatment effects can also arise from variation in treatment timing even if the time- and duration-specific effects of the treatment are identically distributed.

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). The second stage-regression is done without a constant, because the data are already demeaned. Hence, the two-stage difference in difference estimator is

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

We define β as the average treatment effect on the treated (ATT):

$$\beta := \mathbb{E}[\beta_{it} \mid D_{it} = 1]$$

where the expectation of β_{it} is taken over all units that receive treatment and all times during which they receive it, as in Section 2.4. The estimators can be written as the solution to the following GMM problem

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

so we have $K + 1$ moment conditions, with K in the first stage and one in the second stage. The moment condition reduces to that of Section 2.7 as a special case.

Assumption 2. *Assume that:*

1. $\mathbb{E}\left[\left\| \tilde{X}_{0i}' \tilde{\varepsilon}_{0i} \right\|^2 \right] < \infty$, $\mathbb{E}[\tilde{\varepsilon}_{it}^2] < \infty$, *and* $\mathbb{E}\left[(\beta_{it} - \beta)^2 \right] < \infty$.
2. $\mathbb{E}\left[\tilde{X}_{0i}' \tilde{X}_{0i} \right]$ *is invertible and* $\mathbb{E}\left[\left\| \tilde{X}_{0i}' \tilde{X}_{0i} \right\|^2 \right] < \infty$.

The invertibility condition of Assumption 2.2 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.[27]

[27] As an extreme case, suppose there are two treatment cohorts with one state each (2000 and 2001) and we only have data one period before the event. Then, $1\{t = 2000\} + 1\{c = 2000\} = 1$ among the untreated observations, resulting in perfect collinearity and hence noninvertibility. To see this, observe that cohort $c = 2000$ has untreated observations only in year $t = 1999$, and cohort $c = 2001$ only has untreated observations in year $t = 2000$. Hence, all untreated observations satisfy $1\{t = 2000\} + 1\{c = 2000\} = 1$.

Theorem 1. *If Assumptions 1 and 2 hold, then* $\hat{\gamma}$ *and* $\hat{\beta}$ *are asymptotically normal,* $\hat{\gamma} \xrightarrow{p} \gamma$, $\hat{\beta} \xrightarrow{p} \beta$, *and* $\sqrt{NT}\left(\hat{\beta} - \beta\right) \xrightarrow{d} \mathcal{N}(0, V)$, *where* $V = G_\beta^{-1} \mathbb{E}\left[\left(g + G_\gamma \psi\right)\left(g + G_\gamma \psi\right)'\right] G_\beta^{-1\prime}$, *with*

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

and

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

Theorem 1 tells us that the 2SDD estimator is consistent for β, and is asymptotically normal. Hence, using a consistent variance estimator provides valid inference asymptotically. The proof proceeds by the arguments in Section 2 and verifying the conditions in Newey and McFadden (1994).[28]

3.2 Event Studies

As we note in Section 2.5, there are multiple ways to implement event-studies using the two stage approach. All of these variations can be understood from within the following framework. Let $t^*(i)$ denote the time at which individual i becomes treated. Let $W_{rit} = 1\left[t - t^*(i) = r\right]$ denote whether individual i is r periods away from treatment at time t. With slight abuse of notation, our model is:

$$
\begin{aligned}
Y_{it} &= \lambda_i + \sum_{r=-\underline{R}}^{\overline{R}} \eta_{rit} W_{rit} + X_{it}'\gamma + \varepsilon_{it} \\
&= \lambda_i + W_{it}'\eta_{it} + X_{it}'\gamma + \varepsilon_{it}.
\end{aligned}
$$

The second equality comes from stacking the $\underline{R} + \overline{R} + 1$ instances of W_{rit} and η_{rit} objects. Notice now that η_{it} is a vector with η_{rit} as its components. In the first stage, we have $Y_{it} = \lambda_i + X_{it}'\gamma + \varepsilon_{it}$. This regression uses all observations with $t - t^*(i) < -\underline{R}^*$, where $\underline{R}^*$ may be zero. Let $Q_{it} := 1\left[t - t^*(i) < -\underline{R}^*\right]$. Then, by analogy to the case for the overall

[28] Standard errors from this deviations-from-untreated-means estimator are numerically identical to those from an estimator that includes unit indicators.

ATT, define $T_i^Q := \sum_{t=1}^T Q_{it}$. Now,

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

and a similar definition applies 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

$$\tilde{Y}_{Qi} = \begin{bmatrix} \tilde{Y}_{i1}Q_{i1} & \cdots & \tilde{Y}_{iT}Q_{iT} \end{bmatrix}.$$

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

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 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

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

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 coefficient across individuals who are observed r periods away from their treatment. For $r > 0$, η_r can be interpreted as the treatment effect r periods after treatment. If there are no pre-trends, $\eta_r = 0$ for all $r \leq 0$.

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

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

For the asymptotics to go through as before, we want a condition analogous to Assumption 2 that is suited for event studies.

Assumption 3. *Assume that:*

1. $\mathbb{E}\left[\left\|\tilde{X}'_{Qi}\tilde{\varepsilon}_{0i}\right\|^2\right] < \infty$, $\mathbb{E}[\tilde{\varepsilon}^2_{it}] < \infty$, *and* $\mathbb{E}\left[(\eta_{rit} - \eta_r)^2\right] < \infty$.
2. $\mathbb{E}\left[\tilde{X}'_{Qi}\tilde{X}_{Qi}\right]$ *is invertible and* $\mathbb{E}\left[\left\|\tilde{X}'_{Qi}\tilde{X}_{Qi}\right\|^2\right] < \infty$.
3. *For all* i, *there exists some* t *where* $Q_{it} = 0$, *and* $\mathbb{E}[N_{ir}] > 0$ *for all* $r \in \{-\underline{R}, \dots, \overline{R}\}$, *where* $N_{ir} := \sum_{t=1}^{T} 1\left[t - t^*(i) = r\right]$.

Note that in event studies, we are regressing on $1\left[t - t^*(i) = r\right]$ in the second stage. Assumption 3.3 is required for invertibility in the second stage, playing the role of Assumption 1.2.

