Phantom Pre-Trends and Dynamic Effects in Event-Studies with Imbalanced Panels

Abhinav Ramaswamy*
Nicolas Studen†

Abstract

Two-way fixed-effect event studies are routinely used to test for parallel trends and estimate dynamic treatment effects in difference-in-differences designs. We show that when the panel is unbalanced (i.e. when units either enter or exit the sample in different periods) the event-study estimator fails in two ways. First, under treatment-effect heterogeneity, its pre-trend (dynamic effect) coefficients are biased when units enter (exit) the panel in different periods. The estimator may thus fabricate pre-trends and dynamic effects, or conceal genuine ones. Crucially, this bias exists even when all treated units adopt treatment in the same period — a setting long considered safe, because known problems with the estimator only arise under staggered adoption. Second, because the set of observed units changes across periods, the coefficients compare shifting subpopulations and thus target a misleading estimand. We propose a new estimator that, unlike existing heterogeneity-robust estimators, resolves both problems. We illustrate the empirical relevance of our findings by reanalyzing applied papers in this setting, and providing a new application studying the effect of defections in U.S. House speaker elections on campaign fundraising.

*Ph.D. Candidate, Department of Political Science, Stanford University. acr122@stanford.edu.
†Ph.D. Candidate, Department of Political Science, Stanford University. nstuden@stanford.edu.

1 Introduction

In difference-in-differences (DiD) designs, researchers must gauge the plausibility of the key identification assumption: absent treatment, treated and control units' average outcomes would have followed parallel paths. Researchers are also often interested in estimating dynamic treatment effects (whether effects grow or shrink over time).

Popularized by landmark economics applications (Jacobson et al., 1993; Autor, 2003) and recommended by textbooks (Angrist and Pischke, 2009, §5.2), the dynamic event-study regression is the canonical approach across economics and political science for estimating dynamic treatment effects and assessing the parallel trends assumption in panel data. In the simplest setting — a binary treatment with a common onset period t^* — a researcher who observes units $i = 1, \ldots, N$ over periods $t = 1, \ldots, \mathcal{T}$ estimates:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k \neq t^*-1} \beta_k \cdot D_i \cdot \mathbf{1}(t = k) + \varepsilon_{it}, \tag{1}$$

where D_i indicates whether unit i is ever treated, α_i are unit fixed effects, λ_t are period fixed effects, and the period immediately preceding treatment ($k = t^* - 1$) is omitted as the reference. The pre-treatment coefficients $\{\beta_k : k < t^*\}$ are then inspected directly: if they are statistically indistinguishable from zero, the parallel trends assumption is considered justified; if they trend or are jointly significant, parallel trends is called into question. Post-treatment coefficients $\{\beta_k : k > t^*\}$ that differ from the contemporaneous treatment effect β_{t^*} provide evidence of dynamic treatment effects.

What happens, however, when the panel is imbalanced? While most theoretical work on DiD assumes a balanced panel for convenience, many empirical settings feature compositional change: units enter or exit the sample at different times, so the set of observed units varies across periods. Consequently, different units may be observed for a different number of periods. Examples include panels of firms (Liao, 2023; Latura and Weeks, 2023), of legislators and other officeholders (Fouirnaies and Hall, 2022; Weschle, 2021), and of countries and states

(Acemoglu et al., 2019; Paglayan, 2021), where entry and exit can reflect a unit's founding or dissolution, an electoral cycle, or changes in data availability. In each case, the panel is unbalanced, and even when treatment-onset is common, units may be observed for differing numbers of pre- and post-treatment periods.

Applying the canonical event-study estimator (Equation 1), which includes unit fixed effects, to such a panel seems logical. Units entering or exiting at different times likely have different baseline outcome levels. Researchers may want to absorb these differences with unit fixed effects to increase efficiency and address pre-trends or dynamic effects that are driven by the panel's changing composition rather than within-unit changes.[1]

This paper shows that with imbalanced panels, even when treatment onsets in a common period (the simplest case), the canonical estimator is biased under effect-heterogeneity and targets the wrong estimand. The bias stems from how unit fixed effects interact with panel imbalance.

