Causal Panel Analysis under Parallel Trends: Lessons from A Large Replication Study

American Political Science Review

Albert Chiu Xingchen Lan Ziyi Liu Yiqing Xu

A. Online Supplementary Materials

Table of Contents

A.1. HTE-Robust Estimators

Scholars have proposed a number of novel estimators that relax TWFE assumptions and allow for HTE. We discuss several of them below. Broadly, we can categorize these estimators along two dimensions: (1) estimation strategy and (2) applicable settings. Along (1), we divide estimators into two groups. We call one group of methods "DID extensions," which use local, 2×2 DIDs between treated and control observations as building blocks, and the other "imputation methods," which impute counterfactual outcomes using an explicit outcome model (in particular, the TWFE model) that is fit globally on all available control observations. We see the former as direct extensions to DID, while the latter embed DID's functional form assumptions in their outcome models. For (2), estimators either are suited only to the staggered setting (which includes the classic DID setting) where treatment is an absorbing state or can accommodate treatment reversals. Those suited to the latter are also suited to the former, which is just a special case of the latter. The reverse is not true.

In the following subsections, which are organized by this typology, we introduce and compare several recently introduced HTE-robust estimators. Although these estimators all relax the TWFE assumption of homogeneous effects, they do not absolve us of needing the parallel trends (PT) assumption or strict exogeneity. These estimators can, however, estimate dynamic treatment effects, which in turn allow us to assess the validity of parallel trends by testing for pretrends.

A.1.1. DID Extensions for the Staggered Setting

We first introduce a set of estimators, each constructed from local 2×2 DID estimates, that are suitable only for the staggered adoption setting. The general strategy of these estimators is to estimate the dynamic cohort average treatment effect on the treated (CATT), $\tau_{g,l}$, for each cohort g and for each period since treatment adoption l using a valid 2×2 DID. By valid, we mean that the DID consists of (1) a pre-period and a post-period and (2) a treated group and a comparison group. The pre-period is such that all observations in both groups are in control, and the post-period is such that observations from the treated group are in treatment and the observations from the comparison group are in control. The choice of comparison group is what primarily distinguishes estimators in this category. To obtain higher-level averages, we then average over our estimates of $\tau_{g,l}$ using appropriate, non-negative weights.

Sun and Abraham (2021) propose an interaction-weighted (`IW`) estimator that is a weighted average of CATT estimates obtained from a TWFE regression with cohort dummies fully interacted with indicators of relative time to the treatment's onset. Specifically,

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{g \notin \mathcal{C}} \sum_{l \neq 0} \tau_{g,l} \mathbf{1}\{E_i = g\} \cdot \mathbf{1}\{K_{i,t} = l\} + \epsilon_{i,t}, \tag{A1}$$

where $\mathcal{C}$ is some set of reference cohorts and $K_{i,t}$ is similarly defined as in the main text. Equivalently, each estimate of $\tau_{g,l}$ from equation A1 can can be characterized as a difference in the average changes in outcome from some fixed pre-period $s < g$ to l periods since g between the treated cohort g and comparison cohorts in $\mathcal{C}$:

$$\hat{\tau}(g,l) = \frac{1}{|\{i : E_{i,t} = g\}|} \sum_{i:E_{i,t}=g} (Y_{i,g+l} - Y_{i,s}) - \frac{1}{|\{i : E_{i,t} \in \mathcal{C}\}|} \sum_{i:E_{i,t}\in\mathcal{C}} (Y_{i,g+l} - Y_{i,s}),$$

The authors recommend using $\mathcal{C} = \{\sup_i E_{i,t}\}$, which is either the never-treated cohort or (if none exists) the last-treated cohort. The estimator then weights $\hat{\tau}_{g,l}$ by the sample share of each cohort $\hat{w}_g$ before taking some average thereof. For example, the dynamic treatment effects (DTE) from relative period l between $-a$ and b can be estimated from

$$\hat{\tau}_l^{IW} = \sum_g \hat{w}_g \hat{\tau}_{g,l}, \quad a \leq l \leq b,$$

and the ATT up to b periods after the treatment's onset from

$$\hat{\tau}^{IW} = \frac{1}{b} \sum_{1 \leq l \leq b} \sum_g \hat{w}_g \hat{\tau}_{g,l}.$$

The authors note that their estimator can be extended to include covariates, but also that this may require additional functional form assumptions.

Using the same general strategy, Callaway and Sant'Anna (2021) propose doubly robust estimators that directly incorporate pre-treatment covariates. These estimators, which we collectively refer to as `CSDID`, use either never-treated ($\hat{\tau}_{nev}^{CS}$) or not-yet-treated units ($\hat{\tau}_{ny}^{CS}$) as the comparison group. $\hat{\tau}_{nev}^{CS}$ uses the same comparison group as `IW` when a never-treated cohort exists, whereas $\hat{\tau}_{ny}^{CS}$ differs and uses all untreated observations of later adopters (including the never-treated) as potential controls for early adopters. Besides the choice of comparison cohort, these estimators both differ from the `IW` estimator in that they allow the user to condition on pre-treatment covariates using both an explicit outcome model and

inverse propensity score weighting (IPW).[A1] If either the outcome model or the propensity score model is correct, the estimators will be consistent.

A.1.2. DID Extensions for the General Setting

The next group of estimators we discuss also use local DIDs as building blocks, but estimators in this group can accommodate treatment reversals. The general strategy is once again to use valid 2×2 DIDs, but this time the goal is to estimate the DTE τ_l for all treated units some number of periods since treatment l—cohorts are no longer defined, since treatment reversals make it insensible to group units by their time of treatment adoption. The literature has effectively proposed one common strategy of selecting a comparison group, which is to match treated and control observations belonging to units with the same treatment history.

IKW (2023) propose one such estimator. Formally, to estimate the ATT, we first define a matched set for each observation (i,t) satisfying $D_{i,t} = 1$ and $D_{i,t-1} = 0$,

$$\mathcal{M}_{i,t} = \left\{ i' : i' \neq i, D_{i',t} = 0, D_{i',t'} = D_{i,t'} \; \forall t' \in \{t-1, t-2, \ldots, t-a\} \right\},$$

where a is the number of periods on which we wish to match treatment histories. The authors also propose "refining" the matched set to incorporate other pre-treatment covariates and past outcomes. We do not further discuss refinement for a more seamless comparison with other estimators and refer interested readers to the original paper. Without refinement and fixing the number of periods a on which to match, the proposed estimator for the DTE l periods since treatment τ_l is,

$$\hat{\tau}_{l,a}^{PM} = \frac{\sum_{t=a}^{T-l+1} \sum_{i=1}^{N} G_{i,t} \hat{\tau}_l^{(i,t)}}{\sum_{t=a+1}^{T-l} \sum_{i=1}^{N} G_{i,t}},$$

where $G_{i,t} = \mathbf{1}\{|\mathcal{M}_{i,t}| > 0\} D_{i,t}(1 - D_{i,t-1})$ is equal to 1 if and only if the observation (i,t) switches into treatment at time t and has a non-empty matched set (and is zero otherwise) and $\hat{\tau}_l^{(i,t)} = (Y_{i,t-1+l} - Y_{i,t-1}) - \sum_{i' \in \mathcal{M}_{i,t}} \frac{1}{|\mathcal{M}_{i,t}|}(Y_{i',t-1+l} - Y_{i',t-1})$ is the local DID obtained from the pre- and post-periods $t-1$ and $t-1+l$, respectively, the treatment "group" consisting of just (i,t), and the comparison group consisting of the matched set for (i,t). To then obtain an estimate for the DTE $\hat{\tau}_l$, we then average over all $\hat{\tau}_l^{(i,t)}$ such that (i,t) Essentially, the strategy is to average over the estimates of the DTE for all units that switch into treatment

[A1] The IPW estimator proposed by Strezhnev (2018) is similar to $\hat{\tau}_{ny}^{CS}$. One small difference is that $\hat{\tau}_{ny}^{CS}$ allows more complex outcome modeling than a simple before-and-after estimator.

at t (if there are any) for each time period t, and then to average across all time periods for which we can obtain an estimate.[A2] If the goal is to estimate the average effect of treatment reversal (ART), we then analogously defined matched sets for each observation (i,t) satisfying $D_{i,t} = 1$ and $D_{i,t-1} = 0$, $\mathcal{M}_{i,t} = \{i' : i' \neq i, D_{i',t} = 1, D_{i',t'} = D_{i,t'} \; \forall t' \in \{t-1, t-2, \ldots, t-a\}\}$. We use $\hat{\tau}_{l,a}^{PM-ART}$ to denote the resulting estimator.

Interestingly, several DID extensions can be viewed as special cases of `PanelMatch`.

Remark A.1 (Relation between $\hat{\tau}_{1,1}^{PM}$ without refinement and $\hat{\tau}^{M}$). Assume we have a balanced panel of units, i.e. every unit i is observed at every time period t. For the special case when we match on only one period ($a = 1$) and are estimating the contemporaneous treatment effect ($l = 1$), without refinement, a weighted average of the `PanelMatch` estimators for the ATT and ART is equivalent to the multiple DID estimator proposed by de Chaisemartin and D'Haultfœuille (2020), or $\hat{\tau}^{M}$, when there exists a 'stable' group (i.e., whenever there is a unit switching into or out of treatment, there is at least one other unit staying in control or treatment; see the next section for a formal statement of this assumption), where the weights are the proportion of "switchers" that are "joiners" versus "leavers." That is, if we do not refine the matched set, then $\frac{N_J}{N_S}\hat{\tau}_{1,1}^{PM} + \frac{N_L}{N_S}\hat{\tau}_{1,1}^{PM-ART} = \hat{\tau}^{M}$, where N_J, N_L, and N_S are the numbers of joiners, leavers, and switchers. The proof is in the next section. This observation allows us to appeal to the results that de Chaisemartin and D'Haultfœuille (2020) prove about $\hat{\tau}^{M}$. Minor adjustments of their proofs will give us that, under some typical assumptions (the details of which we provide later in this section), $\hat{\tau}_{1,1}^{PM}$ without refinement is asymptotically normal, unbiased, consistent for the average contemporaneous treatment (reversal) effect on the treated.

Remark A.2 (Equivalence of `PanelMatch` and `CSDID` without covariate adjustment). Again assume we have a balanced panel of units. If we use a simple difference in means as the outcome model for `CSDID` and employ uniform propensity score weights (i.e., do not adjust for covariates), then `CSDID` is equivalent to `PanelMatch` with an arbitrary number of lags and without refinement (in the staggered setting). This follows from the facts that in the staggered setting, for any time period t: (1) Any observation belonging to a unit that switches into treatment at time t ('switchers') must have been under control for the periods

[A2] Note that, without refinement, all treated observations with the same treatment history share the same matched set, so we can group these observations together and rewrite the inner sum to instead be over all possible treatment histories. We can thus also express the inner sum of the numerator as a weighted sum of local DIDs using a slightly different treatment group—all treated observations with the treatment history—where the weights are proportional to the size of said group.

$1, \ldots t-1$; and (2) all control observations must belong to units that have been under control for the periods $1, \ldots t-1$ (i.e., they have the same treatment history as switchers). Thus, the matched set will always include all units under control (all "not-yet treated" units).

A.1.3. Imputation Methods for the General Setting

The last class of estimators we discuss no longer *directly* take the difference between differences; instead, they take the difference of the observed outcome and an imputed counterfactual outcome (for treated observations)—the before-and-after difference is embedded in the functional form assumption used to impute treated counterfactuals. Under strict exogeneity or a stronger version of the PT, the imputation method allows researchers to make inferences about the ITE of treated observations, $\tau_{i,t}, \forall (i,t)\ s.t.\ D_{i,t} = 1$, the most fine-grained estimand (e.g., Bai and Ng, 2021).

BJS (2024) propose an "imputation procedure" that first imputes the counterfactual outcomes for treated units based on the outcome model,

$$Y_{i,t} = A'_{i,t}\lambda_i + X'_{i,t}\beta + D_{i,t}\Gamma'_{i,t}\theta + \epsilon_{i,t},$$

and then estimates the treatment effect for treated observations with the difference between their observed and their imputed counterfactual outcomes. That is, first use only the untreated observations $\{(i,t) : D_{i,t} = 0\}$ to estimate λ_i and β (by $\hat{\lambda}_i$ and $\hat{\beta}$) using OLS on the regression $Y_{i,t} = A'_{i,t}\lambda_i + X'_{i,t}\beta + \varepsilon_{i,t}$. Then, for each treated observation, set $\hat{Y}^{BJS}_{i,t}(0) = A'_{i,t}\hat{\lambda}_i + X'_{i,t}\hat{\beta}$ and estimate the ITE as $\hat{\tau}^{BJS}_{i,t} = Y_{i,t} - \hat{Y}^{BJS}_{i,t}(0)$. We can then combine these ITE estimates to estimate aggregate quantities, including the ATT and dynamic effects.

LWX (2024) refer to imputation-based estimators as "counterfactual estimators" and discuss several such estimators. LWX (2024) consider a class of outcome models of the form $Y_{i,t}(0) = f(X_{i,t}) + h(U_{i,t}) + \varepsilon_{i,t}$, where $f(\cdot)$ and $h(\cdot)$ are known parametric functions, $X_{i,t}$ is observed, and $U_{i,t}$ is unobserved (whereas in BJS (2024), both X_{it} and A_{it} are observed). Note that this framework subsumes the TWFE outcome model as we can model $Y_{i,t}(0) = X'_{i,t}\beta + \alpha_i + \xi_t + \varepsilon_{i,t}$. We can then use an estimation procedure similar to the one in BJS (2024). LWX (2024) call this estimator the fixed effect counterfactual (`FEct`) estimator, $\hat{\tau}^{fect}$ (for the ATT) or $\hat{\tau}^{fect}_l$ (for the dynamic effects).

