Supplementary Appendix, Not for Publication: Two-stage Differences in Differences

John Gardner
University of Mississippi

Neil Thakral
Brown University

Linh T. Tô
Boston University

Luther Yap
National University of Singapore

June 2026*

Abstract

This document contains appendix material for Gardner et al. (2026).

*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, National University of Singapore (email: lyap@nus.edu.sg).

Contents

List of Appendix Figures

List of Appendix Tables

A Additional empirical applications

This appendix section illustrates the performance of our two-stage estimator through a variety of empirical applications. In particular, we replicate all papers with variation in treatment timing and a single treatment event listed in Table 1 of SA2021 using the existing heterogeneity-robust estimators that have been published as well as 2SDD.[1] The seven papers that we reanalyze appear in Table A1, 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. However, except for the two-way fixed effects (TWFE) estimator, the different methods tend to produce fairly comparable point estimates with one another (Figure A1), and all of the methods that we compare have at least some theoretical justification for their approach to inference, so we do not strictly need to observe a "true" baseline to learn about their reliability. Our discussion therefore largely focuses on the extent of agreement in terms of confidence intervals and t-statistics across the different methods, assessing which methods tend to yield outliers in repeated applications. The discussion emphasizes notable instances in which conclusions may differ depending on the choice of method, along with broader patterns in standard errors and t-statistics across applications. Several patterns emerge from our analysis of the differences in results across estimators. We first investigate differences in coverage of confidence intervals, including a discussion of differences between our method and that of BJS2024. Next, we examine consistency between the t-statistics and standard error estimates across methods, highlighting instances where they disagree most. We then address differences in the pre-event coefficient estimates. Finally, we discuss how the various estimators compare with TWFE.

A.1 Selection of papers and outcomes

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

- Bailey and Goodman-Bacon (2015)
  - Age-adjusted mortality rate (Figure 5)

[1] Compared to Section 4, the larger datasets and extensive set of covariates commonly used in empirical applications magnify the differences in runtime. Some methods can take many hours (SA2021) or even days (CS2021) to run for a single outcome, while the equivalent TWFE regressions run within minutes.

  - Infant mortality rate (Figure 7.A)
  - Age-adjusted mortality rate: children (1–14) (Figure 7.B)
  - Age-adjusted mortality rate: adults (15–49) (Figure 7.C)
  - Age-adjusted mortality rate: older adults (50+) (Figure 7.D)

- Deryugina (2017)
  - Effect of a hurricane on earnings and transfers (Figure 2)
  - Effect of a hurricane on demographics (Figure 3)
  - Effect of a hurricane on transfer components (Figures 4 and 5)

- He and Wang (2017)
  - Subsidized population (Figure 2.A)
  - Poor-quality housing (Figure 2.B)
  - Registered poor households (Figure 2.C)
  - People with disabilities (Figure 2.D)

- Kuziemko, Meckel and Rossin-Slater (2018)
  - Mortality rates of children born to US-born Black mothers (Figure 2.A)
  - Mortality rates of children born to US-born Hispanic mothers (Figure 2.B)

- Lafortune, Rothstein and Schanzenbach (2018)
  - Mean state revenues in lowest income districts (Figure 3)
  - Mean state revenues in highest income districts (Figure 4)
  - Progressivity of state revenues (Figure 5)
  - Mean total revenues per pupil (Figure A3(a))
  - Mean total revenues per pupil in the lowest income quintile of districts (Figure A3(b))
  - Mean total revenues per pupil in the highest income quintile of districts (Figure A3(c))
  - Difference in mean total revenues per pupil between top and bottom quintile districts (Figure A3(d))

- Tewari (2014)
  - Home ownership (Figure 1)

- Ujhelyi (2014)
  - Share of intergovernmental expenditures in total expenditures (Figure 1)

A.2 Event-study estimates

For the first event-study analysis in each empirical paper, we provide the corresponding estimates using the different methods in Figure A1.[2] Figure A8 reports the average standard error for each method applied to each of these papers (also see regression results in Table A2), which we discuss below. We only consider the default BJS2024 variance estimator because five out of seven of our empirical settings contain treated cohorts with only one unit. The Bailey and Goodman-Bacon (2015) analysis of the effects of increasing access to primary care on mortality rates consists of data on more than 3,000 U.S. counties over a 40-year period, with 15 treatment cohorts and a never-treated group. The large standard errors using the CS2021 estimator create difficulty in visually discerning the differences between the other methods in Figure A1a, but Figure A8 and Table A2 make the comparison clearer. In this case, the SA2021 estimator seems to perform best, yielding the most precise estimates, followed by 2SDD and then the dCDH2024 estimator. Notably, the BJS2024 estimator in some cases produces a standard error over twice as large as that of 2SDD. This possibility can arise in finite samples when the covariance between the residual and the estimated treatment effect for unit i at time t contributes a sufficiently negative component to the BJS2024 variance estimator.[3] Deryugina (2017) studies the effect of hurricanes on government transfers using data from over 1,000 U.S. counties over a 44-year period, with 15 treatment cohorts and a never-treated group. In this case, we find the most precise estimates using the SA2021 and dCDH2024 estimators and the least precise estimates again using the CS2021 estimator. Unlike in the previous example, for all 11 post-treatment periods and all 15 outcome variables, the BJS2024 estimator results in smaller standard errors compared to 2SDD. He and Wang (2017) examine the impact of increased bureaucrat quality on the effectiveness of social assistance programs in rural China. The authors present a case study, including field interviews with local officials and bureaucrats, administrative records, and online surveys, as well as analyze a panel dataset consisting of a representative sample of 255 villages over a 12-year period, with between 1 and 30 villages being treated in each of 8 treatment cohorts. Only 2SDD and the BJS2024 event-study estimates provide evidence that supports the case study results by showing evidence of a significant improvement in the delivery of public services to poor households for all four outcomes. Among the other methods, the estimates from dCDH2024 support the finding of a significant effect on one outcome (increase in subsidized population), shown in Figure A1c. 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

[2]The exception is Kuziemko, Meckel and Rossin-Slater (2018), with estimates presented in Figure A7. The remaining estimates appear in Figures A2 to A6.

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

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 A1d), only the BJS2024 variance estimator indicates a significant increase for the highest-income districts, and it does so for five out of the first nine years following the reform (Figure A2b). On average, the BJS2024 approach generates standard errors that are half the size of those produced by the other methods, and their conservative leave-out variance estimator remains infeasible. Excluding their method, the 2SDD approach results in the smallest standard errors, followed by the CS2021 approach. We note that the estimates in Figure A8 understate the advantage of 2SDD because the table conditions on post-treatment periods when all five methods produce estimates, and the dCDH2024 approach only produces treatment-effect estimates up to 10 years after the event (11 estimates instead of 20) for all outcomes. In the periods when dCDH2024 does not produce an estimate, the difference in standard errors between 2SDD and SA2021 nearly triples, and the difference with CS2021 grows fivefold. Tewari (2014) studies how mortgage access changed following the removal of geographic restrictions on banks using a dataset of 39 states over a 32-year period. The data consist of 20 treatment cohorts, including 13 cohorts each consisting of only a single state and 3 cohorts each consisting of only two states. The SA2021 approach provides extremely precise estimates and implies a treatment effect that fluctuates between a significant positive and significant negative effect, while the other methods always generate positive point estimates. The dCDH2024 and CS2021 approaches show some evidence supporting a significant positive effect of deregulation on homeownership. However, we note that only 2SDD is able to estimate the full set of dynamic treatment effects. In particular, the CS2021, SA2021, and dCDH2024 methods do not yield point estimates for the effect of the treatment 9 years after the event. Additionally, the CS2021 and BJS2024 methods do not yield point estimates for the effect of the treatment 9–11 years before the event, the dCDH2024 method does not yield point estimates for the effect of the treatment 6–11 years before the event. The most apparent feature of Figure A1e is the difference between the 2SDD and BJS2024 approaches in the periods preceding the event, which we discuss in Supp. Appendix A.5. Finally, Ujhelyi (2014) investigates the impact of the state-level adoption of merit-based recruitment systems for civil service on government expenditure patterns. The data consist of 48 states over a 25-year period, with 10 treatment cohorts each consisting of only a single state, 6 treatment cohorts each consisting of only two states, and the 5 remaining treatment cohorts each consisting of only three states. While the SA2021 and dCDH2024 methods give the widest confidence intervals on average, we tend to find the narrowest confidence intervals using the CS2021 method. However, the precision of the CS2021 estimator varies substantially across periods.[4] As Figure A1f shows, for the year of the introduction of the merit system, their standard error is about three times larger than that of the other methods. In comparison with CS2021, the

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

2SDD and BJS2024 approaches give slightly higher, but less variable, standard errors.

A.3 Replication of Kuziemko, Meckel and Rossin-Slater (2018)

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

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

The 2SDD approach is easily implemented by using month fixed effects in the first stage and year-since-treatment event dummies in the second stage. To obtain estimates using `csdid` (Rios-Avila, Sant'Anna and Callaway, 2023), `eventstudyinteract` (Sun, 2021), `did_multiplegt_dyn` (de Chaisemartin et al., 2023), and `jwdid` (Rios-Avila, Nagengast and Yotov, 2022), we must define the cohort as the treatment year (not the exact month) to obtain dynamic effects by year since treatment. We present these results in Figure A7. However, we note that the conceptually correct way to do this exercise using those estimators would be to estimate separate effects for each month and then aggregate them into 12-month bins. This process would be somewhat cumbersome, and if undertaken, would require either assuming the distribution of units in each bin is known, or using a bootstrap, or devising a potentially complicated analytical asymptotic adjustment to account for that uncertainty.