Theorem 2. *If Assumptions 1 and 3 hold, then* $\hat{\gamma}$ *and* $\hat{\eta}$ *are asymptotically normal,* $\hat{\gamma} \xrightarrow{p} \gamma$ *and* $\hat{\eta} \xrightarrow{p} \eta$, *and* $\sqrt{NT}(\hat{\eta} - \eta) \xrightarrow{d} \mathcal{N}(0, V)$, *where* $V = G_\eta^{-1}\mathbb{E}\left[\left(g + G_\gamma\psi\right)\left(g + G_\gamma\psi\right)'\right]G_\eta^{-1\prime}$, *with*

$$
\begin{aligned}
G_\eta &= -\mathbb{E}\left[\sum_{t=1}^{T} W_{it}W'_{it}\right], \\
G_\gamma &= -\mathbb{E}\left[\sum_{t=1}^{T} W_{it}\tilde{X}'_{it}\right], \\
\psi &= \mathbb{E}\left[\tilde{X}'_{Qi}\tilde{X}_{Qi}\right]^{-1}\left(\tilde{X}'_{Qi}\left(\tilde{Y}_{Qi} - \tilde{X}_{Qi}\gamma\right)\right),
\end{aligned}
$$

and

$$
g = \sum_{t=1}^{T} W_{it}\left(\tilde{Y}_{it} - \tilde{X}'_{it}\gamma - W'_{it}\beta_{it}\right).
$$

3.3 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.

Parallel trends assumption. The theoretical results presented above assume parallel trends for every group and between every pair of consecutive time periods, as in de Chaisemartin and d'Haultfoeuille (2020); Sun and Abraham (2021). This assumption is stronger than the one used in Callaway and Sant'Anna (2021), 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 trend 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 only include the never-treated groups in the first-stage regression at the cost of losing some power. In contrast, even if the stronger version of parallel trend holds, the Callaway and Sant'Anna (2021) 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 (7) may include $1\{g(i) = g\}t$, where $g(i)$ denotes the group to which observation i belongs.

Triple differences. In a triple differences setting, when the coefficients on treatment, time, and group are consistently estimable, the relevant indicators can be collected in the X_{it} vector. In this case, γ collects these additional fixed effects, and our two-step procedure still applies.

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 Borusyak, Jaravel and Spiess (2024), 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 (7) is AR(1) with correlation parameter ρ, then Equation (7) 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 $Y_{it} - \hat{\rho} Y_{it-1} - \hat{\tilde{\lambda}}_i - \hat{\tilde{\gamma}}_t$ on D_{it}. Since the OLS estimator coincides with the maximum likelihood estimator, the resulting estimator is efficient, though it identifies a different estimand.

Continuous 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 regressing the outcome on X for $D_{it} = 0$ to obtain $\hat{\gamma}$, and the second stage regressing $Y_{it} - X_{it}'\hat{\gamma}$ on D_{it} to obtain $\hat{\beta}$. When implementing the two-stage regression procedure, due to Yitzhaki (1996, also see Angrist and Krueger, 1999 and Angrist and Pischke, 2009), 2SDD identifies a (positive) weighted average of the derivatives of the causal response function. Inference proceeds through GMM as before.

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 3.2 as a vector of indicators for the treatment path, which is defined as the sequence of treatment indicators since first treatment.[29] 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 B.1 in Appendix B, $\beta = \text{plim}\Big(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\Big)^{-1}\Big(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\beta_{it}\Big)$. Hence, in the setting with staggered treatment adoption, $\beta = \text{plim}\Big(\sum_{i=1}^{N} A_i\Big)^{-1}\Big(\sum_{i=1}^{N} A_i\beta_i\Big)$. 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

[29] 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.

all i. Thus, the estimand becomes $\beta = \frac{1}{aN}\mathbb{E}\left[\sum_{i=1}^{N} A_i\beta_i\right] = \frac{1}{aN}\sum_{i=1}^{N}\mathbb{E}[A_i]\beta_i = \frac{1}{N}\sum_{i=1}^{N}\beta_i$, which is exactly the average treatment effect (ATE).

Arbitrary 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.

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 weaknesses. Lastly, since the Borusyak, Jaravel and Spiess (2024) method shares point estimates with ours, it

becomes essential to assess the distinct approaches to inference. We summarize the differences between our approach and the existing papers in Table 1.

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.[30] 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 Borusyak, Jaravel and Spiess (2024) regarding efficiency, which also hold for our estimator, may not apply (though see our discussion in Section 3.3). 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 Borusyak, Jaravel and Spiess (2024).[31]

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

[30] 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).

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

comparison. In particular, we consider the imputation approach from Borusyak, Jaravel and Spiess (2024), using both their "default" asymptotically conservative standard errors and "leave-out" version with improved finite-sample performance, as well as various alternative estimators (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; de Chaisemartin and d'Haultfoeuille, 2024; Wooldridge, 2021).[32] 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 Borusyak, Jaravel and Spiess (2024) 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 2 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 Borusyak, Jaravel and Spiess (2024) 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 Sun and Abraham (2021)

[32]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). With the exception of `jwdid`, the authors of the respective methodological papers were directly involved in the development of the packages.

method leads to similar rejection rates with a larger standard error (around 0.17) and the Callaway and Sant'Anna (2021), de Chaisemartin and d'Haultfoeuille (2024), and Wooldridge (2021) 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, outperforming 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.

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 Borusyak, Jaravel and Spiess (2024) can no longer be computed. Aside from that, removing the restriction leads to similar results for all methods (Appendix Table 1). 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 3 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 Callaway and Sant'Anna (2021) 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 Sun and Abraham (2021) standard error also changes little and leads to similar rates of under-rejection as before. In contrast, the default Borusyak, Jaravel and Spiess (2024) 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 de Chaisemartin and d'Haultfoeuille (2024) and Wooldridge (2021) 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 Callaway and Sant'Anna (2021) estimator, and the Sun and Abraham (2021) estimator, as Appendix Table 2 show. The default Borusyak, Jaravel and Spiess (2024) variance estimator continues to lead to over-rejection, though with a rejection rate of only around 10–12 percent, while the de Chaisemartin and d'Haultfoeuille (2024) and Wooldridge (2021) 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.[33] This observation also explains why the imputation default performs poorly when the number of groups increases relative to N.[34] 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 Appendix Tables 3 to 5).

