eAppendix

Supplement to: "Advances in difference-in-differences methods for policy evaluation research"

The Basic DiD Design: Identifying the DiD effect

Let $D_{g,t} = 1$ if group g is exposed to the policy in period t (the treated group), and $D_{g,t} = 0$ if group g is not exposed to the policy in period t (the comparison group). To understand the effects of California's PFL policy compared with a comparator state, say Nevada, we can define potential outcomes by group. Let $Y_{g,t}(0)$ be the potential outcome if group g were not exposed to the policy at time t, and $Y_{g,t}(1)$ be the potential outcome if the same group were exposed to the policy at time t.

To identify the DiD effect δ_{DiD}, we adopt three key assumptions. First, we adopt a *consistency assumption*, whereby a group's potential outcomes correspond to the one from its observed treatment status: $Y = (1 - D) \cdot Y(0) + D \cdot Y(1)$. For example, if a Californian is exposed to the PFL policy, their observed outcome is the potential outcome of someone who was treated: $Y = Y(1)$, and their potential outcome $Y(0)$ is unobserved. Similarly, if a Nevadan is unexposed to the PFL policy, their observed outcome is the potential outcome of someone who was not treated: $Y = Y(0)$, and their outcome $Y(1)$ is unobserved. Second, we adopt the *parallel trends assumption* that the change in outcomes in the comparator group is a good counterfactual for the untreated potential outcomes in the treated group: $\mathrm{E}\big[Y_{post}(0) - Y_{pre}(0)|D = 1\big] = \mathrm{E}\big[Y_{post}(0) - Y_{pre}(0)|D = 0\big]$. Third, we adopt a *no-anticipation assumption*, stating that the treatment has no effect prior to its implementation. Formally, we can write this assumption as: $Y_{pre}(0) = Y_{pre}(1)$ for all g with $D_g = 1$.

Our target estimand is the average treatment effect on the treated (ATT), defined for DiD as: $ATT \equiv \mathrm{E}\big[Y_{post}(1) - Y_{post}(0)|D = 1\big]$. In our example, this is the post-policy difference in health outcomes in California with a PFL law versus without a PFL law. Applying the assumptions of *parallel trends* and *no-anticipation*, one can estimate the DiD effect:

$$\begin{aligned} ATT &\equiv \mathrm{E}\big[Y_{post}(1) - Y_{post}(0)|D = 1\big] \\ &= \underbrace{\mathrm{E}\big[Y_{post}|D = 1\big] - \mathrm{E}\big[Y_{pre}|D = 1\big]}_{\text{pre-post change in treated group}} - \underbrace{\mathrm{E}\big[Y_{post}|D = 0\big] - \mathrm{E}\big[Y_{pre}|D = 0\big]}_{\text{pre-post change in comparator group}}. \\ &= \hat{\delta}_{DiD} \end{aligned} \tag{1}$$

We can see that the ATT is the double-difference of the pre-post change in outcomes in the treated group and the pre-post change in outcomes in the comparator group. In this basic setup, one can simply estimate the crude DiD effect by plugging in the sample averages: $\hat{\delta}_{DiD} = \big(\bar{Y}_{treat,post} - \bar{Y}_{treat,pre}\big) - \big(\bar{Y}_{comp,post} - \bar{Y}_{comp,pre}\big)$. eFigure 1 provides a graphical illustration.

Forbidden comparisons

The following derivation shows how TWFE produced biased DiD estimates because of "forbidden comparisons".

Goodman-Bacon (2021) illustrates that the problem of negative weights often originate from "forbidden comparisons" between a group treated in a later period (late-treated) and a group treated in an earlier period (early-treated).[1] The weighted average of 2×2 DiDs that comprises TWFE includes "clean" comparisons between treated and not-yet-treated groups, as well as "forbidden comparisons" that generate bias.

Consider a simple example of PFL policies. We have three waves of data—2000 (Period 1), 2005 (Period 2), and 2010 (Period 3)—and two groups—California as the early-treated (PFL in 2004) and New Jersey as the late-treated (PFL in 2009). In Period 1, both states are untreated. In Period 2, California switches to treated, and in Period 3, New Jersey switches to treated. The ATT is a weighted sum of the effect of California's law ($\delta_{early,2}$) and the effect of New Jersey's law ($\delta_{late,3}$):

$$\mathrm{ATT} = \mathrm{E}[\delta] = \frac{1}{2}\delta_{early,2} + \frac{1}{2}\delta_{late,3} \quad (2)$$

with

$$\hat{\delta}_{early,2} = Y_{early,2} - Y_{early,1} - (Y_{late,2} - Y_{late,1})$$
$$\hat{\delta}_{late,3} = Y_{late,3} - Y_{late,2} - (Y_{early,3} - Y_{early,2})$$

For the treatment effect in early-treated group (i.e., individuals living in California) in period 2, i.e., $\tau_{early,2}$, under the parallel trends assumption, both groups would have experienced the same outcome evolution without the treatment, i.e., $Y_{early,2}(0) - Y_{early,1}(0) = Y_{late,2}(0) - Y_{late,1}(0)$, we have

$$\begin{aligned}\hat{\delta}_{early,2} &= Y_{early,2} - Y_{early,1} - (Y_{late,2} - Y_{late,1}) \\ &= Y_{early,2}(1) - Y_{early,1}(0) - (Y_{late,2}(0) - Y_{late,1}(0)) \\ &= Y_{early,2}(1) - Y_{early,2}(0) \\ &= \tau_{early,2}\end{aligned} \quad (3)$$

Hence, $\hat{\delta}_{early,2}$ is unbiased for the treatment effect in the early-treated group. This is similar to what has been discussed in the aforementioned basic 2x2 DiD design.