The two imputation methods, BJS (2024) and LWX (2024), produce the same post-treatment estimated DTE in event-study plots but differ in their estimation of pre-treatment DTE. The former employs a dynamic TWFE specification using only untreated observations,

with period indicators for pre-treatment periods:

$$Y_{i,t} = \alpha_i + \xi_t + \sum_{l=-\underline{T}+1}^{l=0} \tau_l^{TWFE} \cdot \mathbf{1}\{K_{i,t} = l\} + \varepsilon_{i,t} \quad \text{for } (i,t) \text{ with } D_{it} = 0$$

where $\underline{T}$ represents the earliest pre-treatment period and is treated as the reference period. As shown in Roth (2024), this asymmetry in the calculation of pre- and post-treatment DTE can lead to a discontinuity at the first post-treatment period.

LWX (2024), on the other hand, estimates the pre-treatment DTE by comparing observed and predicted pre-treatment outcomes. Li and Strezhnev (2024) demonstrates that because LWX (2024) fits $\hat{Y}_{it}(0)$ using observations where $D_{it} = 0$ and calculates the DTE as $Y_{it} - \hat{Y}_{it}(0)$, this procedure relies on in-sample errors, which can result in DTE estimates being biased toward zero. To address this issue, they recommend using the leave-one-out method: holding out observations from each pre-treatment period, using observations from other pre-treatment periods to impute the counterfactual for the held-out observations, and basing DTE estimation on out-of-sample errors.

Although both methods estimate pre-treatment DTE asymmetrically compared to post-treatment DTE, we prefer the latter method for two reasons. First, both BJS (2024) and LWX (2024) estimate post-treatment DTE using the average of all pre-treatment periods as the reference. However, BJS (2024) uses the earliest pre-treatment period as the reference when estimating pre-treatment DTE, which can lead to inconsistency in the interpretation of results. Second, we apply the leave-one-out procedure across most replicated samples, and the results closely align with those obtained using in-sample errors, indicating robustness of the latter method.

A.1.4. Strict Exogeneity and Parallel Trends

In this subsection, we clarify the relationship between strict exogeneity, often invoked in TWFE models, and the identifying assumptions in DID research designs, including no anticipation, no carryover, and PT assumptions.

Assumption A1 (Functional form).

$$Y_{i,t} = \tau D_{i,t} + X_{i,t}{}'\beta + \alpha_i + \xi_t + \varepsilon_{i,t}, \quad \forall i, t,$$

in which $Y_{i,t}$ is the outcome variable for unit i at time t; $D_{i,t}$ is the treatment variable; $X_{i,t}$ is a vector of covariates; α_i, ξ_t are unit fixed effects and time fixed effects; and $\varepsilon_{i,t}$ are idiosyncratic errors.

Denote $\mathbf{D_i} = \{D_{i,1}, D_{i,2}, \cdots, D_{i,T}\}$ and $\mathbf{X_i} = \{X_{i,1}, X_{i,2}, \cdots, X_{i,T}\}$. Given Assumption A1, we can simplify the potential outcome to depend solely on the treatment status in period t without loss of generality:

$$Y_{i,t}(\mathbf{d_i}) = Y_{i,t}(d_{i,t}) = \begin{cases} Y_{i,t}(1), & \text{if } d_{i,t} = 1 \\ Y_{i,t}(0), & \text{if } d_{i,t} = 0 \end{cases}.$$

Note that, $Y_{i,t}(0) = X_{i,t}{}'\beta + \alpha_i + \xi_t + \varepsilon_{i,t}$ and $Y_{i,t}(1) = Y_{i,t}(0) + \tau$. This formulation of the potential outcomes directly implies **no anticipation** and **no carryover**, as the potential outcome in period t is unaffected by treatment status in other periods.

Assumption A2 (Strict exogeneity).

$$\mathbb{E}[\varepsilon_{i,t} \mid \mathbf{D_i}, \mathbf{X_i}, \alpha_i, \xi_t] = \mathbb{E}[\varepsilon_{i,t} \mid D_{i,t}, X_{i,t}, \alpha_i, \xi_t] = 0, \quad \forall i, t.$$

adapted from Wooldridge (2010, p. 253).

Note that Assumption A2 invokes the functional form assumption. Given Assumptions A1 and A2, we have:

$$\mathbb{E}[Y_{i,t} \mid \mathbf{D_i}, \mathbf{X_i}, \alpha_i, \xi_t] = \mathbb{E}[Y_{i,t} \mid D_{i,t}, X_{i,t}, \alpha_i, \xi_t] \quad \forall i, t,$$

which means that once $D_{i,t}$, covariates, and fixed effects are accounted for, $D_{i,s}$ has no partial effect on $Y_{i,t}$, for any $s \neq t$; hence it is termed *strict*.

Strict exogeneity forbids **feedback**, defined as past outcomes and covariates influencing current treatment assignment (Imai and Kim, 2019). To see this, if feedback from $Y_{i,t-1}$ to $D_{i,t}$ existed, we would have:

$$\begin{aligned}
&\mathbb{E}[Y_{i,t-1} \mid \mathbf{D_i}, \mathbf{X_i}, \alpha_i, \xi_t] \\
&= \mathbb{E}[Y_{i,t-1} \mid D_{i,1}, \ldots, D_{i,t-1}, \underline{D_{i,t}}, \ldots, D_{i,T}, \mathbf{X_i}, \alpha_i, \xi_t] \\
&\neq \mathbb{E}[Y_{i,t-1} \mid D_{i,t-1}, X_{i,t-1}, \alpha_i, \xi_t].
\end{aligned}$$

Imai and Kim (2019) provide a detailed discussion.

Next, we show that Assumptions A1 and A2 together imply PT, defined below:

Assumption A3 (Parallel trends). For any $i, j, s \neq t$,

$$\begin{aligned}
&\mathbb{E}[Y_{i,t}(0) - Y_{i,s}(0) \mid D_{i,t} = 1, D_{i,s} = 0, X_{i,t} = x_1, X_{i,s} = x_2] \\
&= \mathbb{E}[Y_{j,t}(0) - Y_{i,s}(0) \mid D_{j,t} = 0, D_{j,s} = 0, X_{j,t} = x_1, X_{j,s} = x_2].
\end{aligned}$$

As in a canonical DID setting, the PT assumption states that between any two groups of units, once the change in covariates is controlled for, the change in potential outcomes between any two periods is mean independent of the change in observed treatment status.

Given the functional form assumption, it is sufficient to show

$$\begin{aligned}
&\mathbb{E}[\varepsilon_{i,t} - \varepsilon_{i,s} \mid D_{i,t} = 1, D_{i,s} = 0, X_{i,t} = x_1, X_{i,s} = x_2] \\
&= \mathbb{E}[\varepsilon_{j,t} - \varepsilon_{j,s} \mid D_{j,t} = 0, D_{j,s} = 0, X_{j,t} = x_1, X_{j,s} = x_2].
\end{aligned}$$

which is implied by strict exogeneity (with an application of the tower rule to integrate out α_i and ξ_t).

It is worth noting that this version of the PT assumption depends on the parametric model (Assumption A1), while the PT assumption invoked by many HTE-robust estimators is weaker. However, from a practical standpoint, they have similar empirical implications.

A.1.5. Assumptions for Each Estimator

We first dicuss and compare the key identification assumptions required by each method.

Sun and Abraham (2021) define potential outcomes based on treatment history and assume parallel trends for the never-treated potential outcome $Y_{i,t}(\infty)$ of the comparison group: $\mathbb{E}[Y_{i,t}(\infty) - Y_{i,s}(\infty)|E_{i,t} = e]$ is the same for all $s \neq t$ and $e \in \operatorname{supp} E_{i,t}$.[A3] Call this assumption "parallel trends A." Callaway and Sant'Anna (2021) similarly assume parallel trends for the comparison group, but define potential outcomes based on current treatment status. As a result, the statement of the assumption becomes, for all g, $\mathbb{E}[Y_{i,t}(0) - Y_{i,t-1}(0)|E_i = g] = \mathbb{E}[Y_{i,t}(0) - Y_{i,t-1}(0)|E_{i,t} \in \mathcal{C}]$ for each $t \geq \max\{2, g\}$. Call this version "parallel trends B."

de Chaisemartin and D'Haultfœuille (2020) assume both "strong exogeneity" and "common trends." They define the former as, $\mathbb{E}[Y_{i,t}(d) - Y_{i,t-1}(d)|\{D_{i,t}\}_{t=1}^{T}] = \mathbb{E}[Y_{i,t}(d) - Y_{i,t-1}(d)]$ for all i, all $t \geq 2$, and all $d \in \{0, 1\}$.[A4] The common trends assumption requires that this last quantity — that is, $\mathbb{E}[Y_{i,t}(d) - Y_{i,t-1}(d)]$ — does not vary across i for all $t \geq 2$ and $d \in \{0, 1\}$. Combining these two assumptions, we can instead write that $\mathbb{E}[Y_{i,t}(d) - Y_{i,t-1}(d)|\{D_{i,t}\}_{t=1}^{T}] = \mathbb{E}[Y_{j,t}(d) - Y_{j,t-1}(d)]$ for all j (including $j = i$), all i, all $t \geq 2$, and all $d \in \{0, 1\}$. Call this combined version of the assumptions "parallel trends C." Like Sun and Abraham (2021), IKW (2023) define potential outcomes in terms of treatment histories. IKW (2023) do not, however, assume staggered adoption, and so a much wider range of treatment histories are possible. The comparison group is also substantially different. The latter compares units that switch into treatment with those that stay in control and asks that their respective trends be parallel: $\mathbb{E}[Y_{i,t+l}(D_{i,t} = 0, D_{i,t-1} = 0, \{D_{i,t-s}\}_{s=2}^{a})|D_{i,t} = 1, D_{i,t-1} = 0] = \mathbb{E}[Y_{i,t+l}(D_{i,t} = 0, D_{i,t-1} = 0, \{D_{i,t-s}\}_{s=2}^{a})|D_{i,t} = 0, D_{i,t-1} = 0]$. Call this assumption "parallel trends D."

Recall that the imputation estimators connect to DID in a less direct way, which in turn implies different assumptions: They assume a TWFE model for untreated potential outcomes, which requires mean independence for all pairs of units i, j and all pairs of time periods t, s. For example, BJS (2024) define a version of parallel trends as $\mathbb{E}[Y_{i,t}(0) - Y_{i,s}(0)] = \mathbb{E}[Y_{j,t}(0) - Y_{j,s}(0)]$ for all i, j and all t, s. The estimator from BJS (2024) does not require this to hold, instead requiring a weaker assumption, $\mathbb{E}[Y_{i,t}(0)] = A'_{i,t}\lambda_i + X'_{i,t}\beta$ for all i, t. Note that this assumption implies that each idiosyncratic error is zero in expectation, and

[A3] We state the unconditional versions of these assumptions for simplicity.

[A4] The actual assumption is that this equality holds for all groups g, where g is the level of the fixed effects, which may be at a higher level than the unit level (e.g., if i indexes cities, g might be counties or states/provinces). For consistency and simplicity, we assume that this is equal to the unit level in our discussion (i.e., unit fixed effects).

thus we refer to this assumption as "outcome model and mean-zero errors." $\hat{\tau}^{fect}$ from LWX (2024) requires strict exogeneity and the TWFE outcome model, which together imply the PT defined by BJS (2024).

Next, we provide a fuller account of all assumptions invoked by each method.

Sun and Abraham (2021)

- Parallel trends A; and
- No anticipation for the comparison group: $\mathbb{E}[Y^e_{i,e-l} - Y^\infty_{i,e-l} | E_{i,t} = e] = 0$ for all $l > 0$ and $e \in \mathcal{C}$.

Under the above assumptions, the IW estimator is unbiased and consistent.

Callaway and Sant'Anna (2021)

- Random sampling: $\{Y_{i,g,t}, X_i, D_{i,g,t}\}_{i=1}^N : 1 \leq t \leq T$ is iid;
- Limited anticipation up to a known number of periods s: $\mathbb{E}[Y_{i,g,t}(0) - Y_{i,g,t-1}(0) | X, E_i = g] = \mathbb{E}[Y_{i,g,t}(0) - Y_{i,g,t}(0) | X, C = 0]$ for each $t \geq g - s$;
- Overlap: For each $t \geq 2$ and g, there exists $\epsilon > 0$ such that $\mathbb{P}(G_g = 1)\ p_{g,t}(X) < 1 - \epsilon$ almost surely; and
- Parallel trends B.

Under the above assumptions, $\hat{\tau}^{CS}_{nev}$ and $\hat{\tau}^{CS}_{ny}$ are point-identified when the comparison groups are the never-treated or not-yet-treated cohorts, respectively. Additionally, when there are covariates X, the estimators are consistent and asymptotically normal if we also assume the following (dropping the i subscript):

- For all $g = 2, \ldots, T$, (i) there exists a known function $\Lambda : \mathbb{R} \to [0, 1]$ such that $p_g(X) := \mathbb{P}(G_g = 1 | X, G_g + C = 1) = \Lambda(X'\pi^0_g)$, where C is an indicator variable for whether a unit belongs to the comparison group; (ii) $\pi^0_g \text{int}(\Pi)$, where Π is a compact subset of $\mathbb{R}^k$; (iii) $\text{supp}(X) \subseteq S$ for some compact S, and $\mathbb{E}[XX' | G_g + C = 1] \succ 0$; (iv) for $\mathcal{U} = \{x'\pi : x \in \text{supp}(X), \pi \in \Pi\}$, for all $u \in \mathcal{U}$, there exists $\epsilon > 0$ such that $\Lambda(u) \in [\epsilon, 1 - \epsilon]$, $\Lambda(u)$ is strictly increasing and twice continuously differentiable with first derivatives bounded away from zero and infinity and bound second derivatives; (vi) $\mathbb{E}[Y^t_t] < \infty$ for all $t = 1, \ldots, T$.