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 Borusyak, Jaravel and Spiess (2024) 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 Borusyak, Jaravel and Spiess (2024) 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 Borusyak, Jaravel and Spiess (2024) variance estimator and the Wooldridge (2021) estimator, when decreasing the number of treated states from 40 (Appendix Table 2) to 30

[33]The standard errors for the imputation estimator developed in Borusyak, Jaravel and Spiess (2024) are constructed based on the residuals $\tilde{\varepsilon}_{it} = \hat{\tau}_{it} - \hat{\bar{\tau}}_{it}$, where $\hat{\tau}_{it}$ is the estimated treatment effect for unit i at time t and $\hat{\bar{\tau}}_{it}$ is some average of these estimated individual treatment effects. Their "default" is to use cohort-period averages for $\hat{\bar{\tau}}_{it}$, i.e., $\hat{\bar{\tau}}_{it} = \hat{\bar{\tau}}_{gp}$. However, by using cohort-period averages, they are partway to the degenerate limit of zero variance. Hence, the variance estimator in Borusyak, Jaravel and Spiess (2024) is anti-conservative when the groups are small: in the extreme case, $\hat{\bar{\tau}}_{it} = \hat{\tau}_{it}$, so $\tilde{\varepsilon}_{it} = 0$.

[34]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.

(Appendix Table 6), 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.

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 (see Section 2.2), 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 Appendix Table 7 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).[35] 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.[36]

The other methods, however, are markedly outperformed by TWFE. The Borusyak, Jaravel and Spiess (2024) 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 Borusyak, Jaravel and Spiess

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

[36] 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.

(2024) leave-out variance estimator (standard error 0.14) and the Sun and Abraham (2021) approach (standard error 0.17) reject only 20–40 percent as often as TWFE. The Callaway and Sant'Anna (2021), de Chaisemartin and d'Haultfoeuille (2024), and Wooldridge (2021) 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.

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 4) continues to show rejection rates close to 5 percent for the two-stage approach. The same also holds for the Callaway and Sant'Anna (2021), de Chaisemartin and d'Haultfoeuille (2024), and Wooldridge (2021) estimators, though with standard errors around 30 percent larger. Also as before, the default Borusyak, Jaravel and Spiess (2024) 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 Sun and Abraham (2021) estimator. The same patterns hold in the simple case of homogeneous treatment effects (Appendix Table 8). The comparison between Appendix Tables 9 to 12 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 Borusyak, Jaravel and Spiess (2024) variance estimator, the de Chaisemartin and d'Haultfoeuille (2024) estimator, and the Wooldridge (2021) estimator, for which rejection rates reach as high as 31.7 percent, 21.2 percent, and 13.2 percent, respectively. Analogously, the comparison between Appendix Tables 13 and 14 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 Borusyak, Jaravel and Spiess (2024) 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 Wooldridge (2021) 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 Sun and Abraham (2021) using the existing heterogeneity-robust estimators that have been published as well as 2SDD.[37] The seven papers that we reanalyze appear in Table 5, with the number of treatment cohorts ranging from 5 to 21. The 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. Except for the two-way fixed effects (TWFE) estimator, the different methods tend to produce fairly comparable point estimates with one another (Figure 1). Our discussion therefore largely focuses on the extent of agreement in terms of confidence intervals and t-statistics across the different methods. 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 Borusyak, Jaravel and Spiess (2024). 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.[38] Figure 2 reports the average standard error for each method applied to each of these papers (also see regression results in Table 6), which we discuss below. We only consider the default Borusyak, Jaravel and Spiess (2024)

[37] 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 (Sun and Abraham, 2021) or even days (Callaway and Sant'Anna, 2021) to run for a single outcome, while the equivalent TWFE regressions run within minutes.

[38] The exception is Kuziemko, Meckel and Rossin-Slater (2018), discussed in Appendix D, with estimates presented in Appendix Figure 2. The remaining estimates appear in Appendix Figures 3 to 7.

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 Callaway and Sant'Anna (2021) estimator create difficulty in visually discerning the differences between the other methods in Figure 1a, but Figure 2 and Table 6 make the comparison clearer. In this case, the Sun and Abraham (2021) estimator seems to perform best, yielding the most precise estimates, followed by 2SDD and then the de Chaisemartin and d'Haultfoeuille (2024) estimator. Notably, the Borusyak, Jaravel and Spiess (2024) estimator in some cases produces a standard error over twice as large as that of 2SDD. This 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 Borusyak, Jaravel and Spiess (2024) variance estimator.[39]

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 Sun and Abraham (2021) and de Chaisemartin and d'Haultfoeuille (2024) estimators and the least precise estimates again using the Callaway and Sant'Anna (2021) estimator. Unlike in the previous example, for all 11 post-treatment periods and all 15 outcome variables, the Borusyak, Jaravel and Spiess (2024) 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 Borusyak, Jaravel and Spiess (2024) 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 de Chaisemartin and d'Haultfoeuille

[39]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 Borusyak, Jaravel and Spiess (2024) shrinkage method. The variance estimator of Borusyak, Jaravel and Spiess (2024) uses $\hat{\sigma}^2_{\text{BJS}} = \sum_i \left(\sum_t v_{it}(Y_{it} - X'_{it}\hat{\gamma} - D_{it}\hat{\tau}_{it})\right)^2$, while we use $\hat{\sigma}^2_{\text{2SDD}} = \sum_i \left(\sum_t v_{it}(Y_{it} - X'_{it}\hat{\gamma} - D_{it}\hat{\tau})\right)^2 = \hat{\sigma}^2_{\text{BJS}} + \sum_i \left(\sum_t v_{it} D_{it}(\hat{\tau}_{it} - \hat{\tau})\right)^2 + \sum_i \left(\sum_t v_{it} D_{it}(\hat{\tau}_{it} - \hat{\tau})\right)\left(\sum_t v_{it}(Y_{it} - X'_{it}\hat{\gamma} - D_{it}\hat{\tau})\right)$. Hence, the Borusyak, Jaravel and Spiess (2024) variance estimator can be larger when $\sum_i \left(\sum_t v_{it} D_{it}(\hat{\tau}_{it} - \hat{\tau})\right)^2 + \sum_i \left(\sum_t v_{it} D_{it}(\hat{\tau}_{it} - \hat{\tau})\right)\left(\sum_t v_{it}(Y_{it} - X'_{it}\hat{\gamma} - D_{it}\hat{\tau})\right) < 0$.