Consider a setting with common treatment-onset and where pre- and post-treatment outcomes are constant within-units (i.e. there are no parallel trends violations or dynamic effects). Unit fixed effects demean each unit's outcome with respect to its *entire* observed history, including post-treatment periods. A treated unit's demeaned outcome is therefore contaminated by its treatment effect. The magnitude of this contamination depends on both the unit's effect-size and the share of its observations that fall after treatment. Because this share varies across entrance- or exit-cohorts (units which enter or exit in a given period), treated units' demeaned outcomes differ even when they have identical raw outcomes. As the mix of observed cohorts changes from period to period, the average shifted-outcome drifts across periods, even when average raw outcomes do not. Under differential-entry, the drift appears in the pre-treatment periods, and under differential-exit, it appears in the post-treatment periods.

[1]Because the composition of observed units changes from period to period, level differences between cohorts can themselves generate movement in the pooled outcome means over the pre-treatment periods (in settings with differential-entry) or post-treatment periods (in settings with differential exit), even without any within-unit trends.

When average effects are heterogeneous across cohorts, these mechanical drifts result in "phantom" pre-trend or dynamic effect coefficients. A phantom pre-trend is a non-zero pre-treatment coefficient that appears even when parallel trends hold exactly, while a phantom dynamic effect is a post-treatment coefficient that grows or shrinks relative to the treatment effect estimate, even when no such dynamic effects exist. This distortion can also run in the opposite direction, masking pre-trends and dynamic effects that genuinely exist. For applied researchers, this means that a failed pre-trend diagnostic in an unbalanced panel need not condemn a design. The failure can lie in the estimator, or in the population the coefficients describe, not in the assumption itself. Importantly, the coefficients are unbiased when effects are homogeneous across cohorts. We show these results extend to the staggered treatment-adoption setting as well.

Independent of this bias, the event-study interactions target the wrong estimand for assessing pre-trends and estimating dynamic effects. Under differential-entry,[2] pre-treatment interactions for a period before all cohorts have entered are estimated only from the cohorts present in that period, not the full set of units used to estimate the treatment effect. They therefore do not test parallel trends for all the units that ultimately identify the treatment effect. Symmetrically, under differential-exit,[3] a post-treatment coefficient for a period after some cohorts have exited is identified only from the survivors — a different subset than the contemporaneous coefficient β_{t^*}, which targets all treated units. Later post-treatment coefficients are thus not directly comparable to β_{t^*}, and a researcher may mistake the changing composition of surviving cohorts for growing or shrinking treatment effects relative to β_{t^*}. Liu (2025) shows an analogous version of this estimand issue exists in event-studies with staggered adoption and panel balance — aggregated "lead and lag" coefficients are not directly comparable, because they are identified from changing subsets of units.

Recent "heterogeneity-robust" (hereafter: HTE-robust) estimators (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Borusyak et al., 2024; Liu et al., 2024; Baker et al., 2022)

[2] i.e. when units enter the data in different pre-treatment periods
[3] i.e. when units exit the data in different post-treatment periods

do not fully resolve the issues we identify. These estimators were developed to handle the "forbidden-comparisons" bias,[4] which occurs *only* under staggered adoption. Crucially, this is not the mechanism that drives the bias result in this paper, as evidenced by the fact that our bias extends to the non-staggered treatment-adoption setting. Because these estimators do not directly model compositional change, the estimand issue remains, and we show that in either treatment-timing regime, they may generate spurious pre-trend or dynamic effect coefficients.

Existing approaches cannot resolve both the bias and estimand issues in imbalanced panels. To tackle this, we redefine the estimand as a weighted average of the *within-cohort* treatment effects, where a cohort is the set of units that enter and exit the panel in the same periods. Because a cohort has fixed composition, this restores, within each cohort, a well-defined pre-trend test and dynamic-effect target when the overall sample exhibits differential entry or exit, respectively. It also implies a corresponding estimation approach that is unblemished by the phantom bias: estimating the main effect, dynamic effects, and diagnosing parallel trends, within cohorts.