IKW (2023) The authors discuss several assumptions, including

- Balanced panel;
- No spillover (temporally, or across units);
- Limited carryover; and
- (Conditional) parallel trends $\mathbb{E}[Y_{t+F}(D_t = 0, D_{t-1} = 0) - Y_{t-1}|D_t = 1, D_{t-1} = 0, Z_t] = \mathbb{E}[Y_{t+F}(D_t = 0, D_{t-1} = 0) - Y_{t-1}|D_t = 0, D_{t-1} = 0, Z_t]$ where $Z_t = \left(\{D_{t-l}, Y_{t-l}\}_{l=2}^{L}, \{X_{t-l}\}_{l=0}^{L}\right)$

de Chaisemartin and D'Haultfœuille (2020) Note that in the original paper, de Chaisemartin and D'Haultfœuille (2020) define their estimator in terms of a group level (the level of the fixed effects) that need not be equal to the unit level. For simplicity and ease of comparison, we state their assumptions for the case where the group level is the same as the unit level (i.e., unit fixed effects). $\hat{\tau}^M$ is unbiased, consistent, and asymptotically normal under the following assumptions:

- Balanced panel;
- Independent groups, i.e. $(Y_{i,t}(0), Y_{i,t}(1), D_{i,t})_{1 \leq t \leq T}$ are mutually independent;
- Strong exogeneity;
- Common trends; and
- The existence of stable groups, i.e. whenever there exists a "joiner" $(i,t) : D_{i,t} = 1, D_{i,t-1} = 0$ or a "leaver" $(i,t) : D_{i,t} = 0, D_{i,t-1} = 1$, then there also exists a unit staying in control $(i',t) : D_{i',t} = D_{i',t-1} = 0$ or treatment $(i',t) : D_{i',t} = D_{i',t-1} = 1$, respectively.

BJS (2024) The imputation estimator is unbiased under the following assumptions:

- General model for $Y(0)$ (which subsumes the TWFE model) and zero mean error, for all (i,t), $Y_{i,t}(0) = A'_{it}\lambda_i + X'_{it}\beta + \epsilon_{i,t}$, where $\mathbb{E}[\epsilon_{i,t}] = 0$;
- No anticipation, $Y_{i,t} = Y_{i,t}(0)$ for all (i,t) such that $D_{i,t} = 0$;
- Null model for causal effects (i.e., no restrictions on the ITEs), $(\tau_{i,t})_{(i,t):D_{i,t}=1}$ is some unknown vector of length N_1, where N_1 is the number of treated observations.

Furthermore, if errors are homoskedastic and mutually uncorrelated, $\mathbb{E}[\epsilon\epsilon'] = \sigma^2 I_N$, then the imputation error is efficient. Two additional assumptions ensure that the estimator is consistent:

- Clustered standard errors, $\epsilon_{i,t}$ are uncorrelated accross units and have bounded variance, $Cov(\epsilon_{i,t}, \epsilon_{j,s}) = 0$ for all $i \neq j$, and $Var(\epsilon_{i,t}) < \bar{\sigma}^2$ for some finite $\bar{\sigma}^2$; and
- Herfindahl condition, $\|v\|_H^2 := \sum_i (\sum_t |v_{i,t}|)^2 \to 0$, where $v_{i,t}$ are weights such that $\hat{\tau} = \sum_{i,t} v_{i,t} Y_{i,t}$.

Lastly, asymptotic normality is guaranteed by the following:

- Higher moments of weights, there exists $\delta > 0$ such that $\mathbb{E}[|\epsilon_{i,t}|^{2+\delta}$ is uniformly bounded and $\sum_i \left(\frac{\sum_t |v_{i,t}|}{\|v\|_H} \right)^{2+\delta}$; and
- $\liminf n_H \sigma^2 > 0$, where $n_H = \|v\|_H^{-2}$ and $\sigma^2 = Var(\hat{\tau})$.

LWX (2024) Under the following two assumptions along with some regularity conditions, `FEct` is unbiased and consistent:

- Functional form, $Y_{i,t}(0) = X_{i,t}'\beta + \alpha_i + \xi_t + \varepsilon_{i,t}$; and
- Strict exogeneity, $\varepsilon_{i,t} \perp\!\!\!\perp \{D_{j,s}, X_{j,s}, \alpha_j, \xi_s\}$ for all $i, j = 1, \ldots, N$ and all $s, t = 1, \ldots, T$.

A.1.6. Proof of Remark A.1

First, we note that de Chaisemartin and D'Haultfœuille (2020) define $\hat{\tau}^M$ to allow for 'group' level fixed effects that may be higher up than the unit level. Let $N_{g,t}$ denote the number of observations in group g at time t. We assume a "sharp design," meaning all units in the same cell (g,t) have the same treatment. Let $N_{d,d',t} = \sum_{g:\substack{D_{g,t}=d,\\ D_{g,t-1}=d'}} N_{g,t}$ denote the number of observations with treatment status d in period t and status d' in period $t-1$. Let $Y_{.,g,t} = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} Y_{i,g,t}$ denote the average outcome (across observations) in group g at time t. Define the following quantities:

$$DID_{+,t} = \sum_{g:D_{g,t}=1,D_{g,t-1}=0} \frac{N_{g,t}}{N_{1,0,t}} \big(Y_{.,g,t} - Y_{.,g,t-1}\big) - \sum_{g:D_{g,t}=D_{g,t-1}=0} \frac{N_{g,t}}{N_{0,0,t}} \big(Y_{.,g,t} - Y_{.,g,t-1}\big) \text{ and}$$

$$DID_{-,t} = \sum_{g:D_{g,t}=D_{g,t-1}=1} \frac{N_{g,t}}{N_{1,1,t}} \big(Y_{.,g,t} - Y_{.,g,t-1}\big) - \sum_{g:D_{g,t}=0,D_{g,t-1}=1} \frac{N_{g,t}}{N_{0,1,t}} \big(Y_{.,g,t} - Y_{.,g,t-1}\big),$$

letting $DID_{+,t} = 0$ whenever $\min\{N_{1,0,t}, N_{0,0,t}\} = 0$ and $DID_{-,t} = 0$ whenever $\min\{N_{1,1,t}, N_{0,1,t}\} = 0$. Finally, define

$$\hat{\tau}^M = \sum_{t=2}^{T} \left(\frac{N_{1,0,t}}{N_S} DID_{+,t} + \frac{N_{0,1,t}}{N_S} DID_{-,t} \right),$$

where $N_S := |(g,t) : t \geq 2, D_{g,t} \neq D_{g,t-1}|$ is the number of switchers.

Now, we consider the case where the group level is the same as the unit level. Note that then $N_{g,t} = 1$ always.

We can now write

$$DID_{+,t} = \sum_{i:D_{i,t}=1,D_{i,t-1}=0} \frac{1}{N_{1,0,t}} \big(Y_{i,t} - Y_{i,t-1}\big) - \sum_{i:D_{i,t}=D_{i,t-1}=0} \frac{1}{N_{0,0,t}} \big(Y_{i,t} - Y_{i,t-1}\big)$$

and similarly

$$DID_{-,t} = \sum_{i:D_{i,t}=D_{i,t-1}=1} \frac{1}{N_{1,1,t}} \big(Y_{i,t} - Y_{i,t-1}\big) - \sum_{i:D_{i,t}=0,D_{i,t-1}=1} \frac{1}{N_{0,1,t}} \big(Y_{i,t} - Y_{i,t-1}\big)$$

Now consider $\hat{\tau}^{PM}$ with the choice of $l = 1$, which estimates the contemporaneous treat-

ment effect at the moment of joining treatment,

$$\hat{\tau}_{1,a}^{PM} = \frac{\sum_{i=1}^{N}\sum_{t=a+1}^{T}\mathbf{1}\left\{|\mathcal{M}_{it}|>0\right\}D_{i,t}(1-D_{i,t-1})\left((Y_{i,t}-Y_{i,t-1})-\sum_{i'\in\mathcal{M}_{i,t}}\frac{1}{|\mathcal{M}_{i,t}|}(Y_{i',t}-Y_{i',t-1})\right)}{\sum_{i=1}^{N}\sum_{t=a+1}^{T}\mathbf{1}\left\{|\mathcal{M}_{it}|>0\right\}D_{i,t}(1-D_{i,t-1})}.$$

Now further restrict lags used for matching to $a=1$. Then the matched set $\mathcal{M}_{i,t} = \left\{i' : i'\neq i, D_{i,'t}=0, D_{i,'t-1}=D_{i,t-1}\right\}$ is just units that have the same treatment status in the previous period and are in control in the current period. Under the assumption that a stable group exists, the matched set must be nonempty for any "joiner" $((i,t): J_{i,t}=1)$, where $J_{i,t}=\mathbf{1}\left\{D_{i,t}=1\right\}\mathbf{1}\left\{D_{i,t-1}=0\right\}$, and so $\mathbf{1}\left\{|\mathcal{M}_{it}|>0\right\}D_{i,t}(1-{}_{i,t})=J_{i,t}$. Let $N_J := |(g,t): t\geq 2, J_{i,t}=1|$ be the number of joiners.

Now we have,

$$\begin{aligned}
\hat{\tau}_{1,1}^{PM} =& \frac{\sum_{t\geq 2}\sum_{i:J_{i,t}=1}\left((Y_{i,t}-Y_{i,t-1})-\frac{1}{|\mathcal{M}_{i,t}|}\sum_{i'\in\mathcal{M}_{i,t}}(Y_{i',t}-Y_{i',t-1})\right)}{N_J}\\
=& \frac{1}{N_J}\sum_{t\geq 2}\sum_{i:J_{i,t}=1}\left((Y_{i,t}-Y_{i,t-1})-\frac{1}{|\mathcal{M}_{i,t}|}\sum_{i':D_{i',t}=D_{i',t-1}=0}(Y_{i',t}-Y_{i',t-1})\right)\\
=& \frac{1}{N_J}\sum_{t\geq 2}\sum_{i:J_{i,t}=1}\left((Y_{i,t}-Y_{i,t-1})-\frac{1}{N_{0,0,t}}\sum_{i':D_{i',t}=D_{i',t-1}=0}(Y_{i',t}-Y_{i',t-1})\right)\\
=& \frac{1}{N_J}\sum_{t\geq 2}\left(\sum_{i:J_{i,t}=1}(Y_{i,t}-Y_{i,t-1})-\frac{N_{1,0,t}}{N_{0,0,t}}\sum_{i':D_{i',t}=D_{i',t-1}=0}(Y_{i',t}-Y_{i',t-1})\right)\\
=& \frac{N_{1,0,t}}{N_S}DID_{+,t}.
\end{aligned}$$

We can alter the definition of the matched set to target "leavers" $\{(i,t): D_{i,t}=0, D_{i,t-1}=1\}$ to get an estimate for the contemporaneous effect of leaving $\hat{\tau}_{1,1}^{PM-ART}$ and similarly show that $\hat{\tau}_{1,1}^{PM-ART} = \frac{N_{0,1,t}}{N_L}DID_{-,t}$, where N_L is the number of leavers. Observe that $\hat{\tau}^M = \frac{N_J}{N_S}\hat{\tau}_{1,1}^{PM} + \frac{N_L}{N_S}\hat{\tau}_{1,1}^{PM-ART}$.

A.2. Implementation Details

This section elaborates on our reanalysis procedures, which are documented in the replication Markdown files. For each paper, the reanalysis process consists of five components: (1) a fundamental summary and visualization, (2) point estimates, (3) dynamic treatment effects, (4) diagnostic tests, and (5) sensitivity analyses.

A.2.1. Summary of Context and Visualization

Summary Table. We meticulously document various aspects of each paper, including the outcome variable, treatment variable, unit and time indicators, covariates, treatment patterns, and the fixed effects used.

Researchers typically motivate the TWFE model in one of two ways: by describing it as a form of "difference-in-differences" (DID), or by framing it as a method that exploits "within"-unit (or within-group) variation through fixed effects.

We categorize the treatment pattern into three types:

1. **Classic (including 2×2 and Block):** All treated units receive the treatment simultaneously, resembling a conventional DID design.

2. **Staggered:** Different units adopt the treatment at different time points, with no treatment reversals.

3. **General:** The treatment can have reversals.

Visualizing Treatment Status: We use the `panelView` package (Mou, Liu and Xu, 2023) to visualize each unit's treatment status over time. Treated observations are shown in a deep blue, while control observations appear in a lighter shade of blue. Units are reordered based on the timing of their initial exposure to treatment.

Visualizing the Outcome: Using `panelView`, we depict the outcome variable's trajectory for each unit within the study's time window. Control units are displayed in gray, while treated units are shown in blue. For studies involving staggered adoption, we also plot the average outcome trajectory by cohort.

A.2.2. Point Estimates

Original (Reported) Results: We employ the `fixest` package (Berge, Krantz and McDermott, 2023) to run a fixed-effects regression that includes the treatment indicator, covariates, and fixed effects as specified in the original paper. We present the raw regression output in the Markdown files.

Replicated Results: The replicated estimates match the originally reported estimates except for those in Hall and Yoder (2022) and Sanford (2023). Because the datasets in Hall and Yoder (2022) and Sanford (2023) are very large, we base our analysis on a 1% subsample for Hall and Yoder (2022) and a 0.2% subsample for Sanford (2023). Consequently, our estimates deviate slightly from the reported ones. We provide both unit-level clustered standard errors and standard errors from a 200-round clustered bootstrap, along with their corresponding confidence intervals.

Goodman-Bacon Decomposition: For analyses with a staggered treatment pattern and no additional fixed effects beyond unit and time fixed effects, we use the Goodman-Bacon decomposition (Goodman-Bacon, 2021). This approach decomposes the replicated estimate into a weighted average of all possible 2×2 DID estimates across different cohorts. Because the `bacondecomp` package is designed for balanced panels, the weighted DID estimates may not perfectly align with the replicated estimates. Nevertheless, the decomposition provides a valuable diagnostic for evaluating the possible influence of "invalid" comparisons.

TWFE: While retaining the same regression specification as the replicated estimates, we modify the sample by excluding always-treated units.

`FEct`: We use the `fect` package to implement the imputation methods. When the original specifications include fixed effects at levels higher than the unit or incorporate unit-specific trends, we apply the "cfe" method in `fect`. We estimate uncertainty through a 200-round cluster bootstrap. The software automatically excludes all always-treated observations.

Other HTE-Robust Estimators: For studies that do not include additional fixed effects beyond unit and time, we also implement `PanelMatch` (Imai, Kim and Wang, 2023). For analyses with a staggered treatment pattern, we implement the stacked DID estimator (Cengiz et al., 2019), the `IW` estimator (Sun and Abraham, 2021), the `CSDID` estimator (Callaway and

Sant'Anna, 2021), and the `DID_multiple` estimator (De Chaisemartin and d'Haultfoeuille, 2024). In most cases, we present both unit-level clustered standard errors and, if computing time is manageable, standard errors obtained from a 200-round clustered bootstrap, along with confidence intervals. All always-treated units are automatically dropped.