When estimating the treatment effect in the late-treated group (i.e., individual living in New Jersey) in period 3, i.e., $\delta_{late,3}$, we have

$$Y_{early,3} - Y_{early,2} = (Y_{early,3}(0) + \tau_{early,3}) - (Y_{early,2}(0) + \tau_{early,2})$$
$$Y_{late,3} - Y_{late,2} = (Y_{late,3}(0) + \tau_{late,3}) - Y_{late,2}(0)$$

Under the parallel trends assumption, i.e., $Y_{early,3}(0) - Y_{early,2}(0) = Y_{late,3}(0) - Y_{late,2}(0)$, we have

$$\hat{\delta}_{late,3} = Y_{late,3} - Y_{late,2} - (Y_{early,3} - Y_{early,2})$$

$$= \tau_{late,3} - \tau_{early,3} + \tau_{early,2} \quad (4)$$

Combining (2), (3), and (4), we have

$$E[\delta] = \frac{1}{2}\tau_{late,3} + \tau_{early,2} - \frac{1}{2}\tau_{early,3} \quad (5)$$

As can be seen in Eq.(5), $E[\delta]$ is a weighted sum of three treatment effects, where one treatment effect (i.e., $\tau_{early,3}$) receives a negative weight. This comes from the term $\hat{\delta}_{late,3}$, which compares a late-treated group to an early-treated group, e.g., comparing individuals in New Jersey to individuals in California after California has been treated. If the treatment effect for the early-treated group does not change over time, i.e., $\tau_{early,2} = \tau_{early,3}$, then (5) simplifies to $\frac{1}{2}\tau_{late,3} + \frac{1}{2}\tau_{early,2}$. Hence, $E[\delta]$ estimates the average of treatment effects on the treated (ATT). However, if the treatment effect changes over time, negative weights may occur and $E[\delta]$ will be biased.

Similar issues for event-study designs under heterogeneous treatment effects

Sun and Abraham (2021) show that, in a staggered treatment setting, forbidden comparisons will contaminate the coefficients on leads and lags due to the negative weighting problem in the presence of heterogenous dynamic treatment effects.[2] This can lead to inaccurate estimation of policy effects and unreliable testing results of the parallel trends assumption. Only when the evolution of the policy effects is identical across groups (e.g., between California and New Jersey) will the estimated coefficients for the leads or lags be unbiased.[2] Callaway and Sant'Anna (2022) use a Monte Carlo simulation to illustrate how estimated coefficients for leads and lags using TWFE methods differ from the true effects in the presence of heterogenous treatment effects.[3]

The overall ATT

As mentioned in the main manuscript, there are several ways to summarize $\mathrm{ATT}_{g,t}$'s into an overall ATT. This includes averaging all group-time ATTs, group ATTs, event-time ATTs, or cohort ATTs.[4] Determining the most appropriate method for summarizing an overall ATT requires careful consideration. A simple average of all $\mathrm{ATT}_{g,t}$'s will put more weight on groups that undergo treatment for longer periods. Alternatively, averaging group ATTs, where group ATTs equal the mean of $\mathrm{ATT}_{g,t}$'s across all post-treatment time periods within each group, provides an interpretation akin to the ATT in a standard 2×2 DiD design, i.e., the overall effect of being in the treatment group.

On the other hand, averaging event times or cohort ATTs, where event times or cohort ATTs are the mean of $\mathrm{ATT}_{g,t}$'s for each event time or cohort, may introduce complexities in interpretating of the overall ATT, especially when the sample is unbalanced and potential changes in group composition across different time periods. Using a balanced panel with respect to the event time may alleviate this concern. These nuances were thoroughly discussed in Callaway and Sant'Anna (2021) and Goin and Riddell (2023).[4, 5]

Adjusting covariates in the DiD designs

Zeldow and Hatfield (2021) extensively discussed time-varying confounders in the basic 2x2 DiD regression. They show that, depending on the confounding scenario, regressions that adjust for covariates with constant effects (i.e., by controlling for covariates themselves) or time-varying effects (e.g., by controlling for the interaction term between covariates and time) usually yield more robust DiD estimators. This is the conventional approach adopted by researchers using the basic DiD design, i.e., estimating a version of Eq. (1) in the main manuscript that adjusts for covariates potentially affecting the treated and comparator groups differently.

Likewise, for staggered policy implementation using TWFE regressions, researchers often estimate a version of Eq. (2) in the main manuscript that adjusts for all observable time-varying covariates. Time-invariant covariates are omitted as they are absorbed by the group fixed effects. However, as highlighted in recent econometrics literature,[6] there are several limitations besides the "bad control" problem:

First, TWFE regressions incorporating time-varying covariates do not condition the PTA on time-invariant covariates. If time-invariant covariates have time-varying effects on the outcome, failing to account for them may violate the PTA and generate bias.

Second, TWFE regressions only effectively control for changes in time-varying covariates over time, but not their levels. Hence, TWFE regressions only compares treated units and untreated units with same changes in covariate values, rather than comparing units with similar covariate levels. If the trajectories of untreated potential outcomes also depend on the level of the covariates, TWFE regressions may perform poorly. For example, in examining the effect of the PFL law, a TWFE regression only compares California (a treated state) to an untreated state with equivalent population change. However, if changes in health outcomes within a state depends on its population level (e.g., smaller in a treated state with a larger population), failing to consider this may result in bias.

One approach to address the first issue is to include an interaction term between the time-invariant covariate and time to control for time-varying confounders. For the second issue, some procedures can be employed to match each treated unit with a comparator unit with similar or identical covariate values. Recent heterogeneity-robust DiD approaches can perform these adjustments. For example, Callaway and Sant'Anna (2021) allows using regression adjustment, inverse probability weighting (IPW), and doubly-robust estimators (DR) for matching. Wooldridge (2021) allows using regression adjustment. It is worth noting that this approach will include all potential interactions among groups, time, and covariates, which may result in an extremely large number of estimated coefficients, increasing the challenges for model estimation. In Borusyak, Jaravel, and Spiess (2021), covariates can be incorporated in the first step regression that imputes counterfactual outcome for treated units using control units.

Honest DiD

Rambachan and Roth (2023) propose an approach to provide robust inference and sensitivity analysis under PTA violations.[7] Intuitively, they impose restrictions that bound the extent of post-treatment violations of parallel trends to be a constant $\bar{M}$ times the pre-treatment differences in trends. Then they propose conducting sensitivity analysis that accesses the robustness of treatment effect across a range of $\bar{M}$. Specifically, the sensitivity results can show a set of confidence intervals of ATTs across different $\bar{M}$, and conclude whether there is significant treatment effect when allowing for post-treatment differences in trends up to $\bar{M}$ times the difference in pre-treatment trends. Currently, this procedure can be integrated with the Callaway and Sant'Anna (2021) estimator in R (command: `HonestDiD`) and Stata (command: `honestdid`).