(2024) 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 Borusyak, Jaravel and Spiess (2024) 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 (Appendix Figure 7b). On average, the Borusyak, Jaravel and Spiess (2024) approach generates standard errors that are half the size of those produced by the other methods, and their conservative leave-out variance estimator remains infeasible. Excluding their method, the 2SDD approach results in the smallest standard errors, followed by the Callaway and Sant'Anna (2021) 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 de Chaisemartin and d'Haultfoeuille (2024) 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 de Chaisemartin and d'Haultfoeuille (2024) does not produce an estimate, the difference in standard errors between 2SDD and Sun and Abraham (2021) nearly triples, and the difference with Callaway and Sant'Anna (2021) 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 Sun and Abraham (2021) 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 de Chaisemartin and d'Haultfoeuille (2024) and Callaway and Sant'Anna (2021) 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 Callaway and Sant'Anna (2021), Sun and Abraham (2021), and de Chaisemartin and d'Haultfoeuille (2024) methods do not yield point estimates for the effect of the treatment 9 years after the event. Additionally, the Callaway and Sant'Anna (2021) and Borusyak, Jaravel and Spiess (2024) methods do not yield point

estimates for the effect of the treatment 9–11 years before the event, the de Chaisemartin and d'Haultfoeuille (2024) 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 Borusyak, Jaravel and Spiess (2024) 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 Sun and Abraham (2021) and de Chaisemartin and d'Haultfoeuille (2024) methods give the widest confidence intervals on average, we tend to find the narrowest confidence intervals using the Callaway and Sant'Anna (2021) method. However, the precision of the Callaway and Sant'Anna (2021) estimator varies substantially across periods.[40] 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 Callaway and Sant'Anna (2021), the 2SDD and Borusyak, Jaravel and Spiess (2024) approaches give slightly higher, but less variable, standard errors.

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 Callaway and Sant'Anna (2021) 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 Callaway and Sant'Anna (2021) 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 Callaway and Sant'Anna (2021) 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 7, 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 Sun and Abraham (2021) and de Chaisemartin and d'Haultfoeuille (2024) approaches offer notable advantages when the number of groups is large (Bailey and Goodman-Bacon,

[40] In addition, the Callaway and Sant'Anna (2021) method does not provide estimates for any of the periods before the event in this application.

2015; Deryugina, 2017) but are outperformed by the Callaway and Sant'Anna (2021) 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 Callaway and Sant'Anna (2021) 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 Borusyak, Jaravel and Spiess (2024) estimator produces smaller standard errors than 2SDD does.[41]

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.

Estimates for which a given method's standard error diverges from its counterparts' average standard error appear in Figure 3.[42] 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 Callaway and Sant'Anna (2021) 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 Borusyak, Jaravel and Spiess (2024), Sun and Abraham (2021), and de Chaisemartin and d'Haultfoeuille (2024)

[41]The differences are significant except in two settings where the sample size of estimates is small (Table 6).

[42]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 Callaway and Sant'Anna (2021) estimator in this setting.

methods, most outliers arise because the estimates are unusually precise. For the Sun and Abraham (2021) estimator, as previously noted, this occurs in part due to the overly precise standard errors for the Tewari (2014) paper. For the de Chaisemartin and d'Haultfoeuille (2024) 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 8. 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 de Chaisemartin and d'Haultfoeuille (2024) 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 Sun and Abraham (2021) estimator also leads to significantly larger standard errors for longer-term treatment effects, and the same holds for the Callaway and Sant'Anna (2021) 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 4. The normalized difference between t-

statistics falls in the top or bottom 5 percent of the distribution most often for the Callaway and Sant'Anna (2021) and de Chaisemartin and d'Haultfoeuille (2024) estimators and least often for the 2SDD estimator. Using the top or bottom 1 percent of the distribution as the cutoff, the de Chaisemartin and d'Haultfoeuille (2024) and Sun and Abraham (2021) estimators result in the most outliers. Analyzing absolute t-statistics rather than normalized differences reveals additional insights (Table 9 Panel A). Compared to the 2SDD approach, the Borusyak, Jaravel and Spiess (2024), Sun and Abraham (2021), and de Chaisemartin and d'Haultfoeuille (2024) 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 90^{th} percentile of the distribution (approximately 4.3, $p < 10^{-5}$) and the 99^{th} percentile of the distribution (approximately 7.4, $p < 10^{-13}$) of t-statistics in our sample. The Callaway and Sant'Anna (2021) 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 9 Panel C), the Callaway and Sant'Anna (2021) 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 (Appendix Table 15), as well as when expanding the sample to include estimates that only a subset of methods produce and adding paper-outcome-period fixed effects (Appendix Table 16). 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 4); this moderation places the 2SDD estimator toward the conservative end of the spectrum, evident from its low rate of extreme t-statistics (Table 9).

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 de Chaisemartin and d'Haultfoeuille (2024) and Sun and Abraham (2021) estimators tend to produce greater statistical significance in the post-treatment period (Table 9), we do not find the same pattern in the pre-treatment periods (Appendix Table 17), where greater statistical significance would indicate violations

of parallel trends.[43] In the case of the Callaway and Sant'Anna (2021) estimator, we see significantly fewer significant coefficients in the pre-treatment periods.[44] This suggests that the 2SDD and Borusyak, Jaravel and Spiess (2024) methods may offer a more conservative approach.

The discrepancy in pre-event coefficient estimates between 2SDD and Borusyak, Jaravel and Spiess (2024) 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 Borusyak, Jaravel and Spiess (2024) 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 (Appendix Figure 8). 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 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

[43] These comparisons exclude the period immediately preceding the event because some of the methods (Sun and Abraham, 2021; de Chaisemartin and d'Haultfoeuille, 2024) normalize the effect in this period to zero.

[44] Callaway and Sant'Anna (2021) 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.

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.[45] 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 Appendix Table 7). 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. This procedure transparently handles the complexities of staggered treatment adoption with familiar and straightforward tools, analogous to traditional regression methods. Estimation

[45] 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.