`StackedDID`: Following Bleiberg (2021) and Cengiz et al. (2019), we construct a cohort-specific dataset for each cohort of ever-treated units, including all never-treated units. We stack these datasets and estimate an overall effect via a fixed-effects regression that incorporates the treatment indicator, covariates, and stack-unit and stack-year interaction fixed effects.

`IW`: We use the `sunab()` command in the `fixest` package to implement the `IW` estimator, setting the `att` option to `TRUE` to retrieve the total average treatment effect.

`CSDID`: We employ the `did` package to implement the `CSDID` estimator (Sant'Anna and Callaway, 2021). Specifically, we set `est_method = "reg"` to use only the outcome model, rather than the double-robust model, when estimating the ATT. To compare point estimates under different control groups, we set the `control_group` option to both `"notyettreated"` and `"nevertreated"`.

`DID_multiple`: Using the `DIDmultiplegtDYN` package, we implement the `DID_multiple` estimator (De Chaisemartin and d'Haultfoeuille, 2024). By setting `effects` (the number of event-study effects to be estimated) to its maximum, we compute the average treatment effects for all post-treatment periods.

`PanelMatch`: For analyses that do not incorporate additional fixed effects beyond unit and time, we employ the `PanelMatch` package. We set the `lag` option to the maximum value that avoids errors, ensuring that units are also matched based on missingness patterns by setting `match.missing = TRUE`. We specify `covs.formula = NULL` and `refinement.method = "none"` to guarantee equal weighting of control units within each matched set. The confidence interval is derived from the built-in bootstrap method.

Balanced `FEct`: By specifying `balance.period` in `fect`, we estimate the ATT for a subset of units that have certain non-missing pre-treatment and post-treatment periods. We keep the lag and lead parameters consistent with the `PanelMatch` command.

A.2.3. Dynamic Treatment Effects & Event Study Plots

When estimating dynamic treatment effects (DTEs), we label the last pre-treatment period as relative period 0 and the first post-treatment period as relative period 1. The indexing rule for cases with treatment reversals follows Liu, Wang and Xu (2024). To implement this, we may use the `get.cohort()` command from the `fect` package. We then employ `esplot()` from the same package to visualize the DTEs. All always-treated units are excluded from these estimations.

TWFE (No Reversals): When the treatment has no reversals, we include interaction terms between a dummy denoting whether a unit is treated and each lead or lag indicator relative to the treatment. We then estimate a fixed-effects regression that incorporates the same fixed effects as in the original specification. We set the last pre-treatment period as the reference period and obtain a confidence interval using both clustered standard errors and a 200-round clustered bootstrap.

TWFE (With Reversals): When the treatment has reversals, we first determine each unit's relative periods to the treatment using `get.cohort()`. We then create a binary "treat" indicator as follows: (1) For never-treated units, we set "treat" to 0. (2) For ever-treated units that revert to untreated status, we set "treat" to 1 for all observations before the unit's final treatment exit. For instance, if the treatment path is $0, 0, 0, 1, 1, 0, 0$, then "treat" is set to 1 for the first five observations. (3) If a unit is already treated in the initial period, we exclude observations prior to its first treatment exit because relative periods are not clearly defined. We then interact "treat" with each lead and lag indicator and run a fixed-effects regression. As before, we set the last pre-treatment period as the reference period and obtain confidence intervals via clustered standard errors and a 200-round clustered bootstrap.

`FEct:` The procedure here parallels the steps described in the point estimates section.

For analyses with a staggered treatment pattern, we also apply the previously mentioned HTE-robust estimators to estimate the DTEs.

`stackedDID:` We construct and stack cohort-specific datasets the same way we do when estimating point effects. The difference is that in the regression, we interact a dummy for whether a unit is ever-treated with each lead and lag indicator, and we include stack-unit and stack-year interaction fixed effects. We set the last pre-treatment period as the reference

period and compute confidence intervals using clustered standard errors and a 200-round clustered bootstrap.

`IW:` We use the same `sunab()` command described in the point estimates section, except we set `att = FALSE` to aggregate treatment effects by relative period without binning. The reference period remains the last pre-treatment period.

`CSDID:` We use a similar command to that described in the point estimates section, but specify `type = "dynamic"` to aggregate treatment effects by relative period. We also set `cband = FALSE` to obtain period-wise confidence intervals and designate `base_period = "universal"` so that the last pre-treatment period is used as the base period.

`DID_multiple:` We use a similar command to the one described previously, but set `placebo` to its maximum feasible value, enabling a comparison of pre-treatment trends between treated units and their controls.

`PanelMatch:` We employ the same approach as in the point estimates section, except we specify `placebo.test = TRUE` to obtain pseudo-treatment effects for pre-treatment periods.

Balanced `FEct`: We use the same procedure as in the point estimates section.

A.2.4. Diagnostic Tests

For studies with more than three pre-treatment periods, we use the F-test and a placebo test to evaluate the parallel-trend (PT) assumption. For studies with treatment reversals and more than three post-exit periods, we also perform a carryover-effects test. All these tests rely on estimates obtained from `FEct`, and further details appear in Liu, Wang and Xu (2024). We report the corresponding p-values in a test-results table.

F-Test: We use an F-test to detect the presence of a pretrend. We define "residuals" as the differences between $Y(0)$ and $\hat{Y}(0)$. The null hypothesis posits that the mean of these residuals in each pre-treatment period is (jointly) zero. This test is conducted over those pre-treatment periods for which the number of ever-treated units exceeds 30% of the total treated units. A small p-value, causing rejection of the F-test, suggests a potential failure of the PT assumption.

Placebo Test: Using the `fect` package's placebo feature, we set `placebo.period = c(-2,0)` and exclude the last three pre-treatment periods during model fitting. For studies with only three pre-treatment periods, we set the number of placebo periods in their placebo tests to 2. We then check whether the estimated ATT in these "placebo periods" is significantly different from zero. The null hypothesis is that the mean pseudo-treatment effect in this range is zero. A small p-value, prompting rejection of the placebo test, indicates a potential violation of the PT assumption.

(No) Carryover Effects Test: We utilize the `fect` package's carryover-effects test. By specifying `carryover.period = c(1,2)`, we exclude the first two periods following a unit's return to untreated status from model fitting and assess whether the estimated ATT in these periods is significantly different from zero. The null hypothesis is that the mean pseudo-treatment effect over these periods is zero. A small p-value, leading to rejection of the test, suggests a potential failure of the no-carryover-effects assumption.

A.2.5. Robust Confidence Set & Sensitivity Analysis

Finally, we implement the sensitivity analysis developed by Rambachan and Roth (2023) to evaluate robustness against potential violations of the PT assumption. Specifically, we allow for differences in trends between treatment and control groups but constrain how large those differences can be. Suppose the dynamic treatment effects (DTE) vector, μ, can be decomposed into the true treatment effects, τ, and a violation-of-PT component, δ:

$$\mu = \begin{pmatrix} 0 \\ \tau_{\text{post}} \end{pmatrix} + \begin{pmatrix} \delta_{\text{placebo}} \\ \delta_{\text{post}} \end{pmatrix}.$$

Following Rambachan and Roth (2023), we impose a *relative magnitude* (RM) restriction on δ:

$$|\delta_{t+1} - \delta_t| \;\leq\; \bar{M} \cdot \max\Big\{ \big|\delta_{-1} - \delta_{-2}\big|,\; \big|\delta_0 - \delta_{-1}\big| \Big\} \quad \text{for all } t \geq 0,$$

where $\mathcal{P} = \{-2, -1, 0\}$ is the set of placebo periods. Hence, the maximum observed change in δ between consecutive placebo periods is

$$\max\Big\{ \big|\delta_0 - \delta_{-1}\big|,\; \big|\delta_{-1} - \delta_{-2}\big| \Big\}.$$

We introduce two modifications to the original Rambachan and Roth (2023) framework. First, our placebo periods are estimated with the same imputation method used for the post-

treatment DTEs to maintain comparability and consistency (Roth, 2024). Second, since the imputation method does not fix a specific reference period, we explicitly incorporate the PT violation observed in the last pre-treatment placebo period ($t = 0$). Thus, if $\bar{M} = 0$, the PT violation from the final placebo period extends into all post-treatment periods ($\delta_t = \delta_0$ for $t > 0$). Letting $\bar{M} > 0$ allows the PT violation to change between consecutive post-treatment periods, but keeps that change bounded by $\bar{M}$ times the largest consecutive placebo discrepancy.

Point Estimates: Let the ATT of interest be $\theta = l'_{att}\delta_{post}$, where $l_{att} = \left[\frac{n_1}{\sum_{t=1}^{T_{post}} n_t}, \ldots, \frac{n_{T_{post}}}{\sum_{t=1}^{T_{post}} n_t}\right]'$ and n_t is the number of observations in post-treatment period t. From Lemma 2.1 in Rambachan and Roth (2023), define the set of θ values consistent with a given μ under $\delta \in \Delta^{RM}(\bar{M})$:

$$S(\mu, \Delta^{RM}) \;=\; \left[\theta^{lb}(\mu, \Delta^{RM}),\ \theta^{ub}(\mu, \Delta^{RM})\right],$$

where

$$\theta^{lb}(\mu, \Delta^{RM}(\bar{M})) = l'_{att}\mu_{\text{post}} - \max_{\delta}\Big\{ l'_{att}\delta_{\text{post}} : \delta \in \Delta^{RM}(\bar{M}),\ \delta_{\text{placebo}} = \mu_{\text{placebo}}\Big\},$$

$$\theta^{ub}(\mu, \Delta^{RM}(\bar{M})) = l'_{att}\mu_{\text{post}} - \min_{\delta}\Big\{ l'_{att}\delta_{\text{post}} : \delta \in \Delta^{RM}(\bar{M}),\ \delta_{\text{placebo}} = \mu_{\text{placebo}}\Big\}.$$

Under a finite-sample normal approximation for the estimated DTEs $\hat{\mu}$, i.e., $\hat{\mu} \sim N(\tau+\delta, \Sigma_n)$, Rambachan and Roth (2023) define the confidence set for θ, $\mathcal{C}_n\big(\hat{\mu}_n, \Sigma_n\big)$, by

$$\inf_{\delta \in \Delta^{RM}(\bar{M}),\, \tau}\ \inf_{\theta \in \mathcal{S}(\delta+\tau, \Delta^{RM}(\bar{M}))} \mathbb{P}_{\hat{\mu}_n \sim \mathcal{N}\left(\delta+\tau, \Sigma_n\right)}\Big(\theta \in \mathcal{C}_n\big(\hat{\mu}_n, \Sigma_n\big)\Big) \;\geq\; 1-\alpha.$$

They also show that for a broad class of distributions $\mathcal{P}$ such that $\delta_P \in \Delta^{RM}(\bar{M})$ for all $P \in \mathcal{P}$, one can replace Σ_n with a consistent estimate $\hat{\Sigma}_n$ and obtain

$$\liminf_{n\to\infty}\ \inf_{P \in \mathcal{P}}\ \inf_{\theta \in \mathcal{S}(\delta_P+\tau_P, \Delta^{RM}(\bar{M}))} \mathbb{P}_P\Big(\theta \in \mathcal{C}_n(\hat{\mu}_n, \hat{\Sigma}_n)\Big) \;\geq\; 1-\alpha.$$

We use the `createSensitivityResults_relativeMagnitudes` function from the `Honestdid` package to compute $\mathcal{C}_n(\hat{\mu}_n, \hat{\Sigma}_n)$ for different values of $\bar{M}$. We obtain the estimated DTE $\hat{\mu}_n$ and its variance-covariance matrix $\hat{\Sigma}_n$ from `fect`. When $\bar{M} = 0$, the resulting confidence interval can be viewed as a "de-biased" interval that adjusts for the PT violation observed at $t = 0$. If this confidence set excludes zero, we further evaluate robustness by computing the "breakdown value" $\tilde{M}$, the smallest $\bar{M}$ for which zero enters the confidence interval.

Dynamic Treatment Effects: We further examine the robustness of the estimated DTEs by computing confidence intervals at $\bar{M} = 0$ and $\bar{M} = 0.50$ for each post-treatment period. This follows the same approach described above for point estimates but uses a different weighting vector l.

A.3. More Replication Results

A.3.1. Sample Selection Criteria

We collect our replication sample from three leading journals in political science, *APSR*, *AJPS*, and *JOP*. We screen all full research articles published in these journals during 2017-2023 using the following four criteria:

1. The paper uses panel data analysis as a critical piece of evidence to support a causal argument. Specifically, either the abstract or the introduction of the paper needs to mention the results from the panel analysis.

2. The paper uses at least one linear model to analyze panel data, such as DID, TWFE, or lagged dependent variable (LDV) models, and the treatment variable has to be binary. In other words, papers that use only discrete outcome models or continuous treatments are excluded. We include this criterion because most of the analytical tools the literature has developed so far are designed for linear models with binary treatments.

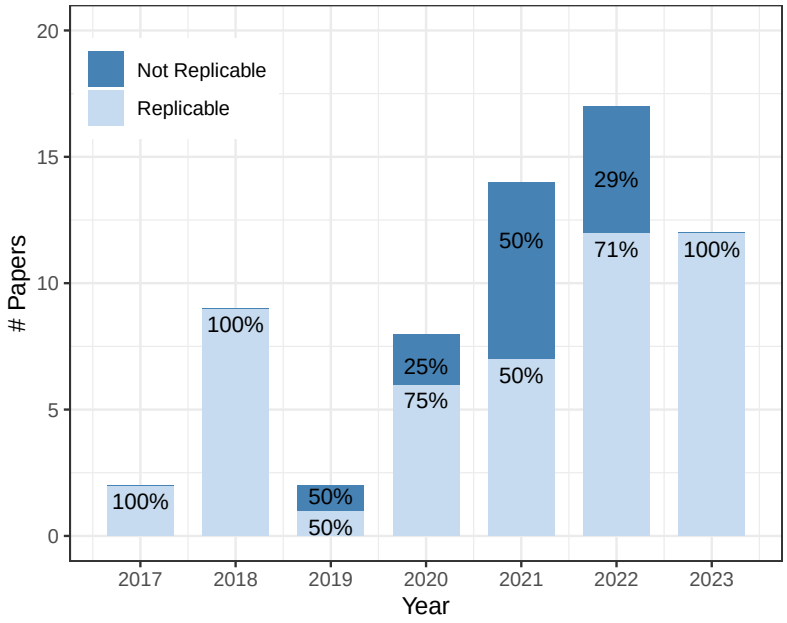
3. We exclude papers that use a regression discontinuity design or an instrumental variables (IVs) design, including Bartik IVs, as their primary identification strategy.