To achieve parallel trends within cohorts, we recommend matching treated to control units in a cohort based on pre-treatment outcome trajectories and covariates, and then estimating an event-study without unit fixed effects (our approach also accommodates inverse-propensity weighting and outcome modeling methods).[5] Beyond resolving the bias and estimand problems, this strategy inherits the usual benefits of matching: it addresses *genuine* pre-trend violations driven by compositional changes or within-unit trends, makes parallel trends in the treatment period more credible by balancing covariates across treated and con-

[4]Under staggered adoption, TWFE estimators identify effects for each treatment-cohort using already-treated and later-treated units as controls ("forbidden comparisons") (Goodman-Bacon, 2021; Sun and Abraham, 2021), which allows the weights that aggregate these effects into a point estimate to be negative (often referred to as the "negative-weighting problem"). Consequently, when effects are heterogeneous across units, it is possible that every unit has a positive treatment effect in each period but the model still reports an overall negative effect (de Chaisemartin and D'Haultfœuille, 2020).

[5]Because each cohort is a balanced sub-panel, one could estimate the canonical event-study within-cohort. However, some sort of matching will likely be necessary to achieve parallel trends, after which, including unit fixed effects is inefficient.

trol units, and reduces residual variance. Furthermore, for pre-trends, we recommend *against* aggregation: parallel trends is clearest assessed cohort by cohort. For dynamic effects, we recommend aggregating the cohort-specific coefficients at each event time, weighting each cohort by its share of the treated units observed in that period — the same weighting as the main effect. Once differential exit sets in, we apply the same rule but report the share of treated units that have left the sample, so the plot is explicit that the estimand localizes to the surviving cohorts from that period onward.[6]

We illustrate the practical stakes of our findings and proposed estimator in two ways. First, we document several published DiD designs that share our setting of an unbalanced panel with differential entry (Liao, 2023; Weschle, 2021; Clarke, 2020; Fouirnaies and Hall, 2022; Latura and Weeks, 2023). We reanalyze one in the main text and two more in the supplemental appendix, and show that in each case a more precisely defined estimand would support a more convincing identification story. In the main-text example, we show how a properly targeted estimation approach changes a published substantive result (Liao, 2023). Second, we apply the estimator to a new study of the effect of defections in U.S. House speaker elections on campaign fundraising, where the pooled diagnostic would have led a careful researcher to abandon a design that within-cohort analysis shows to be sound.

2 The Bias of the Canonical Event-Study Estimator

In this section, we prove the bias of the canonical event-study estimator (hereafter, ES-estimator, see Equation 1) for imbalanced panels. Specifically, we show that when treatment effects are heterogeneous across cohorts, the ES-estimator can generate the appearance of pre-trend violations even when parallel trends holds exactly. The opposite concern will also hold: the ES-estimator may report no pre-trend violations when parallel trends breaks.

[6]This is not equivalent to running an off-the-shelf event study and annotating it with attrition counts: estimating within cohorts first is what keeps composition change from biasing the dynamics. For example, were cohorts to sit at different post-treatment baselines and exit one by one, with no unit actually trending, a pooled estimator would manufacture growing dynamic effects, whereas within-cohort aggregation would not.

2.1 Setup

We work throughout in the *single-treatment-onset* (non-staggered) setting: every treated unit receives treatment at a common period t^*, and the event study takes the form of Equation (1). We develop the result in the *differential-entry* case — units enter the panel in different pre-treatment periods — as this setting threatens identification of the main effect, which is of greatest interest to practitioners. For clarity, we work with *common-exit* — units are observed through a common terminal period. However, analogous results apply to settings with *differential-exit*, as the mechanism will depend only on units being observed for differing numbers of periods.

Notation. We observe units $i = 1, \ldots, N$ over periods $t = 1, \ldots, \mathcal{T}$, with common treatment onset at t^*. Each unit is either treated ($D_i = 1$) or a control ($D_i = 0$). We partition units into C *entry cohorts*, indexed by $c \in \{1, \ldots, C\}$ and defined by the period e_c in which the cohort enters the panel, with

$$e_1 < e_2 < \cdots < e_C < t^*.$$

A cohort thus collects all units — treated and control — that first appear in the panel in the same period.[7] Because all units are observed through the common terminal period $\mathcal{T}$, cohort c is observed for

$$T_{\text{pre}}(c) = t^* - e_c \quad \text{pre-treatment periods}, \qquad T_{\text{post}} = \mathcal{T} - t^* + 1 \quad \text{post-treatment periods},$$

for a total of $T_c = T_{\text{pre}}(c) + T_{\text{post}}$ observed periods. Later-entering cohorts have smaller $T_{\text{pre}}(c)$ and hence smaller T_c (this is the panel imbalance), while T_{post} is common across cohorts.