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. 2

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. 22

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press. 22

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. 3, 6, 22, 24

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. 3, 31, 33, 34, 38, 44, 46, 47, 52

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. , 2, 23, 24, 25

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.* 25

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2021. "Revisiting event study designs: Robust and efficient estimation." *arXiv preprint arXiv:2108.12419.* 2

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting event study designs: Robust and efficient estimation." *Review of Economic Studies.* 2, 3, 4, 6, 7, 8, 12, 13, 14, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 48, 49, 53

Brown, Nicholas, and Kyle Butts. 2023. "Dynamic Treatment Effect Estimation with Interactive Fixed Effects and Short Panels." *Mimeo.* 39

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. 1, 25

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

Caetano, Carolina, Brantly Callaway, Stroud Payne, and Hugo Sant'Anna Rodrigues. 2022. "Difference in differences with time-varying covariates." *arXiv preprint arXiv:2202.02903.* 9, 39

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230. 1, 2, 3, 9, 10, 13, 20, 21, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 48, 49, 53

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. 4

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*. 4

de Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–2996. 1, 3, 6, 7, 8, 13, 20, 48

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. 37, 38

de Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2024. "Difference-in-differences estimators of intertemporal treatment effects." *Review of Economics and Statistics*, 1–45. 2, 3, 13, 25, 26, 27, 29, 31, 32, 33, 34, 35, 36, 37, 48, 49, 53

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*. 25

Deryugina, Tatyana. 2017. "The fiscal cost of hurricanes: Disaster aid versus social insurance." *American Economic Journal: Economic Policy*, 9(3): 168–198. 3, 31, 33, 34, 35, 38, 44, 52

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. 4

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. 2, 4, 39

Dumont, Michel, Glenn Rayp, Olivier Thas, and Peter Willeme. 2005. "Correcting standard errors in two-stage estimation procedures with generated regressands." *Oxford Bulletin of Economics and Statistics*, 67(3): 421–433. 13

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. 52

Gardner, John. 2020. "Two-stage differences in differences." *Mimeo.* , 2

Gardner, John, Neil Thakral, Linh Tô, and Luther Yap. 2024. "Two-stage differences in differences." *Mimeo.* 48

Gibbons, Charles E, Juan Carlos Suárez Serrato, and Michael B Urbancic. 2018. "Broken or fixed effects?" *Journal of Econometric Methods*, 8(1): 20170002. 3

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2): 254–277. 3, 6, 7, 8

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. 4

Hansen, Lars Peter. 1982. "Large sample properties of generalized method of moments estimators." *Econometrica: Journal of the econometric society*, 1029–1054. 14

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. 31, 33, 34, 35, 44, 52

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. 3, 6

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. 30, 52

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. 25, 26, 27, 32, 33, 34, 35, 38, 44, 52

Liu, Licheng, Ye Wang, and Yiqing Xu. 2019. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." 2

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.* 12, 13

Miller, Douglas L. 2023. "An Introductory Guide to Event Study Models." *Journal of Economic Perspectives*, 37(2): 203–230. 25

Newey, Whitney K. 1984. "A method of moments interpretation of sequential estimators." *Economics Letters*, 14(2-3): 201–206. 14

Newey, Whitney K, and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." *Handbook of econometrics*, 4: 2111–2245. 14, 18

Rios-Avila, Fernando, Arne J. Nagengast, and Yoto V. Yotov. 2022. "JWDID: Stata module to estimate Difference-in-Difference models using Mundlak approach." 25

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." 25

Roth, Jonathan. 2024. "Interpreting event-studies from recent difference-in-differences methods." *Mimeo*. 13

Sant'Anna, Pedro HC, and Jun Zhao. 2020. "Doubly robust difference-in-differences estimators." *Journal of Econometrics*, 219(1): 101–122. 9

Sun, Liyang. 2021. "EVENTSTUDYINTERACT: Stata module to implement the interaction weighted estimator for an event study." *Statistical Software Components, Boston College Department of Economics*. 25

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, 225(2): 175–199. 2, 3, 11, 12, 20, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 48, 49, 52, 53

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. 32, 33, 34, 35, 38, 44, 52

Thakral, Neil, and Linh Tô. 2020. "Anticipation and consumption." *Available at SSRN 3756188*. 2

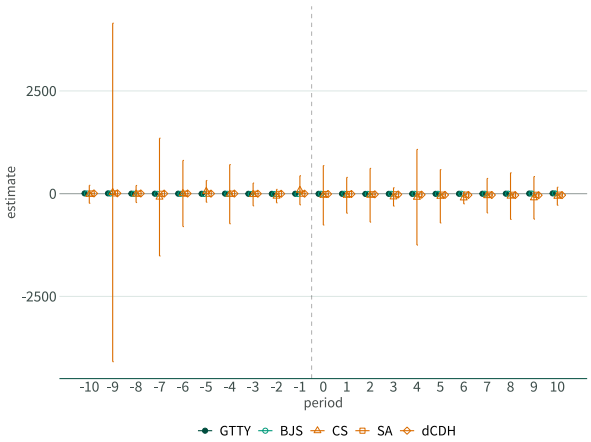
Thakral, Neil, and Linh T Tô. 2023. "When Are Estimates Independent of Measurement Units?" *Mimeo*. 24

Ujhelyi, Gergely. 2014. "Civil service rules and policy choices: evidence from US state governments." *American Economic Journal: Economic Policy*, 6(2): 338–380. 33, 34, 35, 44, 52

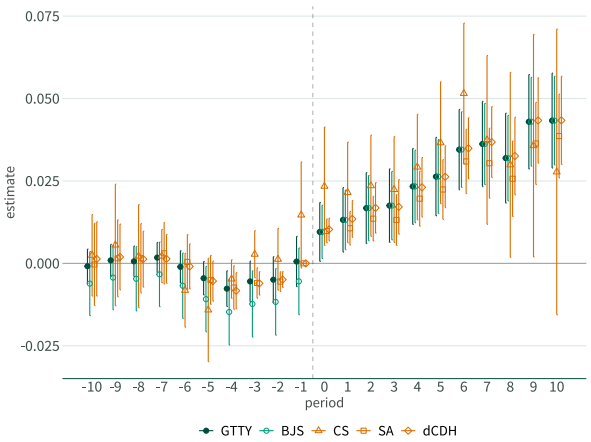
Wooldridge, Jeffrey M. 2021. "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators." *Available at SSRN 3906345*. 2, 4, 25, 26, 27, 29, 48, 49