4. We exclude papers that do not exploit within-unit variation despite the longitudinal structure of the data. These designs are drastically different from the rest of the panel studies in their estimand, their identification assumptions, and the properties of their estimators and are worth investigating separately.

A.3.2. Replicability

For papers that meet our four screening criteria, we try to find replication materials from public data-sharing platforms, such as the *Harvard Dataverse*, and the authors' personal websites. For each paper, we choose one model that we think can best represent the paper's central claim. Specifically, we sequentially go through the following two criteria: (1) the authors claim that it is the preferred model; and (2) the model uses the most complete dataset (i.e., with the least missing values). Using data and code from the replication materials, we are able to successfully replicate the main results of 49 of 64 papers that meet our criteria. By successful replication, we mean that we can replicate the point estimate of the chosen specification up to the second decimal point. Figure A1 shows the number of replicable and non-replicable papers by year.

FIGURE A1. REPLICABILITY BY YEAR

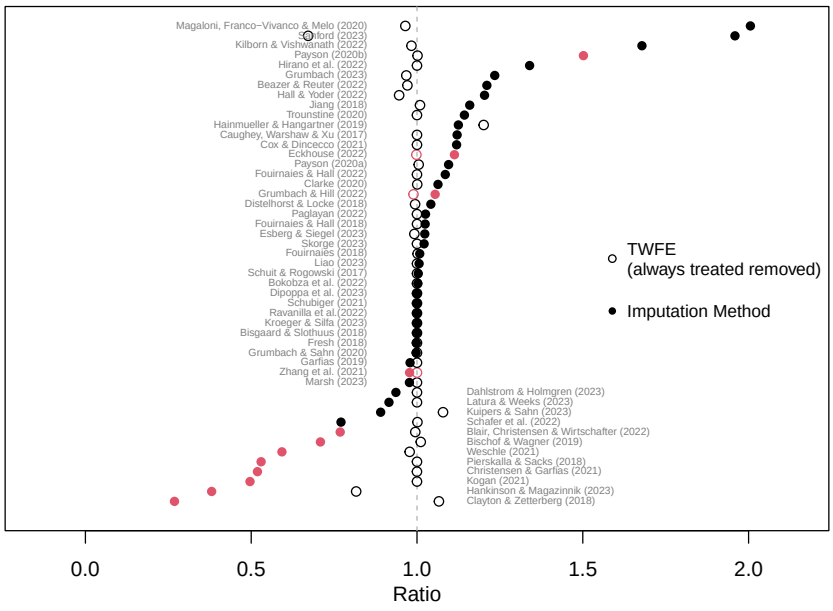


Note: The above figure shows the number of papers that meet our criteria. The grey and black bars represent the number of replicable papers and the number of papers that cannot be replicated.

A.3.3. Reported vs TWFE and `FEct` Estimates

To assess how the imputation estimator differs from TWFE, we depict the ratio of the `FEct` estimates to the TWFE estimates, after excluding the always-treated units (ensuring identical sample sets), using solid circles in Figure A2. We also juxtapose the cited TWFE estimates with those omitting the always-treated units, represented by hollow circles. The red circles denote studies where the `FEct` estimates are not statistically significant at the 5% level. Notably, both the mean and median of these ratios are close to one. This finding suggests that, although there are noticeable differences in individual cases, TWFE does not systematically under- or over-estimates the ATT. Figure A2 also shows that the presence of always-treated units is not the primary driver of these differences. When these units are excluded, the TWFE estimates align closely with the reported estimates in most cases.

FIGURE A2. TWFE VS. IMPUTATION METHOD: ESTIMATES

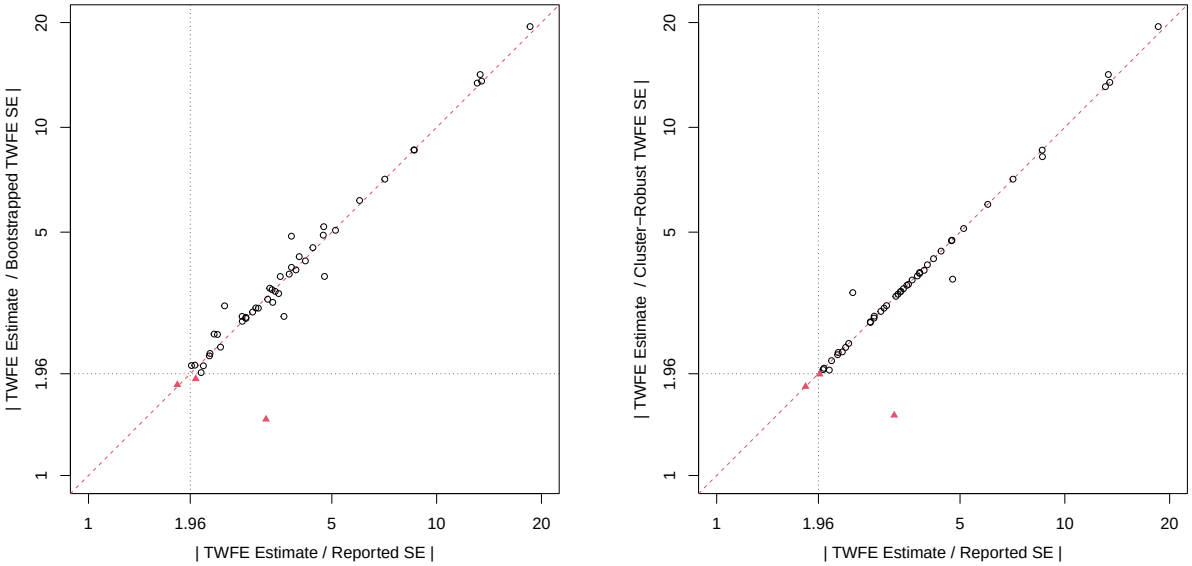


Note: In the above figure, solid circles represent the ratios of the estimates from the imputation method (`FEct`) to TWFE coefficients with always-treated units removed; hollow circles represent the ratios of reported TWFE coefficients to TWFE coefficients with always-treated units removed. Statistically insignificant `FEct` estimates at the 5% level are painted in red.

A.3.4. Inferential Methods

The following figures show that cluster-robust SEs, which were used by almost all authors in the original studies, yield SE estimates similar to those obtained from cluster-bootstrap procedures in the majority of studies. One exception is Cox and Dincecco (2021) in which the number of units is very small ($N = 10$).

FIGURE A3. ROBUSTNESS TO INFERENTIAL METHODS

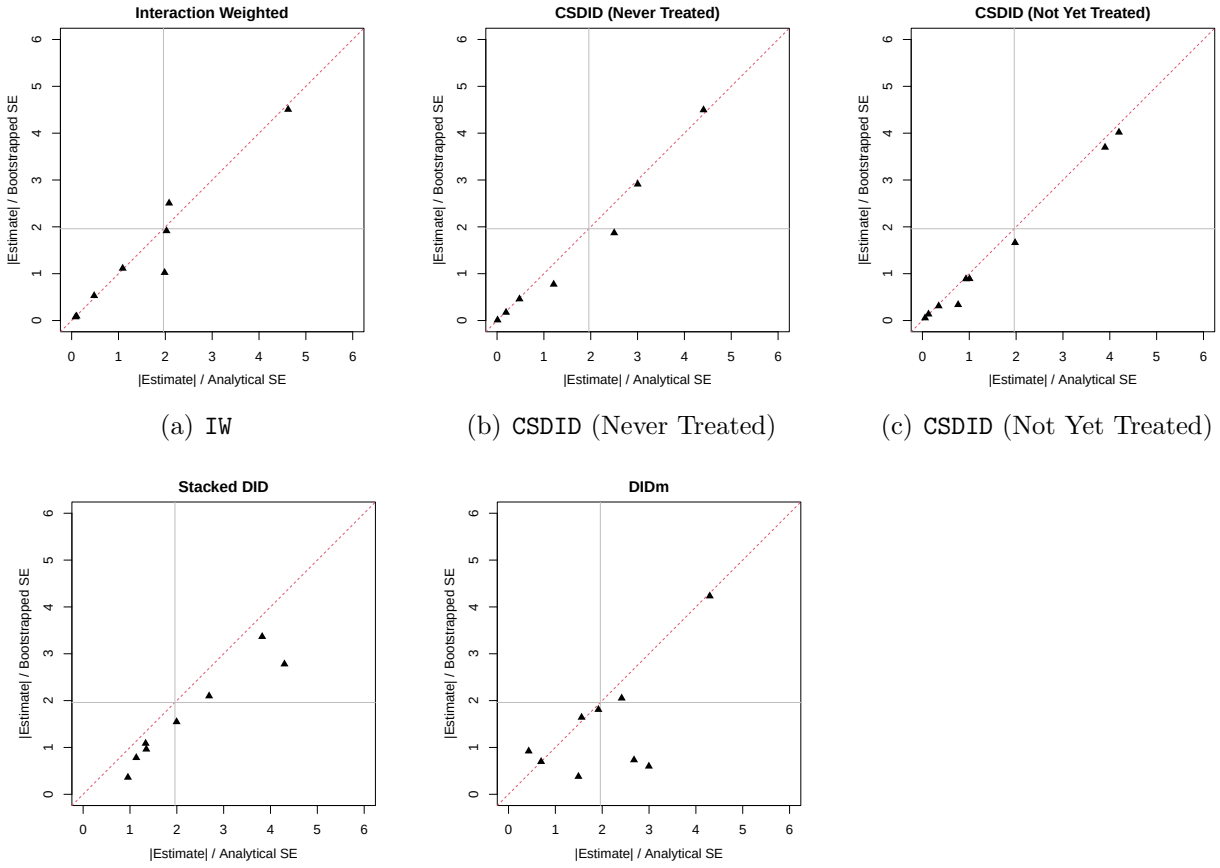


(a) TWFE w/ cluster-robust SEs (b) TWFE w/ cluster-bootstrapped SEs

Note: The left panel compares the absolute values of the original z scores and replicated z scores using cluster-robust SEs. The right panel compares the absolute values of the original z scores and replicated z scores using cluster-bootstrapped SEs. Both axes are on log scales. The original estimate in Zhang et al. (2021) is statistically insignificant at the 5% level. Our replication analysis finds that, additionally, Eckhouse (2022) and Grumbach and Hill (2022) are statistically insignificant at the 5% with the cluster-robust SE; Bischof and Wagner (2019) and Blair, Christensen and Wirtschafter (2022) are statistically insignificant at the 5% with cluster-bootstrapped SEs. Bootstrap percentile methods yield similar findings.

Figure A4 shows that, for studies with staggered adoption settings in our sample, cluster-bootstrapped SEs are generally larger than analytically derived SEs for five HTE-robust estimators (with smaller z-scores). We use cluster-bootstrapped SEs throughout the main text.

FIGURE A4. ANALYTICAL VS. BOOTSTRAPPED SES



(a) `IW` (b) `CSDID` (Never Treated) (c) `CSDID` (Not Yet Treated)

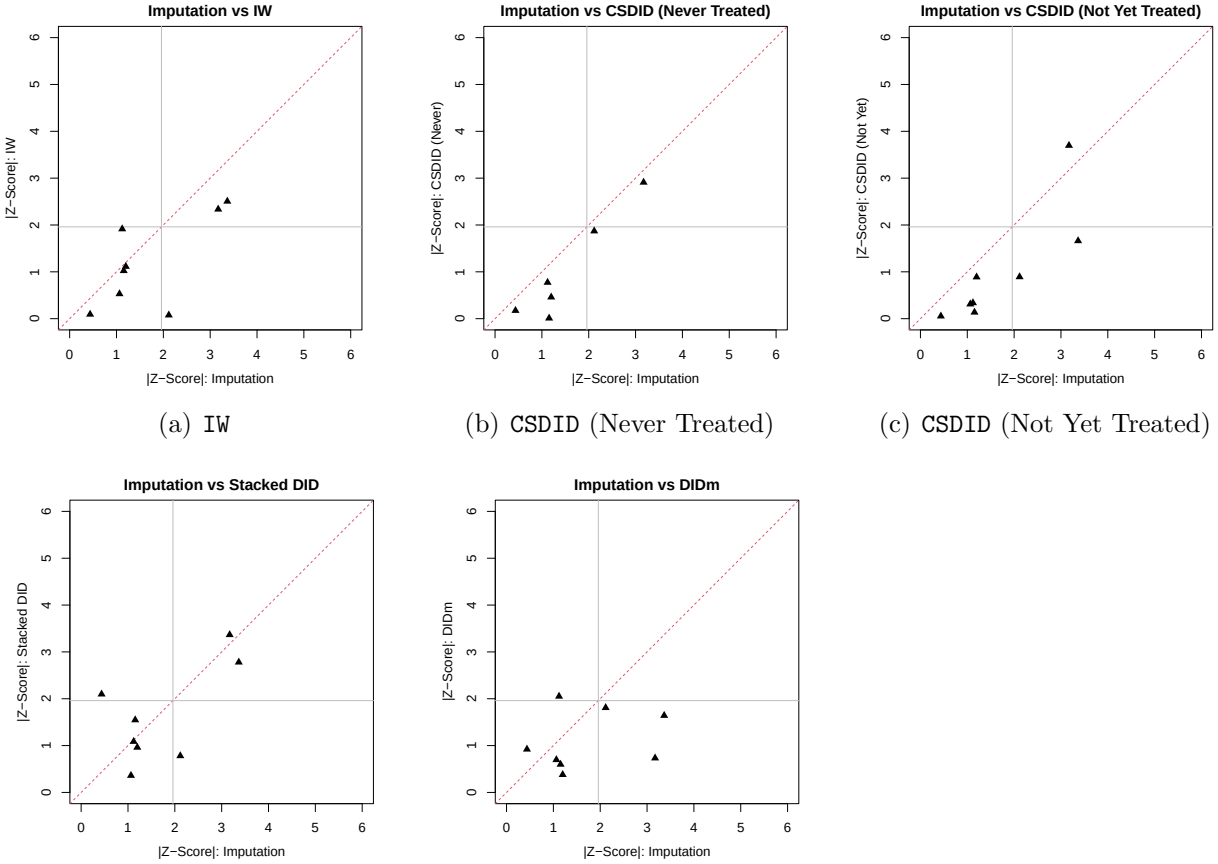
(d) `StackDID` (e) `DID_multiple`

Note: In each figure, the x-axis represents the absolute value of z-scores calculated using analytical SEs, while the y-axis represents the absolute value of z-scores calculated using cluster-bootstrapped SEs.