Simulation study: Data generating process (DGP)

We created 500 datasets, each comprising 30 units and 36 time periods for each unit. We divided these units to 6 cohorts that receive treatment at different times ("staggered" design). The DGP includes the following base model:

$$Y_{it} = \beta D_{it} + \mu_i + \gamma_t + e_{it}$$

where Y_{it} is the outcome of interest in unit i at time t, μ_i is the unit fixed effect, γ_t is the time fixed effect, and e_{it} is the error term. μ_i, γ_t , and e_{it} follow a normal distribution. D_{it} is the treatment indicator (1 in treated unit-time, and 0 otherwise), and β is the treatment effect, which was set based on scenarios:

- Across units, β can either be homogeneous (a random number between 2 and 10, same across units), random (each unit can have a random number between 2 and 10), or larger for those treated earlier (the first treated unit was assigned a random number between 6 and 10, the second treated unit was then assigned a number equal the effect of the first-treated unit minus 1, and so on)
- Across time, β is either constant, or increases linearly over time, i.e., $\beta \cdot (t - first_{treat} + 1)/36$ if $D_{it} = 1$, where $first\_treat$ is the initial treatment timing.

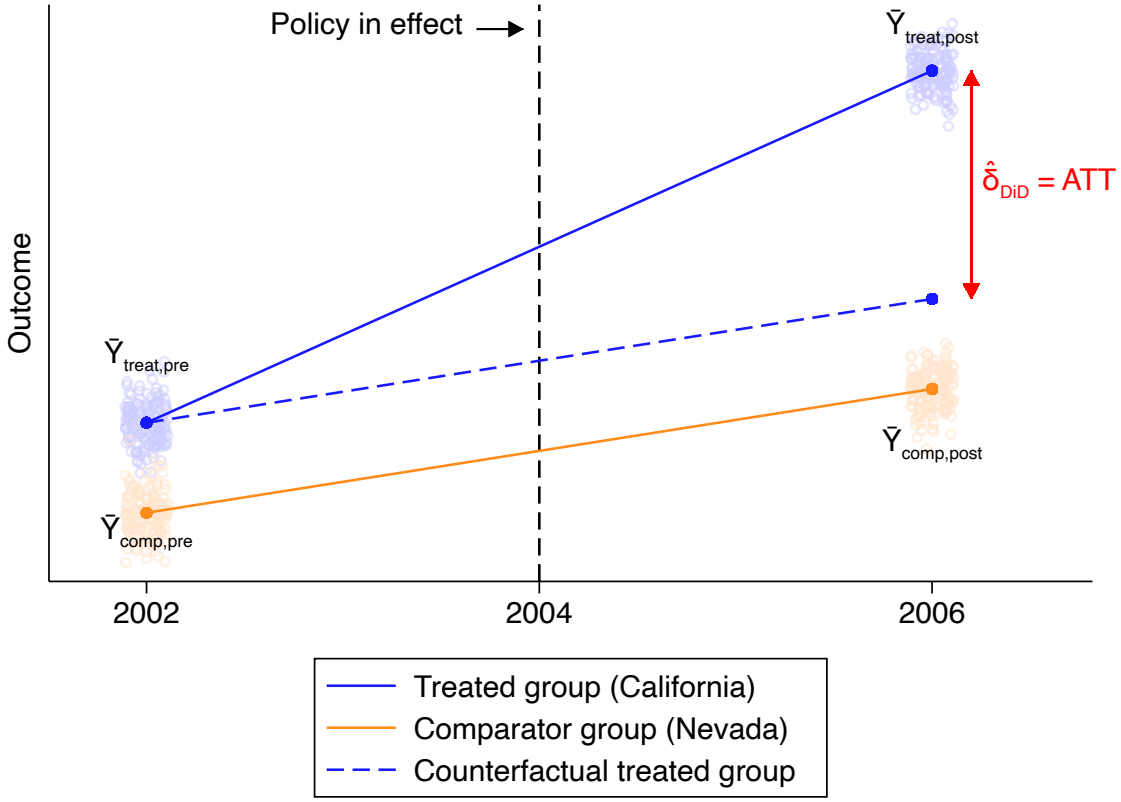
The base model does not allow for time-varying confounders, and thus no PTA violations. In scenarios with PTA violations, we also include a covariate with time-varying effects:

$$Y_{it} = \beta D_{it} + \lambda_t X_i + \mu_i + \gamma_t + e_{it}$$

where X_i follows a normal distribution. We allow the time-varying effect λ_t to be smaller among early treated units.

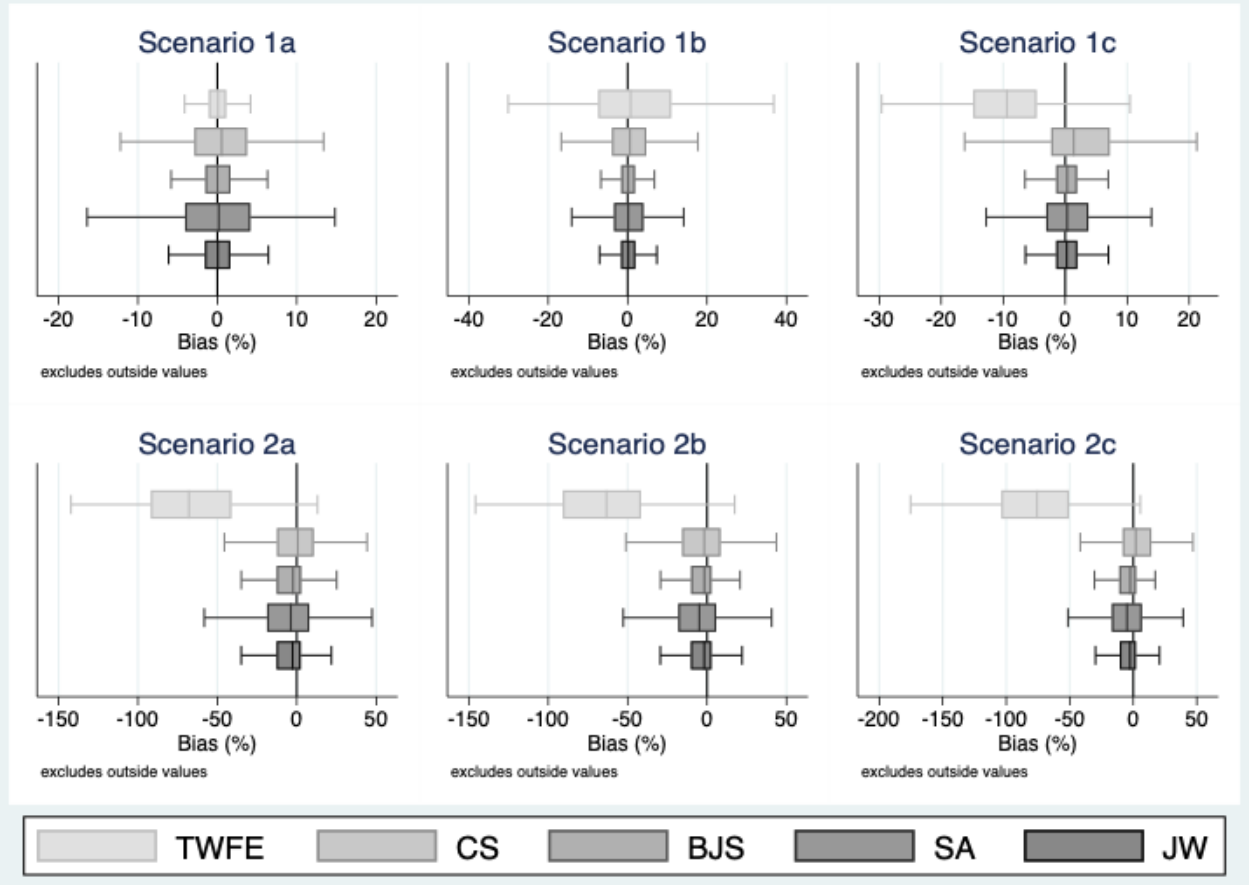
In sensitivity analyses, we increase the number of units from 30 to 50 (eFigures 2-3, eTable 1), the number of time periods from 36 to 60 (eFigures 4-5, eTable 2), and the number of simulation runs from 500 to 1000 (eFigures 6-7, eTable 3). The results from all of these sensitivity analyses were consistent with those from the base model.