Yitzhaki, Shlomo. 1996. "On using linear regressions in welfare economics." *Journal of Business & Economic Statistics*, 14(4): 478–486. 22

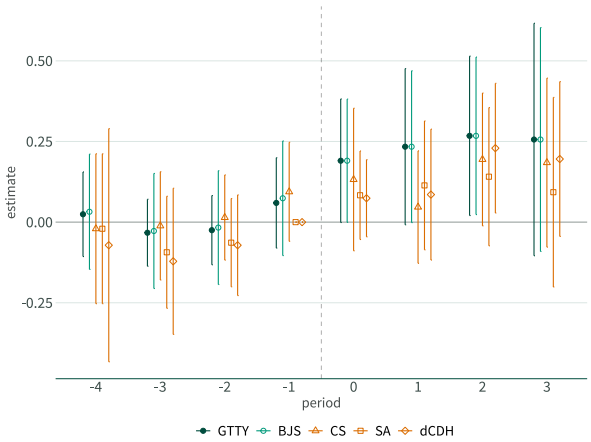
Figure 1: Empirical applications: Event-study estimates



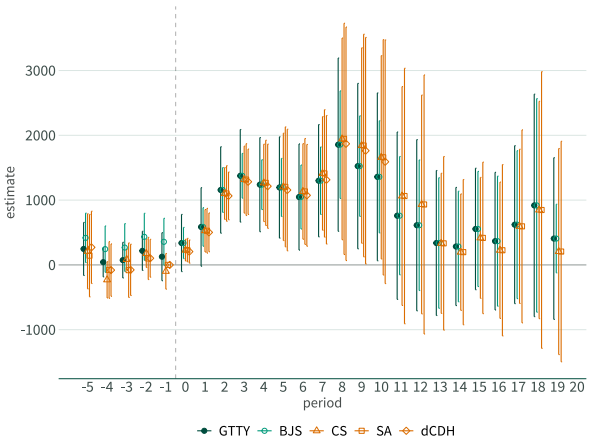
(a) Bailey and Goodman-Bacon (2015)



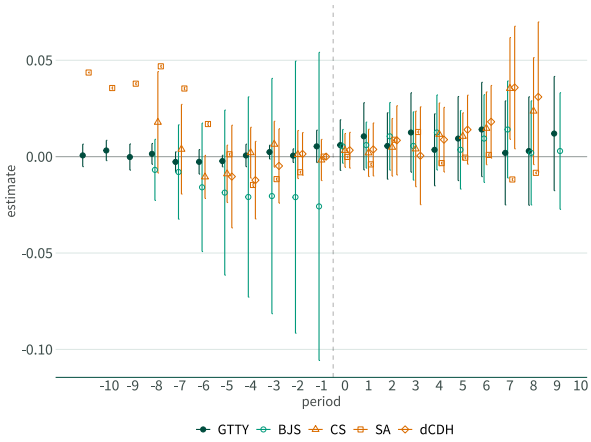
(b) Deryugina (2017)



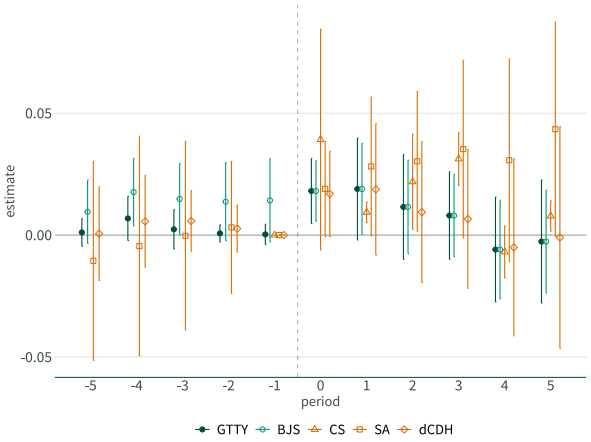
(c) He and Wang (2017)



(d) Lafortune, Rothstein and Schanzenbach (2018)



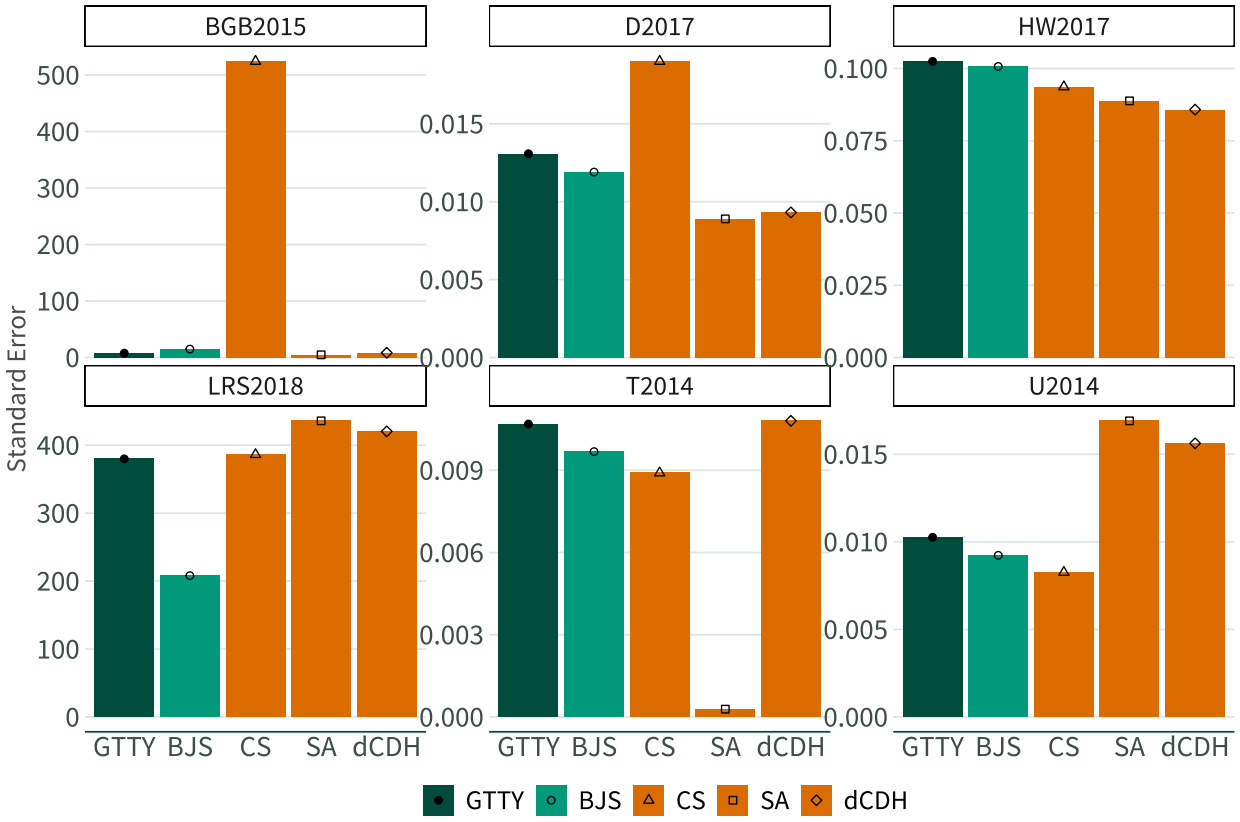
(e) Tewari (2014)



