Causal Panel Analysis under Parallel Trends: Lessons from a Large Reanalysis Study*

Albert Chiu (Stanford) Xingchen Lan (NYU) Ziyi Liu (Berkeley) Yiqing Xu (Stanford)

American Political Science Review
Vol. 120, Iss. 1, Feb. 2026, pp. 245–266

Abstract

Two-way fixed effects (TWFE) models are widely used in political science to establish causality, but recent methodological discussions highlight their limitations under heterogeneous treatment effects (HTE) and violations of the parallel trends (PT) assumption. This growing literature has introduced numerous new estimators and procedures, causing confusion among researchers about the reliability of existing results and best practices. To address these concerns, we replicated and reanalyzed 49 studies from leading journals that employ TWFE models for causal inference using observational panel data with binary treatments. Using six HTE-robust estimators, diagnostic tests, and sensitivity analyses, we find: (i) HTE-robust estimators yield qualitatively similar but highly variable results; (ii) while a few studies show clear signs of PT violations, many lack evidence to support this assumption; and (iii) many studies are underpowered when accounting for HTE and potential PT violations. We emphasize the importance of strong research designs and rigorous validation of key identifying assumptions.

Keywords: panel data, two-way fixed effects, parallel trends, heterogeneous treatment effects, pretrend, event-study plot, difference-in-differences, robust confidence set

*Albert Chiu, PhD student, Department of Political Science, Stanford University. Email: altchiu@stanford.edu. Xingchen Lan, PhD student, Wilf Family Department of Politics, New York University. Email: xingchenlan@nyu.edu. Ziyi Liu, PhD student, Haas School of Business, University of California, Berkeley. Email: zyliu2023@berkeley.edu. Yiqing Xu, Assistant Professor, Department of Political Science, Stanford University. Email: yiqingxu@stanford.edu. We thank Quintin Beazer, Kirill Borusyak, Kirk Bansak, Gary Cox, Avi Feller, Anthony Fowler, Francisco Garfias, Justin Grimmer, Jens Hainmueller, Erin Hartman, Guido Imbens, Julia Payson, Annamaria Prati, Jonathan Roth, Luwei Ying, Ye Wang, and participants at the 39th PolMeth, the 11th Asian PolMeth, Seminars at UCSD, UCLA, UC Berkeley, and University of Chicago for helpful comments and suggestions. We are also grateful to Neil Malhotra, Andrew Baker, Anton Strezhnev, and the Alethsia platform for providing pre-publication reviews of this paper. Further thanks go to two anonymous reviewers and the APSR editor, Andrew Eggers, for their helpful comments. We appreciate the authors of the studies we reviewed for their constructive feedback and for making their data publicly available.

1. Introduction

Over the past decade, political scientists have increasingly relied on panel data to draw causal conclusions (Xu, 2023). A favored method for such analyses is the two-way fixed effects (TWFE) model because of its ability to control for unobserved time-invariant confounders and common time trends. In our survey of 102 articles published from 2017 to 2023 in three top political science journals using observational panel data with *binary* treatments, 64 studies (63%) assume a TWFE model with the following functional form or a close variant:[1]

$$Y_{i,t} = \tau^{TWFE} D_{i,t} + X_{i,t}{}'\beta + \alpha_i + \xi_t + \varepsilon_{i,t}, \quad \text{for all } i, t, \tag{1}$$

where $Y_{i,t}$ and $D_{i,t}$ are the outcome and treatment variables for unit i at time t; $X_{i,t}$ is a vector of time–varying covariates; α_i and ξ_t are unit and time fixed effects; and $\varepsilon_{i,t}$ is idiosyncratic errors.[2] Researchers typically interpret τ^{TWFE} as the treatment effect and estimate the model using ordinary least squares. The resulting estimator for τ^{TWFE} is commonly known as the TWFE estimator. Moreover, researchers frequently conflate this model with a difference-in-differences (DID) design, and use the two terms interchangeably.[3]

Recent methodological discussions have raised concerns about the validity of TWFE models and the associated identifying assumptions, leaving many researchers in a quandary. First, existing findings based on the TWFE models may not hold given recent developments. Second, with the introduction of numerous new estimators and diagnostics, there is confusion about the current best practices. This paper seeks to bridge this gap by reviewing new

[1] The remaining 38 studies can be categorized into five groups: studies focusing on interaction effects (8 studies), studies using nonlinear links such as logit and Poisson (5 studies), studies employing instrumental variables or regression discontinuity designs (8 studies), and studies using other linear specifications, such as only one-way fixed effects or lagged dependent variables (17 studies).

[2] In some studies classified as using TWFE models, "unit" fixed effects are specified at the group level g, where multiple units i are nested (e.g., county fixed effects when i indexes cities), or time fixed effects are at a higher level p (e.g., year fixed effects when t indexes days). For simplicity, we use the notation α_i and ξ_t rather than the more general α_g and ξ_p.

[3] We use the phrase "DID design" in reference to DID research design, which differs from the typical usage in the statistics literature that refers to treatment assignment mechanism (Xu, Zhao and Ding, 2024).

estimation, inference, and diagnostics methods from the methodological literature and by reanalyzing published studies using both new estimators and the TWFE estimator. Based on the findings, we offer several practical recommendations for researchers.

These criticisms of the use of TWFE models mainly come from two directions. First, causal identification using TWFE models require the *strict exogeneity* assumption, which critics argue is stronger than many researchers realize and is often unrealistic in real-world settings (e.g., Imai and Kim, 2019). Strict exogeneity states that:

$$\text{(Strict exogeneity)} \qquad \mathbb{E}[\varepsilon_{i,t} \mid \mathbf{D_i}, \mathbf{X_i}, \alpha_i, \xi_t] = \mathbb{E}[\varepsilon_{i,t} \mid D_{i,t}, X_{i,t}, \alpha_i, \xi_t] = 0, \quad \forall i, t,$$

in which $\mathbf{D_i} = \{D_{i,1}, D_{i,2}, \cdots, D_{i,T}\}$ and $\mathbf{X_i} = \{X_{i,1}, X_{i,2}, \cdots, X_{i,T}\}$. It means that once current treatment status, covariates, and fixed effects are accounted for, treatment status in any other periods has no additional effect on $Y_{i,t}$ (Wooldridge, 2010, p. 253). Under Equation (1), strict exogeneity implies a parallel trends (PT) assumption:

$$\begin{aligned}\text{(Parallel trends)} \qquad & \mathbb{E}[Y_{i,t}(0) - Y_{i,s}(0) \mid D_{i,t} = 1, D_{i,s} = 0, X_{i,t} - X_{i,s} = x_0] \\ & = \mathbb{E}[Y_{j,t}(0) - Y_{j,s}(0) \mid D_{j,t} = 0, D_{j,s} = 0, X_{j,t} - X_{j,s} = x_0],\end{aligned}$$

in which $Y_{i,t}(0) = Y_{i,t}(d_{i,t} = 0)$ represents the untreated potential outcome for unit i at time t. It states that the change in untreated potential outcomes between any two periods is mean independent of the change in observed treatment status during those periods, once changes in covariate values are controlled for. Threats to PT, such as the presence of time-varying confounders and feedback from past outcomes to current treatment assignment, also invalidate strict exogeneity. Therefore, throughout the rest of the article, we use the term "PT violations" to encompass violations of strict exogeneity.[4]

The second group of criticisms concerns the consequences of heterogeneous treatment

[4]We discuss the relationship between strict exogeneity and PT, as well as other assumptions, under Equation (1) in Section A.1 of the Supplementary Materials (SM). Note that the PT assumption invoked by many HTE-robust estimators does not depend on Equation (1).

effects (HTE), that is, τ^{TWFE} is not a constant (e.g., Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020; Strezhnev, 2018; Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Athey and Imbens, 2022; Borusyak, Jaravel and Spiess, 2024). Researchers have shown that, under HTE, TWFE estimates in general do not converge to a convex combination of the individual treatment effects for observations under the treatment condition, even when the PT assumption is valid. The so-called "negative weighting" problem, as described in de Chaisemartin and D'Haultfœuille (2020), is an alarming theoretical result because it implies that a TWFE estimand can be negative (positive) even when all individual treatment effects are positive (negative). To address this issue, researchers have proposed many new estimators that are "HTE-robust"—that is, estimators that converge to some convex combinations of individual treatment effects under their identifying assumptions.

This article thus pursues two goals. First, we explain and compare six recent proposals to amend TWFE models, including the interaction weighted (IW) estimator (Sun and Abraham, 2021), stacked DID (Baker, Larcker and Wang, 2022), CSDID (Callaway and Sant'Anna, 2021), DID multiple (de Chaisemartin and D'Haultfœuille, 2020, 2024), PanelMatch (Imai, Kim and Wang 2023, hereafter IKW, 2023), and the imputation method (Borusyak, Jaravel and Spiess 2024, hereafter BJS, 2024; Liu, Wang and Xu 2024, hereafter LWX, 2024). These estimators produce causally interpretable estimates under HTE and PT (or its variants). Second, we replicate and reanalyze 49 studies published in the *American Political Science Review* (APSR), *American Journal of Political Science* (AJPS), and *The Journal of Politics* (JOP) from 2017 to 2023 which rely on a TWFE model to draw causal conclusions.[5] Our aim is to assess the consequences of using or not using HTE-robust estimators and shed light on the severity of PT violations in political science research.

Our reanalysis shows that, in most studies, the HTE-robust estimators yield qualita-

[5]Replication materials for this article are available for download at the APSR Harvard Dataverse (Chiu et al., 2025).

tively similar estimates to TWFE models. However, there is considerable variation in how closely these estimators align with TWFE. In three cases, at least one HTE-robust estimator produces an estimate with an opposite sign to the TWFE estimate; in one of these cases, the opposite-sign estimate is also statistically significant at the 5% level. There is also a more widespread problem of power: HTE-robust estimators tend to have larger measures of uncertainty, which, combined with even small fluctuations in point estimates, can weaken statistical confidence. This is especially relevant for results that originally teeter on the brink of significance.

The primary concern, however, is the validity of the PT assumption. While only a few studies show clear signs of PT violations, which likely lead to spurious findings, most studies lack the power to rule out that realistic PT violations could explain a nonzero estimated causal effect. In such cases, even mild PT violations (informed by pre-treatment estimates) prevent us from concluding that the original treatment effect is nonzero. This does not mean that these studies are wrong; rather, it indicates that the available data do not have sufficient power to reject the null hypothesis of no effects when the PT assumption is not perfectly met.

Overall, we find that about a small minority of the studies in our sample meet our criteria of being highly credible. In these studies, we can statistically distinguish the treatment effect from zero using an HTE-robust estimator, even when allowing for mild PT violations benchmarked against placebo estimates using pre-treatment data. We recognize this as a high standard, as most researchers do not account for the power needed for the sensitivity analysis we perform. To be clear, our intent is not to criticize the authors of the studies we have replicated, since many of the methods we used were not available at the time their studies were conducted. Our goal is to guide and improve future research.

In light of these findings, we urge researchers to prioritize a strong research design and sufficient power in causal panel studies. Credible observational studies should feature a

well-defined treatment-outcome pair, shock-induced variation in treatment assignment, and sufficient power to ensure results are not undermined by small perturbations of key identifying assumptions. Research design has often been overlooked in causal panel analyses, likely because researchers have become accustomed to accepting the strong parametric and exogeneity assumptions behind TWFE models. Recent studies have emphasized the importance of (quasi-)randomness in treatment assignment for the robustness of DID findings (e.g., Roth and Sant'Anna, 2023).

This article makes several contributions. First, we propose a typology of various estimators for causal panel analysis. Our typology is based on the settings in which an estimator can be used and how controls are chosen. We also provide a comprehensive comparison of these estimators and show how several proposals are equivalent in some circumstances. We hope this discussion helps researchers deepen their understanding of these estimators. Second, we adapt the robust confidence set approach for sensitivity analysis proposed by Rambachan and Roth (2023) to the setting of imputation estimators. We find it highly useful as it avoids the issue of conditional inference–where hypothesis testing conditional on passing a pretest (e.g., a pretrend test) can distort estimation and inference (Roth, 2022). Third, our reanalysis instills confidence in existing political science research that uses TWFE models correctly while also cautioning against potential risks, such as the failure of the PT assumption and insufficient power. Based on these findings, we provide recommendations to improve practices, including the choice of estimators and the use of proper diagnostics. Finally, we contribute to the ongoing conversation on replication and reproducibility in political science (e.g., Eggers et al., 2015; Lall, 2016; Hainmueller, Mummolo and Xu, 2019; Lal et al., 2024).

Our work is closely related to Baker, Larcker and Wang (2022), who evaluate the credibility of a handful of studies with staggered adoption treatments in finance and accounting. It differs in that: (i) we use a wider range of estimators and diagnostic tests on a larger and

more diverse set of empirical applications, many of which involve treatment reversals; (ii) our review suggests that while the weighting issue under HTE is important, the main threats to causal inference with panel data are PT violations and insufficient power. Our work also relates to Roth et al. (2023), Xu (2023), de Chaisemartin and D'Haultfœuille (2023), Arkhangelsky and Imbens (2024), and Baker et al. (2025), who review and synthesize the recent methodological advancements in the DID literature. What sets this paper apart is our application of these innovations to data, allowing us to evaluate the practical relevance of the theoretical critiques.

This research has a few limitations. First, we do not examine methods based on sequential ignorability, an alternative identification framework that assumes no unobserved confounding but allows for dynamic treatment selection up to the current time period. Second, our analysis does not encompass studies that use continuous treatments, which is common in political science research. Finally, as the methodological literature continues to evolve rapidly, our recommendations should be regarded as reflecting current best practices.

2. TWFE and Its Pitfalls

In this section, we review the pitfalls of TWFE models identified in the literature. In the classic two-group and two-period case, the TWFE estimator $\hat{\tau}^{TWFE}$ is equivalent to the DID estimator, which consistently estimates the average treatment effect on the treated (ATT) under no anticipation and PT even with HTE. These results do not hold more generally in more complex settings with differential treatment adoption times (known as staggered adoption) or treatment reversal, as we will discuss below.

Our survey of the top journals reveals that the TWFE model under Equation (1) is the most commonly adopted approach for estimating causal effects using panel data in political science. Fixed effects models began their rise to prominence in political science in the early 2000s, and criticism promptly followed. In a debate with Green, Kim and Yoon (2001), Beck

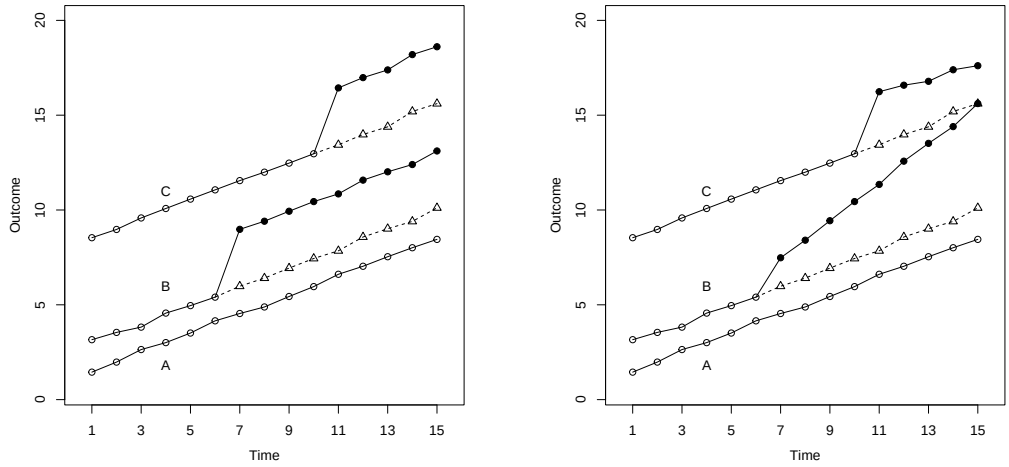
and Katz (2001) and King (2001) argue that linear fixed effects models often lead to misleading findings because they throw away valuable information, ignore rich temporal dynamics, and are incapable of capturing complex time-varying heterogeneities. Moreover, since both treatment and outcome variables are often serially correlated in panel data, researchers have advised against using standard error (SE) estimators suitable for cross-sectional data, such as Huber-White robust SEs (Bertrand, Duflo and Mullainathan, 2004). Scholars also recommend using bootstrap procedures to more effectively control Type I error rates when the number of clusters (units) is small (Cameron, Gelbach and Miller, 2008).