This highlights the flexibility and simplicity advantages of 2SDD. With 2SDD, implementing the conceptually correct approach is straightforward: Simply include month fixed effects in the first stage and years-since-treatment indicators in the second stage.[5]

A.4 Comparison of performance across methods

A.4.1 Synthesizing results on confidence interval coverage

In the simulations from Section 4, the 2SDD and CS2021 estimators both deliver rejection rates closest to 5 percent, albeit with a larger standard error for the latter. In Figure A8, we obtain comparable standard error estimates using the 2SDD and CS2021 estimators in four settings (He and Wang, 2017; Lafortune, Rothstein and Schanzenbach, 2018; Tewari, 2014; Ujhelyi, 2014). However, we find substantially larger standard errors using the CS2021 estimator in the remaining settings (Bailey and Goodman-Bacon, 2015; Deryugina, 2017), highlighting the merits of our approach. We present a concise summary of the points discussed in the preceding section in Table A3, which

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

primarily focuses on comparing standard errors as a measure of performance. While these results derive from a limited sample of empirical papers, the 2SDD estimator stands out as a practical choice for applied researchers. The SA2021 and dCDH2024 approaches offer notable advantages when the number of groups is large (Bailey and Goodman-Bacon, 2015; Deryugina, 2017) but are outperformed by the CS2021 and 2SDD estimators with a relatively large number of small cohorts (Lafortune, Rothstein and Schanzenbach, 2018; Tewari, 2014; Ujhelyi, 2014). On the other hand, the CS2021 estimator seems to perform particularly poorly, yielding larger standard errors, when the number of groups is large. With a medium-sized number of groups (He and Wang, 2017), all of the methods seem to perform adequately. In five empirical applications (with Bailey and Goodman-Bacon, 2015 as the exception), the BJS2024 estimator produces smaller standard errors than 2SDD does.[6]

A.4.2 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 A9.[7] Each row in the figure corresponds to a single event-study coefficient for which the normalized difference falls in the top or bottom 5 percent of the distribution, along with the normalized differences for all five methods. A negative normalized difference indicates that a method produces a more precise estimate than its counterparts, while a positive normalized difference indicates the opposite. This representation shows several striking patterns. First, we find the greatest number of outliers for the CS2021 estimator, despite excluding estimates for the Bailey and Goodman-Bacon (2015) paper. Nearly all of the outliers using this method fall in the imprecise end of the distribution. 2SDD results in the fewest outliers, mostly in cases where it produces more conservative standard error estimates than other methods, rather than for producing overly precise estimates. On the other hand, for the BJS2024, SA2021, and dCDH2024 methods, most outliers arise because the estimates are unusually precise. For the SA2021 estimator, as previously noted, this occurs in part due to the overly precise standard errors for the Tewari (2014) paper. For the dCDH2024 estimator, the issue relates to its high precision in estimating short-term treatment effects, and relatively low precision in estimating longer-term treatment effects, which we highlight next.

[6]The differences are significant except in two settings where the sample size of estimates is small (Table A2).

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

A.4.3 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 A4. A positive value for a given method indicates that it produces relatively less precise estimates of longer-term treatment effects. In four of the empirical applications (Deryugina, 2017; Lafortune, Rothstein and Schanzenbach, 2018; Tewari, 2014; Ujhelyi, 2014), the dCDH2024 estimator results in significantly lower precision for longer-term effects compared to 2SDD. In three of these cases Deryugina (2017); Lafortune, Rothstein and Schanzenbach (2018); Ujhelyi (2014), the SA2021 estimator also leads to significantly larger standard errors for longer-term treatment effects, and the same holds for the CS2021 estimator in the first two cases. Other than cases in which other methods yield overly precise standard errors (i.e., Lafortune, Rothstein and Schanzenbach, 2018 for Borusyak, Jaravel and Spiess, 2024, and Tewari, 2014 for Sun and Abraham, 2021), we find only two instances in which another method yields relatively greater precision for long-run treatment effects compared to 2SDD (He and Wang, 2017 and Ujhelyi, 2014 for Callaway and Sant'Anna, 2021) at the 10 percent significance level.

A.4.4 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 A10. The normalized difference between t-statistics falls in the top or bottom 5 percent of the distribution most often for the CS2021 and dCDH2024 estimators and least often for the 2SDD estimator. Using the top or bottom 1 percent of the distribution as the cutoff, the dCDH2024 and SA2021 estimators result in the most outliers. Analyzing absolute t-statistics rather than normalized differences reveals additional insights (Table A5 Panel A). Compared to the 2SDD approach, the BJS2024, SA2021, and dCDH2024 estimators produce larger absolute t-statistics on average (column 1) and a higher share of statistically significant event-study coefficients (column 2). Moreover, those estimators produce a higher share of estimates with extreme levels of statistical significance, defined using as thresholds the 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 CS2021 estimator leads to significantly smaller absolute t-statistics and a significantly smaller share of significant event-study coefficients, but no significant reduction in extreme levels of statistical significance. These conclusions continue to hold if we use weights to adjust for differences in the number of periods for each outcome variable. In fact, when weighting by the inverse of the

number of outcomes for each paper (Table A5 Panel C), the CS2021 estimator produces higher absolute t-statistics, more significant event-study coefficients, and a greater proportion of extremely statistically significant estimates. We also find similar results when restricting the sample to the subset of estimates that all five estimators agree are statistically significant (Table A6), as well as when expanding the sample to include estimates that only a subset of methods produce and adding paper-outcome-period fixed effects (Table A7). 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 A10); this moderation places the 2SDD estimator toward the conservative end of the spectrum, evident from its low rate of extreme t-statistics (Table A5).

A.5 Differences in pre-event coefficients

One of the most noticeable features of the event-study graphs is the difference in estimates in the periods leading up to the event. While the dCDH2024 and SA2021 estimators tend to produce greater statistical significance in the post-treatment period (Table A5), we do not find the same pattern in the pre-treatment periods (Table A8), where greater statistical significance would indicate violations of parallel trends.[8] In the case of the CS2021 estimator, we see significantly fewer significant coefficients in the pre-treatment periods.[9] This suggests that the 2SDD and BJS2024 methods may offer a more conservative approach. The discrepancy in pre-event coefficient estimates between 2SDD and BJS2024 requires further discussion. These differences do not stem from a fundamental distinction in the methodologies. Instead, they reflect different choices about what pre-event coefficients to estimate, with both methods accommodating either choice. One approach, which BJS2024 advocate, is to estimate the pre-event coefficients in the first stage of estimation, which uses only untreated observations. This approach results in more outlier standard errors (Figure A11). 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.

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

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

A.6 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 provides 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 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.[10] This is not a problem of the heterogeneity-robust estimators simply being imprecise: In 29 percent of these instances, all of the heterogeneity-robust estimators are statistically significant with the same sign while the TWFE estimate is not significantly different from zero. In 58 percent of these discrepancies, TWFE is statistically significant while all of the heterogeneity-robust estimators are 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, as discussed in Section 4. 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.

B Serial Correlation

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

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

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

[10] 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.

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

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

and the following second-stage regression:

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

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

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

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

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

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

This implies

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

or equivalently

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

which allows us to back out the parameters of interest.

Table A1: 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		✓

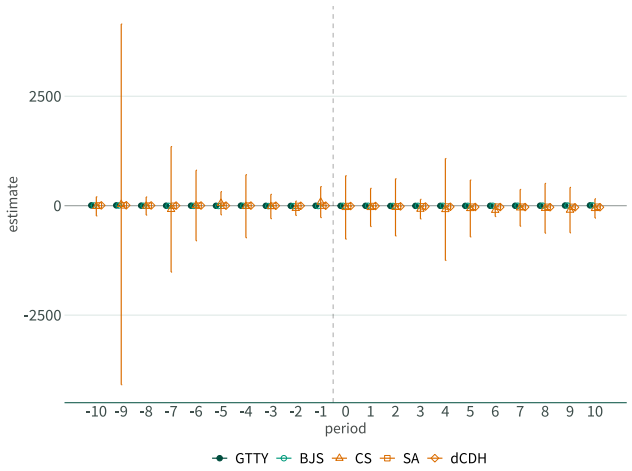
Note: This table describes the set of empirical papers that we reexamine using publicly available data and code. One paper in the broader replication set reports treatment effect estimates at the yearly level while the timing of treatment is at the monthly level (Kuziemko, Meckel and Rossin-Slater, 2018); its event-study estimates appear in Figure A7. 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.

References

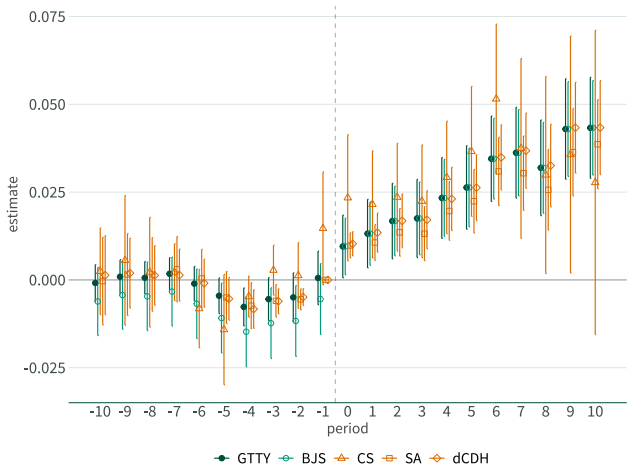
Bailey, Martha J, and Andrew Goodman-Bacon. 2015. "The War on Poverty's experiment in public medicine: Community health centers and the mortality of older Americans." *American Economic Review*, 105(3): 1067–1104.