[7] Both treated and control units belong to cohorts; what defines a cohort is entry timing, not treatment status. Each cohort contains n_c treated and m_c control units.

Two notions of parallel trends. Because the composition of observed units changes across periods, "parallel trends" admits two distinct meanings that coincide in a balanced panel but must be separated here:

(PT-W) **Within-cohort parallel trends.** For every cohort c, treated and control units in that cohort share a common counterfactual trend: $\mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0) \mid D_i = 1, c] = \mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0) \mid D_i = 0, c]$ for all $t \in \{e_c + 1, ..., t^*\}$. This is the identification condition relevant to the treatment effect each cohort contributes.

(PT-P) **Pooled parallel trends.** Averaging over all units observed in each period, the mean treated–control outcome difference is constant across pre-treatment periods: $\mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0) \mid D_i = 1] = \mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0) \mid D_i = 0]$ for all $t \in \{1, ..., t^*\}$.

Data-generating process. To prove the bias, we study a generic DGP in which both notions of parallel trends hold exactly by construction:

$$Y_{it} = \begin{cases} \mu_0 & \text{if } D_i = 0, \\ \mu_1 & \text{if } D_i = 1 \text{ and } t < t^*, \\ \mu_1 + \tau_c & \text{if } D_i = 1 \text{ and } t \geq t^*, \end{cases} \tag{2}$$

where μ_1 and μ_0 are scalar constants so there is a fixed level difference between treated and control units, and τ_c is a *cohort-specific* treatment effect. Outcomes are flat in the pre-period (allowing for common time trends does not change the result, see Appendix A), and the treatment effect is static (no dynamics). Thus, the only departure from a textbook two-group design is that the panel is unbalanced. Under this DGP both PT-W and PT-P hold exactly: within any cohort the treated–control gap is $\mu_1 - \mu_0$ in every pre-period, and the pooled treated and control pre-period means are flat at μ_1 and μ_0 regardless of which cohorts happen to be observed. The result below is therefore striking: the canonical estimator manufactures pre-trends even though there is no pre-trend to detect under either

notion.

2.2 How the Bias Emerges

By the Frisch-Waugh-Lovell (FWL) theorem, the event-study coefficients $\hat{\beta}_k$ in Equation (1) are identical to those from regressing the *unit-demeaned* outcome $\tilde{Y}_{it} \equiv Y_{it} - \bar{Y}_i$ on the unit-demeaned interaction regressors and demeaned time dummies, where $\bar{Y}_i$ averages over *all* of unit i's observed periods. The key observation is that this average runs over the unit's *entire* history — including post-treatment periods.

Demeaned outcome. For a treated unit in cohort c, the time-average outcome is

$$\bar{Y}_i = \frac{T_{\text{pre}}(c)\,\mu_1 + T_{\text{post}}\,(\mu_1 + \tau_c)}{T_c} = \mu_1 + \frac{T_{\text{post}}}{T_c}\,\tau_c, \tag{3}$$

so the treatment effect contaminates the mean. The demeaned outcome is therefore nonzero in every period:

$$\tilde{Y}_{it} = \begin{cases} -\dfrac{T_{\text{post}}}{T_c}\,\tau_c, & t < t^*, \\[2ex] \dfrac{T_{\text{pre}}(c)}{T_c}\,\tau_c, & t \geq t^*, \end{cases} \tag{4}$$

with $\tilde{Y}_{it} = 0$ for all control units. Crucially, the pre-period value depends on T_c, so it *varies across cohorts*. As later pre-periods admit later-entering cohorts (larger T_{post}/T_c), the cross-sectional average of $\tilde{Y}$ shifts from period to period — a "trend" on the left-hand side that is purely an artifact of demeaning.