2.1. Recent Criticisms

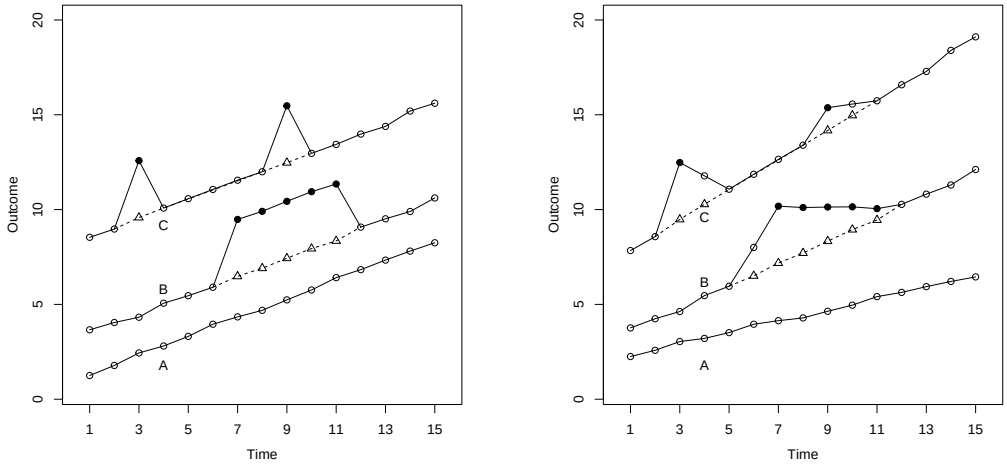
In the past few years, a surge of studies has renewed investigation into the properties of the TWFE estimator and the assumptions it requires to achieve causal identification. One group of work urges researchers to better understand TWFE models (and their assumptions) from a design-based perspective, with a focus on restrictions on treatment assignment mechanisms. For example, Imai and Kim (2019) point out that the strict exogeneity assumption not only implies the well-known no time-varying confounder requirement, but it also forbids a "feedback" effect from past outcomes to treatment assignment. Blackwell and Glynn (2018) clarify that such an assumption is closely related to baseline randomization in which the treatment vector is generated prior to, or independent of, the realization of the outcome. Another body of work cautions researchers that the PT assumption, even in a simple 2×2 DID setting, is sensitive to functional form. For example, Kahn-Lang and Lang (2020) emphasize the implicit functional form restrictions imposed by PT, encouraging researchers to justify it from a (quasi-)experimental perspective and address pre-treatment covariate imbalance. Roth and Sant'Anna (2023) point out that strong assumptions are needed for PT to be scale-independent, ensuring that a monotonic transformation of the outcome does not invalidate it. In practice, we find that many political science studies do not provide

strong evidence to justify strict exogeneity or PT.

FIGURE 1. TOY EXAMPLES: TWFE ASSUMPTIONS SATISFIED VS. VIOLATED



(a) Staggered DID: without Treatment Reversal



(b) General Setting: with Treatment Reversal

Note: The above panels show outcome trajectories of units in a staggered adoption setting (a) and in a general setting (b). Solid and hollow circles represent observed outcomes under the treatment and control conditions during the current period, respectively, while triangles represent counterfactual outcomes (in the absence of the treatment across all periods), $Y_{i,t}(\mathbf{d_i} = \mathbf{0})$. The data on the *left* panels in both (a) and (b) are generated by DGPs that satisfy TWFE assumptions while the data on the *right* are not. The divergence between hollow circles and triangles in the right panel of (b), both of which are under the control condition, is caused by anticipation or carryover effects.

A second body of research explores the implications of HTE with binary treatments within TWFE models (e.g., Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020; Strezhnev, 2018; Callaway and Sant'Anna, 2021; Athey and Imbens, 2022; Borusyak, Jaravel

and Spiess, 2024). Most of this literature assumes staggered adoption, but the insights from that setting are still relevant when there are treatment reversals. In Figure 1, we present two simplified examples for the staggered adoption and general settings. The left panel of Figure 1(a) represents outcome trajectories in line with standard TWFE assumptions, which not only include PT but also require that the treatment effect be contemporaneous and unvarying across units and over time. In contrast, the right panel portrays a scenario where PT holds, but the constant treatment effect assumption is not met. Various decompositions by the aforementioned researchers reveal that even under PT, when treatments begin at different times (such as in staggered adoption) and treatment effects evolve over time and vary across units, the TWFE estimand is generally not a convex combination of the individual treatment effects for observations subjected to the treatment. The basic intuition behind this theoretical result is that TWFE models use post-treatment data from units who adopt treatment earlier in the panel as controls for those who adopt the treatment later (e.g., Goodman-Bacon, 2021). HTE-robust estimators capitalize on this insight by avoiding these "invalid" comparisons between two treated observations.

A third limitation of the canonical TWFE specification is its presumption of no temporal and spatial interference. In most uses of TWFE models, researchers assume that there are no spatial spillovers and that treatment effects occur contemporaneously, hence no anticipation or carryover effects. No anticipation effects means that future treatments do not affect today's potential outcomes, while no carryover effects means that today's treatment does not affect future potential outcomes: For any i, t,

$$\text{(No anticipation effects)} \qquad Y_{i,t}(\mathbf{d}_i) = Y_{i,t}(d_{i,1}, d_{i,2}, \ldots, d_{i,t})$$

$$\text{(No carryover effects)} \qquad Y_{i,t}(\mathbf{d}_i) = Y_{i,t}(d_{i,t}, d_{i,t+1} \ldots, d_{i,T})$$

in which $\mathbf{d}_i = (d_{i,1}, d_{i,2}, \ldots, d_{i,T})$ is unit i's vector of potential treatment conditions in all periods (from period 1 to T) and $Y_{i,t}(\mathbf{d}_i)$ is the potential outcome in period t given $\mathbf{d}_i$. The

TWFE model specified in Equation (1) satisfies these two assumptions because $Y_{i,t}(\mathbf{d}_i) = Y_{i,t}(d_{i,t})$. These assumptions are obviously very strong, but they are rarely questioned or tested in practice (Imai and Kim, 2019; Athey and Imbens, 2022; Wang, 2021). Although some recent methods permit arbitrary carryover effects in staggered adoption settings, these effects are not distinguishable from contemporaneous effects.[6] This limitation becomes more complex when treatment reversal is possible, as demonstrated in Figure 1(b). In Figure 1(b), data in the left panel are consistent with TWFE assumptions, while the right panel illustrates deviations from PT, constant treatment effect, and the absence of anticipation or carryover effects. Real-world data often encounter the challenging scenarios depicted in the right panel rather than the idealized conditions in the left. Scholars have proposed methods to handle limited carryover effects in the general setting (IKW, 2023; LWX, 2024). The challenge of addressing spatial spillover effects without strong structural assumptions still persists (Wang et al., 2025; Wang, 2021), but its resolution is beyond the scope of this article.

2.2. Causal Estimands

To define the estimands clearly, consider the panel setting where multiple units $i \in \{1, \ldots, N\}$ are observed at each time period $t \in \{1, \ldots, T\}$. Each unit-time pair (i, t) uniquely identifies an observation. Define $E_{i,t}$ as unit i's "event time" at time t. For each i, let $E_{i,t} = \max\{t' : t' \leq t, D_{i,t'} = 1, D_{i,t'-1} = 0\}$ if $\exists s \leq t : D_{i,s} = 1$, and $E_{i,t} = \min\{t' : D_{i,t'} = 1, D_{i,t'-1} = 0\}$ otherwise. That is to say, $E_{i,t}$ is the most recent time at which unit i switched into treatment or, if i has not yet been treated at any point up until time t, the first time i switches into treatment. If i is never treated, we let $E_{i,t} = \infty$. In the staggered setting, the event time for each unit is constant, $E_i = E_{i,t}$, and $D_{i,t} = \mathbf{1}\{t \geq E_{i,t}\}$, where $\mathbf{1}\{\cdot\}$ is the indicator function. In such settings, we can partition units into distinct "cohorts" $g \in \{1, \ldots, G\}$ according to the timing of treatment adoption E_i. Units transitioning to treatment at period g $(i : E_{i,t} = g)$

[6] See Section A.4 of the SM in Liu, Wang and Xu (2024) for more details.

form cohort g, whereas units that never undergo treatment belong to the "never-treated" cohort ($i : E_{i,t} = \infty$). $Z_{i,t}$ ($Z_{i,g,t}$) represents the variable Z for unit i (part of cohort g) at time t. We use $Y_{i,t}(1)$ and $Y_{i,t}(0)$ to denote the potential outcomes under treatment and control, respectively, and $Y_{i,t} = D_{i,t}Y_{i,t}(1)+(1-D_{i,t})Y_{i,t}(0)$ to denote the observed outcome.[7]

The finest estimand is the individual treatment effect, $\tau_{i,t} = Y_{i,t}(1) - Y_{i,t}(0)$, of which there exists one for each observation (i, t).[8] Most political science research, however, typically focuses on estimating a single summary statistic. Commonly, this is the ATT, which represents individual treatment effects averaged over all observations exposed to the treatment condition. In between these extremes of granularity and coarseness are time-varying dynamic treatment effects, which are across-unit averages of individual treatment effects at each time period relative to treatment adoption. In the staggered adoption setting, we can further subdivide by cohort. We use τ_l ($\tau_{g,l}$) to denote the dynamic treatment effect l periods after treatment adoption (for treatment cohort g), with $l = 1$ representing the period immediately after treatment adoption.[9] $\tau_{g,l}$ is also what some authors refer to as a cohort average treatment effect on the treated (Strezhnev, 2018; Sun and Abraham, 2021) or group-time average treatment effect (Callaway and Sant'Anna, 2021).

Each of the estimators we discuss can be used to estimate τ_l. The outcome model analogous to TWFE for estimating dynamic effects is a lags-and-leads specification. For simplicity, we first describe the staggered setting. Let $K_{i,t} = (t - E_{i,t} + 1)$ be the number of periods until (when $K_{i,t} \leq 0$) or since unit i's event time at time t (e.g., $K_{i,t} = 1$ if unit i switches into treatment at time t). Consider a regression based on the following specification:

[7]The current notation will not cause confusion because we do not allow feedback or temporal spillover. In some of the studies we refer to, potential outcomes are defined in terms of treatment history, as opposed to current treatment status. We adopt similar notations for these frameworks. For instance, we use $Y_{i,t}(D_{i,t} = 1, \{D_{i,s}\}_{s<t} = 0)$ to represent the potential outcome under the specified treatment history.

[8]This is without loss of generality when feedback and interference are excluded. In staggered DID designs, carryover effects are permissible. When potential outcomes are defined in terms of treatment history, $\tau_{i,t}$ is defined as $Y_{i,t}(1) - Y_{i,t}(\infty)$ where $Y_{i,t}(\infty)$ signifies the untreated potential outcome when unit i never undergoes treatment.

[9]Some of the authors we reference denote this first post-treatment period with $l = 0$.

$$Y_{i,t} = \alpha_i + \xi_t + X'_{i,t}\beta + \sum_{\substack{l=-a \\ l\neq 0}}^{b} \tau_l^{TWFE} \cdot \mathbf{1}\{K_{i,t} = l\} + \tau_{b+}^{TWFE} \mathbf{1}\{K_{i,t} > b\} \cdot D_{i,t} + \varepsilon_{i,t}, \qquad (2)$$

where a and b are the number of lag and lead terms (BJS, 2024). In the social science literature, the typical practice is to exclude $l = 0$, which corresponds to the time period immediately before the transition into the treatment phase, and use it as a reference period (Roth, 2022). Conventionally, $\hat{\tau}_l^{TWFE}$ is interpreted as an estimate of τ_l or as a meaningful weighted average of pertinent individual treatment effects. Meanwhile, $\hat{\tau}_{b+}^{TWFE}$ is viewed as an estimate for the long-term effect.

3. HTE-Robust Estimators

In this section, we offer a brief overview and comparison of several recently introduced HTE-robust estimators. We use the term "HTE" to refer to individual treatment effects that are arbitrarily heterogeneous, that is, $\tau_{i,t} \neq \tau_{j,s}$ for some i, j, s, t. HTE-robust estimators are defined as those that produce causally interpretable estimates under their respective identifying assumptions. For a more comprehensive discussion on these estimators, please refer to the SM.

3.1. Summary of HTE-Robust Estimators

Table 1 summarizes the estimators we discuss in this article. The primary difference resides in the mechanics of their estimation strategies: there are methods based on canonical DIDs and methods based on imputation. We refer to the former as *DID extensions* and the latter as *imputation methods*. DID extensions use dynamic treatment effects, estimated from local, 2×2 DIDs between treated and control observations, as building blocks. Imputation methods use individual treatment effects, estimated as the difference between an imputed outcome under control and the observed outcome (under treatment), as building blocks.

The imputation estimator we use in this article employs TWFE, fitted with all observations under the control condition, to impute treated counterfactuals. Different strategies also entail different assumptions. Each DID extension, for example, relies on a particular type of PT assumption, whereas imputation methods presuppose a TWFE model for untreated potential outcomes and require a zero mean for the error terms, which is implied by strict exogeneity.

Another noteworthy difference lies in the settings in which these estimators are applicable: Some estimators can only be used in settings with staggered treatment adoption, while others can accommodate scenarios with treatment reversals. In the latter setting, all estimators we discuss also require no anticipation and no or limited carryover effects. Furthermore, the estimators diverge in terms of (1) how they select untreated observations as controls for treated units, (2) how they incorporate pre-treatment or exogenous covariates, and (3) the choice of the reference period. We discuss these details further below and in Section A.1 of the SM.

TABLE 1. SUMMARY OF HTE-ROBUST ESTIMATORS

Type	DID Extensions: uses 2×2 DIDs as building blocks					Imputation Methods	
Setting	Staggered: treatment reversals not allowed			General: treatment reversals allowed			
Research article	Sun and Abraham (2021)	Callaway and Sant'Anna (2021)	Baker, Larcker and Wang (2022)	de Chaisemartin and D'Haultfœuille (2020, 2024)	IKW (2023)	BJS (2024)	LWX (2024)
Method known as	`IW`	`CSDID`	`StackedDID`	`did_mulitple`	`PanelMatch`	$\texttt{DID}_{impute}$	`FEct`
Key ID assumption	Parallel trends	Parallel trends	Parallel trends	Parallel trends	Parallel trends	Strict exogeneity	Strict exogeneity
Finest estimand	$\tau_{g,l}$	$\tau_{g,l}$	τ_l^{vw}	$\tau_{g,l}$	τ_l	$\tau_{i,t}$	$\tau_{i,t}$
Common aggregated estimand	ATT	ATT	ATT	ATT for switchers for l periods		ATT	ATT
Comparison group	Never-treated or last-treated	Never-treated or not-yet-treated	Never-treated	Matched stable group (not-yet-treated)	Matched stable group (not-yet-treated)	Imputed counterfactuals (not-yet-treated)	Imputed counterfactuals (not-yet-treated)
Reference period(s)	Period 0	An arbitrary pre-treatment period	Period 0	Period 0	Period 0	All pre-treatment periods	All pre-treatment periods
Covariate adjustment	Possible extension	Outcome & propensity score modeling	Outcome modeling	Possible extension	Refined matched set and outcome modeling	Outcome modeling	Outcome modeling

3.2. DID Extensions

DID extensions are all built from local, 2×2 DID estimates. The overarching strategy for these estimators is to estimate the dynamic treatment effects, τ_l (or $\tau_{g,l}$ for each cohort g in the staggered setting), for each period since the most recent initiation of treatment, l, using one or more *valid* 2×2 DIDs. By "valid," we mean that the DID includes (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, whereas the post-period is such that observations from the treated group are in treatment and those from the comparison group are in control. The choice of the comparison group is the primary distinction between estimators in this category. To obtain higher-level averages such as the ATT, we then average over our estimates of τ_l (or $\tau_{g,l}$), typically employing appropriate, convex weights.