(f) Ujhelyi (2014)

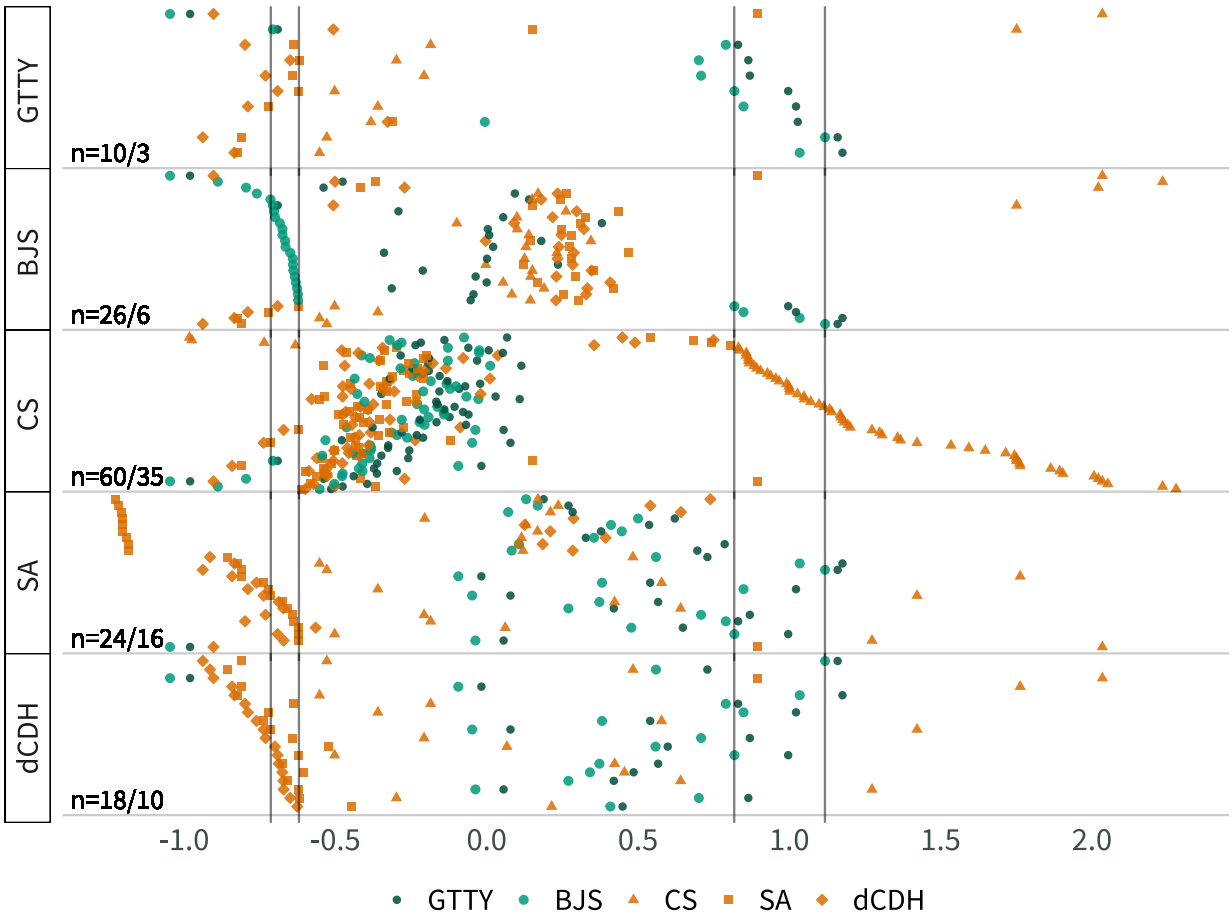
Note: This table reports event-study estimates from applying each estimator to the first event-study specification for each of the main empirical settings in Table 5.