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.*

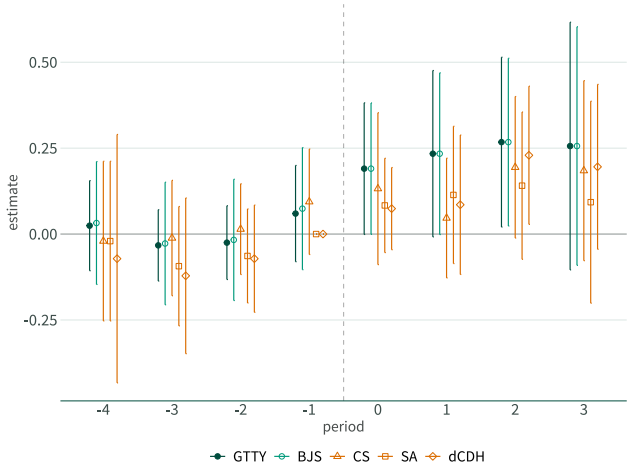
Figure A1: Empirical applications: Event-study estimates



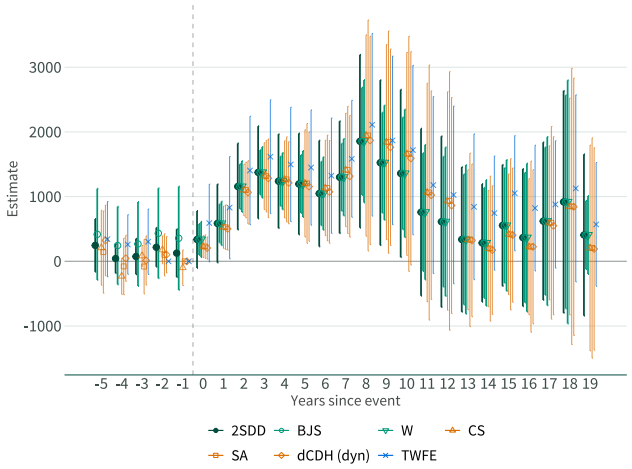
(a) Bailey and Goodman-Bacon (2015)



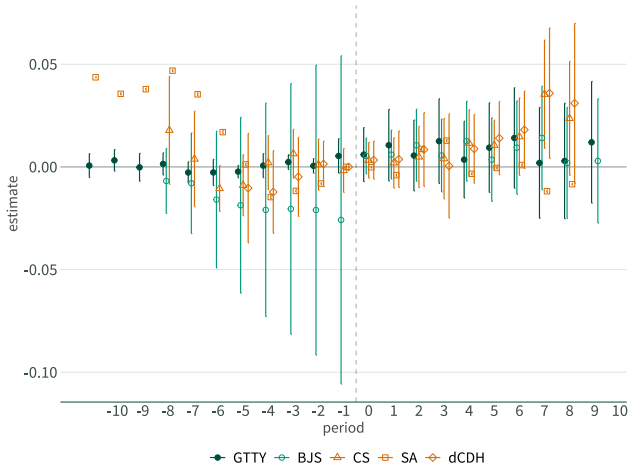
(b) Deryugina (2017)



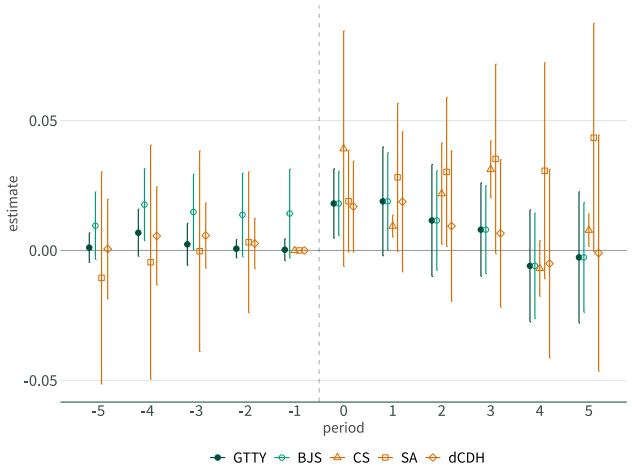
(c) He and Wang (2017)



(d) Lafortune, Rothstein and Schanzenbach (2018)



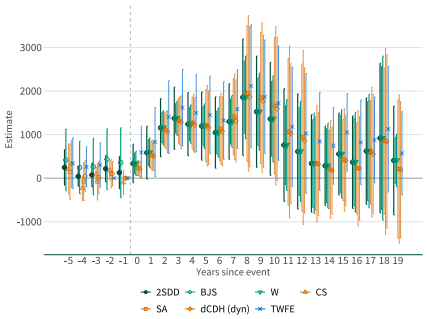
(e) Tewari (2014)



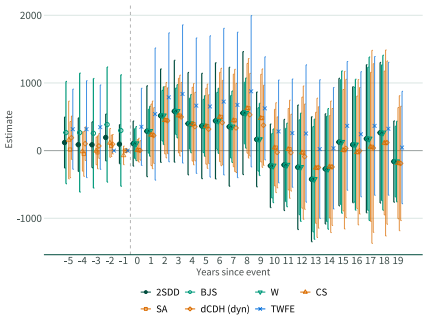
(f) Ujhelyi (2014)

Note: This figure reports event-study estimates from applying each estimator to the first event-study specification for each of the main empirical settings in Table A1.

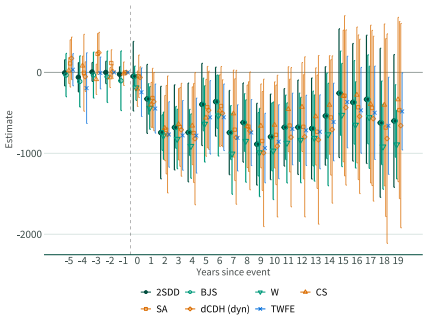
Figure A2: Empirical applications: Lafortune, Rothstein and Schanzenbach (2018) event study estimates



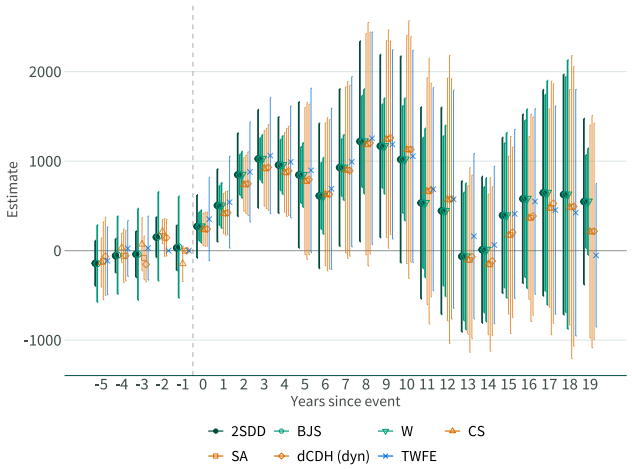
(a) Figure 3



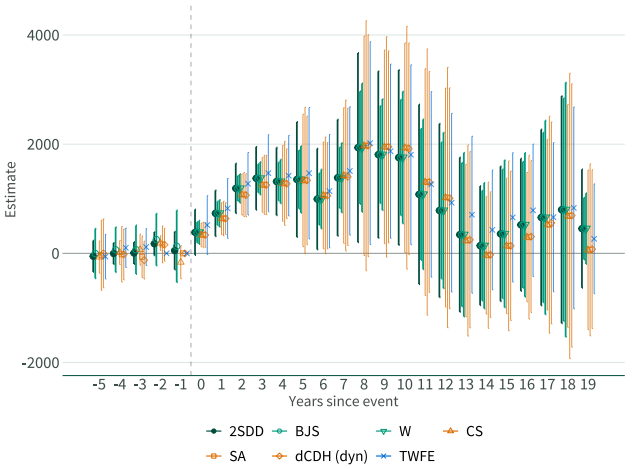
(b) Figure 4



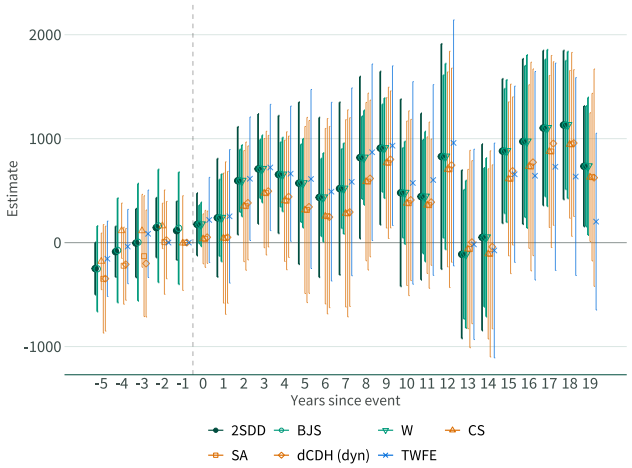
(c) Figure 5



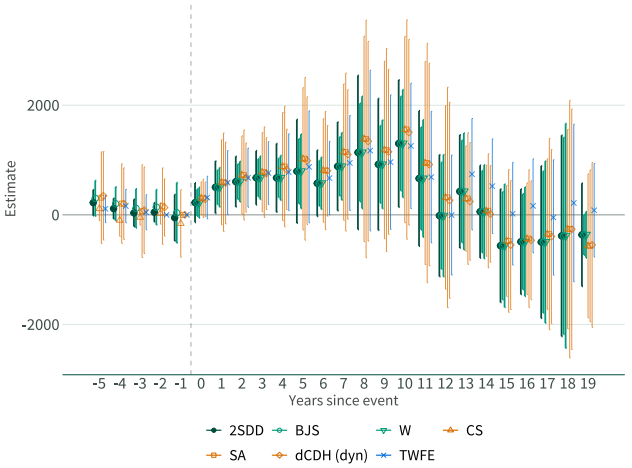
(d) Figure A.3(a)



(e) Figure A.3(b)



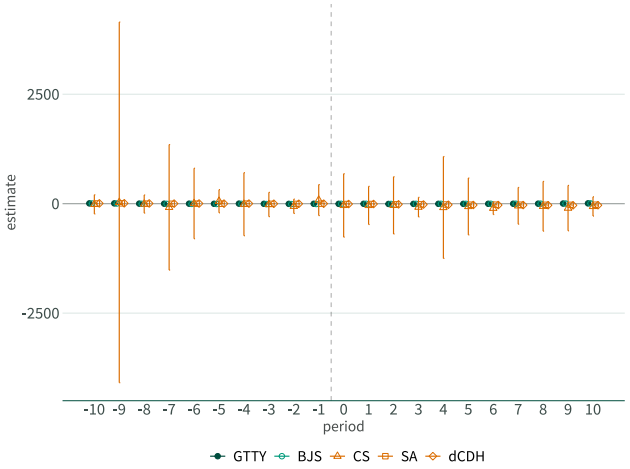
(f) Figure A.3(c)



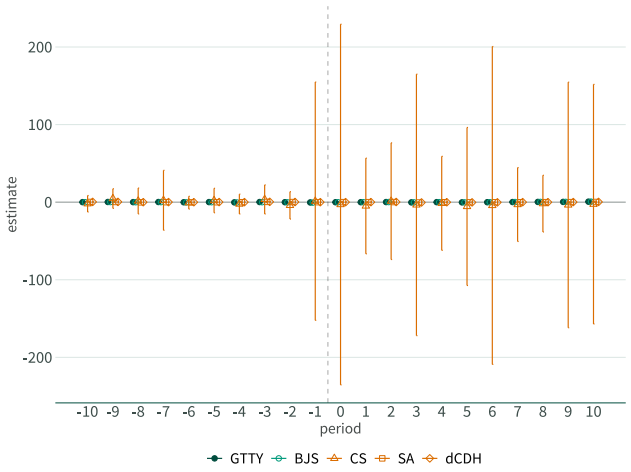
(g) Figure A.3(d)

Note: This figure reports event-study estimates from applying each estimator to the event-study specifications in Lafortune, Rothstein and Schanzenbach (2018).