Three estimators in this category are appropriate only for the staggered setting. Sun and Abraham (2021) propose an interaction-weighted (`IW`) estimator, which is a weighted average of $\tau_{g,l}$ estimates obtained from a TWFE regression with cohort dummies fully interacted with indicators of relative time to the onset of treatment. They demonstrate that, in a balanced panel, each resulting estimate of $\tau_{g,l}$ can be characterized as a difference in the change in average outcome from a fixed pre-period $s < g$ to a post-period l periods since g between the treated cohort g and the comparison cohort(s) in some set $\mathcal{C}$. The authors recommend using $\mathcal{C} = \sup_i E_i$, which is either the never-treated cohort or, if no such cohort exists, the last-treated cohort. By default, `IW` uses $l = 0$ as the reference period and can accommodate covariates in the TWFE regression.

Employing the same general approach, Callaway and Sant'Anna (2021) propose two estimators, one of which uses never-treated units ($\hat{\tau}^{CS}_{nev}$) and the other not-yet-treated units ($\hat{\tau}^{CS}_{ny}$) as the comparison group. We label these estimators collectively as `CSDID`. Note that

$\hat{\tau}^{CS}_{nev}$ uses the same comparison group as `IW` when a never-treated cohort exists,[10] whereas $\hat{\tau}^{CS}_{ny}$ uses all untreated observations of not only never-treated units but also later adopters as controls for earlier adopters. Besides the choice of comparison cohort, `CSDID` estimators differ from `IW` in that they allow users to condition on pre-treatment covariates using both an explicit outcome model and inverse probability weighting simultaneously; consistency of the estimators requires at least one of these to be correct. By default, both `IW` and `CSDID` use the period immediately before the treatment's onset as the reference period for estimating the ATT.

Stacked DID, first formally introduced by Cengiz et al. (2019), is another related estimator sometimes used to address HTE concerns. As described by Baker, Larcker and Wang (2022), it involves creating separate sub-datasets for each treated cohort by combining data from that cohort (around treatment adoption) and data from the never-treated cohort from the same periods. These cohort-specific datasets are then "stacked" to form a single dataset. An event-study regression akin to Equation (2) with the addition of sub-dataset specific unit and time dummies is then run. This method uses the same comparison group as `IW` and the never-treated version of `CSDID` without covariates, but stacked DID estimates a single dynamic treatment effect for a given relative period rather than separate estimates for each cohort. Essentially, stacked DID is a special case of `IW` that uses immutable weights selected by OLS. We denote the corresponding estimand τ^{vw}_l to reflect the fact that it is variance-weighted. These weights are generally neither proportional to cohort sizes nor guaranteed to sum to one (Wing, Freedman and Hollingsworth, 2024). Thus, while stacked DID avoids the "negative weighting" problem and meets our criteria for HTE-robustness, its estimands are not the same as τ_l or the ATT.

In settings with treatment reversals, separate groups of researchers have converged on

[10] This equivalence holds when there are no missing data. Otherwise, `IW` from the saturated regression differs from one that directly estimates local DIDs, including the never-treated version of `CSDID`. These estimates are typically close but can differ substantially, as in Kuipers and Sahn (2023).

the same strategy for choosing a comparison group: matching treated and control observations that belong to units with identical treatment histories. IKW (2023) suggest one such estimator, `PanelMatch`, which begins by constructing a "matched set" for each observation (i, t) such that unit i transitions into treatment at time t. This matched set includes units that both (1) are not under treatment at time t and (2) share the same treatment history as i for a fixed number of periods leading up to the treatment onset. For each treated observation (i, t) and for every post-period $(t + l - 1)$ such that unit i is still under treatment, it then estimates a local DID using the same pre-period $s < t$. The treatment "group" comprises solely of observation (i, t), and the members of the matched set for (i, t) that are still under control during $t + l - 1$ serve as the comparison group. To obtain τ_l for a given l, it then averages over the corresponding local DID estimates from all treated observations. IKW (2023) propose incorporating covariates by "refining" matched sets and use $l = 0$ as the reference period.

Using a similar strategy, de Chaisemartin and D'Haultfœuille (2020) propose a "multiple DID" estimator, `DID_multiple`. A notable difference is that they include local DIDs for units leaving the treatment and not just those joining the treatment; when there are no treatment reversals or covariates, `DID_multiple` is a special case of `PanelMatch`. The original proposal for `DID_multiple` also only considers the case where we match on a single period and where $l = 1$, but since it has been extended (de Chaisemartin and D'Haultfœuille, 2024). Consequently, the target estimand is not the ATT but rather an average of the contemporaneous effects of "switching" (i.e., the effect of joining or the negative of the effect of leaving at the time of doing so). In the staggered setting, the `PanelMatch` estimator aligns with the not-yet-treated version of `CSDID` (without covariate adjustment). We delve into details on the connections between these three estimators in the SM.

All DID extensions are built using local, 2×2 DIDs, and their assumptions reflect this. Specifically, they each rely on a form of the PT assumption—that is, the expected changes

in untreated potential outcomes from one period to the other are equal between the treated and the chosen comparison groups. We defer readers to the SM for a fuller account of each method's assumptions.

3.3. The Imputation Method

Imputation estimators do not explicitly estimate local DIDs. Instead, they take the difference between the observed outcome and an imputed counterfactual outcome for each treated observation. The connection to the TWFE model is in the functional form assumption used to impute counterfactual outcomes. Specifically, an imputation estimator first fits a parametric model for the potential outcome under control $Y_{i,t}(0)$—in our case, $Y_{i,t}(0) = X_{i,t}'\beta + \alpha_i + \xi_t + \varepsilon_{i,t}$—using only control observations $\{(i,t) : D_{i,t} = 0\}$. It is also through this outcome model that one can adjust for time-varying covariates. Then, it imputes $Y_{i,t}(0)$ for all treated observations $\{(i,t) : D_{i,t} = 1\}$ using the estimated parameters. Finally, it estimates the individual treatment effect, $\tau_{i,t}$, for each treated observation (i,t) by calculating the difference between the observation's observed outcome $Y_{i,t} = Y_{i,t}(1)$ and its imputed counterfactual outcome $\hat{Y}_{i,t}(0)$. Inference for the estimated $\hat{\tau}_{i,t}$ is possible, although uncertainty estimates need to be adjusted to account for the presence of idiosyncratic errors (e.g., Bai and Ng, 2021). BJS (2024) and LWX (2024) each propose estimators in this category. Each paper proposes a more general framework that nests many models, including TWFE. The latter also introduces several specific imputation estimators, one of which uses the TWFE model, and the authors refer to the resulting estimator as the fixed effects counterfactual estimator, or `FEct`.

Although DID extensions and imputation methods rely on slightly different identification assumptions, these assumptions usually lead to similar observable implications. Researchers commonly use the presence or absence of pretrends to judge how plausible the PT assumption is. In the classic two-group setting, if there are data from multiple pre-treatment periods,

researchers can plot the time series of average outcomes of each group and visually inspect whether they indeed trend together. The intuition is that if PT holds and the average outcome trends of the treated and control groups are indeed parallel in pre-treatment periods when $Y(0)$'s are observed for all units, then it is plausible that PT also holds in the post-treatment periods, when $Y(0)$'s are no longer observable for units in the treatment group. Conversely, differential trends in the pre-treatment periods should make us suspicious of PT. In more complex settings or where we wish to control for observed confounders, researchers often use dynamic estimates before and after the onset of treatment, τ_l, to construct so-called "event-study plots" to judge the presence of pretrends. If PT holds, then pre-treatment dynamic estimates should be around zero. We provide a more thorough discussion and an example of the event-study plot in the next section when we introduce our procedure.

3.4. Choice of Estimators

In general, we believe the credibility of identifying assumptions is more important than the choice of estimator. After all, in the staggered setting, when assumptions hold, `IW`, `CSDID`, `DID_multiple`, `PanelMatch`, and the imputation estimator all converge to the same or a similar estimand. However, there are a few reasons to favor the imputation estimator. First, it can handle complex settings, including those with treatment reversal—which account for over half of the studies in our sample—and can accommodate time-varying covariates, additional fixed effects, and unit- or group-specific time trends commonly seen in social science research. The imputation estimator connects to TWFE through a shared outcome model, and thus any of the aforementioned modifications to the outcome model can be directly mirrored. DID extensions relates to TWFE through their shared connection to DID in the two-group, two period setting. Classic DID's inability to naturally accommodate these complexities limits DID extensions on this front. Just like TWFE, the imputation estimator risks misspecification bias, and adding more redundant terms may significantly increase variance.

However, we still consider the added flexibility to be an advantage. Second, imputation estimators are the most efficient under homoskedastic errors (BJS, 2024).[11] Moreover, by using the average of all pre-treatment periods as the reference point rather than a single pre-period, as the default in DID extensions, they provide greater power in hypothesis testing for pretrends. The main drawback of imputation estimators is that their current implementations (either `FEct` or `DID_impute`) do not allow for automated adjustment of time-invariant covariates, an advantage offered by `CSDID` and `PanelMatch`. Adjusting for pre-treatment characteristics can improve credibility of research, as conditional PT may be more plausible than the unconditional one (Sant'Anna and Zhao, 2020).

4. Data and Procedure

Next, we assess the robustness of empirical findings from causal panel analyses in political science and compare results obtained using the different methods we have discussed. We will explain our sample selection rules, describe standard practices in the field, and outline our reanalysis approach. Readers can find a more detailed explanation of our sample selection criteria and replication and reanalysis procedure in Section A.2 of the SM.

4.1. Data

Our replication sample comprises studies from three leading political science journals, *APSR*, *AJPS*, and *JOP*, published over a recent 7-year span from 2017 to 2023. We initially include all studies, including both long and short articles, that employ panel data analyses with a binary treatment as a key component of their causal argument, resulting in a total of 102 studies. After a careful review, as explained in footnote 1, we find that 64 studies employ a

[11] In Section A.3 of the SM, we demonstrate that the imputation estimator tends to yield larger z-scores, based on data from our sample. In the majority of cases, the imputation estimator has a smaller standard error, and the difference can be especially dramatic for `IW` and the never-treated version of `CSDID`, which often discard the vast majority of untreated observations.

TWFE model similar to Equation (1). We then attempt to replicate the main results of these 64 studies and are successful in 49 cases (76.6%). Though a significant proportion of studies failed to replicate, we note that the success rate is still higher than that of Hainmueller, Mummolo and Xu (2019) at 55% and is comparable to that of Lal et al. (2024) at 67%. We credit this to the new replicability standards set by journals. A detailed explanation of how we select the "main model" is provided in Section A.3 of the SM. Table 2 depicts the distribution of successful replications, along with reasons for replication failures, across the various journals.

Table 2. Sample Selection and Replicability of Qualified Studies

Journal	All	TWFE (attempted)	Incomplete data	Replication error	Replicable	Success rate%
APSR	22	13	2	1	10	76.9
AJPS	31	21	3	3	15	71.4
JOP	49	30	6	0	24	80.0
Total	102	64	11	4	49	76.6

4.2. Settings and Common Practices

Table 3 presents an overview of the standard practices and settings in the studies that we successfully replicated. The majority of studies in our sample (67.3%) use the DID design/method/approach to justify the use of the TWFE model, while the remaining studies advocate for the model's ability to exploit "within" variations in the data. Out of the 49 studies, nine (18.4%) employ a classic block DID setting, which includes two-group, two-period designs (three studies) and multi-period block DID designs (six studies). Thirteen studies (26.5%) use a staggered DID design, while the remaining 27 studies (55.1%) fall into the "general" category, meaning they allow for treatment reversals. Except for five, all studies feature a continuous outcome of interest. Most use cluster-robust SEs or panel-corrected SEs (Beck and Katz, 1995), and eight studies employ bootstrap procedures for estimating uncertainties. A subset of studies explore alternative model specifications by

adding lagged dependent variables (eight studies), unit-specific linear time trends (fifteen studies), and higher-than-unit-level time trends (one study). Notably, 32 studies use some type of visual inspection—either average outcomes over time, event-study plots, or both—to evaluate the plausibility of PT. Four studies published in 2023 (33%) employ HTE-robust estimators, compared to none before 2023, indicating rapid adoption of these methods. Of these, two use `CSDID`, one `PanelMatch`, and one the imputation estimator.

Table 3. Settings and Common Practice

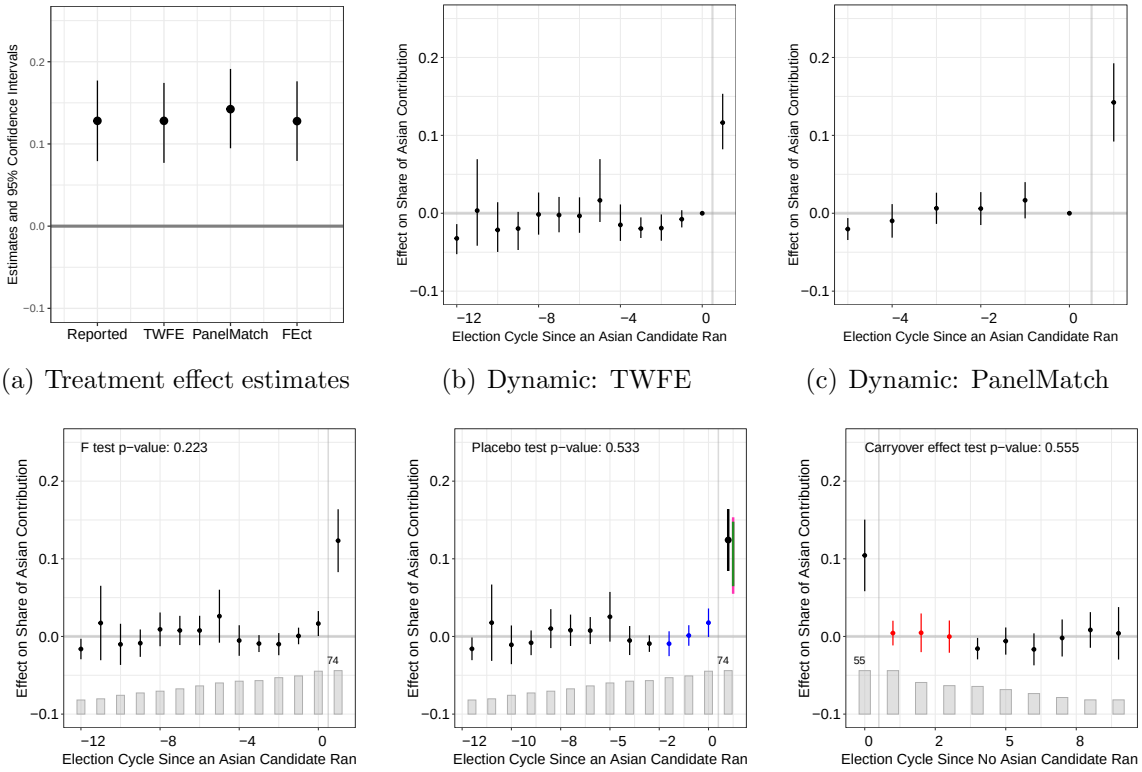
Motivations for TWFE			*Variance Estimator*		
"Difference-in-differences"	33	67.3%	Cluster-robust SE or PCSE	48	98.0%
"Within" variations	16	32.7%	Cluster-bootstrapped procedures	8	16.3%
Treatment setting			*Variants in specifications*		
Classic 2×2 DID	3	6.1%	Lagged dependent variables	8	16.3%
Multi-period block DID	6	12.2%	Higher-than-unit-level time trends	1	2.0%
Staggered DID	13	26.5%	Unit-specific linear time trends	15	30.6%
General	27	55.1%			
			Data visualization		
Outcome variable			Group average outcomes	19	38.8%
Continuous	44	89.8%	Event-study plots	23	46.9%
Binary	5	10.2%	Neither	17	34.7%

4.3. Procedure

We use data from Grumbach and Sahn (2020) to illustrate our process for replication and reanalysis. The authors investigate coethnic mobilization by minority candidates during U.S. congressional elections. To simplify our analysis, we focus on the impact of the presence of an Asian candidate on the proportion of general election contributions from Asian donors. To begin, we aim to understand the research setting and data structure. We visualize the patterns of treatment and outcome variables using plots, which are shown in the SM. In this application, treatment reversals clearly take place. Some data are missing (due to redistricting), but the issue does not seem to be severe. We record important details such as the number of observations, units, and time periods, the type of variance estimator, and

other specifics of the main model. Next, we replicate the main finding, employing both the original variance estimator and a cluster-bootstrap procedure.