A.3.5. Imputation vs. Other Methods

Figure A5 demonstrates that, for studies with staggered adoption settings in our sample, the imputation estimator (`FEct`) is generally more efficient than other HTE-robust estimators (with bigger z-scores). All SEs are calculated using cluster-bootstrapping.

FIGURE A5. IMPUTATION AND OTHER HTE-ROBUST ESTIMATORS: Z-SCORES



(a) `IW` (b) `CSDID` (Never Treated) (c) `CSDID` (Not Yet Treated)

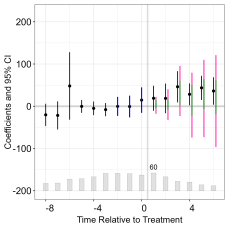
(d) `StackDID` (e) `DID_multiple`

Note: In each figure, the x-axis represents the absolute value of z-scores from `FEct`, while the y-axis represents the absolute value of z-scores from an alternative HTE-robust estimator.

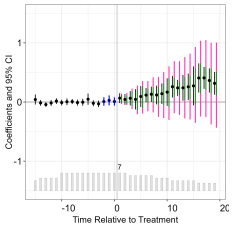
A.3.6. Placebo Tests & Robust Confidence Sets

Figure A6. Placebo Tests & Robust CSs

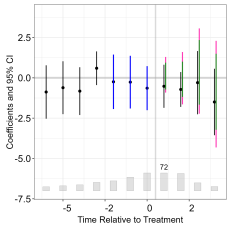
Beazer and Reuter (2022)
p-value $= 0.72$, $\tilde{M} = 0.00$



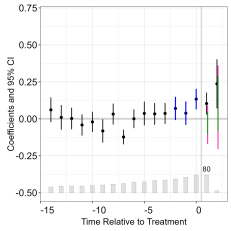
Bischof and Wagner (2019)
p-value $= 0.71$, $\tilde{M} = 0.14$



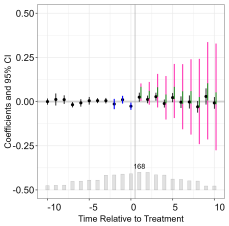
Blair, Christensen and Wirtschafter (2022)
p-value $= 0.52$, $\tilde{M} = 0.00$



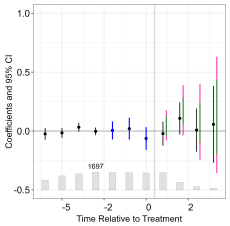
Bokobza et al. (2022)
p-value $= 0.00$, $\tilde{M} = 0.00$



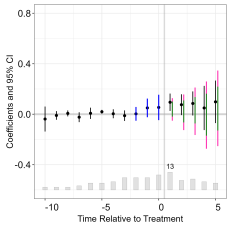
Caughey, Warshaw and Xu (2017)
p-value $= 0.41$, $\tilde{M} = 0.12$



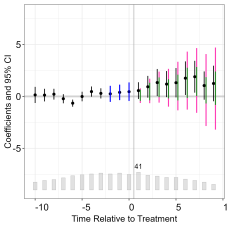
Christensen and Garfias (2021)
p-value $= 0.76$, $\tilde{M} = 0.00$



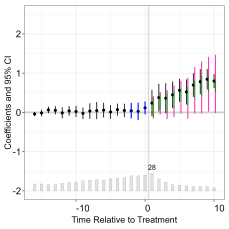
Clarke (2020)
p-value $= 0.16$, $\tilde{M} = 0.00$



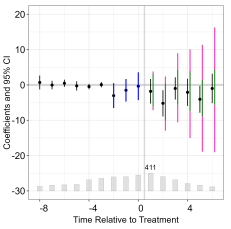
Clayton and Zetterberg (2018)
p-value $= 0.24$, $\tilde{M} = 0.00$



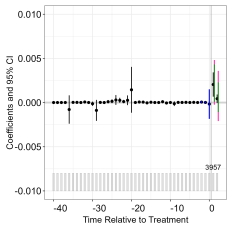
Cox and Dincecco (2021)
p-value $= 0.53$, $\tilde{M} = 0.30$



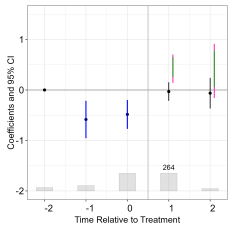
Dahlström and Holmgren (2023)
p-value $= 0.17$, $\tilde{M} = 0.00$



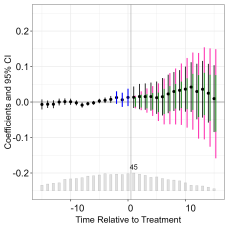
Dipoppa, Grossman and Zonszein (2023)
p-value $= 0.91$, $\tilde{M} = 0.00$



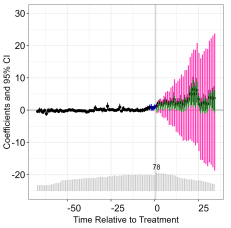
Distelhorst and Locke (2018)
p-value $= 0.00$, $\tilde{M} = 0.40$



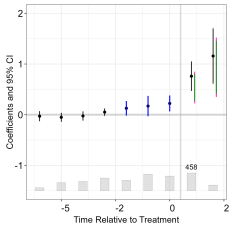
Eckhouse (2022)
p-value $= 0.22$, $\tilde{M} = 0.00$



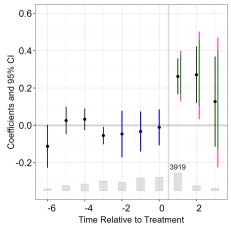
Esberg and Siegel (2023)
p-value $= 0.02$, $\tilde{M} = 0.03$



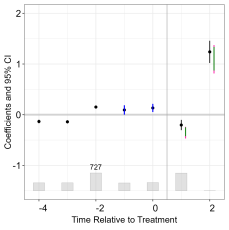
Fouirnaies (2018)
p-value $= 0.01$, $\tilde{M} = 1.75$



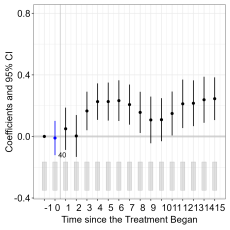
Fouirnaies and Hall (2018)
p-value $= 0.47$, $\tilde{M} = 1.00$



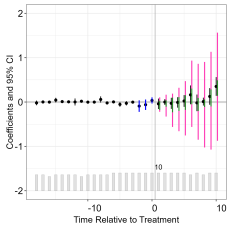
Fouirnaies and Hall (2022)
p-value $= 0.00$, $\tilde{M} = 1.20$



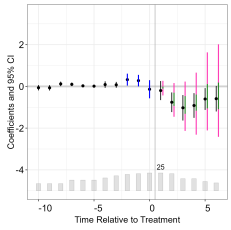
Fresh (2018)
p-value $= 0.86$



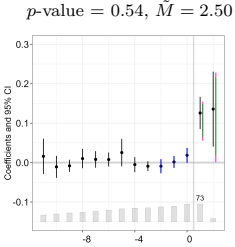
Garfias (2019)
p-value $= 0.32$, $\tilde{M} = 0.00$



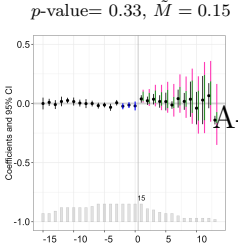
Grumbach (2023)
p-value $= 0.03$, $\tilde{M} = 0.00$



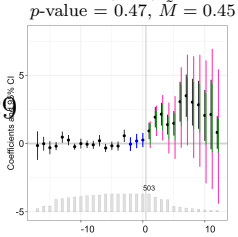
Grumbach and Sahn (2020)
p-value $= 0.54$, $\tilde{M} = 2.50$



Grumbach and Hill (2022)
p-value$= 0.33$, $\tilde{M} = 0.15$



Hainmueller and Hangartner (2019)
p-value $= 0.47$, $\tilde{M} = 0.45$



Hankinson and Magazinnik (2023)
p-value $=0.29$, $\tilde{M} = 0.00$

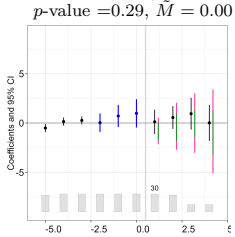
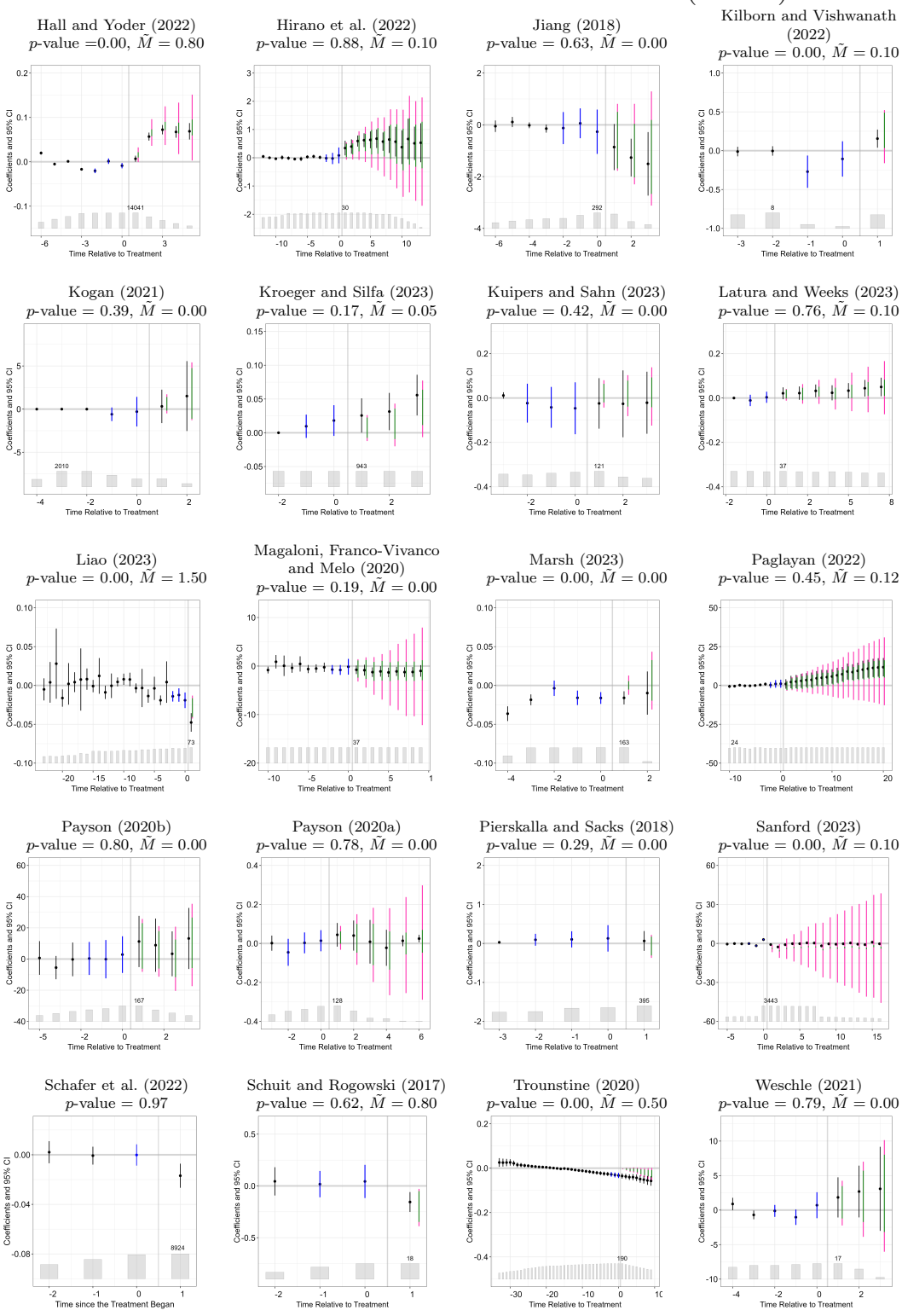


Figure A6. Placebo Tests & Robust CSs (Cont.)



Note: We report p-values from the placebo test (hiding two or three pre-treatment periods for each switch from the control condition to the treatment condition). Four cases with only one pre-treatment period are excluded. $\tilde{M}$, the breakdown value for $\bar{M}$ is calculated for studies with more than three pre-treatment periods; two additional studies, Fresh (2018) and Schafer et al. (2022), are not included.