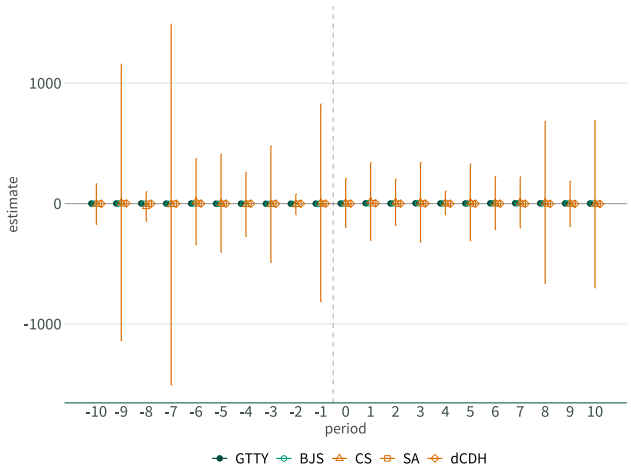
Figure A3: Empirical applications: Bailey and Goodman-Bacon (2015) event study estimates



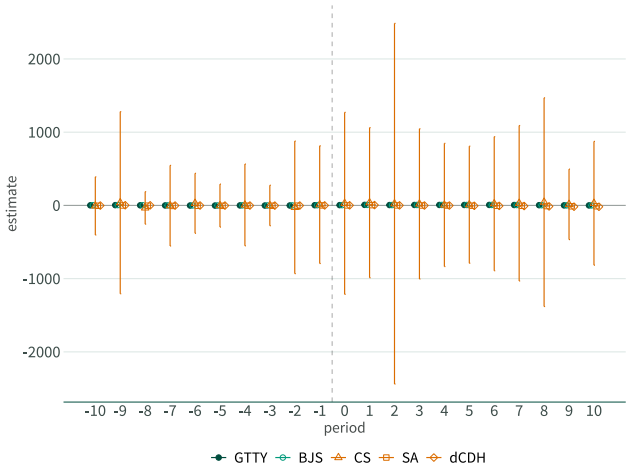
(a) Figure 5



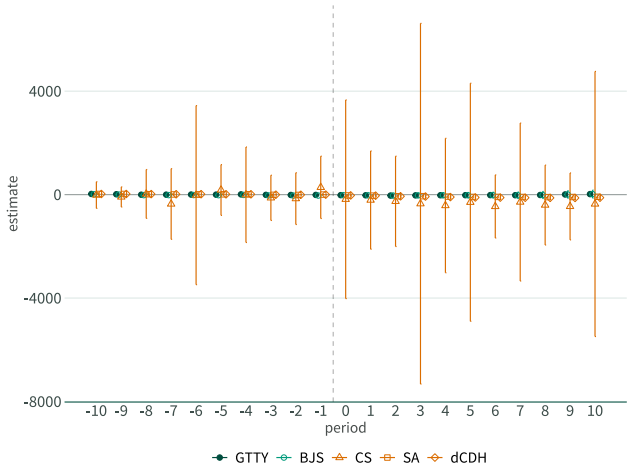
(b) Figure 7(a)



(c) Figure 7(b)



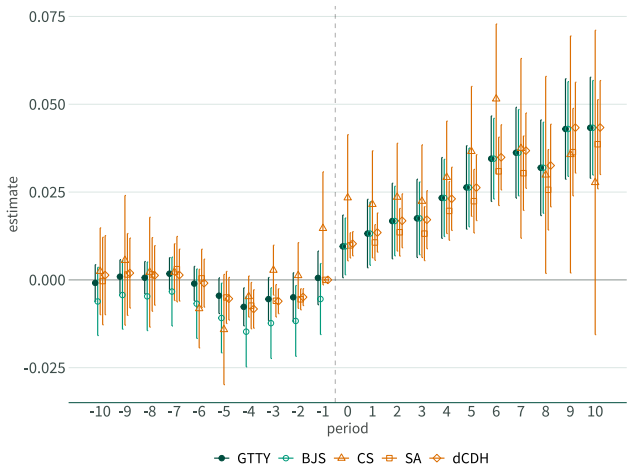
(d) Figure 7(c)



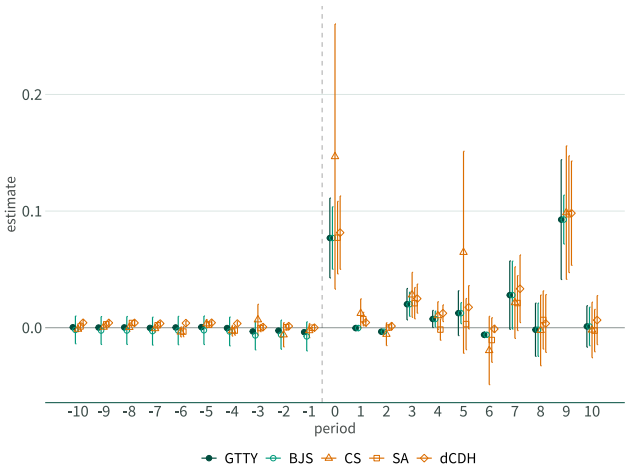
(e) Figure 7(d)

Note: This figure reports event-study estimates from applying each estimator to the event-study specifications in Bailey and Goodman-Bacon (2015).

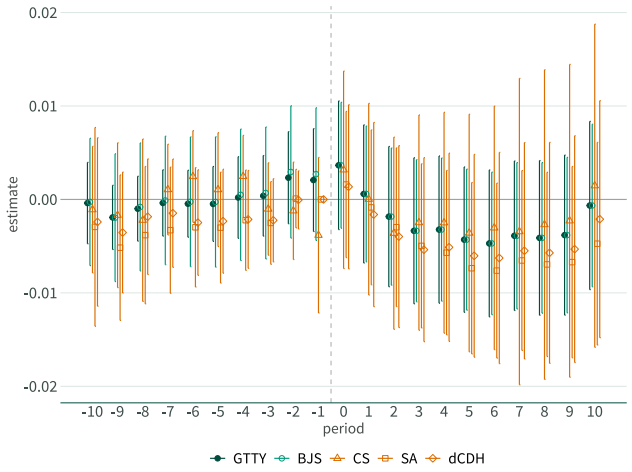
Figure A4: Empirical applications: Deryugina (2017) event study estimates (part 1)



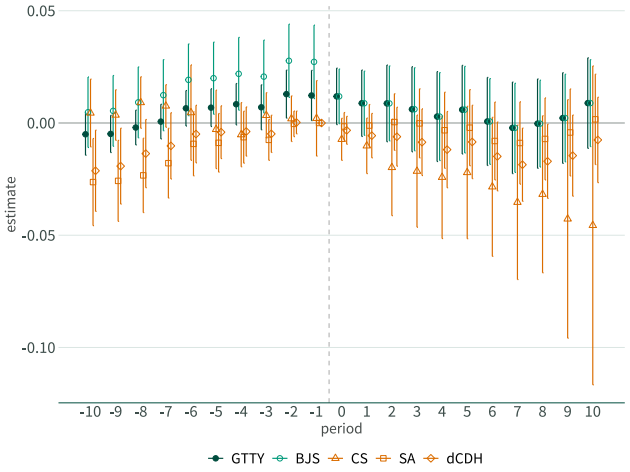
(a) Figure 2(a)



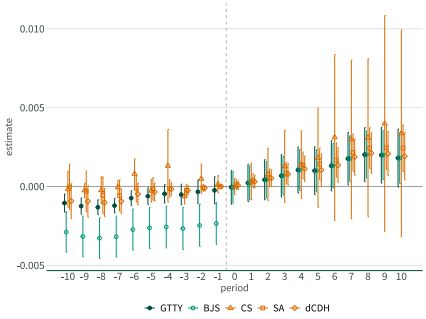
(b) Figure 2(b)



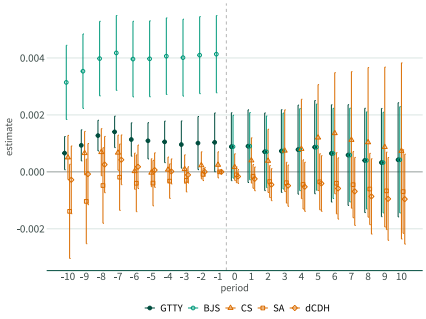
(c) Figure 2(c)



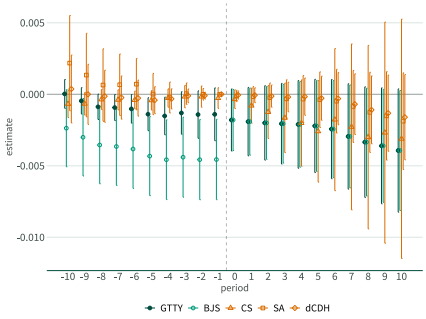
(d) Figure 2(d)



(e) Figure 3(b)



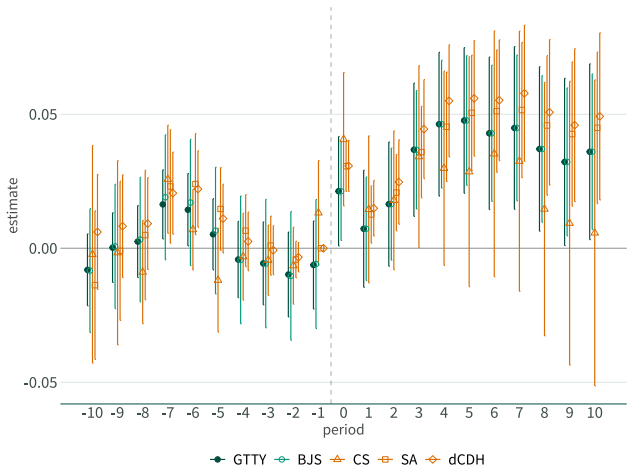
(f) Figure 3(c)



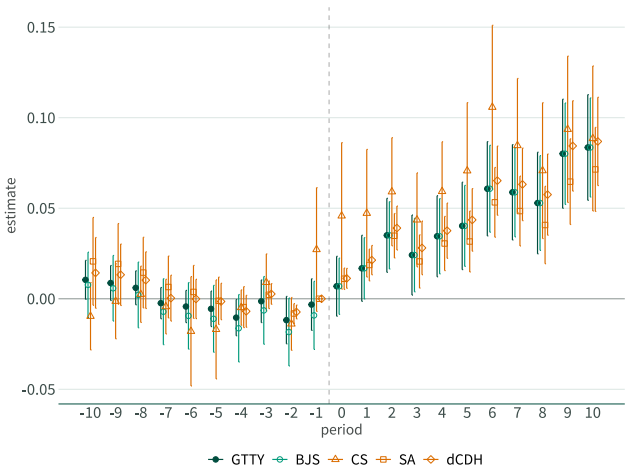
(g) Figure 3(d)

Note: This figure reports event-study estimates from applying each estimator to the event-study specifications in Deryugina (2017); see Figure A5 for the remaining estimates.