FIGURE 2. REANALYSIS OF GRUMBACH AND SAHN (2020)



(a) Treatment effect estimates (b) Dynamic: TWFE (c) Dynamic: PanelMatch

(d) Dynamic: Imputation (e) Placebo & sensitivity tests (f) Test for carryover effects

Note: Reanalysis of data from Grumbach and Sahn (2020). Subfigures (a): Treatment effect or ATT estimates from multiple methods; subfigures (b)-(d): event-study plots using TWFE, `PanelMatch`, and the imputation estimator (`FEct`); Subfigure (e)-(f): results from the placebo test (and robust CS) and test for carryover effects using `FEct`—the blue points in (e) and red points in (f) represent the holdout periods in the respective tests. In (e), the green and pink bars represent the 95% robust CSs when $\bar{M} = 0$ and $\bar{M} = 0.5$, respectively. CIs in all subfigures—excepted for the reported estimate in (a)—are produced by bootstrap percentile methods.

We then re-estimate the ATT and dynamic treatment effects using estimators discussed in the previous section. For staggered adoption treatment cases, we apply seven estimators: TWFE (with always treated units removed for easier comparisons with other estimators), the imputation estimator (`FEct`), `PanelMatch`, `DID_multiple`, `StackedDID`, `IW`, and `CSDID` (both not-yet-treated and never-treated versions). For applications with treatment reversals

like Grumbach and Sahn (2020), we use the first three estimators only.[12] The comparison between the TWFE estimate and the other estimates sheds light on whether original findings are sensitive to relaxing the constant treatment effect assumption. Figures 2(a)-(d) show the results from this example. The similarity between estimates for the ATT in panel (a) suggests that the original finding is robust to the choice of estimators. The event-study plots from HTE-robust estimators in panels (c) and (d) are broadly consistent with the event-study plot from TWFE in panel (b).

Next, we conduct diagnostic tests based on the imputation estimator, including the F test and the placebo test, to further assess the plausibility of PT and, in applications with treatment reversal, the no-carryover-effect assumption. We use the imputation estimator because it is applicable across all studies in our replication sample, can incorporate time-varying covariates, and remains highly efficient. Figures 2(d)–(f) show the results from the F test, placebo test, and test for no carryover effects on our running example, respectively. Both a visual inspection and the formal tests suggest that PT and no-carryover-effect assumptions are quite plausible.

Finally, we compute the robust confidence sets proposed by Rambachan and Roth (2023), which account for potential PT violations when testing the null hypothesis of no post-treatment effect. Specifically, we employ the relative magnitude restriction, with two modifications to accommodate the imputation method. First, we use estimates from the placebo test to ensure that benchmark pre-treatment estimates are obtained using the same approach as post-treatment ATT estimates. This alignment prevents potential asymmetry in testing and treatment effect estimation (Roth, 2024). Second, since the imputation method does not rely on a single reference period, we explicitly incorporate the placebo estimate from the last

[12] A recent development of `DID_multiple`, DIDmultiplegtDYN, allows for the estimation of dynamic and long-term effects (de Chaisemartin and D'Haultfœuille, 2024). It defines an estimand that accounts for carryover effects and targets the first onset of the treatment. To remain consistent with the imputation framework, which allows multiple treatment onsets, we apply `DID_multiple` only to cases featuring staggered treatment timing.

pre-treatment period ($\hat{\delta}_0$) to account for deviations of post-treatment estimates from earlier reference periods. Mathematically, we decompose each dynamic estimate, μ_t, into the true treatment effect, τ_t, and a trend (bias) component, δ_t, such that: $\mu_t = \tau_t + \delta_t$. Our modified relative magnitude restriction then requires that, for all $t \geq 0$,

$$\left|\delta_{t+1} - \delta_t\right| \leq \bar{M} \cdot \max_{s \in \mathcal{P} \setminus \{0\}} \left|\delta_{s+1} - \delta_s\right|, \tag{3}$$

where $\mathcal{P}$ is the set of placebo periods. In our application, we set $\mathcal{P} = \{-2, -1, 0\}$, so the maximum violation among placebo periods is: $\max\left\{\left|\delta_0 - \delta_{-1}\right|, \left|\delta_{-1} - \delta_{-2}\right|\right\}$.

When $\bar{M} = 0$, the relative magnitude restriction reported in Equation (3) implies that $\delta_t = \delta_0$ for all $t > 0$, meaning that the PT violation remains fixed at the same level as in the last pre-treatment placebo period. In this case, the robust CS obtained at $\bar{M} = 0$ acts as a debiased confidence interval, using the placebo estimate from the last pre-treatment period as the benchmark for bias. Allowing $\bar{M} > 0$ permits PT violations to vary over time, but constrains the change in magnitude of violations between consecutive post-treatment periods to remain within $\bar{M}$ times the largest consecutive discrepancy observed during the placebo periods.

Rambachan and Roth (2023, 2653) suggest using $\bar{M} = 1$ as a "natural benchmark" when the number of placebo periods is roughly equal to the number of post-treatment periods, treating any potential PT violations as no worse than those already observed.[13] In our reanalysis, we first construct 95% robust confidence sets for each post-treatment dynamic effect and the ATT at $\bar{M} = 0$ and $\bar{M} = 0.5$. Figure 2(e) illustrates these robust confidence sets for the estimated ATT using the imputation method. The center of the robust confidence sets is smaller than the point estimate because $\hat{\delta}_0 > 0$. If the confidence sets for $\bar{M} = 0$ does not include zero, as in this case, we conduct a sensitivity analysis by varying $\bar{M}$ over a wider range to determine the "breakdown value" $\tilde{M}$, which is the smallest value of $\bar{M}$

[13] In our setting, the number of post-treatment periods typically exceeds the number of placebo periods, which means the criterion is even more lenient than the authors have suggested.

at which the robust confidence set first includes zero. In Grumbach and Sahn (2020), the breakdown value is $\tilde{M} = 2.5$, which means that the estimated coethnic mobilization effect remains statistically distinguishable from zero unless PT violations are more than 2.5 times the largest discrepancy observed during the placebo periods.

Overall, the results from Grumbach and Sahn (2020) appear highly robust, regardless of the choice of point and variance estimators. The PT and no-carryover-effect assumptions seem plausible. The study also has sufficient power to distinguish the ATT from zero, even under potential, realistic PT violations.

5. Systematic Assessment

We perform the replication and reanalysis procedure described above for all 49 studies in our sample. This section offers a summary of our findings, with complete results for each article available in the SM. We organize our results around two main questions: (1) Are existing empirical findings based on TWFE models robust to HTE-robust estimators? (2) Is the PT assumption plausible, and do original findings remain robust to mild PT violations informed by pretrends? We also discuss other issues observed in the replicated studies, including the presence of carryover effects and sensitivity to model specifications.

HTE-robust estimators yield qualitatively similar but highly variable estimates. To examine the impact of the weighting problem caused by HTE associated with TWFE models, we first compare the estimates obtained from the imputation estimator, `FEct`, for all studies to those originally reported. We choose the imputation estimator for the reason mentioned earlier. Most importantly, it is applicable to all studies in our sample, including those with treatment reversals and those with additional time trends. Figure 3 plots the comparison. The horizontal axis represents the originally reported TWFE estimates, and the vertical axis represents `FEct` estimates, both normalized using the same originally reported

SEs. If the point estimates are identical, then the corresponding point should lie exactly on the 45-degree line. Red triangles represent studies where the imputation estimates are statistically insignificant at the 5% level, based on cluster-bootstrapped SEs.

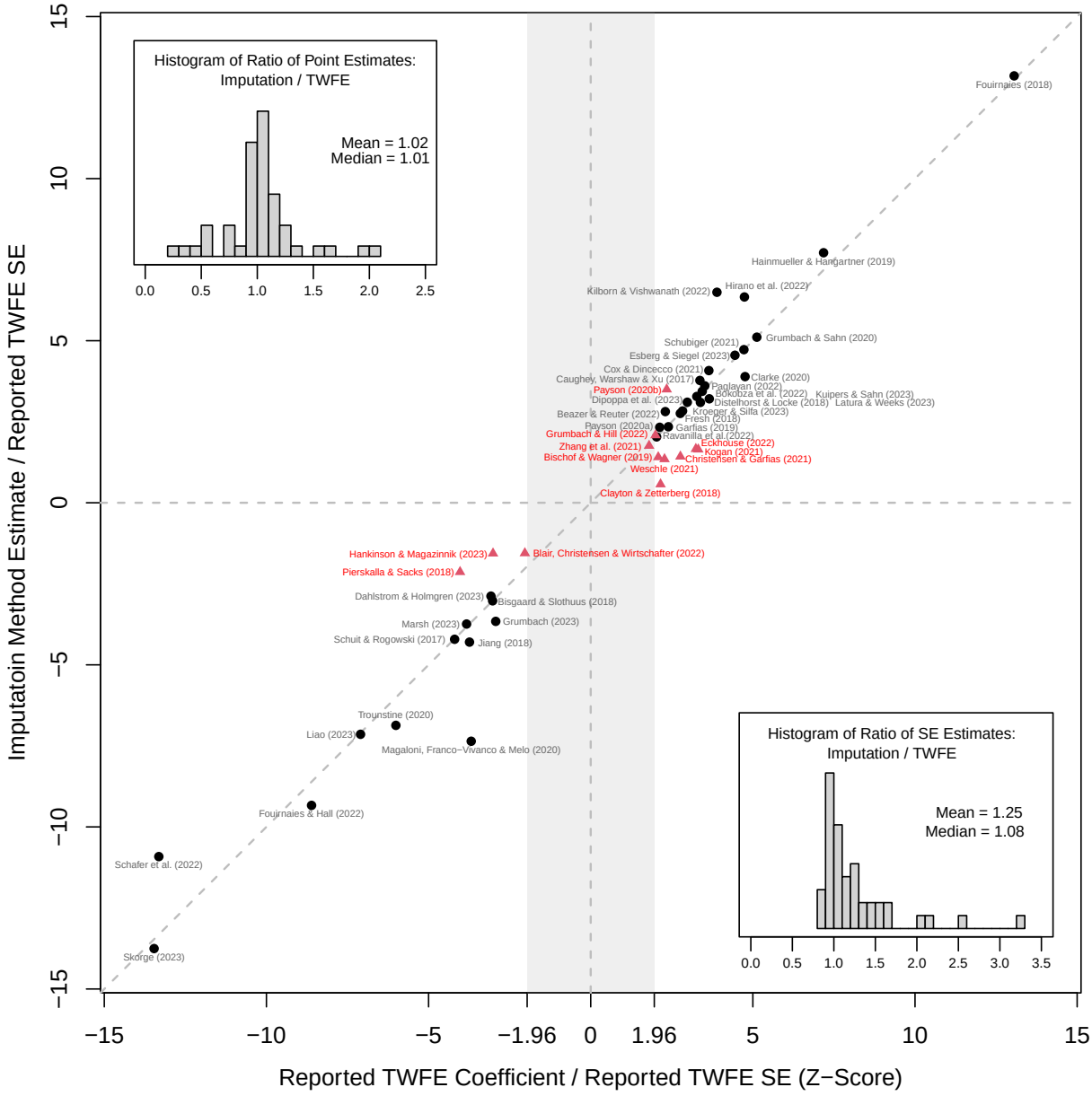
We observe several patterns. First, TWFE coefficients are statistically significant at the 5% level in all but one study, and the absolute values of z-scores for a significant minority of studies cluster around 1.96, indicating possibly a file-drawer problem and potential publication bias. Second, the points largely follow the 45-degree line, with the imputation estimates *always* having the same sign as the original estimates. This suggests that while scenarios where accounting for HTE completely reverses the empirical findings are theoretically possible, they are rare.

However, results sometimes deviate significantly. In the top left corner of Figure 3, we plot the histogram of the ratio of imputation to TWFE estimates. Although the mean and median of the ratio are close to one, at the extremes, we observe imputation estimates as small as one-fourth or as large as more than double the TWFE estimates. The most consequential deviations occur in studies that were originally near the margins of statistical significance. Additionally, we plot the ratio of SE estimates from the imputation method to TWFE in the bottom right corner of Figure 3. The median is 1.08, meaning that in the majority of cases the SE estimate from the imputation method is at least 8% larger than the SE from TWFE. The mean is 1.25, and the distribution is right-skewed; in the extreme, the ratio was almost as high as three. Combined, drops in point estimates and increases in uncertainty lead to the third pattern: When we switch from TWFE to the imputation estimator, the number of studies that are statistically insignificant at the 5% level increases from one to twelve (24%).

If we restrict our attention to eleven studies with staggered treatments, we can broaden our comparison set to include more HTE-robust estimators discussed earlier.[14] Figure 4

[14] Kogan (2021) and Magaloni, Franco-Vivanco and Melo (2020) are excluded because the original spec-

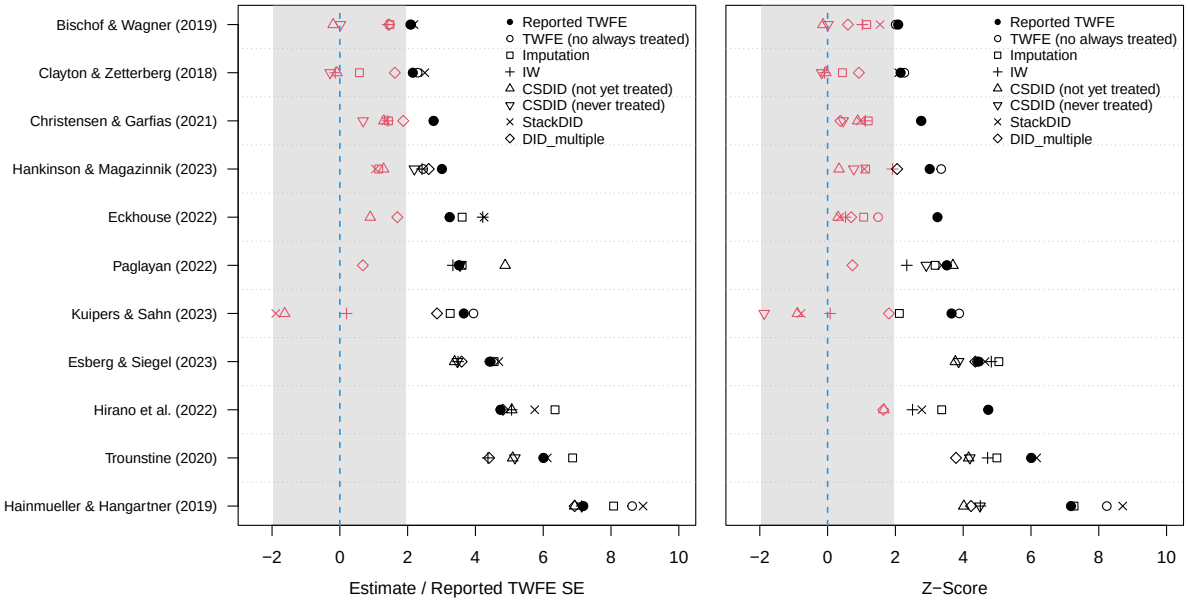
FIGURE 3. TWFE VS. THE IMPUTATION ESTIMATOR: ALL CASES



Note: The above figure compares reported TWFE coefficients with imputation method (`FEct`) estimates. Both estimates for each application are normalized by the same reported TWFE SE. Fouirnaies and Hall (2018) and Hall and Yoder (2022) are close to the 45-degree line but are not included in the figure as their TWFE z-scores exceed 15. Black circles (red triangles) represent studies whose imputation method estimates for the ATT are statistically significant (insignificant) at the 5% level, based on cluster-bootstrapped SEs. The top-left (bottom-right) corners display histograms of the ratio of point (SE) estimates based on the imputation method and TWFE. These plots show that changes in point estimates, combined with the efficiency loss from using the imputation method, contribute to the loss of statistical significance in some studies.

ifications include additional time trends, which are not supported by HTE-robust estimators except the imputation estimator.