eFigure 1. The basic difference-in-differences design



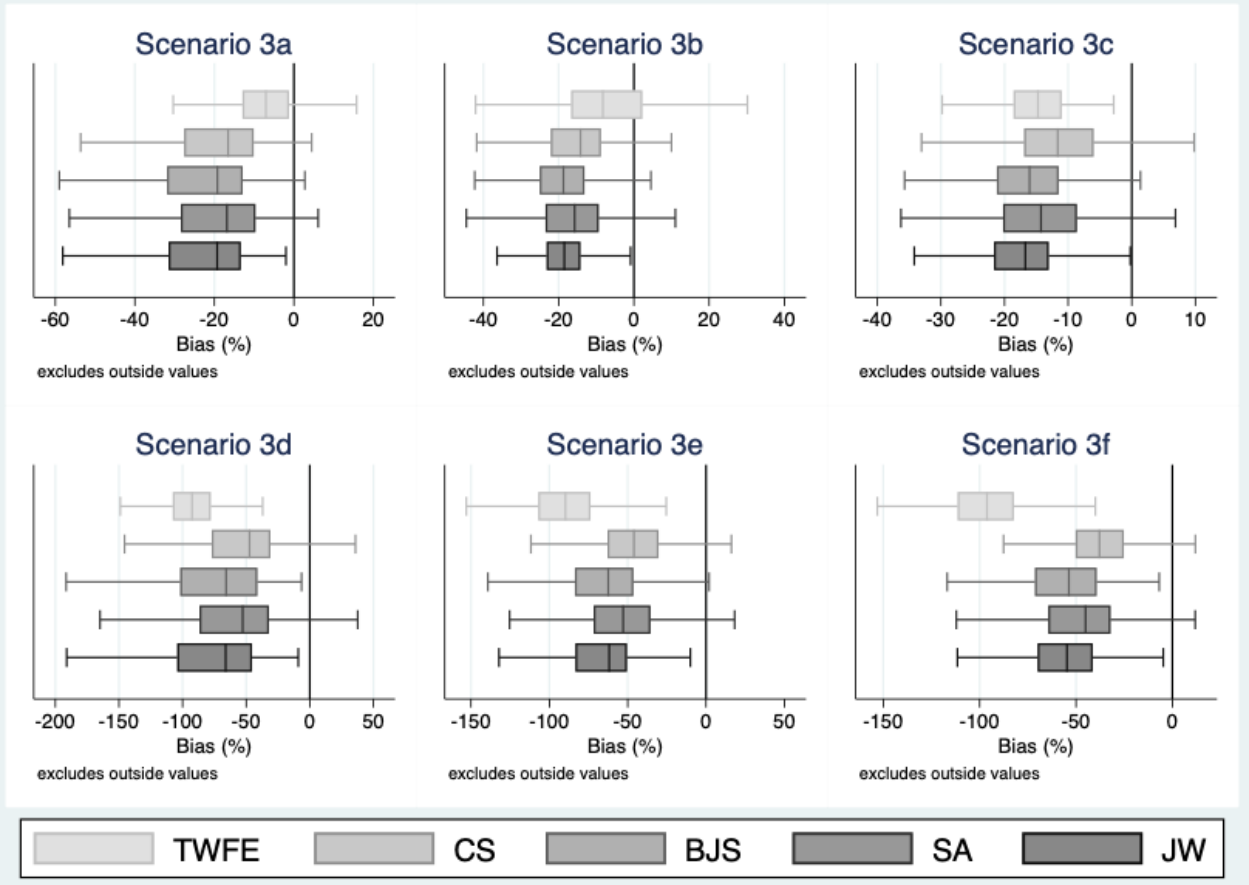
Note: Each dot represents a hypothetical data point from an individual living in California (light blue) or Nevada (light orange). The blue dotted line (counterfactual treated group) denotes what would have happened to the treated group (California) in the absence of treatment.

eFigure 2. Monte Carlo simulation results for Scenarios 1 and 2 increasing the number of units to 50