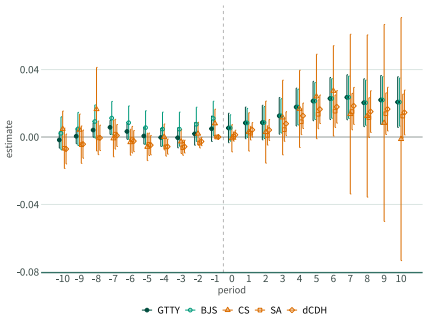
Figure A5: Empirical applications: Deryugina (2017) event study estimates (part 2)



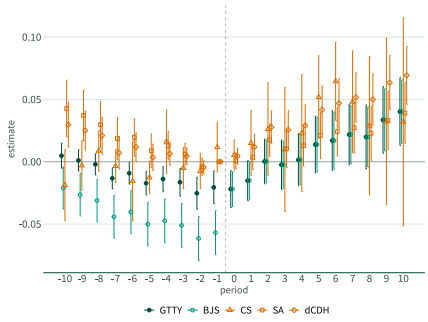
(a) Figure 4(a)



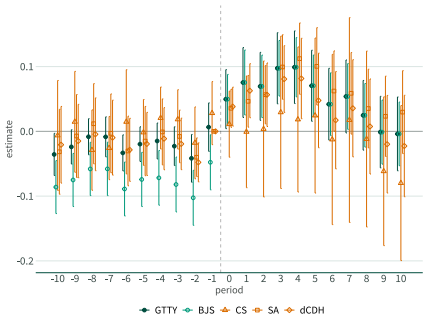
(b) Figure 4(b)



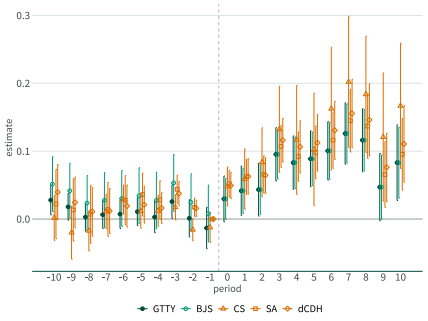
(c) Figure 4(c)



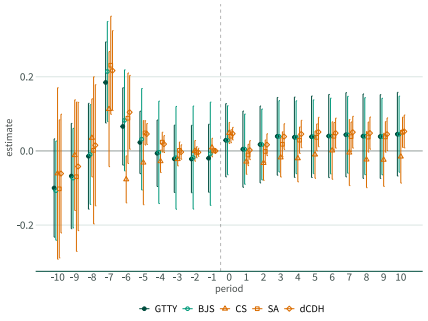
(d) Figure 4(d)



(e) Figure 5(a)



(f) Figure 5(b)



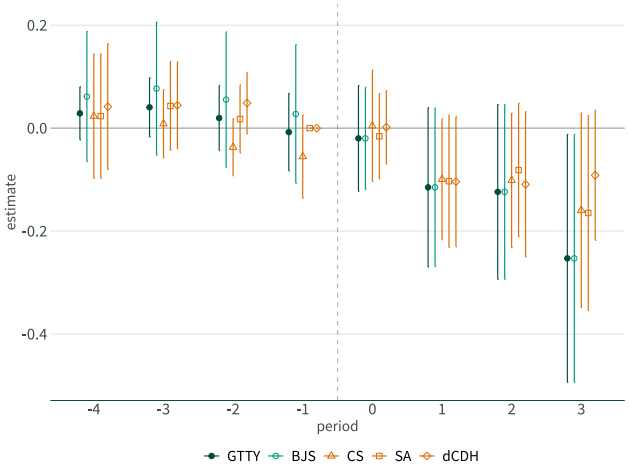
(g) Figure 5(c)

(h) Figure 5(d)

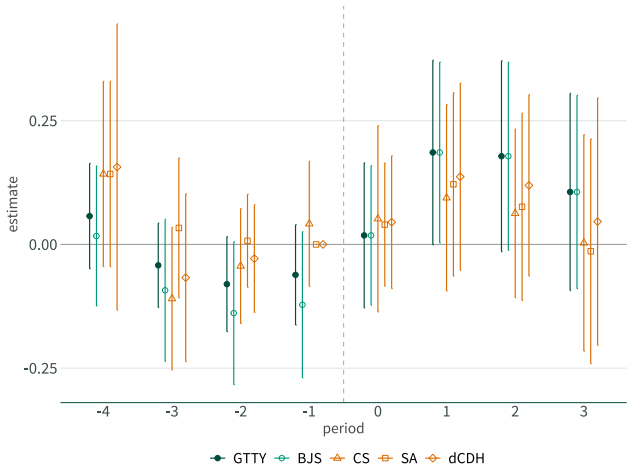
Note: This figure reports event-study estimates from applying each estimator to the event-study specifications in Deryugina (2017); see Figure A4 for the remaining estimates.

Figure A6: Empirical applications: He and Wang (2017) event study estimates

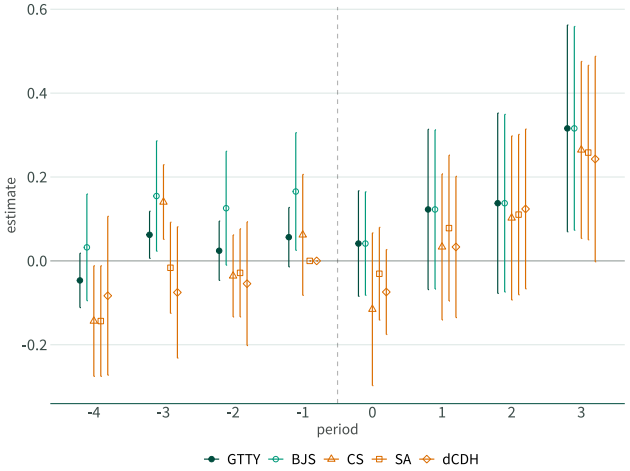
(a) Figure 2(a)



(b) Figure 2(b)



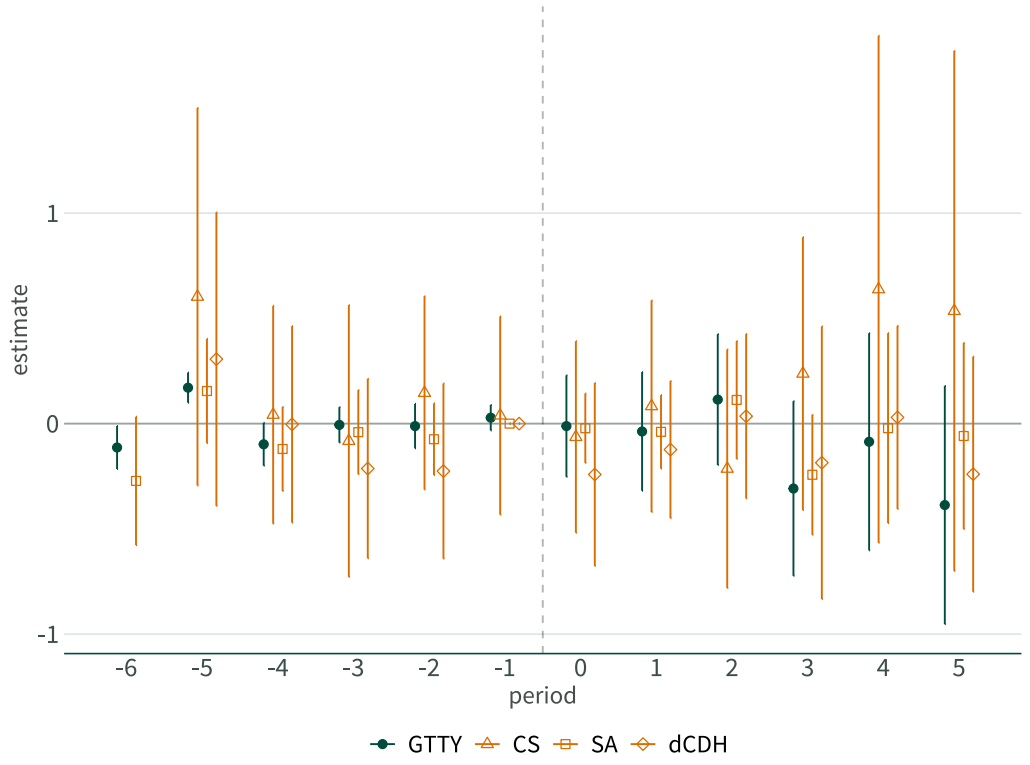
(c) Figure 2(c)