FIGURE 4. COMPARISON OF ESTIMATES: THE STAGGERED SETTING



Note: The above figures compare reported TWFE coefficients with estimates from various alternative estimators. In the left panel, all eight estimates for each application are normalized by the same reported SE to highlight changes resulting from the use of alternative estimators. In the right panel, the estimates are divided by their respective bootstrapped SEs. To facilitate visualization, we multiply all estimates by the sign of the reported coefficient. In both figures, black (red) symbols represent estimates that would be statistically significant (insignificant) at the 5% level, assuming they were treated as z-scores. The normalized `CSDID` (never treated) estimate for Kuipers and Sahn (2023), -5.36, falls out of plotting area and is therefore not shown in the left panel. Kogan (2021) and Magaloni, Franco-Vivanco and Melo (2020) are excluded because the authors' original TWFE specifications include unit-specific linear time trends, which are not supported by most HTE-robust estimators except the imputation estimator. Some estimates are missing because of too few never-treated units. `PanelMatch` is excluded because it targets a different estimand.

visually compares the points estimates (left panel) and z-scores (right panel). In the left panel, all point estimates are normalized by the same reported TWFE SEs for each study. In all but three studies, the estimates from all HTE-robust estimators share the same sign, though there is a noticeable amount of variation in the estimated effect size.[15]

As in Figure 3, TWFE does not appear to be systematically upward or downward biased

[15] Although a sign change is observed in Clayton and Zetterberg (2018) when using `CSDID`, `IW`, and `DID_multiple`, these estimates are negligibly small and statistically insignificant. Similarly, the not-yet-treated version of `CSDID` is the opposite sign but miniscule and statistically insignificant in Bischof and Wagner (2019). The estimates from `IW` and `CSDID` are also of the opposite sign for Kuipers and Sahn (2023), and they are of a much larger magnitude. The never-treated version of `CSDID` is also statistically significant.

compared to HTE-robust estimators. Another observation that carries over is that HTE-robust estimators generally require more power to reject the null hypothesis of no effect. In five of the eleven studies, at least four HTE-robust estimates per study are statistically insignificant. The left panel shows that the changes in point estimate alone are often sufficient to render the results statistically insignificant. The comparison of z-scores in the right panel highlights that increased uncertainties can be substantial. Combined with earlier evidence, these findings from the staggered cases suggest that the HTE issue regarding TWFE is empirically significant and warrants careful consideration by researchers.

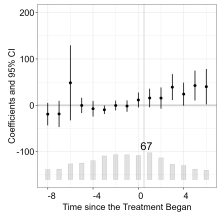
It is worth noting that while only a small fraction of studies in our sample (eight studies, 16.3%) employ a bootstrap procedure to estimate SEs or CIs, the widely practiced cluster-robust SE typically performs adequately. This is because the number of units (clusters) is generally large, with a median of 317. However, ten studies have fewer than 50 units; among them, two studies that were significant at the 5% level using cluster-robust SEs fell below this threshold when using cluster-bootstrapped SEs, both of which were already marginally significant. We provide comparisons of reported, cluster-robust, and cluster-bootstrapped SEs in the SM.

Relaxing PT renders most studies unable to reject the null. Although the recent methodological literature has heavily focused on HTE, PT violations—long recognized as a potential pitfall—remain a primary concern in practice. In Figure 5, we present the event-study plot based on estimates from the imputation estimator for each study in our sample. We also report the ATT estimates and their bootstrapped SEs. Due to space limitations, we present the event-study plots from other estimators, as well as results from the placebo tests, robust CSs, and sensitivity analyses, in the SM.

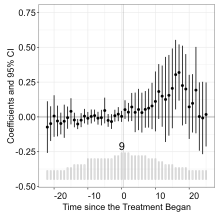
What is concerning is that, in at least six studies (12%), the PT assumption seems highly implausible. In these studies, the dynamic estimates in the pre-treatment periods deviate

Figure 5. Dynamic Treatment Effects w/ Imputation Estimator

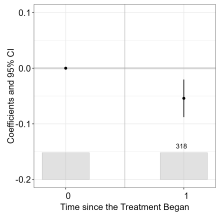
Beazer and Reuter (2022)
ATT: 24.38 (8.79)



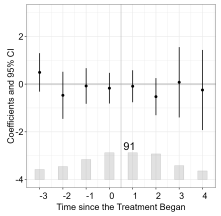
Bischof and Wagner (2019)
ATT: 0.08 (0.07)



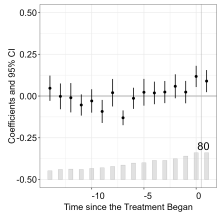
Bisgaard and Slothuus (2018)
ATT: -0.05 (0.02)



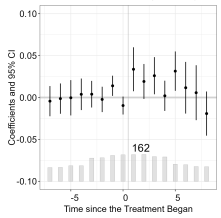
Blair, Christensen and Wirtschafter (2022)
ATT: -0.50 (0.40)



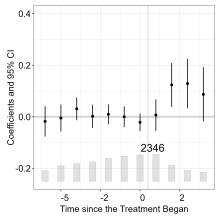
Bokobza et al. (2022)
ATT: 0.11 (0.03)



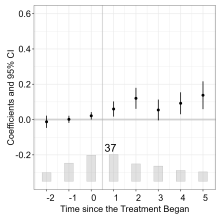
Caughey, Warshaw and Xu (2017)
ATT: 0.014 (0.005)



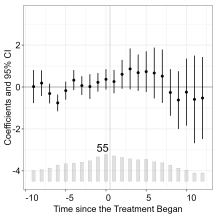
Christensen and Garfias (2021)
ATT: 0.04 (0.04)



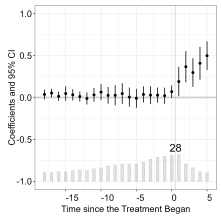
Clarke (2020)
ATT: 0.08 (0.03)



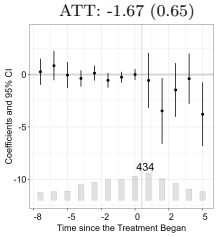
Clayton and Zetterberg (2018)
ATT: 0.19 (0.44)



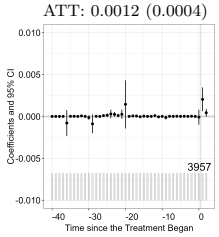
Cox and Dincecco (2021)
ATT: 0.44 (0.11)



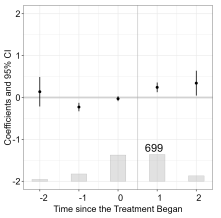
Dahlström and Holmgren (2023)
ATT: -1.67 (0.65)



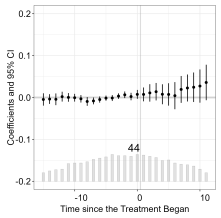
Dipoppa, Grossman and Zonszein (2023)
ATT: 0.0012 (0.0004)



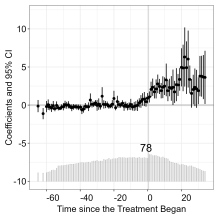
Distelhorst and Locke (2018)
ATT: 0.15 (0.05)



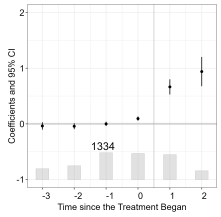
Eckhouse (2022)
ATT: 0.015 (0.014)



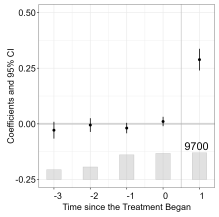
Esberg and Siegel (2023)
ATT: 2.89 (0.57)



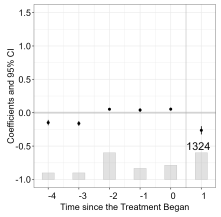
Fouirnaies (2018)
ATT: 0.87 (0.11)



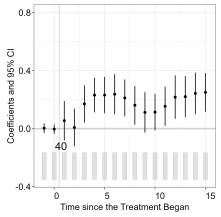
Fouirnaies and Hall (2018)
ATT: 0.29 (0.02)



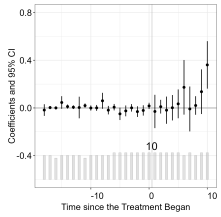
Fouirnaies and Hall (2022)
ATT: -0.26 (0.03)



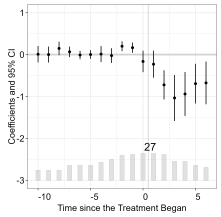
Fresh (2018)
ATT: 0.17 (0.06)



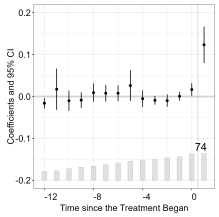
Garfias (2019)
ATT: 0.07 (0.03)



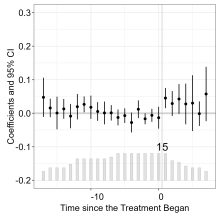
Grumbach (2023)
ATT: -0.57 (0.15)



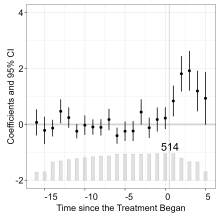
Grumbach and Sahn (2020)
ATT: 0.13 (0.03)



Grumbach and Hill (2022)
ATT: 0.033 (0.019)



Hainmueller and Hangartner (2019)
ATT: 1.51 (0.21)



Hankinson and Magazinnik (2023)
ATT: -0.51 (0.45)

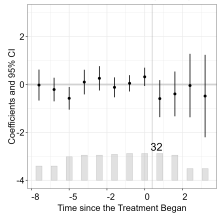


Figure 5. Dynamic Treatment Effects w/ Imputation Estimator (Cont.)



Note: We report the estimated ATT and corresponding bootstrap SEs (in parentheses) using `FEct`. For Skorge (2023), we use an "exit" plot becuase all treated units receives the treatment in the first period.

substantially from zero compared with those in the post-treatment period, and the F test rejects the null.[16] For the remaining studies, the CIs of pre-treatment estimates mostly cover zero, and the F test and placebo test do not reject. However, this could be simply due to a lack of statistical power. Therefore, to assess the robustness of the findings, we need additional tools that simultaneously account for both the estimated pretrend and statistical power. The sensitivity analysis proposed by Rambachan and Roth (2023) addresses this issue. Therefore, we conduct such an analysis with a modified relative magnitude Equation (3) for 42 studies that have at least three pre-treatment periods.[17]

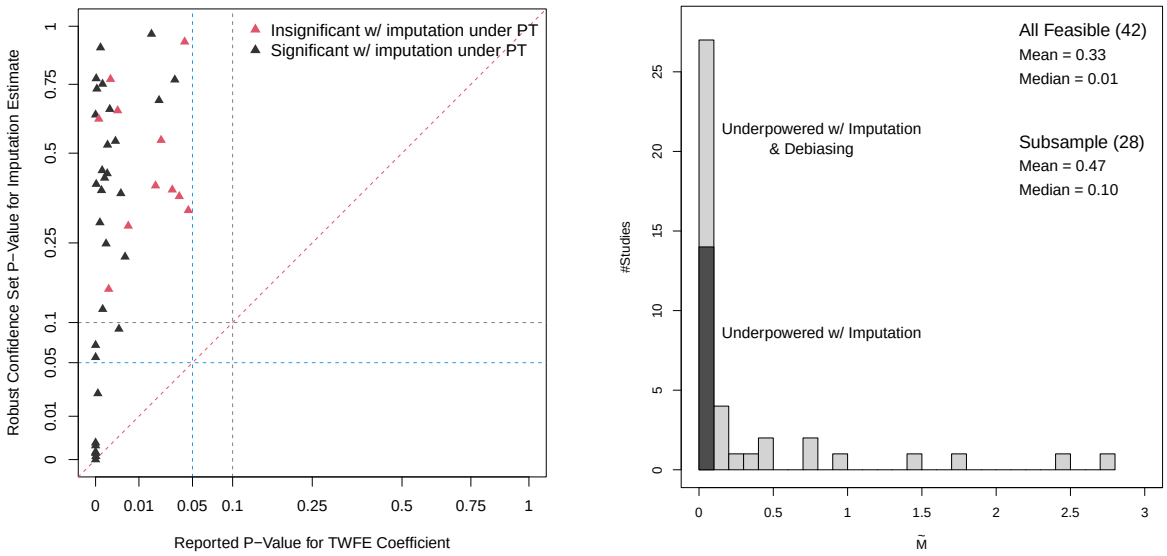
Figure 6(a) shows that when we construct a robust confidence set with $\bar{M} = 0.5$, the null hypothesis of no effect is rejected at the 5% level in only eight (19%) of the 42 studies. Many studies that were robust to HTE now appear underpowered. Figure 6(b) displays the distribution of breakdown values $\tilde{M}$. The spike at 0 consists of two types of studies: those that are statistically insignificant with the imputation estimator when three placebo periods are considered (dark gray) and those that become statistically insignificant due to debiasing using the placebo estimate $\hat{\delta}_0$. In other words, many results are not robust when we take into account $\hat{\delta}_0$ even without a relative magnitude shift. Among the 42 studies, the median is close to 0 and the mean is 0.33. Focusing only on the studies that remain statistically significant with the imputation estimator, the median and mean are still as low as 0.10 and 0.47, respectively.

These results suggest that although most statistically significant findings are robust to the imputation estimator, in the vast majority of these studies, accounting for potential PT violations—even very mild ones based on estimated pretrends during the placebo periods—prevents us from rejecting the null. In other words, in most studies, the ATT estimates

[16] There are other cases where the F test rejects, but we do not consider them highly problematic because, with a large sample size, a confounder inconsequential to the ATT estimate can still produce a small p-value in the F test. This is why the sensitivity analysis approach is particularly useful.

[17] For studies with only three pre-treatment periods, we set the number of placebo periods in their placebo tests to 2.

FIGURE 6. ALLOWING PT VIOLATIONS WITH ROBUST CONFIDENCE SETS



(a) Robust infernece under PT violations (b) Distribution of breakdown threhsold $\tilde{M}$

Note: The above figures present findings from the sensitivity analysis. The sample consists of 42 studies with more than three pre-treatment periods, allowing for such an analysis. Subfigure (a) displays the p-values of partially identified ATT estimates using the imputation method under restricted relative magnitude PT violations with $\bar{M} = 0.5$, compared to reported p-values for TWFE coefficients assuming PT. A square root scale is used to facilitate visualization. Black (red) triangles represent studies that are statistically significant (insignificant) at the 5% level when using the imputation method under PT. Subfigure (b) shows a histogram of $\tilde{M}$, the breakdown values of $\bar{M}$; the dark gray bar represents studies whose ATT estimates are statistically insignificant at the 5% level when using the imputation method.

are not substantial enough to be differentiated from realistic PT violations or from estimation and sampling errors.

Other issues. Our reanalysis highlights several additional issues. First, the presence of missing values is widespread. Although most methodological work presumes balanced panels without missing data, in reality, many empirical studies encounter varying degrees of data missingness. Substantial differences in results for estimators that are numerically identical in balanced panels suggest that such violations may have important implications. During replication, we generate plots that display the patterns of treatment status for each study (Mou, Liu and Xu, 2023). Based on these plots, we also see that in some studies the pattern of missingness is either seemingly nonrandom or extremely prevalent, which weakens our

confidence in the respective empirical findings.

Second, we perform tests for carryover effects for all studies with treatment reversals. If this test is rejected, it suggests that the treatment effects persist beyond the treatment periods. Among 27 studies, five reject the null hypothesis of no carryover effects at 5%. Part of the concern is that the imputation method and DID extensions will use control observations from previously treated units to fit the potential outcome model or as comparisons for treated observations, and if there are carryover effects, then the comparisons will become tainted. LWX (2024) note that the presence of carryover effects for a limited number of periods is less concerning, as researchers can recode treatment to persist for some time after a unit transitions out of treatment. Despite its prevalence, we observe that carryover effects do not substantially alter the findings in most studies. Specifically, in six studies that reject the null of no carryover effects, when we exclude two periods after the treatment switches back to zero, the ATT estimates remain similar in magnitude, and statistically significant results remain significant (Figure A8 in the SM). Nevertheless, we recommend that researchers make it a practice to check for potential carryover effects, considering the low cost of conducting such tests and adjustments.