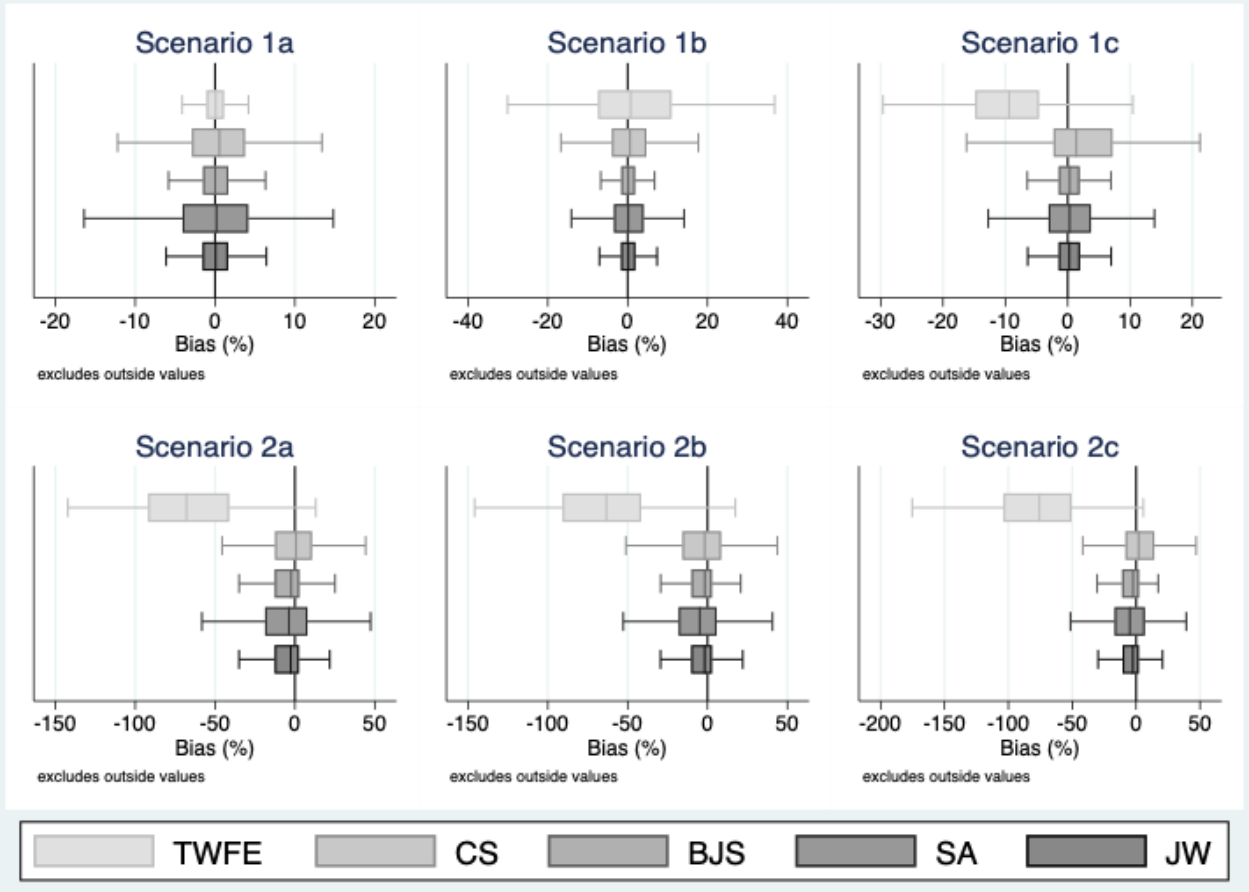
Abbreviations: TWFE: Two-way fixed effects; CS: Callaway-Sant'Anna; BJS: Borusyak-Jaravel-Spiess; SA: Sun-Abraham; JW: Wooldridge
Note: Scenarios 1a-1c have constant effects, and Scenarios 2a-2c have dynamic (linear trend) effects. Scenarios 1a and 2a have homogeneous effects across groups; Scenarios 1b and 2b have heterogeneous (at random) effects across groups; and Scenarios 1c and 2c have heterogeneous (large first) effects across groups. Each scenario is listed in Table 4 in the main article.

eFigure 3. Monte Carlo simulation results for Scenario 3 increasing the number of units to 50



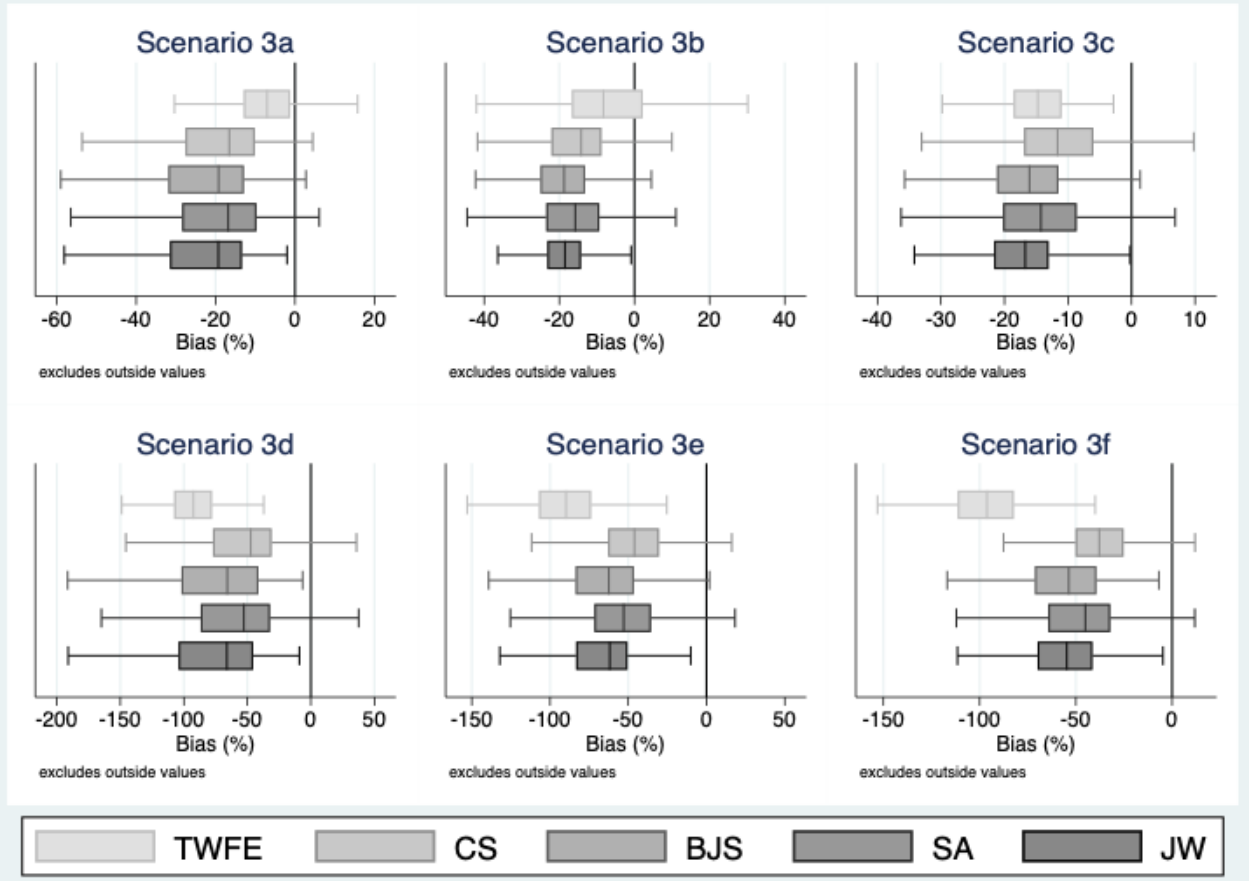
Abbreviations: TWFE: Two-way fixed effects; CS: Callaway-Sant'Anna; BJS: Borusyak-Jaravel-Spiess; SA: Sun-Abraham; JW: Wooldridge
Note: Scenarios 3a-3c have constant effects, and Scenarios 3d-3e have dynamic (linear trend) effects. Scenarios 3a and 3d have homogeneous effects across groups; Scenarios 3b and 3e have heterogeneous (at random) effects across groups; and Scenarios 3c and 3f have heterogeneous (large first) effects across groups. Each scenario is listed in Table 4 in the main article.