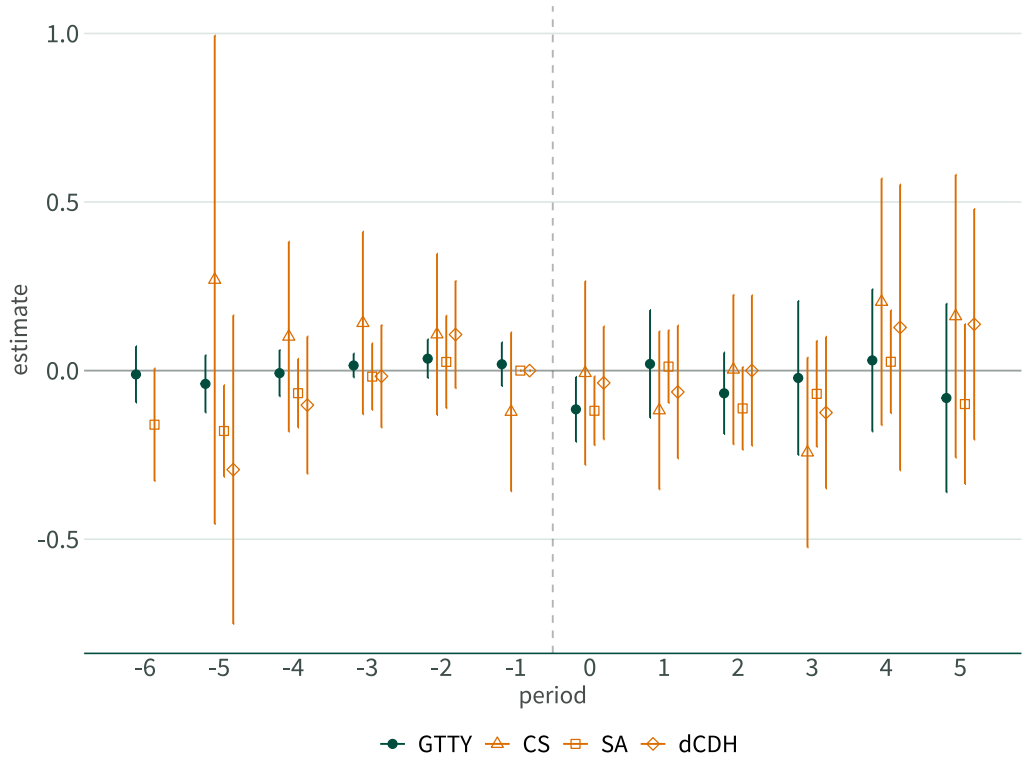
(d) Figure 2(d)

Note: This figure reports event-study estimates from applying each estimator to the event-study specifications in He and Wang (2017).

Figure A7: Empirical applications: Kuziemko, Meckel and Rossin-Slater (2018) event study estimates



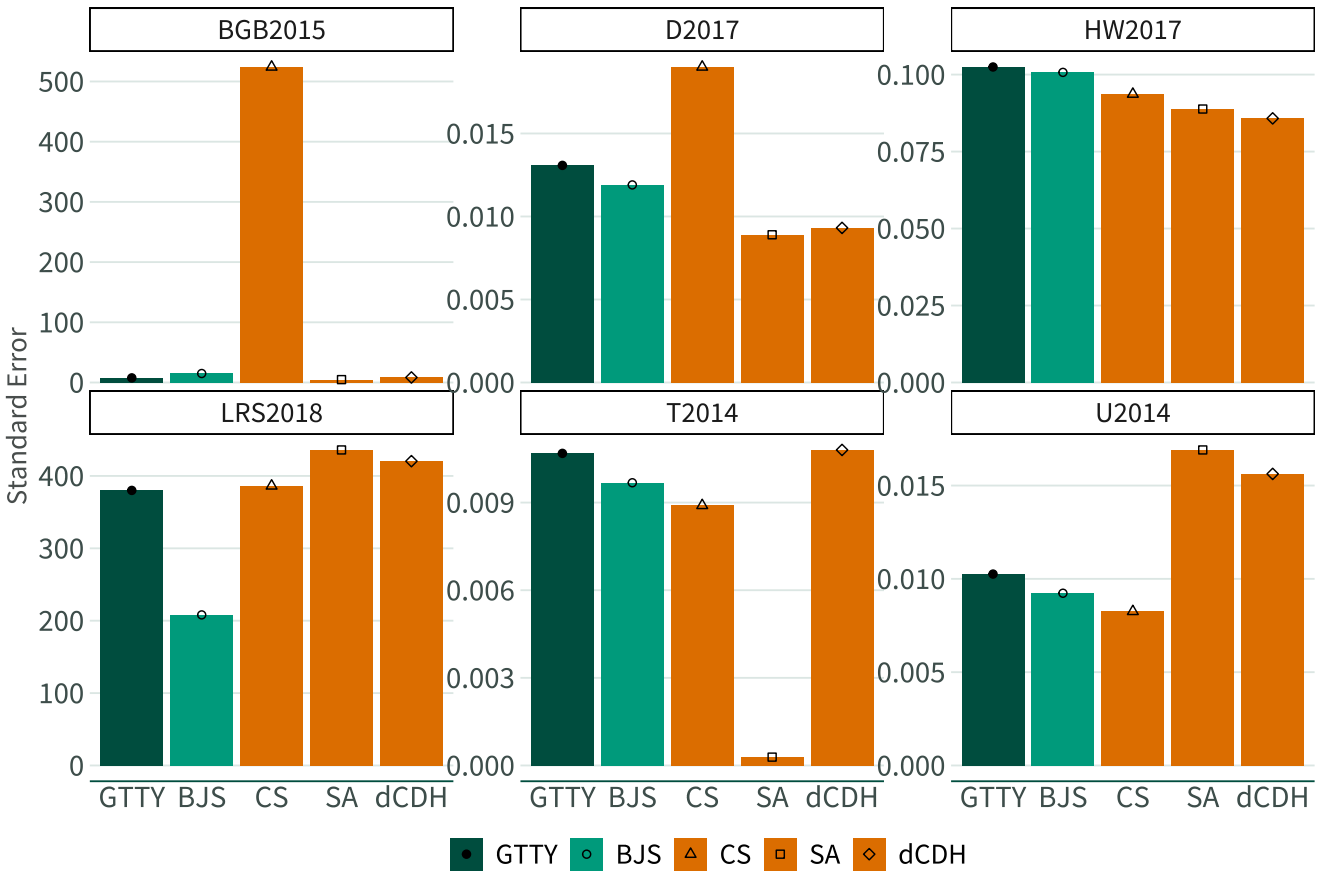
(a) Figure 2(a)



(b) Figure 2(b)

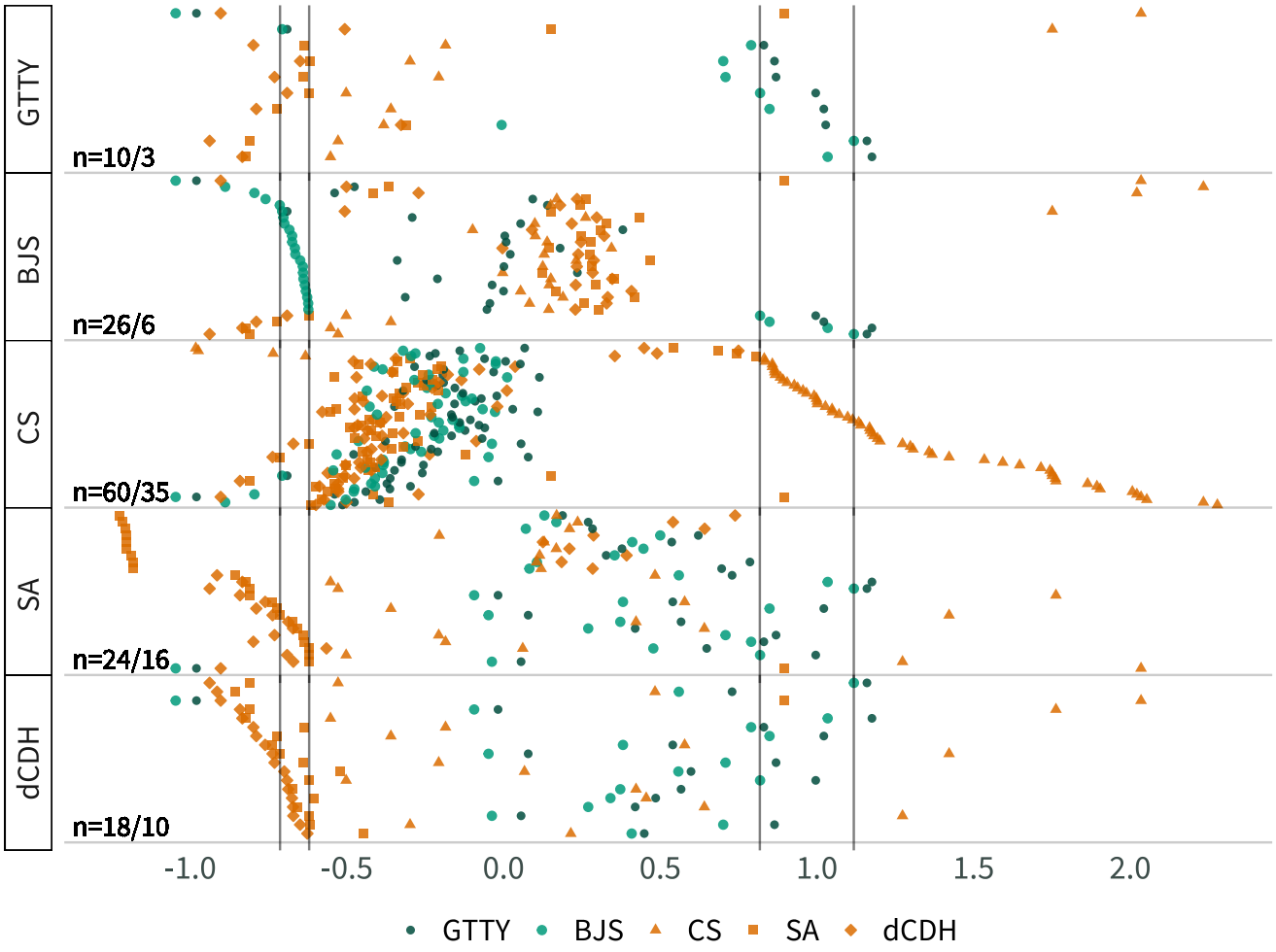
Note: This figure reports event-study estimates from applying each estimator to the event-study specifications in Kuziemko, Meckel and Rossin-Slater (2018).