Finally, many findings are sensitive to model specifications. Some studies that we exclude from our sample employ one-way fixed effects or fixed effects at a level different from that at which treatment is assigned. Many of these findings do not hold when we reanalyze them using a TWFE model. We should clarify that this does not imply that the original results are wrong; rather, these models implicitly operate under different identifying assumptions, and there is substantial variation in how much consideration authors give to this point. Some studies do not provide a rationale for their choice to use one-way fixed effects, while others explicitly outline the type of unobserved confounders they intend to control for. In one instance, the authors inaccurately label their specification as a DID design. The TWFE and DID estimators are generally not equivalent, and we emphasize that this difference becomes

even more pronounced when fixed effects are not assigned at the level of treatment. In such cases, a TWFE specification does not correspond to even a broadly defined DID research design.

Summary. In Table 4 below, we summarize the main findings of our reanalysis. The numbers represent the proportion of studies in a given journal or across all journals that fall under the respective category (hence, a smaller number is better). The first two rows relate to the ratio of the imputation estimate to the TWFE estimate, which proxies the consequences of the weighting problem due to HTE. Across all journals, this ratio is less than 0.8 in 18% of studies and less than 0.5 in 6% of studies, indicating that using an HTE-robust estimator does sometimes have a substantively significant impact and is important to use at least as a robustness check. The third and fourth rows indicate the proportion of studies with evidence of PT violations based on the F test and placebo tests, respectively. Across all journals, around a quarter of studies reject the null hypothesis at the 5% level using the F test, with slightly fewer rejecting when using the placebo test due to the loss in power from excluding data from the placebo periods. Note, though, that failure to reject

Table 4. Summary of Findings

	APSR	*AJPS*	*JOP*	All	n
Imputation to TWFE Ratio < 0.8	0.00	0.13	0.30	0.18	49
Imputation to TWFE Ratio < 0.5	0.00	0.00	0.13	0.06	49
F test reject null	0.30	0.21	0.30	0.27	44
Placebo test reject null	0.40	0.29	0.15	0.25	44
TWFE *not* reject null	0.00	0.00	0.04	0.02	49
Imputation *not* reject null	0.00	0.13	0.44	0.25	49
Imputation *not* reject null w/ $\bar{M} = 0$	0.40	0.23	0.74	0.50	42
Imputation *not* reject null w/ $\bar{M} = 0.5$	0.70	0.77	0.90	0.81	42

Note: Entries (except in the last column) are proportions of studies satisfying each set of conditions. A null is deemed rejected if $p < 0.05$. Five studies with a single pre-treatment period are not included in the summary statistics for the F test and placebo test. Seven studies are not included in the last two rows due to too few pre-treatment periods.

can result from insufficient power and is not sufficient to support that the PT holds.

The fifth and sixth rows display the proportion of studies in which the null hypothesis of no effect is not rejected using the TWFE and imputation estimators, respectively. When TWFE is employed, this occurs in only one study (2%); however, when the imputation estimator is used, this number increases to 24%, suggesting that many studies in our sample are potentially underpowered with an HTE-robust estimator. The last two rows show that in 50% and 81% of the studies, the robust confidence sets for the ATT include zero with $\bar{M} = 0$ (debiasing only) and $\bar{M} = 0.5$ (debiasing plus restricted relative magnitude in biases), respectively. The comparison of results in the last four rows highlights that the main source of fragility in the existing literature is potential PT violations, rather than concerns of HTE.

Our reanalysis is not meant to criticize existing studies, many of which were conducted before recent methodological advances. In fact, we have observed rapid adoption of these new methods and greater statistical power in more recently published studies. However, we want to emphasize two key points: (1) many studies do not adequately assess the plausibility of PT, the key identifying assumption of their research designs; and (2) given recent methodological developments, causal panel analysis under PT requires significantly more statistical power to account for HTE and potential PT violations than previously believed.

6. Recommendations

Based on the findings of the reanalysis, we provide the following recommendations.

Research design is the key. An important component of a strong research design is a clear understanding of how treatment is assigned. The PT assumption, which both TWFE models and most modern methods rely on, is silent on the assignment mechanism. As a result, many researchers assume that these methods can be applied when the assignment mechanism is unknown and that the absence of a pretrend is sufficient to make PT credible.

Another perspective suggests that (quasi-)randomization is required for PT to be credible (e.g., Kahn-Lang and Lang, 2020). Our view aligns more closely with the latter, but is less stringent. We argue that for PT to be credible, researchers must justify the following assumption:

$$\Delta_{s,t} Y_{i,t}(0) \perp\!\!\!\perp D_{i,t}, \ \forall s, t$$

where $\Delta_{s,t} Y_{i,t}(0) = Y_{i,t}(0) - Y_{i,s}(0)$ is the before-and-after difference in untreated potential outcomes across any two periods (assuming no covariates, though this naturally extends to include them). This assumption is slightly stronger than PT, which only assumes mean independence, but is more intuitive. It demonstrates both the strength and limitation of causal panel analysis under PT: while the panel structure helps account for time-invariant unobserved confounding through before-and-after differences, the introduction of the treatment must act as a shock in so much as it is orthogonal to the evolution of untreated potential outcomes; hence, any dynamic relationships between past outcomes or covariates and treatment are ruled out. In other words, under PT, a strong causal panel design still requires some (quasi-)random element in treatment assignment. Importantly, the research design issues cannot be resolved simply by applying the novel estimators surveyed in this article.

Inspecting raw data helps spot obvious issues. The research design phase should also include inspection of the raw data. We encourage researchers to plot the data at hand to better understand patterns of the treatment status, missingness, and outliers (Mou, Liu and Xu, 2023). Treatment status should vary both by unit and time. If the majority of variation occurs over time (across units) with little or no variation between units at any given time period (or across time within a given unit), the TWFE estimand will be likely dominated by impermissible comparisons and thus susceptible to larger biases. Moreover, HTE-robust estimators will estimate the treatment effect using very little data and thus be underpowered. Equally important is the need for researchers to understand the degree and possible origins

of data missingness prior to initiating statistical analysis. If missingness does not seem to be random, or if it is too prevalent, leaving insufficient meaningful variation in the data, researchers should consider halting the analysis at this stage. Just as in the cross-sectional case, plotting the raw data can also help researchers to spot outliers and highly skewed distributions, which may require additional preprocessing. At this stage, researchers can also trim the data to make the units in comparison more similar in terms of pre-treatment or time-invariant covariates (Sant'Anna and Zhao, 2020).

When estimates diverge, understand why. At the estimation stage, we recommend using at least one HTE-robust estimator alongside a benchmark TWFE model. While TWFE is often more efficient, its constant treatment effects assumption is too restrictive in many contexts. As shown in this article, TWFE does not produce systematic upward or downward bias compared to the imputation estimator, but it can severely bias causal estimates in individual cases. We recommend the imputation estimator in most settings primarily for its flexibility, though other estimators also have their advantages. The more critical question, however, is why results differ between estimators when they do. If TWFE and an HTE-robust estimator diverge, it could be due to HTE or PT violations. If HTE-robust estimators themselves diverge, it is often because the data are too sparse, leading to high variability, or the PT assumption fails differentially, causing estimators weighting control units differently to produce varying estimates. Plotting raw data and using diagnostics (such as the Goodman-Bacon decomposition) typically clarifies these issues. We also recommend keeping the benchmark TWFE model for its transparency.

For uncertainty estimates, researchers should use cluster-robust SEs when the number of clusters is large (e.g., exceeds 50) and opt for cluster-bootstrap or cluster-jackknife procedures when the number of clusters is relatively small. The clustering level should match the higher of either the time-series units or the level of treatment assignment. This follows the rule

of thumb to cluster at the level of potential outcome input index, taking into account both treatment assignment and potential temporal spillover (Fu, Samii and Wang, 2025). For novel HTE-robust estimators, cluster-bootstrap or jackknife is generally safer than relying on various analytical SEs. In the SM, we show that cluster-bootstrapped SEs are typically larger than analytically derived SEs for five HTE-robust estimators using data from our sample.

Conduct diagnostics to assess key assumptions and robustness of findings. With a clear research design, researchers should critically evaluate key identification assumptions and test the robustness of findings when these assumptions are violated. Event-study plots, available for TWFE and most HTE-robust estimators, are valuable tools for assessing whether the no-anticipation and PT assumptions hold. We recommend creating an event-study plot using the chosen estimator(s), followed by both visual inspection and statistical tests to assess how plausible the PT assumption is. Importantly, the absence of statistical significance in pretrend coefficients should not be taken as conclusive evidence for the validity of PT. To avoid the conditional inference problem, we recommend performing a sensitivity analysis with robust confidence sets across different values of $\bar{M}$, regardless of the pretest results. As shown in this article, such tests require more statistical power than rejecting the null that the average treatment effect is zero. Researchers should, therefore, allocate sufficient statistical power for these diagnostic tests during the research design phase.

Panel data provide valuable opportunities for social scientists to tackle complex causal questions; however, these data, especially when analyzed under PT, present distinct challenges. Our findings are not intended to dissuade researchers from employing PT-based research designs in causal panel analysis. Rather, our aim is to guide researchers in conducting their analyses more transparently and credibly. To facilitate this, we have integrated all the procedures described in this paper into the open-source `R` package `fect`, and we offer

detailed tutorials for these methods.

Data Availability Statement

Research documentation and/or data that support the findings of this study are openly available at the *American Political Science Review* Dataverse: https://doi.org/10.7910/DVN/9RJFZF.

Conflict of Interest

The authors declare no ethical issues or conflicts of interest in this research.

Ethical Standards

The authors affirm that this research did not involve human participants.

References

Arkhangelsky, Dmitry and Guido Imbens. 2024. "Causal Models for Longitudinal and Panel Data: A Survey." *The Econometrics Journal* 27(3):C1–C61.

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

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.

Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon and Pedro HC Sant'Anna. 2025. "Difference-in-Differences Designs: A Practitioner's Guide." *arXiv preprint arXiv:2503.13323* .

Baker, Andrew C, David F Larcker and Charles CY Wang. 2022. "How Much Should We Trust Staggered Difference-in-Differences Estimates?" *Journal of Financial Economics* 144(2):370–395.

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.

Beck, Nathaniel and Jonathan N Katz. 1995. "What to Do (and Not to Do) with Time-Series Cross-Section Data." *American Political Science Review* 89(3):634–647.

Beck, Nathaniel and Jonathan N Katz. 2001. "Throwing Out the Baby with the Bath Water: A Comment on Green, Kim, and Yoon." *International Organization* 55(2):487–495.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119(1):249–275.

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.

Blackwell, Matthew and Adam Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data Under Selection on Observables." *American Political Science Review* 112(2):1067–1082.

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.

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.

Cameron, A Colin, Jonah B Gelbach and Douglas L Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90(3):414–427.

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.

Chiu, Albert, Xingchen Lan, Ziyi Liu and Yiqing Xu. 2025. "Replication Data for: Causal Panel Analysis under Parallel Trends: Lessons from A Large Reanalysis Study." https://doi.org/10.7910/DVN/9RJFZF. Harvard Dataverse.

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, C and X D'Haultfœuille. 2023. "Credible Answers to Hard Questions: Differences-in-Differences for Natural experiments." *Working Paper, SSRN* .

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'Haultfœuille. 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.

Eggers, Andrew C, Anthony Fowler, Jens Hainmueller, Andrew B Hall and James M Snyder Jr. 2015. "On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races." *American Journal of Political Science* 59(1):259–274.

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.

Fu, Jiawei, Cyrus Samii and Ye Wang. 2025. "Inference for Group Interaction Experiments.". Conference Paper, presented at thh 2024 Annual Conference of the American Political Science Association.

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.

Green, Donald P, Soo Yeon Kim and David H Yoon. 2001. "Dirty Pool." *International Organization* 55(2):441–468.

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.

Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27(2):163–192.

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.

Kahn-Lang, Ariella and Kevin Lang. 2020. "The Promise and Pitfalls of Differences-in-Differences: Reflections on *16 and Pregnant* and Other Applications." *Journal of Business & Economic Statistics* 38(3):613–620.

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.

King, Gary. 2001. "Proper Nouns and Methodological Propriety: Pooling Dyads in International Relations Data." *International Organization* 55(2):497–507.

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.

Lal, Apoorva, Mackenzie Lockhart, Yiqing Xu and Ziwen Zu. 2024. "How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on 67 Replicated Studies." *Political Analysis* 32(4):521–540.

Lall, Ranjit. 2016. "How Multiple Imputation Makes a Difference." *Political Analysis* 24(4):414–433.

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.

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. 2022. "Pretest with caution: Event-study estimates after testing for parallel trends." *American Economic Review: Insights* 4(3):305–322.

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

Roth, Jonathan and Pedro HC Sant'Anna. 2023. "When is Parallel Trends Sensitive to Functional Form?" *Econometrica* 91(2):737–747.

Roth, Jonathan, Pedro HC Sant'Anna, Alyssa Bilinski and John Poe. 2023. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." *Journal of Econometrics* 235(2):2218–2244.

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

Sant'Anna, Pedro H C and Jun B Zhao. 2020. "Doubly Robust Difference-in-Differences Estimators." *Journal of Econometrics* 219(1):101–122.

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.

Wang, Ye. 2021. "Causal Inference under Temporal and Spatial Interference." Working Paper, New York University.

Wang, Ye, Cyrus Samii, Haoge Chang and PM Aronow. 2025. "Design-based Inference for Spatial Experiments under Unknown Interference." *The Annals of Applied Statistics* 19(1):744–768.

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.

Wing, Coady, Seth M Freedman and Alex Hollingsworth. 2024. "Stacked difference-in-differences." *Working Paper, NBER* .

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

Xu, Yiqing. 2023. Causal Inference with Time-Series Cross-Sectional Data: A Reflection. In *The Oxford Handbook of Methodological Pluralism in Political Science*, ed. Janet M. Box-Steffensmeier, Dino P. Christenson and Valeria Sinclair-Chapman. Oxford: Oxford University Press chapter 30.

Xu, Yiqing, Anqi Zhao and Peng Ding. 2024. "Factorial Difference-in-Differences." *arXiv preprint arXiv:2407.11937* .

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.

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

A.1.1. DID Extensions for the Staggered Setting

A.1.2. DID Extensions for the General Setting

A.1.3. Imputation Methods for the General Setting

A.1.4. Strict Exogeneity and Parallel Trends

A.1.5. Assumptions for Each Estimator

A.1.6. Proof of Remark A.1

A.1.7. Choice of the Reference Period(s)

A.2. Implementation Details

A.2.1. Summary of Context and Visualization

A.2.2. Point Estimates

A.2.3. Dynamic Treatment Effects & Event Study Plots

A.2.4. Diagnostic Tests

A.2.5. Robust Confidence Set & Sensitivity Analysis

A.3. More Replication Results

A.3.1. Sample Selection Criteria

A.3.2. Replicability

A.3.3. Reported vs TWFE and Imputation Estimates

A.3.4. Inferential Methods

A.3.5. Imputation vs. Other Methods

A.3.6. Placebo Tests & Robust Confidence Sets

A.3.7. Carryover Effects

A.3.8. Summary of Findings

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.

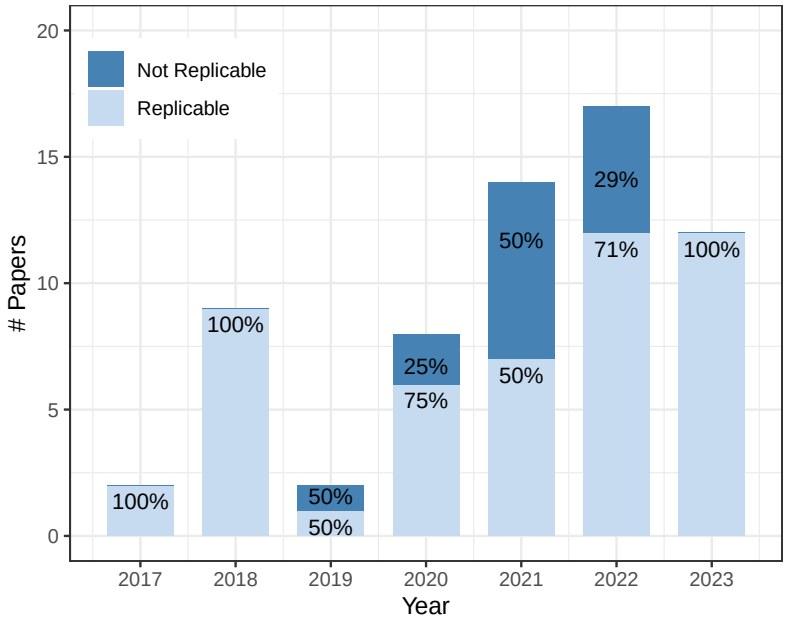
3. We exclude papers that use a regression discontinuity design or an instrumental variables (IVs) design, including Bartik IVs, as their primary identification strategy.