eFigure 4. Monte Carlo simulation results for Scenarios 1 and 2 increasing the number of time periods to 60



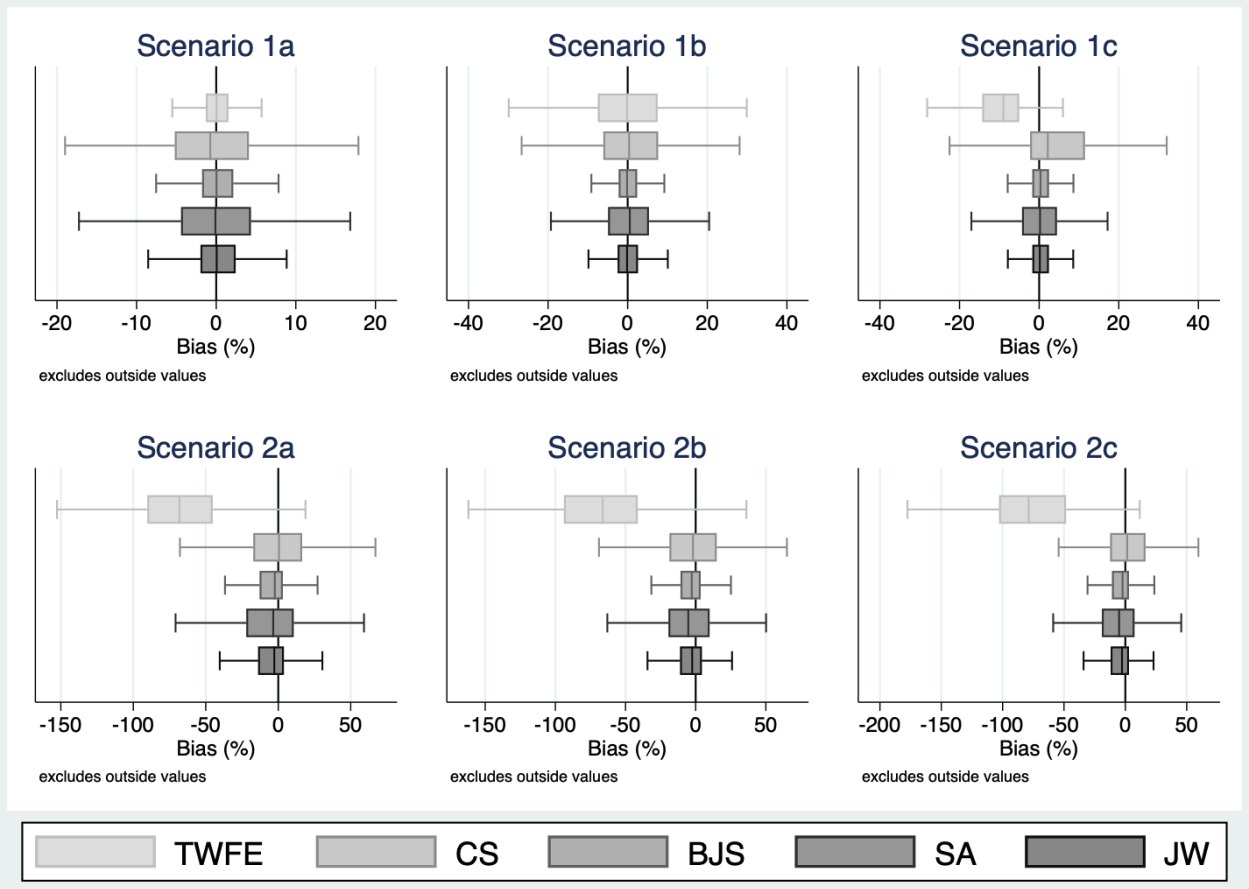
Abbreviations: TWFE: Two-way fixed effects; CS: Callaway-Sant'Anna; BJS: Borusyak-Jaravel-Spiess; SA: Sun-Abraham; JW: Wooldridge
Note: Scenarios 1a-1c have constant effects, and Scenarios 2a-2c have dynamic (linear trend) effects. Scenarios 1a and 2a have homogeneous effects across groups; Scenarios 1b and 2b have heterogeneous (at random) effects across groups; and Scenarios 1c and 2c have heterogeneous (large first) effects across groups. Each scenario is listed in Table 4 in the main article.

eFigure 5. Monte Carlo simulation results for Scenario 3 increasing the number of time periods to 60