Demeaned regressors. Consider a post-treatment interaction $X_{it}^{(k)} = D_i \cdot \mathbf{1}(t = k)$ for $k \geq t^*$. For a treated unit in cohort c (observed in every post-period, since $e_c < t^* \leq k$), demeaning gives $\tilde{X}_{it}^{(k)} = 1 - 1/T_c$ at $t = k$ and $-1/T_c$ otherwise. Summing over all post-

treatment interactions,

$$\sum_{k\geq t^*} \tilde{X}_{it}^{(k)} = \begin{cases} -\dfrac{T_{\text{post}}}{T_c}, & t < t^*, \\ \dfrac{T_{\text{pre}}(c)}{T_c}, & t \geq t^*, \end{cases} \tag{5}$$

which mirrors $\tilde{Y}_{it}$ in (4) period by period, governed by the *same* factor T_{post}/T_c.

Homogeneous effects: no bias. If $\tau_c = \tau$ for all c, then comparing (4) and (5) gives

$$\tilde{Y}_{it} = \tau \cdot \sum_{k\geq t^*} \tilde{X}_{it}^{(k)} \qquad \text{for every observed } (i,t). \tag{6}$$

The demeaning distortion on the left is absorbed exactly by the post-treatment regressors on the right with the common coefficient τ. Nothing is left over for the pre-treatment interactions.

Heterogeneous effects: bias emerges. If τ_c varies across cohorts, the ratio $\tilde{Y}_{it}/\sum_{k\geq t^*} \tilde{X}_{it}^{(k)} = \tau_c$ is cohort-specific, but OLS constrains the post-treatment coefficients to be common, so the post-treatment interactions cannot reproduce $\tilde{Y}$ for all cohorts at once. Fitting the post-interactions with a common value $\bar{\tau}$ therefore leaves, for a treated unit in cohort c in a pre-period, the residual

$$\tilde{Y}_{it} - \bar{\tau} \sum_{k\geq t^*} \tilde{X}_{it}^{(k)} = -\frac{T_{\text{post}}}{T_c}\,\delta_c, \qquad \delta_c \equiv \tau_c - \bar{\tau}, \tag{7}$$

which is constant within a cohort but differs across cohorts. Averaging this residual over the treated units observed in pre-period k — the cohorts that have entered by then, $\{c : e_c \leq k\}$ — gives

$$\bar{r}_k \;=\; -\frac{\sum_{c:\, e_c \leq k} n_c \dfrac{T_{\text{post}}}{T_c}\,\delta_c}{\sum_{c:\, e_c \leq k} n_c}. \tag{8}$$

As k runs from the earliest entry e_1 toward $t^* - 1$, the set $\{c : e_c \leq k\}$ expands as later cohorts enter, so $\bar{r}_k$ generically changes from one pre-period to the next. This period-

to-period movement is the phantom: among treated units, the pre-treatment interactions $D_i \cdot \mathbf{1}(t = k)$ are the only regressors that distinguish one pre-period from another, so they must absorb the shifting residual, yielding $\hat{\beta}_k \neq 0$. Only when effects are homogeneous ($\delta_c = 0$ for all c) is $\bar{r}_k$ constant at zero and the phantom absent, recovering the cancellation of the previous paragraph.

2.3 Formal Results

The mechanics above imply that in imbalanced panels with differential-entry, phantom pre-trends emerge under effect heterogeneity across cohorts.

Proposition 1 (No phantom pre-trends under homogeneous effects)**.** *When $\beta_k = 0$ for every pre-treatment period $k < t^*$ (parallel trends holds exactly both in the pooled (PT-P) and within-cohort (PT-W) senses) and $\tau_c = \tau$ for all c (treatment effects are homogeneous across cohorts c), then $\hat{\beta}_k = \beta_k = 0$ for each $k < t^*$, regardless of the panel composition (the cohort entry times e_c and sizes n_c, m_c).*

Proof. By (4) and (5), $\tilde{Y}_{it} = \tau \sum_{k \geq t^*} \tilde{X}_{it}^{(k)}$ for every observed (i, t) (for treated units by direct computation; for controls both sides are zero). Hence the coefficient vector $\beta_k = \tau$ for $k \geq t^*$, $\beta_k = 0$ for $k < t^*$, and $\lambda_t = 0$ produces a perfect fit (residual sum of squares equal to zero). As this is the global minimum and the OLS solution is unique under full column rank, $\hat{\beta}_k = 0$ for all $k < t^*$. □