Figure 2: Empirical applications: Comparison of standard errors



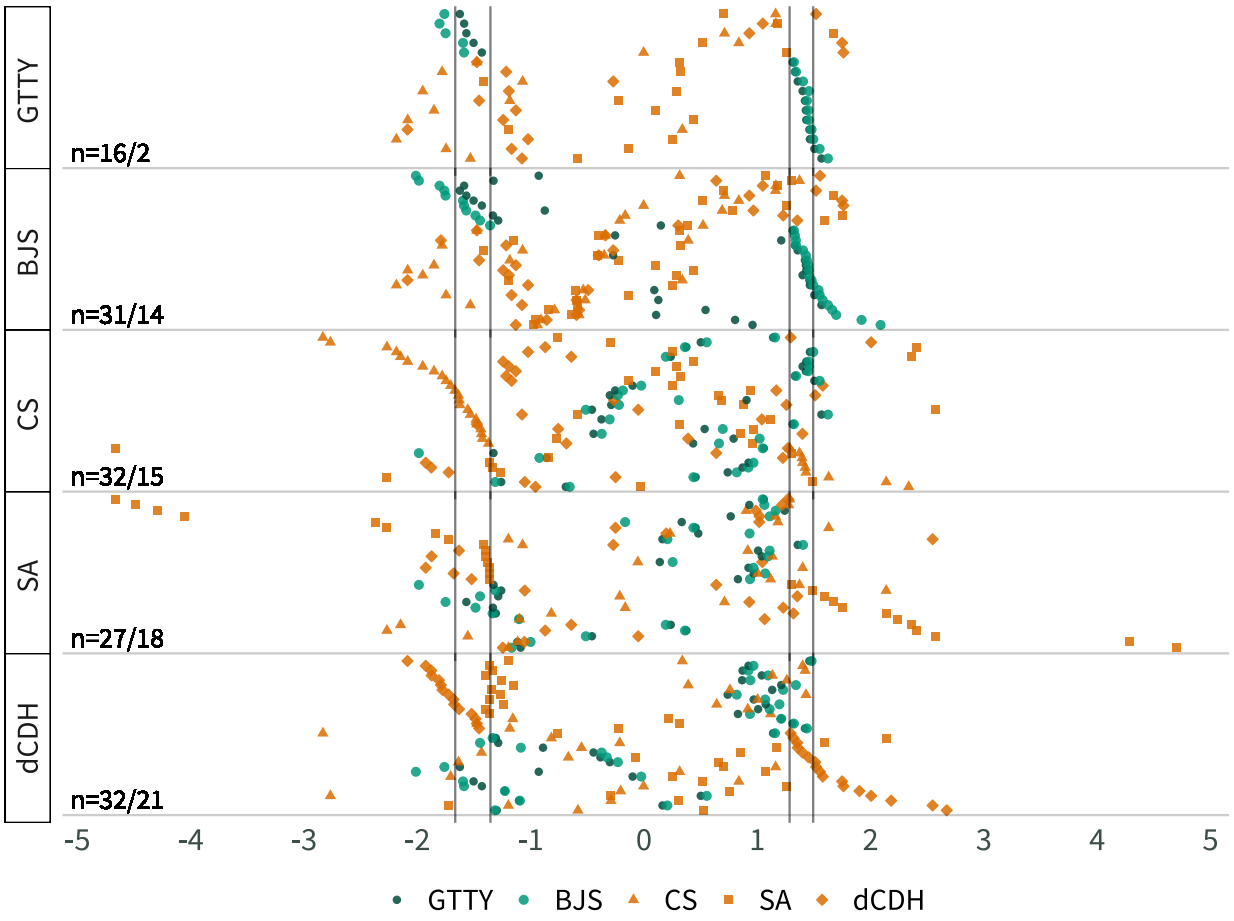
Note: This figure reports the average standard error across all dynamic treatment effect estimates for each replicated paper and each estimation method. The set of papers corresponds to the main empirical settings from Table 5.

Figure 3: 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 4: 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.

Table 1: Comparison of approaches to estimation and inference

	Group-independent estimation	Group-independent inference	Analytical standard errors
Gardner et al. (2024)	✓	✓	✓
Borusyak, Jaravel and Spiess (2024)	✓	×	✓
Callaway and Sant'Anna (2021)	×	×	×
Sun and Abraham (2021)	×	×	✓
de Chaisemartin and d'Haultfoeuille (2024)	×	×	✓
de Chaisemartin and d'Haultfoeuille (2020)	×	×	×
Wooldridge (2021)	×	×	✓

Note: In the first two columns, procedures that estimate group effects or variances separately are denoted × while procedures that do not are denoted ✓. In the third column, procedures that use analytical standard errors are denoted ✓ while procedures that bootstrap in their implementation are denoted ×.

Table 2: 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. 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 Borusyak, Jaravel and Spiess (2024). CS, SA, dCDH, and W refer to the methods proposed by Callaway and Sant'Anna (2021), Sun and Abraham (2021), de Chaisemartin and d'Haultfoeuille (2024), and Wooldridge (2021), respectively. Average speed per simulation using the corresponding Stata package for each method is reported in seconds.

Table 3: 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. See the note accompanying Table 2 for further information.

Table 4: 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 2. See the note accompanying Table 2 for further information.

Table 5: 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 D 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 6: Empirical applications: Comparison of standard errors

	BGB2015	D2017	HW2017	LRS2018	T2014	U2014
BJS	7.2748	-0.0012	-0.0018	-171.9702	-0.0010	-0.0010
	(4.7166)	(0.0004)	(0.0005)	(11.5239)	(0.0013)	(0.0011)
CS	516.3992	0.0059	-0.0088	6.2230	-0.0018	-0.0020
	(272.8267)	(0.0029)	(0.0065)	(30.8968)	(0.0014)	(0.0033)
SA	-2.8849	-0.0042	-0.0137	55.9069	-0.0104	0.0067
	(1.6728)	(0.0025)	(0.0054)	(37.9005)	(0.0008)	(0.0021)
dCDH	0.6955	-0.0038	-0.0167	40.3952	0.0001	0.0054
	(0.4478)	(0.0025)	(0.0090)	(37.5220)	(0.0018)	(0.0021)
Number of outcomes	5	15	4	7	1	1
Number of periods	11	11	4	11	9	6
Covariates	✓	✓				✓
Weights	✓			✓	✓	

Note: The data consist of standard error estimates from applying the estimator listed in each row to the empirical settings in Table 5. Each column reports the results from regressing standard error estimates for the specified paper on indicators for each method, with 2SDD as the omitted category. BJS refers to the imputation estimator with default asymptotic standard errors from Borusyak, Jaravel and Spiess (2024), and CS, SA, and dCDH refer to the methods proposed by Callaway and Sant'Anna (2021), Sun and Abraham (2021), and de Chaisemartin and d'Haultfoeuille (2024), respectively. The number of periods denotes the number of post-treatment coefficients each paper estimates (common across all outcome variables). The last two rows of the table indicate whether the event-study specification in each paper includes covariates and uses sample weights (common across all outcome variables). We report heteroskedasticity-robust standard errors in parentheses, and standard errors are adjusted for clustering at the outcome level for papers which contain estimates for more than one outcome variable.

Table 7: 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 appendix figures 3 to 7.

Table 8: 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 5. The regression omits the linear period trend for the 2SDD estimator. See Table 6 for additional details.

Table 9: Empirical applications: Comparison of t-statistics (post-treatment periods)

	$\lvert t\rvert$	$\mathbb{1}_{\{\lvert t\rvert>1.96\}}$	$\mathbb{1}_{\{\lvert t\rvert>p_{90}\}}$	$\mathbb{1}_{\{\lvert t\rvert>p_{99}\}}$
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 5. 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.