Figure A8: Empirical applications: Comparison of standard errors



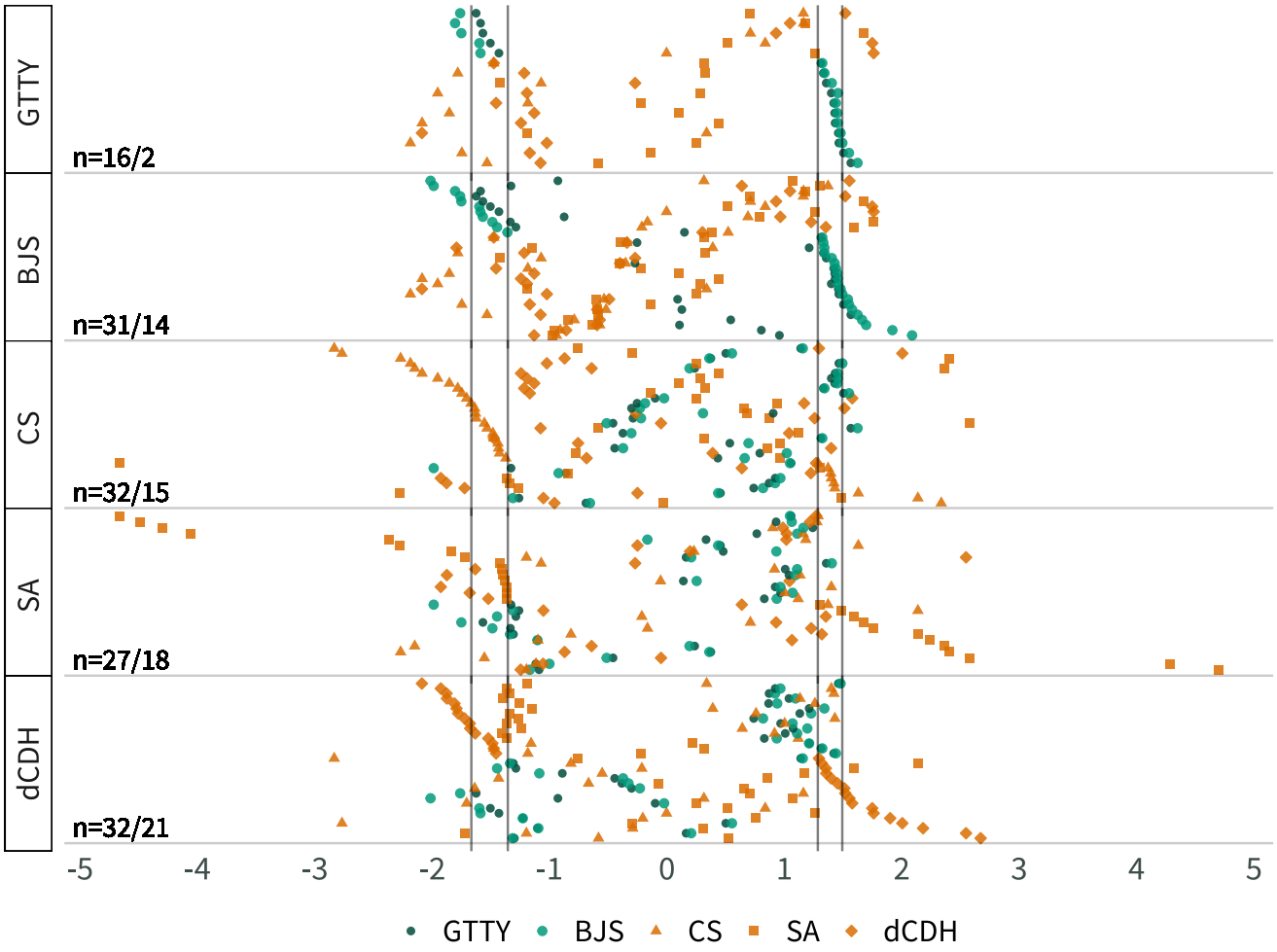
Note: This figure reports the average standard error across all dynamic treatment effect estimates for each replicated paper and each estimation method. The set of papers corresponds to the main empirical settings from Table A1.

Figure A9: Empirical applications: Outlier post-treatment normalized standard error differences



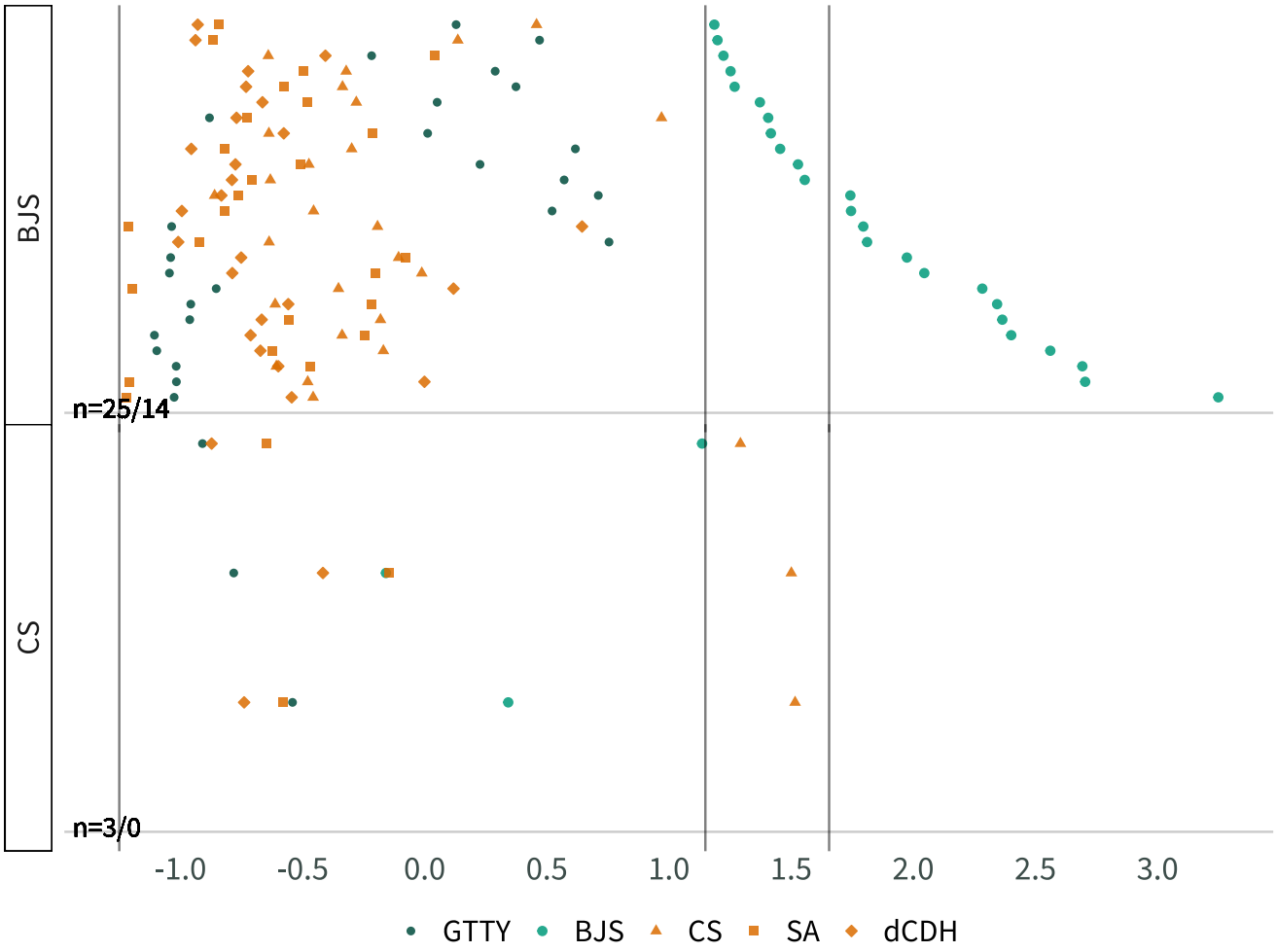
Note: Each panel of this figure corresponds to one of the five estimators we investigate. Each entry for a given estimator corresponds to an estimate (associated with a particular post-treatment period, outcome variable, and empirical setting) for which that estimator's standard error significantly deviates 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 A10: Empirical applications: Outlier post-treatment normalized t-statistic differences



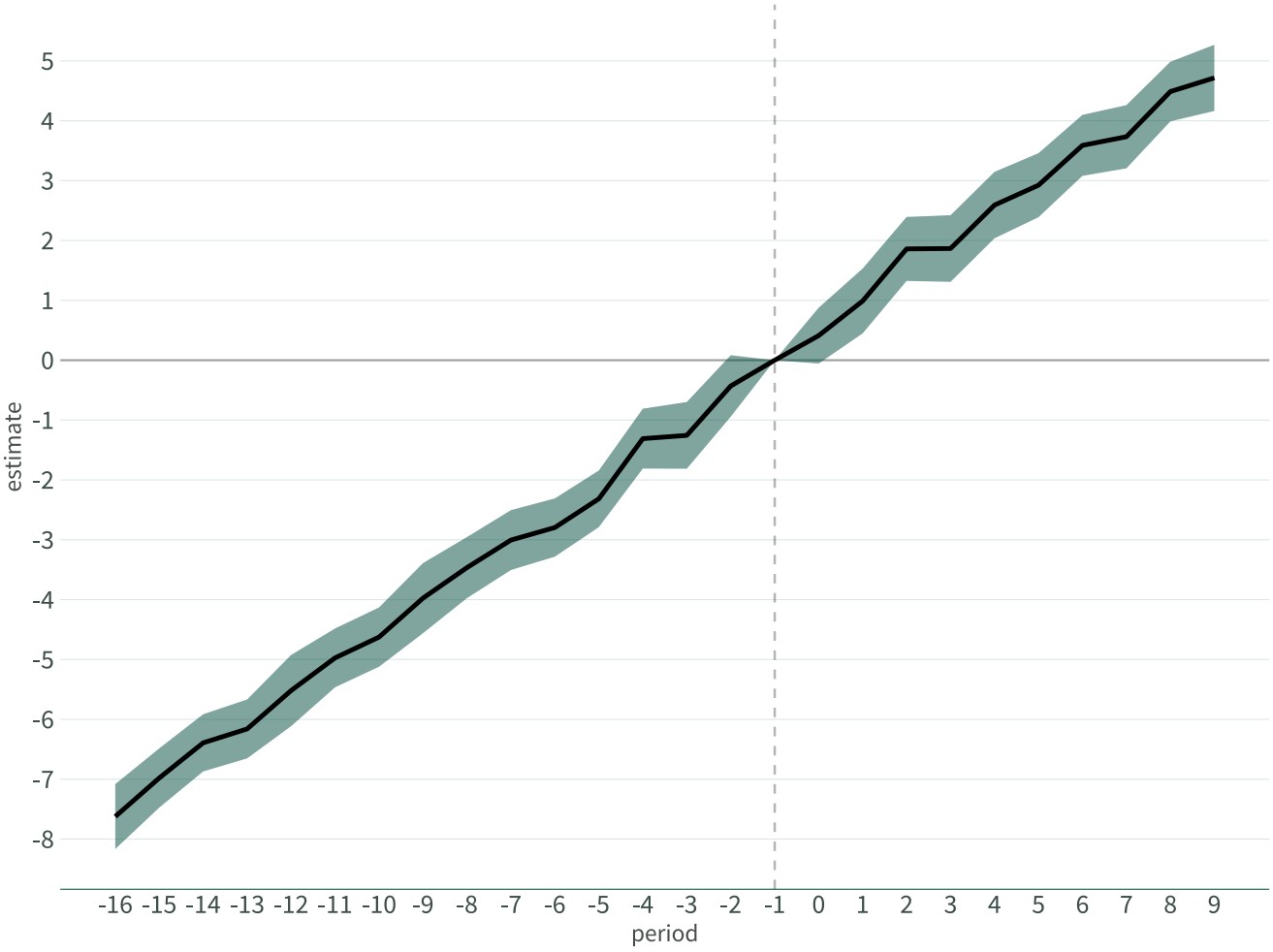
Note: Each panel of this figure corresponds to one of the five estimators we investigate. Each entry for a given estimator corresponds to an estimate (associated with a particular post-treatment period, outcome variable, and empirical setting) for which that estimator's t-statistic significantly deviates 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.