Proposition 2 (Phantom pre-trends under heterogeneous effects)**.** *When $\beta_k = 0$ for every pre-treatment period $k < t^*$ (parallel trends holds exactly in both the pooled (PT-P) and within-cohort (PT-W) senses) and $\tau_c \neq \tau_{c'}$ for some $c \neq c'$ (treatment effects are heterogeneous across cohorts c), then $\exists k < t^*$ such that $\hat{\beta}_k \neq \beta_k = 0$.*

Proof Sketch. See Section 2.2. In short, it shows that the average residual among treated units in the pre-periods ($\bar{r}_k$ in (8)) varies across pre-periods whenever effects are heterogeneous across cohorts. Among treated units the pre-treatment interactions are the only

regressors that distinguish one pre-period from another, so they must absorb this period-varying residual, giving $\hat{\beta}_k \neq 0$ for some $k < t^*$.

Formal Proof. See Appendix A which derives the exact coefficients $\hat{\beta}_k = \bar{a} - \bar{a}_{S_k}$ (which is also the bias, since $\beta_k = 0$) , where $\bar{a}$ is the average treated cohort effect and $\bar{a}_{S_k}$ its average over the cohorts observed at period k.

Corollary 1 (Obscuring genuine pre-trends)**.** *Proposition 2 implies that the event-study estimator can also obscure genuine pre-trends.*

Proof. Decompose the outcome into its untreated potential outcome and the realized effect,

$$Y_{it} = Y_{it}(0) + D_i \, \mathbf{1}(t \geq t^*) \big[Y_{it}(1) - Y_{it}(0) \big].$$

Let $\mathrm{ES}_k(w)$ denote the kth-period coefficient that the canonical event-study estimator returns when run on outcome w. Since OLS is linear in the outcome,

$$\hat{\beta}_k \;=\; \mathrm{ES}_k(Y) \;=\; \underbrace{\mathrm{ES}_k\big(Y(0)\big)}_{\equiv\ \beta_k} \;+\; \underbrace{\mathrm{ES}_k\big(D_i \mathbf{1}(t \geq t^*)(Y_{it}(1) - Y_{it}(0))\big)}_{\equiv\ b_k},$$

where β_k is the genuine pre-trend the test aims to detect. In our setup $Y_{it}(1) - Y_{it}(0) = \tau_{c(i)}$, so b_k is the event study applied to the outcome $W_{it} \equiv D_i \mathbf{1}(t \geq t^*) \tau_{c(i)}$, which is zero in every pre-period and for every control unit, and equals the treatment effect $\tau_{c(i)}$ only for treated units in post-treatment periods. By Proposition 2, this produces nonzero pre-treatment coefficients ($b_k \neq 0$ for some $k < t^*$) whenever effects are heterogeneous across cohorts c. Because b_k depends only on the treatment effects and the panel structure — not on $Y_{it}(0)$ — it is *added* to whatever pre-trend is present. When $\beta_k = 0$ it manufactures a violation ($\hat{\beta}_k = b_k \neq 0$); when $\beta_k \neq 0$ it can inflate, attenuate, or — if $b_k \approx -\beta_k$ — cancel it, so the estimator can report a null precisely when parallel trends fails. □

The canonical estimator therefore does not merely add spurious pre-trends; it is biased for the pre-trend by the fixed amount b_k, in either direction. We derive this exact bias in

Appendix A. Crucially, we take the pooled target PT-P ($\beta_k = \mathrm{ES}_k(Y(0))$) for concreteness, but composition change makes even this pooled coefficient an imperfect summary of the underlying trends; we return to a cohort-specific target when we introduce our estimator.

3 Simulation Evidence

This section provides numerical illustrations of the phantom biases across different data generating processes.

3.1 Phantom Pre-Trends in the Proof's Data Generating Process

We simulate the DGP of Equation (2) over a panel of six periods with treatment onset in the fourth (event time $k = 0$), so the omitted reference is event time -1. Three entry cohorts enter in the first three periods and are observed through the end of the panel; they are thus observed for $T_{\text{pre}} = 3, 2, 1$ pre-treatment periods, respectively, with a common $T_{\text{post}} = 3$. Each cohort contains 30 treated and 30 control units. We set the control mean to $\mu_0 = 1$ and the treated pre-treatment mean to $\mu_1 = 2$. To give unit fixed effects a genuine efficiency rationale, we add a persistent unit-level intercept $\alpha_i \sim N(0, \sigma_\alpha^2)$ with $\sigma_\alpha = 1$ (constant across each unit's periods) and small idiosyncratic noise $\varepsilon_{it} \sim N(0, 0.1^2)$; both within-cohort (PT-W) and pooled (PT-P) parallel trends continue to hold in expectation, so there is no pre-trend in the data. We fit the canonical event study of Equation (1) and plot the estimated coefficients $\hat{\beta}_k$ against event time.