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

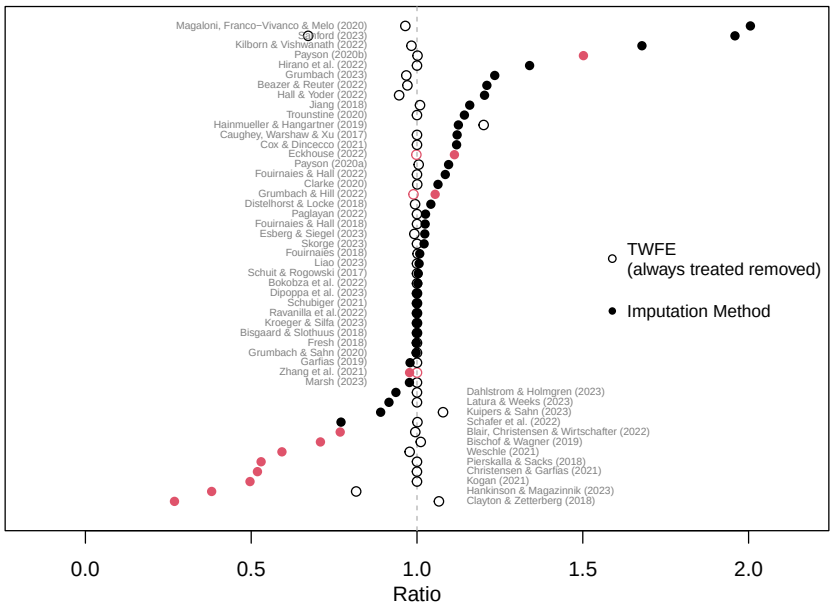


Note: The above figure shows the number of papers that meet our criteria. The grey and black bars represent the number of replicable papers and the number of papers that cannot be replicated.

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

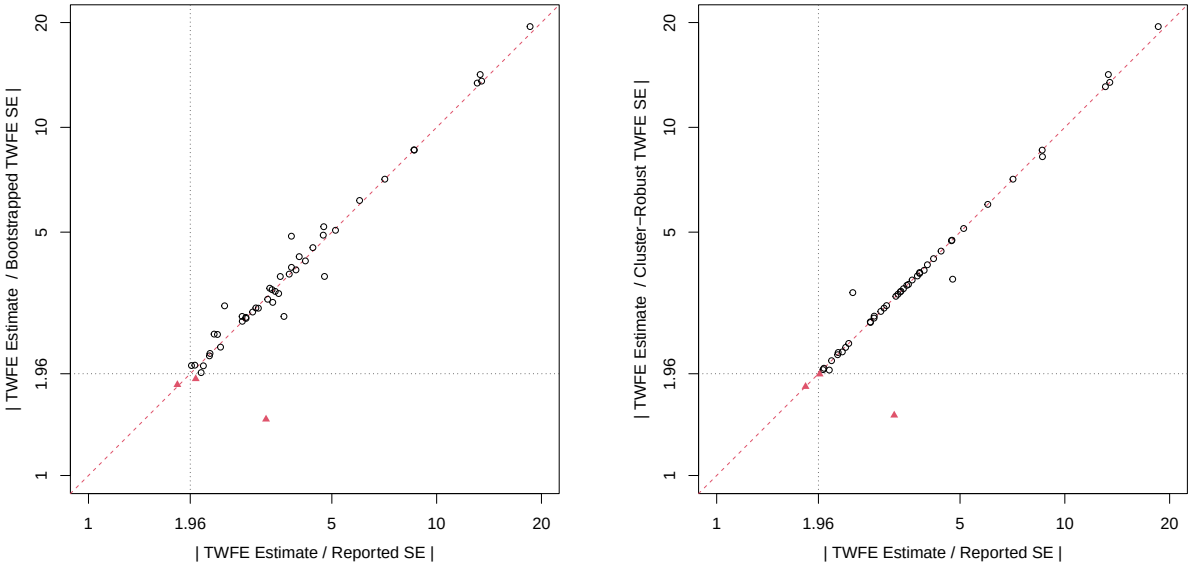


Note: In the above figure, solid circles represent the ratios of the estimates from the imputation method (`FEct`) to TWFE coefficients with always-treated units removed; hollow circles represent the ratios of reported TWFE coefficients to TWFE coefficients with always-treated units removed. Statistically 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

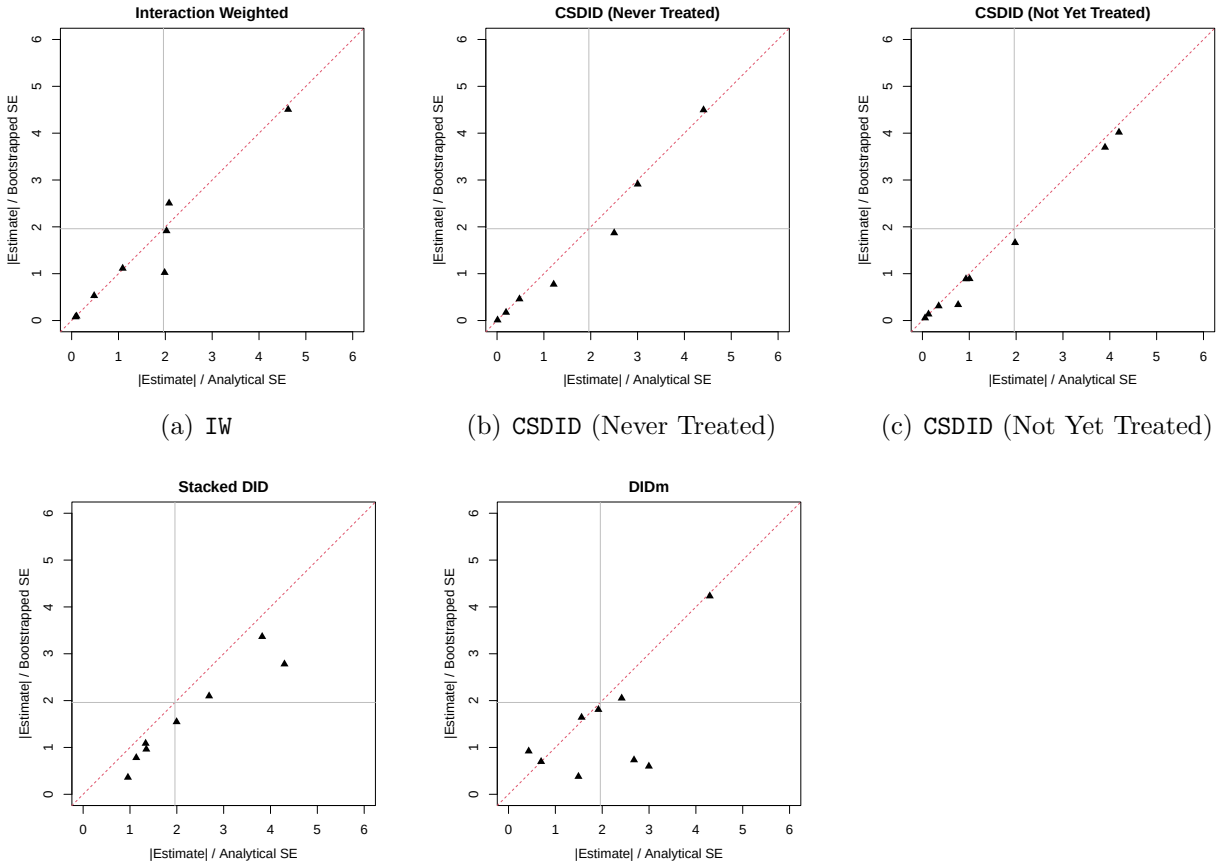


(a) TWFE w/ cluster-robust SEs (b) TWFE w/ cluster-bootstrapped SEs

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)

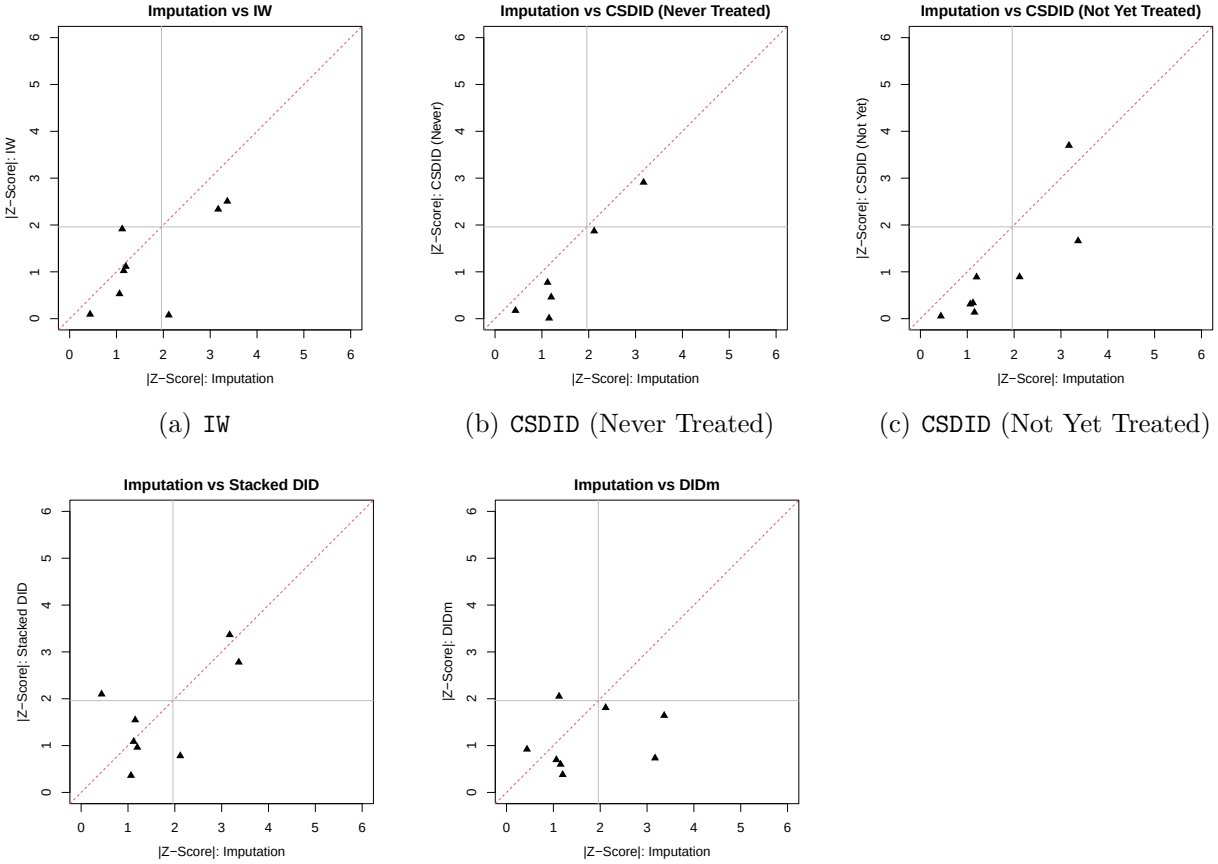
(d) `StackDID` (e) `DID_multiple`

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)

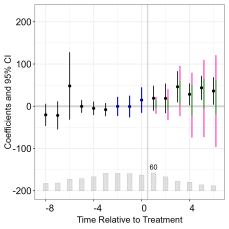
(d) `StackDID` (e) `DID_multiple`

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.

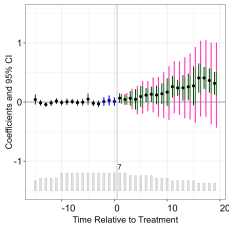
A.3.6. Placebo Tests & Robust Confidence Sets

Figure A6. Placebo Tests & Robust CSs

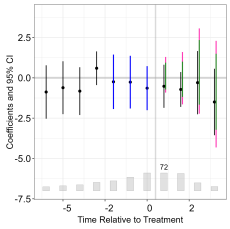
Beazer and Reuter (2022)
p-value $= 0.72$, $\tilde{M} = 0.00$



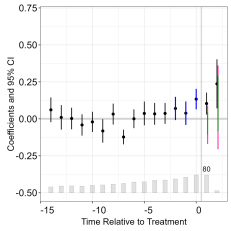
Bischof and Wagner (2019)
p-value $= 0.71$, $\tilde{M} = 0.14$



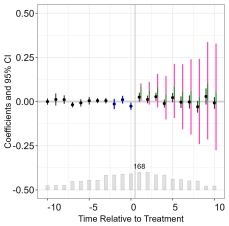
Blair, Christensen and Wirtschafter (2022)
p-value $= 0.52$, $\tilde{M} = 0.00$



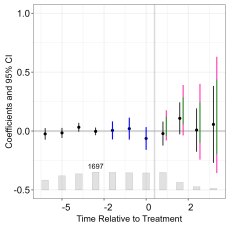
Bokobza et al. (2022)
p-value $= 0.00$, $\tilde{M} = 0.00$



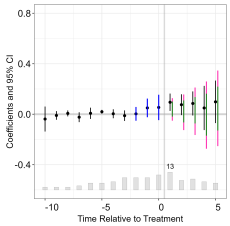
Caughey, Warshaw and Xu (2017)
p-value $= 0.41$, $\tilde{M} = 0.12$



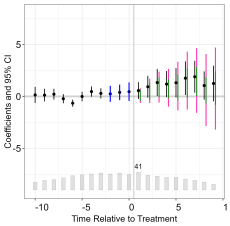
Christensen and Garfias (2021)
p-value $= 0.76$, $\tilde{M} = 0.00$



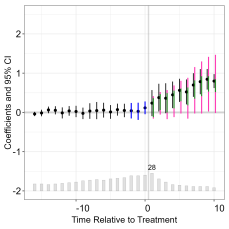
Clarke (2020)
p-value $= 0.16$, $\tilde{M} = 0.00$



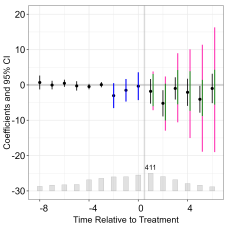
Clayton and Zetterberg (2018)
p-value $= 0.24$, $\tilde{M} = 0.00$



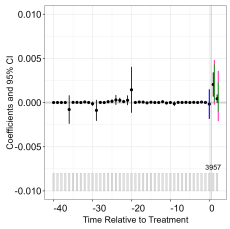
Cox and Dincecco (2021)
p-value $= 0.53$, $\tilde{M} = 0.30$



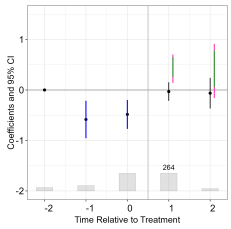
Dahlström and Holmgren (2023)
p-value $= 0.17$, $\tilde{M} = 0.00$



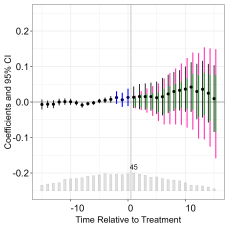
Dipoppa, Grossman and Zonszein (2023)
p-value $= 0.91$, $\tilde{M} = 0.00$



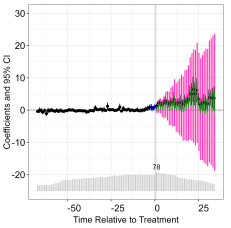
Distelhorst and Locke (2018)
p-value $= 0.00$, $\tilde{M} = 0.40$