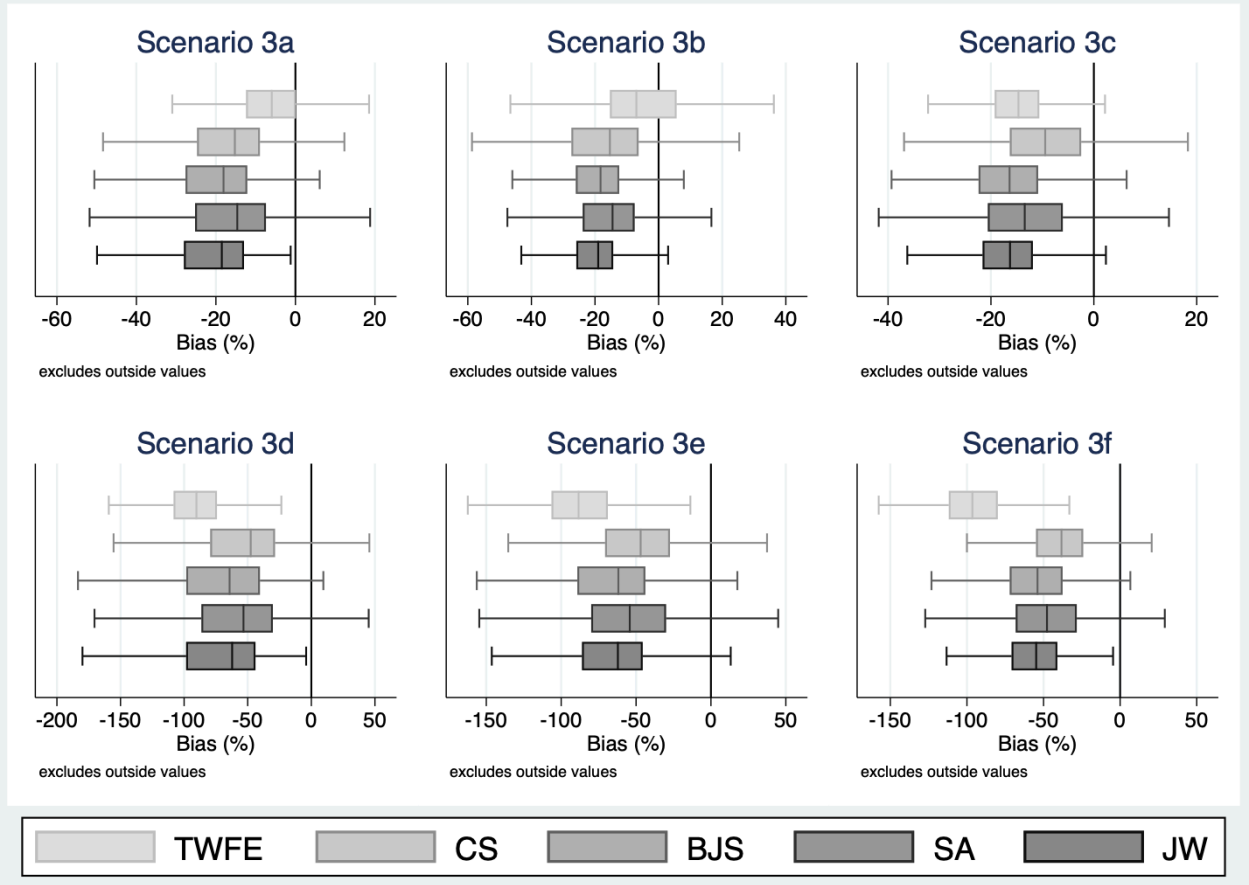
Abbreviations: TWFE: Two-way fixed effects; CS: Callaway-Sant'Anna; BJS: Borusyak-Jaravel-Spiess; SA: Sun-Abraham; JW: Wooldridge
Note: Scenarios 3a-3c have constant effects, and Scenarios 3d-3e have dynamic (linear trend) effects. Scenarios 3a and 3d have homogeneous effects across groups; Scenarios 3b and 3e have heterogeneous (at random) effects across groups; and Scenarios 3c and 3f have heterogeneous (large first) effects across groups. Each scenario is listed in Table 4 in the main article.

eFigure 6. Monte Carlo simulation results for Scenarios 1 and 2 increasing the number of simulation runs to 1000



Abbreviations: TWFE: Two-way fixed effects; CS: Callaway-Sant'Anna; BJS: Borusyak-Jaravel-Spiess; SA: Sun-Abraham; JW: Wooldridge
Note: Scenarios 1a-1c have constant effects, and Scenarios 2a-2c have dynamic (linear trend) effects. Scenarios 1a and 2a have homogeneous effects across groups; Scenarios 1b and 2b have heterogeneous (at random) effects across groups; and Scenarios 1c and 2c have heterogeneous (large first) effects across groups. Each scenario is listed in Table 4 in the main article.

eFigure 7. Monte Carlo simulation results for Scenario 3 increasing the number of simulation runs to 1000



Abbreviations: TWFE: Two-way fixed effects; CS: Callaway-Sant'Anna; BJS: Borusyak-Jaravel-Spiess; SA: Sun-Abraham; JW: Wooldridge
Note: Scenarios 3a-3c have constant effects, and Scenarios 3d-3e have dynamic (linear trend) effects. Scenarios 3a and 3d have homogeneous effects across groups; Scenarios 3b and 3e have heterogeneous (at random) effects across groups; and Scenarios 3c and 3f have heterogeneous (large first) effects across groups. Each scenario is listed in Table 4 in the main article.

eTable 1. Monte Carlo simulation estimates of the ATT increasing the number of units to 50

		No PTA violation (Scenarios 1 and 2)		No PTA violation (Scenario 3)	
Scenario	Methods	Bias (%)	RMSE	Bias (%)	RMSE
Constant, homogeneous effects		Scenario 1a		Scenario 3a	
	TWFE	0.09	0.09	-7.92	0.57
	CS	0.39	0.31	-19.66	1.10
	BJS	0.14	0.13	-23.05	1.19
	SA	-0.03	0.31	-19.86	1.11
	JW	0.07	0.13	-23.07	1.18
Constant, random HTE		Scenario 1b		Scenario 3b	
	TWFE	1.18	0.81	-6.31	0.99
	CS	1.09	0.88	-16.14	1.42
	BJS	0.45	0.33	-19.19	1.22
	SA	0.48	0.39	-16.59	1.11
	JW	0.47	0.27	-19.07	1.17
Constant, large-first HTE		Scenario 1c		Scenario 3c	
	TWFE	-9.72	0.77	-15.17	1.05
	CS	4.40	0.77	-10.88	0.94
	BJS	0.20	0.26	-16.84	1.21
	SA	0.32	0.32	-14.63	1.12
	JW	0.20	0.23	-17.43	1.21
Dynamic, homogeneous HTE		Scenario 2a		Scenario 3d	
	TWFE	-66.12	1.74	-94.90	1.94
	CS	0.60	0.39	-58.99	1.07
	BJS	-5.57	0.32	-76.25	1.29
	SA	-5.37	0.42	-63.08	1.19
	JW	-5.72	0.29	-77.19	1.28
Dynamic, random HTE		Scenario 2b		Scenario 3e	
	TWFE	-65.23	1.78	-89.86	2.01
	CS	-1.03	0.57	-49.70	1.10
	BJS	-5.05	0.35	-65.76	1.32
	SA	-6.12	0.41	-55.65	1.20
	JW	-5.05	0.33	-66.85	1.33
Dynamic, large-first HTE		Scenario 2c		Scenario 3f	
	TWFE	-75.84	2.22	-94.76	2.43
	CS	4.05	0.54	-37.75	1.02
	BJS	-5.55	0.39	-56.01	1.38
	SA	-5.62	0.46	-47.96	1.28
	JW	-5.61	0.35	-56.00	1.36