Figure 1 shows the homogeneous treatment effect case, $\tau_c = -6$ for all three cohorts. The pre-treatment coefficients are zero despite the unbalanced panel, confirming Proposition 1: panel imbalance alone does not generate a phantom pre-trend.

Figure 2 introduces heterogeneous effects: $\tau = (-10, -6, -2)$ for the cohorts observed for $T_{\text{pre}} = 3, 2, 1$ pre-periods, respectively. The same estimator now produces large, statistically significant pre-treatment coefficients that decline monotonically toward the reference period

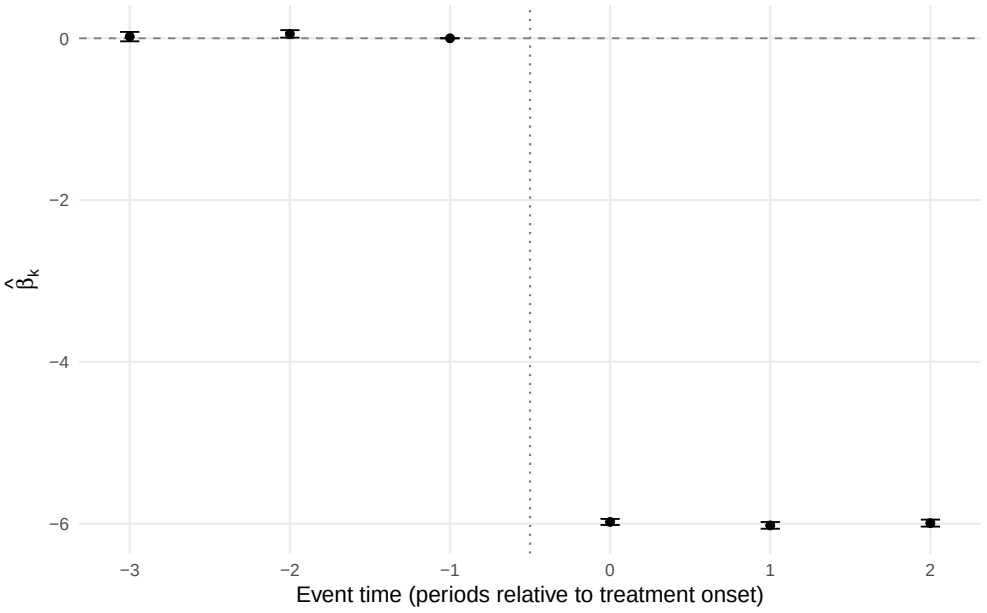

Figure 1: Homogeneous treatment effects: no phantom pre-trend.

Notes: Event-study coefficients $\hat{\beta}_k$ from the canonical specification of Equation (1). Cohort effects are homogeneous at $\tau_c = -6$. The DGP satisfies parallel trends exactly, so the true value of every pre-treatment coefficient is zero. Bars are 95% confidence intervals. The dotted line marks treatment onset, and event time -1 is the omitted reference.

($\approx 2.7 \to 1.5 \to 0$) — a downward phantom pre-trend — even though the data contain no pre-trend whatsoever, confirming Proposition 2. This pattern encourages the researcher to wrongly reject parallel trends or to dismiss a real effect as spurious.

3.2 Alternative DGPs

The phantom is not an artifact of the specific data-generating process used in the proof. The subsections below each modify that DGP along one dimension — adding a common time trend, moving the imbalance to the post-period, and letting cohorts differ in baseline level — and show the distortion persists. Throughout, we keep the within-cohort unit-level variance of Section 3.1, so unit fixed effects always carry an efficiency rationale.