Figure A11: Empirical applications: Outlier pre-treatment normalized standard error differences



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

Figure A12: Event-study in non-staggered setting with pre-trend



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

Table A2: 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 A1. 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 A3: 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 Supp. Appendix A.2. GTTY refers to the method proposed in the current paper. The full set of event-study estimates appear in Figures A1 to A6.

Table A4: 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 A1. The regression omits the linear period trend for the 2SDD estimator. See Table A2 for additional details.

Table A5: 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 A1. 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.

Table A6: Empirical applications: Comparison of t-statistics (always-significant effects)

	$\lvert t\rvert$	$\mathbb{1}_{\{\lvert t\rvert>4\}}$	$\mathbb{1}_{\{\lvert t\rvert>8\}}$
Panel A: Unweighted			
BJS	1.5334	0.3793	0.1379
	(0.2769)	(0.0866)	(0.0457)
CS	-0.1498	-0.0690	0.0000
	(0.1832)	(0.0818)	(0.0000)
SA	0.5620	0.1724	0.0172
	(0.2163)	(0.0894)	(0.0172)
dCDH	1.0068	0.3103	0.0690
	(0.2346)	(0.0885)	(0.0336)
Panel B: Weighted (outcomes)			
BJS	1.6800	0.4359	0.1568
	(0.3428)	(0.1076)	(0.0558)
CS	-0.1863	-0.0207	0.0000
	(0.2274)	(0.0778)	(0.0000)
SA	0.4857	0.1495	0.0101
	(0.2821)	(0.0953)	(0.0103)
dCDH	0.8941	0.2965	0.0384
	(0.2772)	(0.1063)	(0.0199)
Panel C: Weighted (papers)			
BJS	1.7600	0.4461	0.1605
	(0.2837)	(0.0818)	(0.0521)
CS	-0.0756	-0.0282	0.0000
	(0.1837)	(0.0811)	(0.0000)
SA	0.4884	0.1654	0.0147
	(0.2146)	(0.0869)	(0.0147)
dCDH	0.8940	0.2892	0.0588
	(0.2350)	(0.0883)	(0.0290)

Note: This table describes the relationship between each estimator and the absolute t-statistics of the dynamic treatment effect estimates for the subsample of coefficients for which all five methods yield a statistically significant effect. See Table A5 for further information.

Table A7: Empirical applications: Comparison of t-statistics (all estimates)

	$\lvert t\rvert$		$\mathbb{1}_{\{\lvert t\rvert>1.96\}}$		$\mathbb{1}_{\{\lvert t\rvert>4\}}$		$\mathbb{1}_{\{\lvert t\rvert>8\}}$	
Panel A: Unweighted								
BJS	0.5045	0.4715	0.0835	0.0747	0.0882	0.0843	0.0204	0.0189
	(0.1118)	(0.0914)	(0.0338)	(0.0223)	(0.0203)	(0.0165)	(0.0072)	(0.0096)
CS	-0.2720	-0.2753	-0.0811	-0.0820	-0.0074	-0.0074	0.0025	0.0025
	(0.0846)	(0.0777)	(0.0311)	(0.0213)	(0.0143)	(0.0133)	(0.0025)	(0.0067)
SA	0.7074	0.7041	0.0628	0.0619	0.0965	0.0965	0.0198	0.0198
	(0.2036)	(0.1791)	(0.0333)	(0.0218)	(0.0204)	(0.0162)	(0.0069)	(0.0088)
dCDH	0.5937	0.4312	0.1347	0.0944	0.2351	0.2351	0.1683	0.1683
	(0.1109)	(0.0906)	(0.0355)	(0.0236)	(0.0248)	(0.0207)	(0.0186)	(0.0175)
Panel B: Weighted (outcomes)								
BJS	0.3672	0.2885	0.0710	0.0514	0.0612	0.0529	0.0134	0.0110
	(0.1043)	(0.0957)	(0.0355)	(0.0232)	(0.0183)	(0.0135)	(0.0049)	(0.0072)
CS	-0.2803	-0.2837	-0.0780	-0.0789	-0.0089	-0.0089	0.0028	0.0028
	(0.0876)	(0.0909)	(0.0318)	(0.0231)	(0.0146)	(0.0140)	(0.0028)	(0.0054)
SA	0.7915	0.7882	0.0798	0.0789	0.0982	0.0982	0.0223	0.0223
	(0.2286)	(0.2021)	(0.0353)	(0.0228)	(0.0210)	(0.0156)	(0.0078)	(0.0082)
dCDH	0.4872	0.4079	0.1111	0.0926	0.1663	0.1663	0.1018	0.1018
	(0.1153)	(0.0959)	(0.0370)	(0.0245)	(0.0224)	(0.0168)	(0.0126)	(0.0119)
Panel C: Weighted (papers)								
BJS	0.3034	0.0534	0.0847	0.0478	0.0429	0.0233	0.0093	-0.0006
	(0.1200)	(0.3690)	(0.0520)	(0.0477)	(0.0105)	(0.0213)	(0.0034)	(0.0184)
CS	0.0817	0.0713	0.0189	0.0161	0.0555	0.0555	0.0138	0.0138
	(0.1601)	(0.3473)	(0.0537)	(0.0515)	(0.0338)	(0.0340)	(0.0137)	(0.0152)
SA	2.6647	2.6543	0.1360	0.1332	0.1538	0.1538	0.0974	0.0974
	(0.9733)	(0.9094)	(0.0532)	(0.0496)	(0.0355)	(0.0343)	(0.0341)	(0.0314)
dCDH	0.1893	0.1144	0.0313	0.0206	0.1040	0.1040	0.0792	0.0792
	(0.1188)	(0.3357)	(0.0424)	(0.0403)	(0.0194)	(0.0208)	(0.0165)	(0.0175)
Controls		X		X		X		X

Note: This table describes the relationship between each estimator and the absolute t-statistics of the dynamic treatment effect estimates for the full set of event-study coefficients, including those which only a subset of methods can estimate. The estimates in columns 2, 4, 6, and 8 include paper-outcome-period fixed effects. See Table A5 for further information.

Table A8: Empirical applications: Comparison of t-statistics (pre-treatment periods)

	$\lvert t\rvert$	$\mathbb{1}_{\{\lvert t\rvert>1.96\}}$	$\mathbb{1}_{\{\lvert t\rvert>4\}}$	$\mathbb{1}_{\{\lvert t\rvert>8\}}$
Panel A: Unweighted				
BJS	0.5183	0.1429	0.1250	0.0089
	(0.1256)	(0.0407)	(0.0339)	(0.0063)
CS	-0.4403	-0.1295	-0.0759	0.0000
	(0.0813)	(0.0300)	(0.0214)	(0.0000)
SA	0.3031	-0.0268	-0.0179	0.0134
	(0.2843)	(0.0355)	(0.0264)	(0.0077)
dCDH	-0.0004	-0.0134	-0.0000	0.0000
	(0.0950)	(0.0361)	(0.0276)	(0.0000)
Panel B: Weighted (outcomes)				
BJS	0.4278	0.1198	0.0972	0.0069
	(0.1131)	(0.0378)	(0.0282)	(0.0049)
CS	-0.3218	-0.0851	-0.0434	0.0000
	(0.0823)	(0.0307)	(0.0219)	(0.0000)
SA	0.6721	-0.0035	0.0035	0.0234
	(0.4774)	(0.0341)	(0.0249)	(0.0134)
dCDH	-0.0188	-0.0130	0.0000	0.0000
	(0.0860)	(0.0318)	(0.0222)	(0.0000)
Panel C: Weighted (papers)				
BJS	0.2539	0.0913	0.0563	0.0030
	(0.1149)	(0.0372)	(0.0185)	(0.0021)
CS	-0.2345	-0.0194	0.0013	0.0000
	(0.1087)	(0.0328)	(0.0208)	(0.0000)
SA	5.3443	0.1852	0.1881	0.1500
	(2.6215)	(0.0803)	(0.0805)	(0.0751)
dCDH	-0.0237	0.0008	0.0000	0.0000
	(0.0967)	(0.0269)	(0.0109)	(0.0000)

Note: This table presents results analogous to those in Table A7, but for the subsample of event-study coefficients corresponding to pre-treatment periods.

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

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

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

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

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

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

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

Gardner, John, Neil Thakral, Linh Tô, and Luther Yap. 2026. "Two-stage differences in differences." *Mimeo*.

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.

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

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

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

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

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

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

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

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

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