Abbreviations: TWFE: Two-way fixed effects; CS: Callaway-Sant'Anna; BJS: Borusyak-Jaravel-Spiess; SA: Sun-Abraham; JW: Wooldridge
Note: Each scenario is listed in Table 4.

eTable 2. Monte Carlo simulation estimates of the ATT increasing the number of time periods to 60

		No PTA violation (Scenarios 1 and 2)		PTA violation (Scenario 3)	
Scenario	Methods	Bias (%)	RMSE	Bias (%)	RMSE
Constant, homogeneous effects		Scenario 1a		Scenario 3a	
	TWFE	-0.04	0.09	-9.30	0.93
	CS	0.19	0.42	-36.21	2.02
	BJS	-0.11	0.13	-38.93	2.03
	SA	0.25	0.35	-33.84	1.89
	JW	0.08	0.16	-38.41	1.98
Constant, random HTE		Scenario 1b		Scenario 3b	
	TWFE	2.18	0.83	-6.73	1.22
	CS	-1.37	1.50	-31.55	2.51
	BJS	0.30	0.29	-32.26	2.04
	SA	1.06	0.43	-26.19	1.79
	JW	0.32	0.26	-32.65	1.99
Constant, large-first HTE		Scenario 1c		Scenario 3c	
	TWFE	-10.81	0.83	-17.78	1.33
	CS	10.79	1.16	-17.04	1.50
	BJS	1.21	0.28	-27.79	1.99
	SA	1.51	0.43	-22.89	1.77
	JW	1.29	0.27	-28.66	1.97
Dynamic, homogeneous HTE		Scenario 2a		Scenario 3d	
	TWFE	-73.96	1.92	-109.45	2.20
	CS	0.93	0.59	-114.85	2.00
	BJS	-6.26	0.33	-131.35	2.20
	SA	-7.99	0.48	-111.88	1.99
	JW	-6.50	0.34	-134.46	2.18
Dynamic, random HTE		Scenario 2b		Scenario 3e	
	TWFE	-73.46	1.84	-104.68	2.17
	CS	1.68	0.75	-91.55	2.03
	BJS	-7.54	0.36	-113.75	2.23
	SA	-7.54	0.49	-95.57	2.02
	JW	-7.62	0.34	-113.13	2.15
Dynamic, large-first HTE		Scenario 2c		Scenario 3f	
	TWFE	-85.75	2.36	-111.24	2.64
	CS	9.25	0.76	-72.06	1.76
	BJS	-5.87	0.34	-94.31	2.18
	SA	-6.27	0.49	-79.94	1.99
	JW	-5.87	0.33	-96.68	2.17

Abbreviations: TWFE: Two-way fixed effects; CS: Callaway-Sant'Anna; BJS: Borusyak-Jaravel-Spiess; SA: Sun-Abraham; JW: Wooldridge
Note: Each scenario is listed in Table 4.

eTable 3. Monte Carlo simulation estimates of the ATT increasing the simulation runs to 1000

		No PTA violation (Scenarios 1 and 2)		PTA violation (Scenario 3)	
Scenario	Methods	Bias (%)	RMSE	Bias (%)	RMSE
Constant, homogeneous effects		Scenario 1a		Scenario 3a	
	TWFE	0.13	0.12	-6.82	0.61
	CS	-0.33	0.43	-18.68	1.24
	BJS	0.17	0.16	-21.02	1.18
	SA	0.08	0.35	-17.38	1.11
	JW	0.22	0.18	-21.37	1.16
Constant, random HTE		Scenario 1b		Scenario 3b	
	TWFE	0.40	0.77	-4.82	1.00
	CS	0.73	1.20	-18.12	1.73
	BJS	-0.01	0.34	-19.52	1.25
	SA	0.51	0.44	-16.28	1.16
	JW	-0.01	0.28	-20.24	1.24
Constant, large-first HTE		Scenario 1c		Scenario 3c	
	TWFE	-9.64	0.78	-15.05	1.07
	CS	5.24	0.86	-8.18	0.97
	BJS	0.17	0.32	-16.85	1.25
	SA	0.03	0.43	-13.69	1.13
	JW	0.16	0.29	-16.95	1.20
Dynamic, homogeneous HTE		Scenario 2a		Scenario 3d	
	TWFE	-68.08	1.77	-92.32	2.01
	CS	0.29	0.60	-56.70	1.15
	BJS	-5.65	0.35	-72.87	1.34
	SA	-4.46	0.50	-61.83	1.28
	JW	-5.60	0.34	-72.99	1.31
Dynamic, random HTE		Scenario 2b		Scenario 3e	
	TWFE	-67.77	1.82	-88.34	1.94
	CS	-0.55	0.68	-51.29	1.21
	BJS	-5.15	0.36	-66.74	1.34
	SA	-5.33	0.48	-55.83	1.26
	JW	-5.07	0.34	-66.81	1.31
Dynamic, large-first HTE		Scenario 2c		Scenario 3f	
	TWFE	-76.82	2.24	-94.57	2.44
	CS	3.77	0.69	-38.85	1.11
	BJS	-5.51	0.41	-56.08	1.40
	SA	-6.09	0.51	-48.86	1.35
	JW	-5.62	0.37	-56.50	1.37

Abbreviations: TWFE: Two-way fixed effects; CS: Callaway-Sant'Anna; BJS: Borusyak-Jaravel-Spiess; SA: Sun-Abraham; JW: Wooldridge
Note: Each scenario is listed in Table 4.

Supplemental References

1. Goodman-Bacon A. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*. 2021;225(2):254-277.
2. Sun L, Abraham S. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*. 2021;225(2):175-199.
3. Callaway B, Sant'Anna PH. Problems with two-way fixed-effects event-study regressions. Accessed March 13, 2023. https://cran.r-project.org/web/packages/did/vignettes/TWFE.html
4. Callaway B, Sant'Anna PH. Difference-in-differences with multiple time periods. *Journal of Econometrics*. 2021;225(2):200-230.
5. Riddell CA, Goin DE. Guide for Comparing Estimators of Policy Change Effects on Health. *Epidemiology*. 2023;34(3):e21-e22.
6. Caetano C, Callaway B, Payne S, Rodrigues HSA. Difference in differences with time-varying covariates. *arXiv preprint arXiv:220202903*. 2022;
7. Rambachan A, Roth J. A more credible approach to parallel trends. *Review of Economic Studies*. 2023:rdad018.