A.3.7. Carryover Effects

FIGURE A7. TEST FOR (NO) CARRYOVER EFFECTS

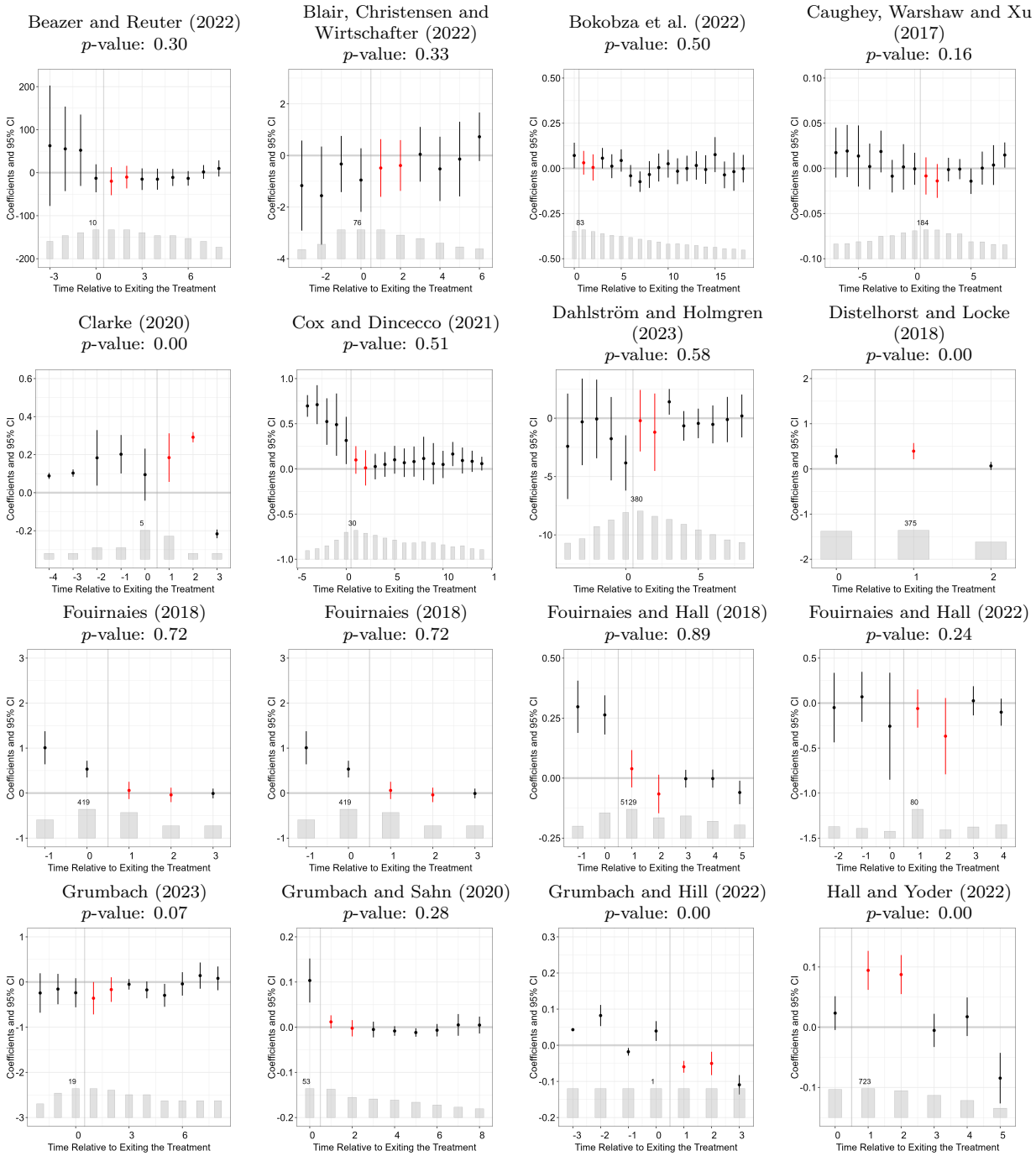
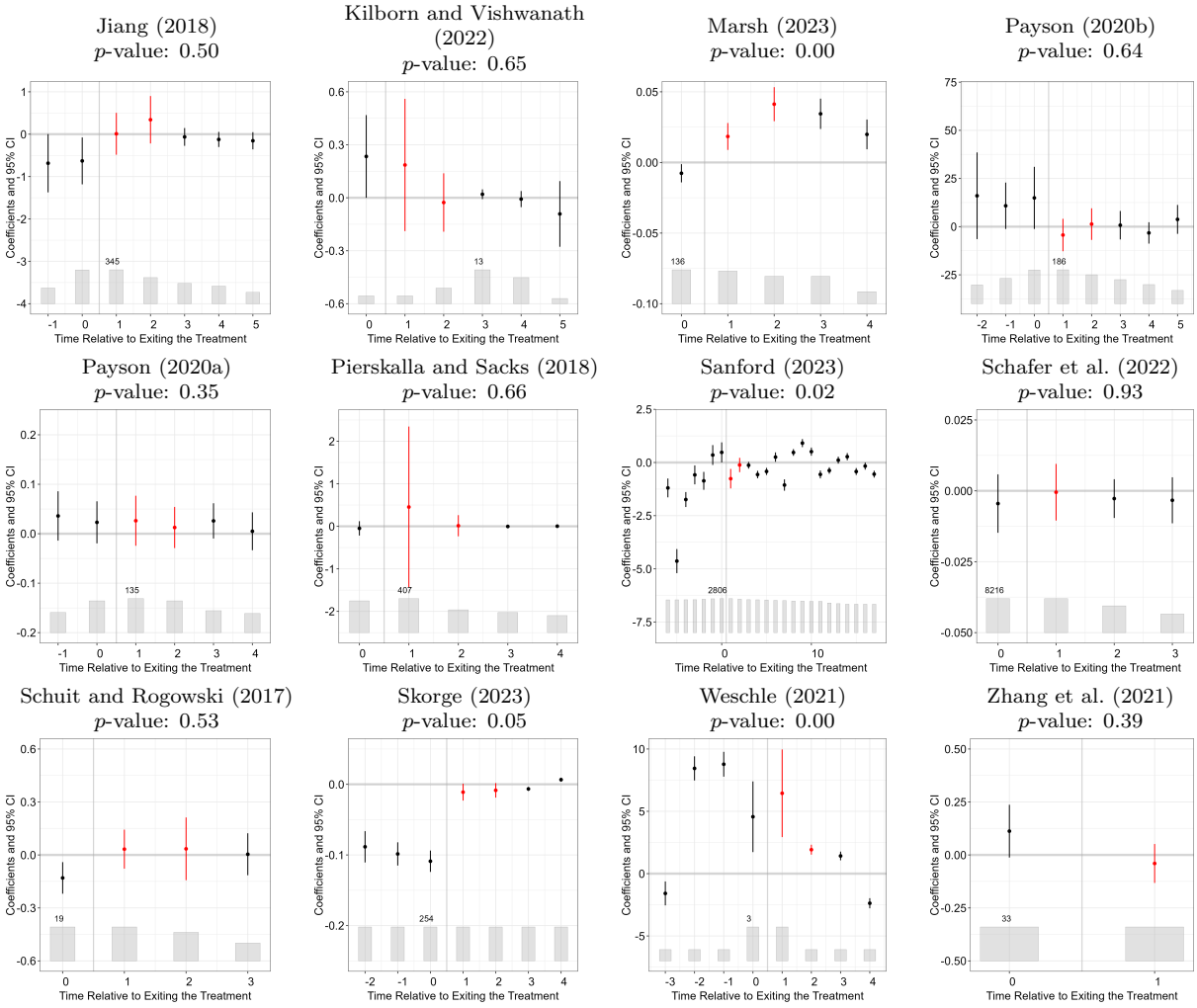


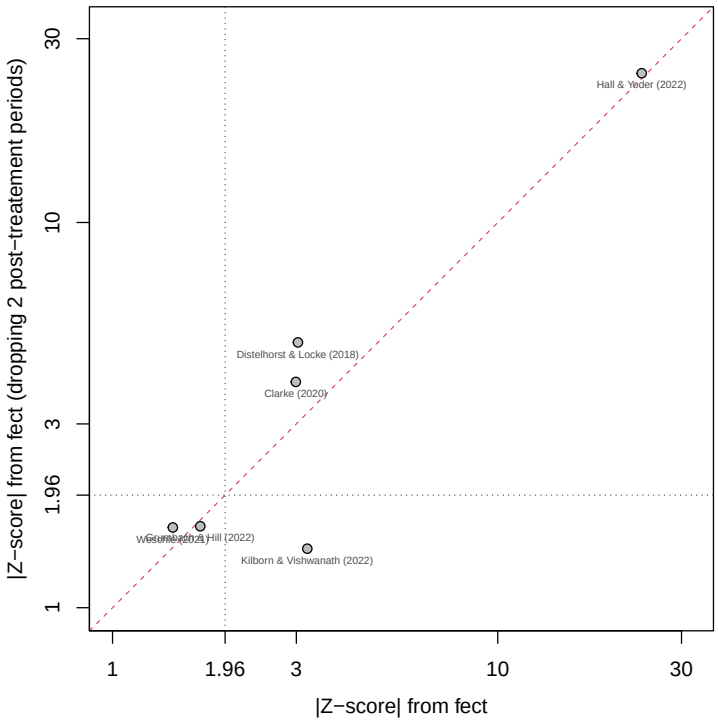
Figure A7. Test for (No) Carryover Effects (Cont.)



Note: We report p-values from the test for no carryover effects for 22 studies with treatment reversal. The test hides two posttreatment periods for each exiting from the treatment condition to the control condition.

The figure below illustrates that the substantive findings obtained from `FEct` remain unchanged even after excluding two periods following the treatment's reversion to untreated status in six studies that reject the no-carryover-effects test. This suggests that, while carryover effects are commonly observed in applied settings, the cost of addressing them—such as by excluding a few potentially affected periods after the treatment reverts to untreated status—is typically minimal.

FIGURE A8. ROBUSTNESS TO REMOVING TWO POSTTREATMENT PERIODS



Note: The figure compares the z-scores from `FEct` using all data and z-scores from `FEct` after removing two posttreatment periods in six studies that reject the no carryover effects test. We observe no sign flipping. Both axes are on log scales.

A.3.8. Summary of Findings

Table A1. Summary of Findings

Paper	Journal	Subfield	T	N	#Obs	Setting	Specification	ATT $p < 0.05$	F Test $p > 0.05$	Placebo test $p > 0.05$	Carryover effect test $p > 0.05$	$\bar{M}$
Beazer and Reuter (2022)	JOP	CP	11	199	2,027	General	u+t	✓	✓	✓	✓	0.00
Bischof and Wagner (2019)	AJPS	CP	42	17	534	Staggered	u+ht		✓	✓	n.a.	0.14
Bisgaard and Slothuus (2018)	AJPS	CP	2	570	1,140	2×2	u+t	✓	n.a.	n.a.	n.a.	n.a.
Blair, Christensen and Wirtschafter (2022)	JOP	IR	18	177	3,186	General	u+t		✓	✓	✓	0.00
Bokobza et al. (2022)	JOP	CP	50	115	3,715	General	u+t	✓			✓	0.00
Caughey, Warshaw and Xu (2017)	JOP	AP	79	50	3,586	General	u+t	✓	✓	✓	✓	0.12
Christensen and Garfias (2021)	JOP	CP	12	3,289	25,536	Staggered	u+t		✓	✓	n.a.	0.00
Clarke (2020)	AJPS	AP	17	702	3,603	General	u+t	✓	✓	✓		0.00
Clayton and Zetterberg (2018)	JOP	CP	17	139	2,227	Staggered	u+t			✓	n.a.	0.00
Cox and Dincecco (2021)	JOP	CP	259	10	1,361	General	u+t	✓	✓	✓	✓	0.30
Dahlström and Holmgren (2023)	JOP	CP	43	371	5,316	General	u+t	✓	✓	✓	✓	0.00
Dipoppa, Grossman and Zonszein (2023)	JOP	CP	159	7,922	1,259,477	Block	u+t	✓	✓	✓	n.a.	0.00
Distelhorst and Locke (2018)	AJPS	CP	4	2,447	6,915	General	u+t	✓				0.40
Eckhouse (2022)	AJPS	AP	24	47	1,023	Staggered	u+t		✓	✓	n.a.	0.00
Esberg and Siegel (2023)	APSR	CP	89	357	22,669	Staggered	u+t	✓			n.a.	0.030
Fouirnaies (2018)	AJPS	AP	23	16,404	45,639	General	u+hu*t	✓			✓	1.75
Fouirnaies and Hall (2018)	AJPS	AP	20	161,820	443,490	General	u+hu*t	✓	✓	✓	✓	1.00
Fouirnaies and Hall (2022)	APSR	AP	130	4,642	11,109	General	u+hu*t	✓			✓	2.75
Fresh (2018)	JOP	AP	17	100	1,695	Block	u+t	✓	✓	✓	n.a.	n.a.
Garfias (2019)	JOP	CP	29	17	445	Block	u+t	✓	✓	✓	n.a.	0.00
Grumbach (2023)	APSR	AP	17	49	833	General	u+t	✓	✓	✓	✓	0.15
Grumbach and Hill (2022)	JOP	AP	20	49	980	General	u+t		✓	✓		0.00
Grumbach and Sahn (2020)	APSR	AP	17	489	6,847	General	u+t	✓	✓	✓	✓	2.50
Hainmueller and Hangartner (2019)	AJPS	CP	21	1,209	22,971	Staggered	u+t	✓	✓	✓	n.a.	0.45
Hall and Yoder (2022)	JOP	AP	9	9,888,539	88,996,851	General	u+t	✓				0.80
Hankinson and Magazinnik (2023)	JOP	AP	10	40	397	Staggered	u+t		✓	✓	n.a.	0.00
Hirano et al. (2022)	JOP	AP	26	33	769	Staggered	u+t	✓	✓	✓	n.a.	0.10
Jiang (2018)	AJPS	CP	12	326	3,891	General	u+hu*t	✓	✓	✓	✓	0.00
Kilborn and Vishwanath (2022)	AJPS	AP	7	347	1,062	General	u+t	✓	✓		✓	0.10
Kogan (2021)	JOP	AP	8	3,005	23,610	Staggered	u+hu*t+ult		✓	✓	n.a.	0.00
Kroeger and Silfa (2023)	JOP	AP	6	2,835	17,010	Block	u+t+ult	✓	✓	✓	n.a.	0.05
Kuipers and Sahn (2023)	APSR	CP	9	294	1,604	Staggered	u+t	✓	✓	✓	n.a.	0.00
Latura and Weeks (2023)	AJPS	CP	10	90	761	Block	u+t	✓	✓	✓	n.a.	0.10
Liao (2023)	JOP	AP	26	384,462	981,096	Block	u+t	✓			n.a.	1.50
Magaloni, Franco-Vivanco and Melo (2020)	APSR	CP	138	286	36,956	Staggered	u+t+ult	✓	✓	✓	n.a.	0.00
Marsh (2023)	APSR	AP	9	2,889	23,952	General	u+t	✓				0.000
Paglayan (2022)	APSR	CP	40	183	2,882	Staggered	u+t	✓	✓	✓	n.a.	0.12
Payson (2020a)	APSR	AP	9	738	6,307	General	u+t	✓	✓	✓	✓	0.00
Payson (2020b)	JOP	AP	13	467	5,982	General	u+t		✓	✓	✓	0.00
Pierskalla and Sacks (2018)	JOP	CP	9	455	2,524	General	u+t			✓	✓	0.00
Ravanilla, Sexton and Haim (2022)	JOP	CP	2	189	378	2×2	u+t	✓	n.a.	n.a.	n.a.	n.a.
Sanford (2023)	AJPS	CP	35	4,633,413	158,423,948	General	u+t	✓				0.10
Schafer et al. (2022)	AJPS	CP	4	381,256	1,163,307	General	u+t	✓	✓	✓	✓	n.a.
Schubiger (2021)	JOP	CP	2	11,958	23,916	2×2	u+t	✓	n.a.	n.a.	n.a.	n.a.
Schuit and Rogowski (2017)	AJPS	AP	5	261	902	General	u+t	✓	✓	✓	✓	0.80
Skorge (2023)	AJPS	CP	7	569	3,983	General	u+t	✓	n.a.	n.a.		n.a.
Trounstine (2020)	APSR	AP	43	4,568	182,809	Staggered	u+t	✓	✓		n.a.	0.50
Weschle (2021)	JOP	CP	7	845	4,714	General	u+t			✓		0.00
Zhang et al. (2021)	JOP	CP	3	61	166	General	u+t		n.a.	✓	✓	n.a.