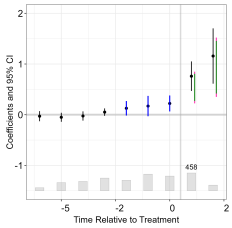
Eckhouse (2022)
p-value $= 0.22$, $\tilde{M} = 0.00$



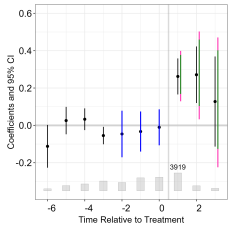
Esberg and Siegel (2023)
p-value $= 0.02$, $\tilde{M} = 0.03$



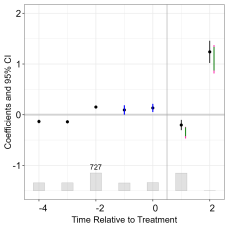
Fouirnaies (2018)
p-value $= 0.01$, $\tilde{M} = 1.75$



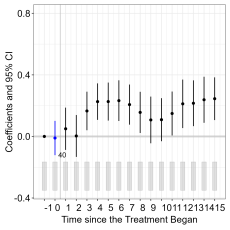
Fouirnaies and Hall (2018)
p-value $= 0.47$, $\tilde{M} = 1.00$



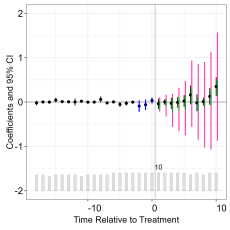
Fouirnaies and Hall (2022)
p-value $= 0.00$, $\tilde{M} = 1.20$



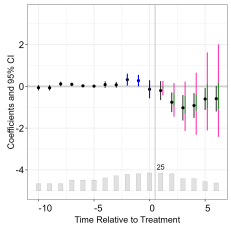
Fresh (2018)
p-value $= 0.86$



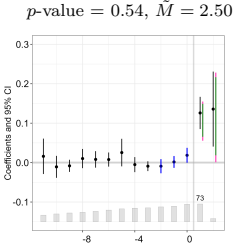
Garfias (2019)
p-value $= 0.32$, $\tilde{M} = 0.00$



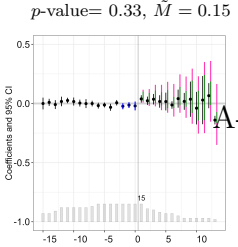
Grumbach (2023)
p-value $= 0.03$, $\tilde{M} = 0.00$



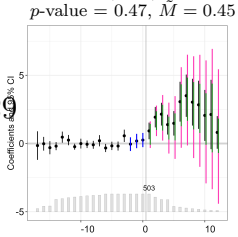
Grumbach and Sahn (2020)
p-value $= 0.54$, $\tilde{M} = 2.50$



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 $=0.29$, $\tilde{M} = 0.00$

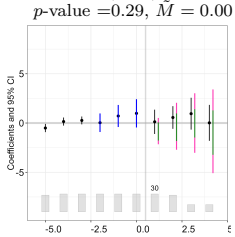
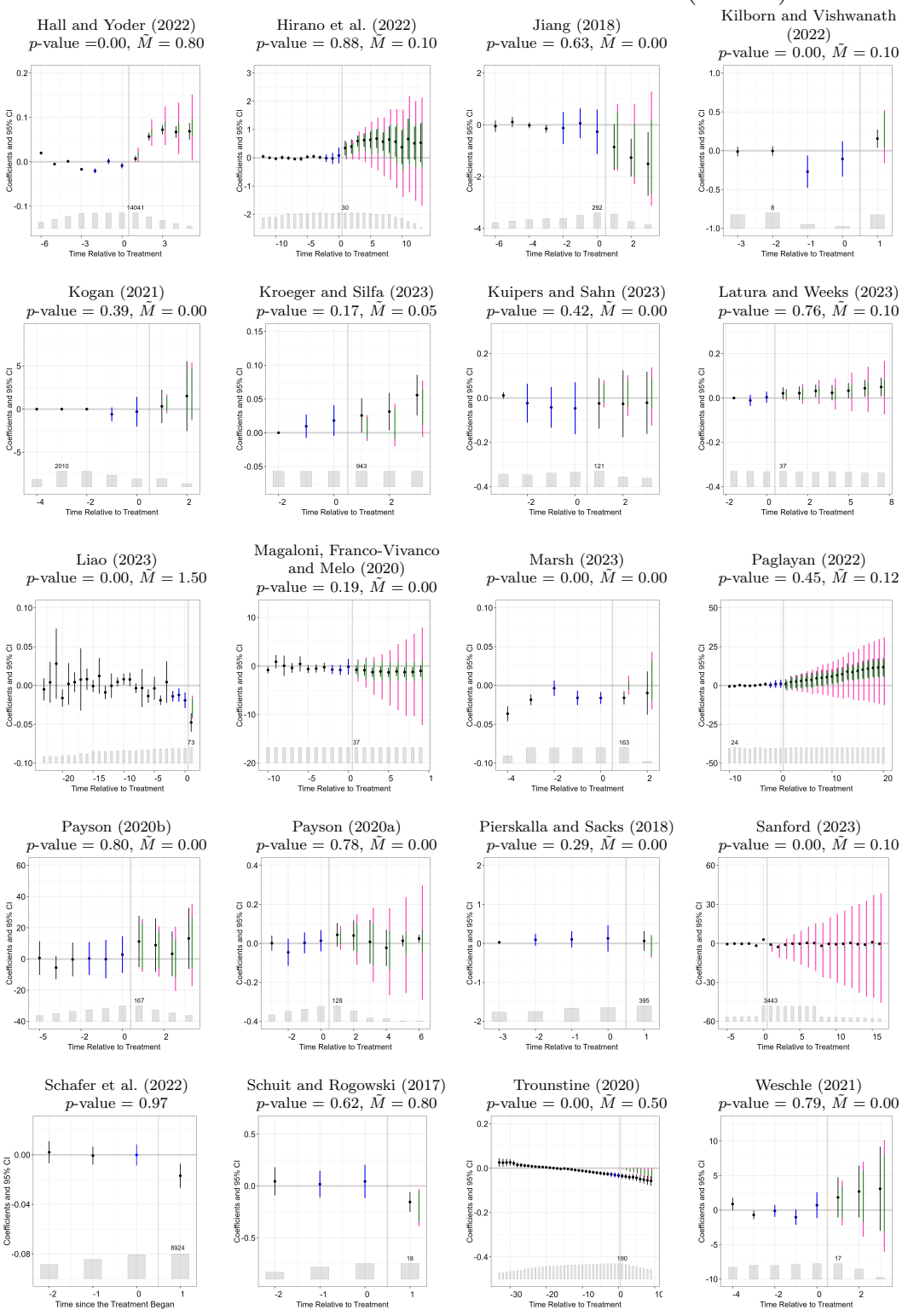


Figure A6. Placebo Tests & Robust CSs (Cont.)



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

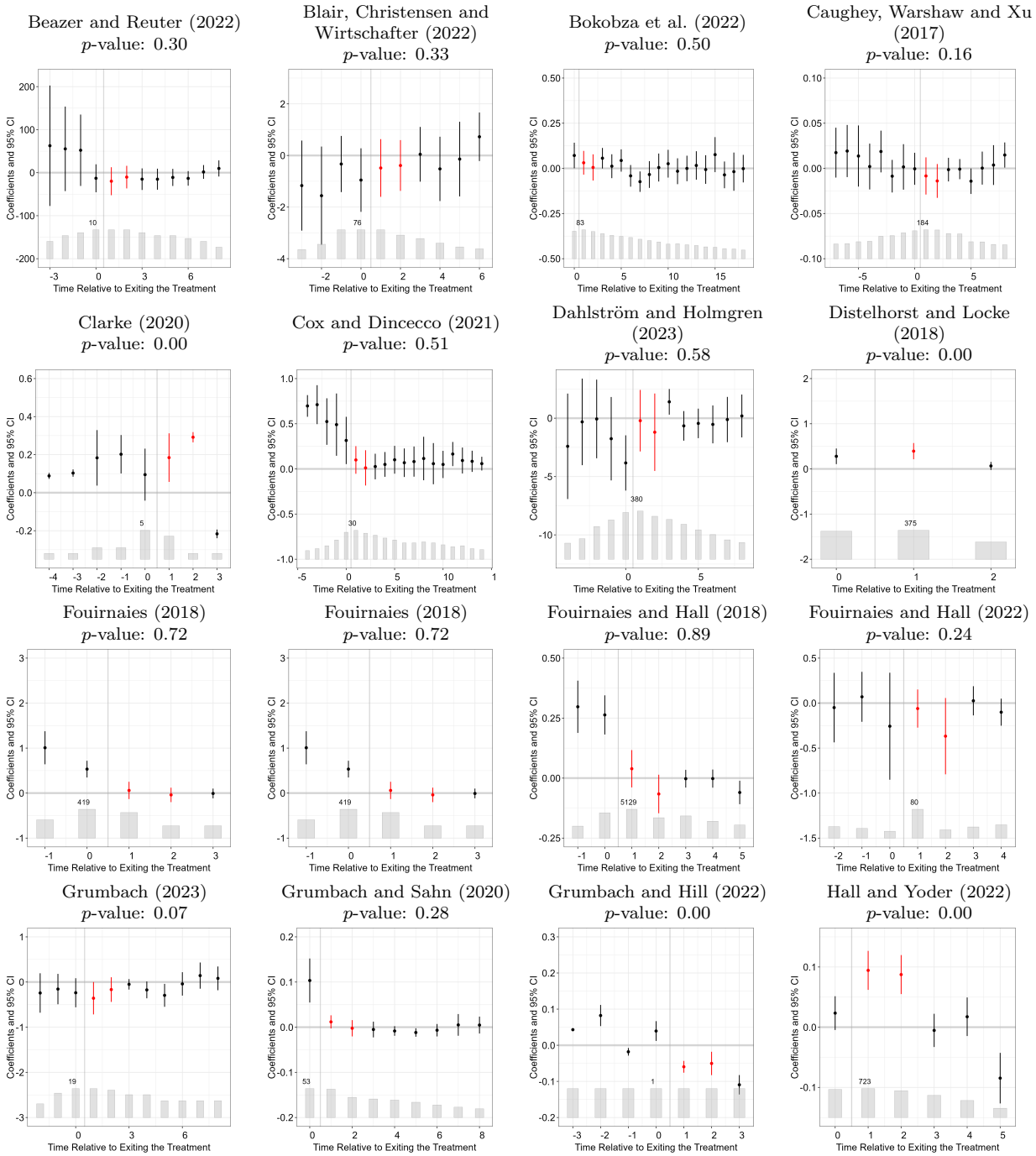
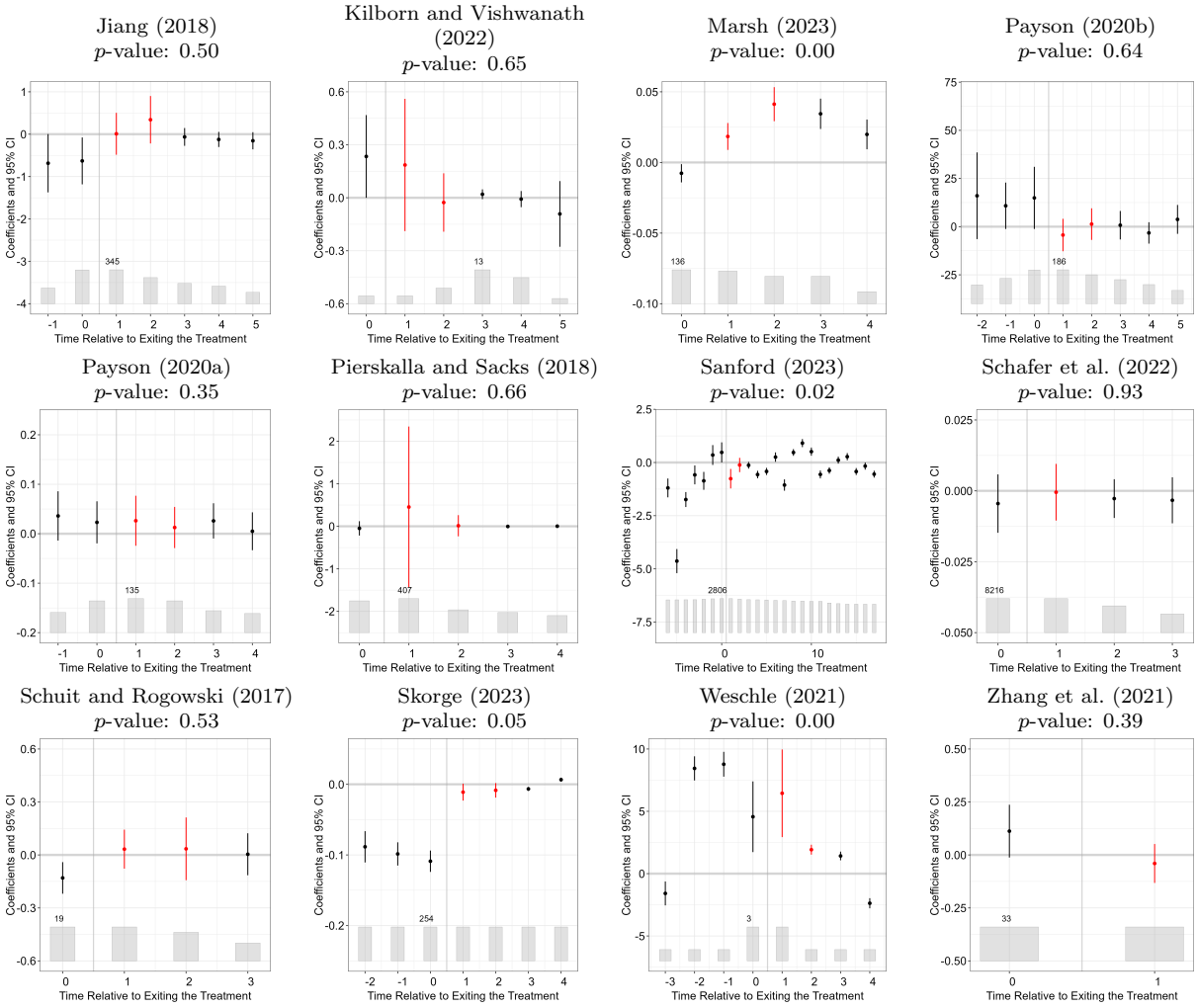


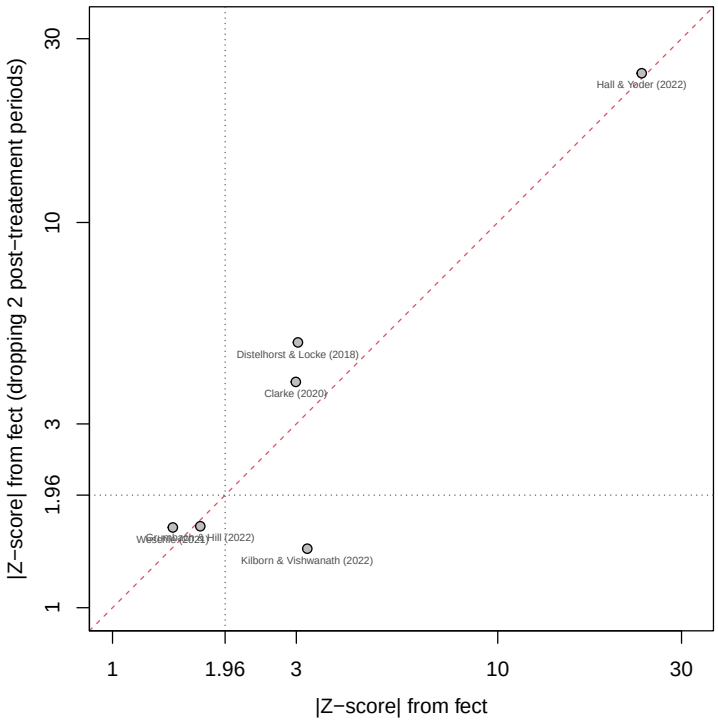
Figure A7. Test for (No) Carryover Effects (Cont.)



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 $\bar{M} = 0$.