Note: ✓ and `n.a.` stand for "true" and "not applicable," respectively. The strongest case for the validity of the design is when we have ✓ in all five columns (or the first four columns with staggered adoption). In the "Specification" column, "u" and "t" represent unit and time fixed effects, respectively; "ht" represents time effects higher than the basic time level; "hu*t" represents group-specific time effects (group is at a higher level than unit); "ult" represents unit-specific linear time trends. $\bar{M}$ represents breakdown $\bar{M}$ value in a sensitivity analysis; a bigger number means the result is more robust to potential PT violations. In Trounstine (2020), the differential trends are likely linear, making the smoothness restriction (SM) on the confidence set more suitable than the relative magnitudes (RM) restriction; nonetheless, we report the breakdown value $\bar{M}$ obtained using the RM restriction in the table for consistency. However, under the SM restriction, the result is not robust at $\bar{M} = 0$.

References

Bai, Jushan and Serena Ng. 2021. "Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data." *Journal of the American Statistical Association* 116(536):1746–1763.

Beazer, Quintin H. and Ora John Reuter. 2022. "Do Authoritarian Elections Help the Poor? Evidence from Russian Cities." *The Journal of Politics* 84(1):437–454.

Berge, Laurent, Sebastian Krantz and Grant McDermott. 2023. *fixest: Fast Fixed-Effects Estimations*. R package version 0.11.1.
URL: *https://lrberge.github.io/fixest/, https://github.com/lrberge/fixest*

Bischof, Daniel and Markus Wagner. 2019. "Do Voters Polarize When Radical Parties Enter Parliament?" *American Journal of Political Science* 63(4):888–904.

Bisgaard, Martin and Rune Slothuus. 2018. "Partisan Elites as Culprits? How Party Cues Shape Partisan Perceptual Gaps." *American Journal of Political Science* 62(2):456–469.

Blair, Graeme, Darin Christensen and Valerie Wirtschafter. 2022. "How Does Armed Conflict Shape Investment? Evidence from the Mining Sector." *The Journal of Politics* 84(1):116–133.

Bleiberg, Joshua. 2021. "STACKEDEV: Stata module to implement stacked event study estimator." *Working Paper* .

Bokobza, Laure, Suthan Krishnarajan, Jacob Nyrup, Casper Sakstrup and Lasse Aaskoven. 2022. "The Morning After: Cabinet Instability and the Purging of Ministers after Failed Coup Attempts in Autocracies." *The Journal of Politics* 84(3):1437–1452.

Borusyak, Kirill, Xavier Jaravel and Jann Spiess. 2024. "Revisiting event-study designs: robust and efficient estimation." *Review of Economic Studies* 91(6):3253–3285.

Callaway, Brantly and Pedro HC Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225(2):200–230.

Caughey, Devin, Christopher Warshaw and Yiqing Xu. 2017. "Incremental Democracy: The Policy Effects of Partisan Control of State Government." *The Journal of Politics* 79(4):1342–1358.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner and Ben Zipperer. 2019. "The Effect of Minimum Wages on Low-Wage Jobs." *Quarterly Journal of Economics* 134(3):1405–1454.

Christensen, Darin and Francisco Garfias. 2021. "The Politics of Property Taxation: Fiscal Infrastructure and Electoral Incentives in Brazil." *The Journal of Politics* 83(4):1399–1416.

Clarke, Andrew J. 2020. "Party Sub-Brands and American Party Factions." *American Journal of Political Science* 64(3):452–470.

Clayton, Amanda and Pär Zetterberg. 2018. "Quota Shocks: Electoral Gender Quotas and Government Spending Priorities Worldwide." *The Journal of Politics* 80(3):916–932.

Cox, Gary W and Mark Dincecco. 2021. "The Budgetary Origins of Fiscal-Military Prowess." *The Journal of Politics* 83(3):851–866.

Dahlström, Carl and Mikael Holmgren. 2023. "Loyal Leaders, Affluent Agencies: The Budgetary Implications of Political Appointments in the Executive Branch." *The Journal of Politics* 85(2):640–653.

de Chaisemartin, Clément and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110(9):2964–2996.

De Chaisemartin, Clément and Xavier d'Haultfoeuille. 2024. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." *Review of Economics and Statistics* pp. 1–45.

Dipoppa, Gemma, Guy Grossman and Stephanie Zonszein. 2023. "Locked Down, Lashing Out: COVID-19 Effects on Asian Hate Crimes in Italy." *The Journal of Politics* 85(2):389–404.

Distelhorst, Greg and Richard M. Locke. 2018. "Does Compliance Pay? Social Standards and Firm-Level Trade." *American Journal of Political Science* 62(3):695–711.

Eckhouse, Laurel. 2022. "Metrics Management and Bureaucratic Accountability: Evidence from Policing." *American Journal of Political Science* 66(2):385–401.

Esberg, Jane and Alexandra A. Siegel. 2023. "How Exile Shapes Online Opposition: Evidence from Venezuela." *American Political Science Review* 117(4):1361–1378.

Fouirnaies, Alexander. 2018. "When Are Agenda Setters Valuable?" *American Journal of Political Science* 62(1):176–191.

Fouirnaies, Alexander and Andrew B. Hall. 2018. "How Do Interest Groups Seek Access to Committees?" *American Journal of Political Science* 62(1):132–147.

Fouirnaies, Alexander and Andrew B Hall. 2022. "How Do Electoral Incentives Affect Legislator Behavior? Evidence from U.S. State Legislatures." *American Political Science Review* 116(2):662–676.

Fresh, Adriane. 2018. "The Effect of the Voting Rights Act on Enfranchisement: Evidence from North Carolina." *The Journal of Politics* 80(2):713–718.

Garfias, Francisco. 2019. "Elite Coalitions, Limited Government, and Fiscal Capacity Development: Evidence from Bourbon Mexico." *The Journal of Politics* 81(1):95–111.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225(2):254–277.

Grumbach, Jacob M. 2023. "Laboratories of Democratic Backsliding." *American Political Science Review* 117(3):967–984.

Grumbach, Jacob M. and Alexander Sahn. 2020. "Race and Representation in Campaign Finance." *American Political Science Review* 114(1):206–221.

Grumbach, Jacob M and Charlotte Hill. 2022. "Rock the Registration: Same Day Registration Increases Turnout of Young Voters." *The Journal of Politics* 84(1):405–417.

Hainmueller, Jens and Dominik Hangartner. 2019. "Does Direct Democracy Hurt Immigrant Minorities? Evidence from Naturalization Decisions in Switzerland." *American Journal of Political Science* 63(3):530–547.

Hall, Andrew B. and Jesse Yoder. 2022. "Does Homeownership Influence Political Behavior? Evidence from Administrative Data." *The Journal of Politics* 84(1):351–366.

Hankinson, Michael and Asya Magazinnik. 2023. "The Supply-Equity Trade-Off: The Effect of Spatial Representation on the Local Housing Supply." *The Journal of Politics* 85(3):1033–1047.

Hirano, Shigeo, Jaclyn Kaslovsky, Michael P. Olson and James M. Snyder. 2022. "The Growth of Campaign Advertising in the United States, 1880–1930." *The Journal of Politics* 84(3):1482–1496.

Imai, Kosuke and In Song Kim. 2019. "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 63(2):467–490.

Imai, Kosuke, In Song Kim and Erik H Wang. 2023. "Matching Methods for Causal Inference with Time-Series Cross-Sectional Data." *American Journal of Political Science* 67(3):587–605.

Jiang, Junyan. 2018. "Making Bureaucracy Work: Patronage Networks, Performance Incentives, and Economic Development in China." *American Journal of Political Science* 62(4):982–999.

Kilborn, Mitchell and Arjun Vishwanath. 2022. "Public Money Talks Too: How Public Campaign Financing Degrades Representation." *American Journal of Political Science* 66(3):730–744.

Kogan, Vladimir. 2021. "Do Welfare Benefits Pay Electoral Dividends? Evidence from the National Food Stamp Program Rollout." *The Journal of Politics* 83(1):20–70.

Kroeger, Mary and Maria Silfa. 2023. "Motivated Corporate Political Action: Evidence from an SEC Experiment." *The Journal of Politics* 85(3):1139–1144.

Kuipers, Nicholas and Alexander Sahn. 2023. "The Representational Consequences of Municipal Civil Service Reform." *American Political Science Review* 117(1):200–216.

Latura, Audrey and Ana Catalano Weeks. 2023. "Corporate Board Quotas and Gender Equality Policies in the Workplace." *American Journal of Political Science* 67(3):606–622.

Li, Zikai and Anton Strezhnev. 2024. "A Guide to Dynamic Difference-in-Differences Regressions for Political Scientists." *Working Paper* .

Liao, Steven. 2023. "The Effect of Firm Lobbying on High-Skilled Visa Adjudication." *The Journal of Politics* 85(4):1416–1429.

Liu, Licheng, Ye Wang and Yiqing Xu. 2024. "A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data." *American Journal of Political Science* 68(1):160–176.

Magaloni, Beatriz, Edgar Franco-Vivanco and Vanessa Melo. 2020. "Killing in the Slums: Social Order, Criminal Governance, and Police Violence in Rio de Janeiro." *American Political Science Review* 114(2):552–572.

Marsh, Wayde ZC. 2023. "Trauma and Turnout: The Political Consequences of Traumatic Events." *American Political Science Review* 117(3):1036–1052.

Mou, Hongyu, Licheng Liu and Yiqing Xu. 2023. "Panel Data Visualization in R (panelView) and Stata (panelview)." *Journal of Statistical Software* 107(7):1–20.

Paglayan, Agustina S. 2022. "Education or Indoctrination? The Violent Origins of Public School Systems in an Era of State-Building." *American Political Science Review* 116(4):1242–1257.

Payson, Julia A. 2020a. "The Partisan Logic of City Mobilization: Evidence from State Lobbying Disclosures." *American Political Science Review* 114(3):677–690.

Payson, Julia A. 2020b. "Cities in the Statehouse: How Local Governments Use Lobbyists to Secure State Funding." *The Journal of Politics* 82(2):403–417.

Pierskalla, Jan H. and Audrey Sacks. 2018. "Unpaved Road Ahead: The Consequences of Election Cycles for Capital Expenditures." *The Journal of Politics* 80(2):510–524.

Rambachan, Ashesh and Jonathan Roth. 2023. "A More Credible Approach to Parallel Trends." *The Review of Economic Studies* 90(5):2555–2591.

Ravanilla, Nico, Renard Sexton and Dotan Haim. 2022. "Deadly Populism: How Local Political Outsiders Drive Duterte's War on Drugs in the Philippines." *The Journal of Politics* 84(2):1035–1056.

Roth, Jonathan. 2024. "Interpreting Event-Studies from Recent Difference-in-Differences Methods." *Working Paper* .

Sanford, Luke. 2023. "Democratization, Elections, and Public Goods: The Evidence from Deforestation." *American Journal of Political Science* 67(3):748–763.

Sant'Anna, P and B Callaway. 2021. "DID: Treatment effects with multiple periods and groups in R." *URL: https://cran. r-project. org/web/packages/did/index. html* .

Schafer, Jerome, Enrico Cantoni, Giorgio Bellettini and Carlotta Berti Ceroni. 2022. "Making Unequal Democracy Work? The Effects of Income on Voter Turnout in Northern Italy." *American Journal of Political Science* 66(3):745–761.

Schubiger, Livia Isabella. 2021. "State Violence and Wartime Civilian Agency: Evidence from Peru." *The Journal of Politics* 83(4):1383–1398.

Schuit, Sophie and Jon C. Rogowski. 2017. "Race, Representation, and the Voting Rights Act." *American Journal of Political Science* 61(3):513–526.

Skorge, Øyvind Søraas. 2023. "Mobilizing the Underrepresented: Electoral Systems and Gender Inequality in Political Participation." *American Journal of Political Science* 67(3):538–552.

Strezhnev, Anton. 2018. "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs." Mimeo, New York University.

Sun, Liyang and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225(2):175–199.

Trounstine, Jessica. 2020. "The Geography of Inequality: How Land Use Regulation Produces Segregation." *American Political Science Review* 114(2):443–455.

Weschle, Simon. 2021. "Parliamentary Positions and Politicians' Private Sector Earnings: Evidence from the UK House of Commons." *The Journal of Politics* 83(2):706–721.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT press.

Zhang, Qi, Dong Zhang, Mingxing Liu and Victor Shih. 2021. "Elite Cleavage and the Rise of Capitalism under Authoritarianism: A Tale of Two Provinces in China." *The Journal of Politics* 83(3):1010–1023.