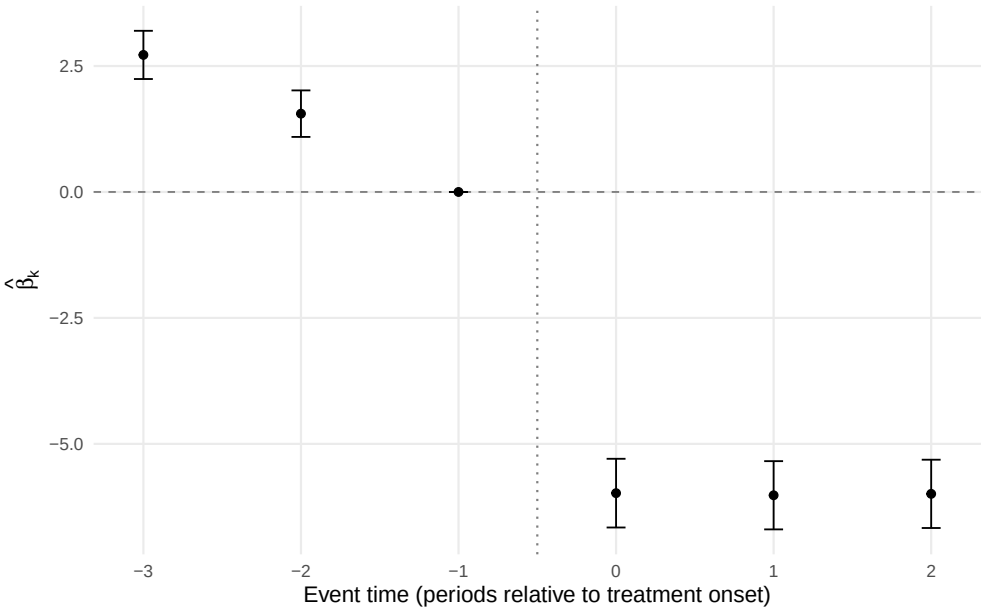

Figure 2: Heterogeneous treatment effects: a downward phantom pre-trend.

Notes: As Figure 1, with heterogeneous cohort effects $\tau = (-10, -6, -2)$. The earliest-entering cohort has the largest-magnitude effect. The DGP again satisfies parallel trends exactly, so the true value of every pre-treatment coefficient is zero. The estimated coefficients decline toward the reference period and are individually distinguishable from zero. Bars are 95% confidence intervals. The dotted line marks treatment onset, and event time -1 is the omitted reference.

3.2.1 Common Time Trend

The flat DGP of Equation (2) held outcomes constant in the pre-period. We now relax this, adding a common time effect γ_t shared by treated and control units:

$$Y_{it} = \begin{cases} \mu_0 + \gamma_t & \text{if } D_i = 0, \\ \mu_1 + \gamma_t & \text{if } D_i = 1 \text{ and } t < t^*, \\ \mu_1 + \gamma_t + \tau_c & \text{if } D_i = 1 \text{ and } t \geq t^*. \end{cases} \tag{9}$$

The time effect γ_t may take any shape (e.g. linear or non-monotone) and is common to both groups, so the treated–control difference remains $\mu_1 - \mu_0$ in every pre-period: both within-cohort (PT-W) and pooled (PT-P) parallel trends continue to hold exactly. We do not allow γ_t to vary across cohorts, as cohort-specific time effects would break PT-P and so confound the exercise; here outcomes simply move over time, with no parallel-trends violation of any kind.

We simulate (9) with an arbitrary non-monotone γ_t, keeping all other features of Section 3.1 fixed — the same three cohorts, the same heterogeneous effects $\tau = (-10, -6, -2)$, and the same unit-level variance. Figure 3 reports the result. Panel (a) confirms the trend is genuinely present in the raw data: both groups move sharply over time, yet track each other in parallel through the pre-period (a constant gap of $\mu_1 - \mu_0 = 1$). Panel (b) shows that the event study — whose period fixed effects λ_t absorb γ_t exactly — produces pre-treatment coefficients identical to the flat-DGP case ($\approx 2.7 \to 1.5 \to 0$). The phantom downward pre-trend is entirely unaffected by the time trend, as established analytically in Appendix A.

3.2.2 Same-Entry, Differential-Exit

The mechanism depends on variation in panel *length*, not on entry timing specifically, and it is symmetric across the two ends of the panel: the estimator's coefficients are distorted wherever the composition of observed cohorts changes. To show this, we move the imbalance entirely

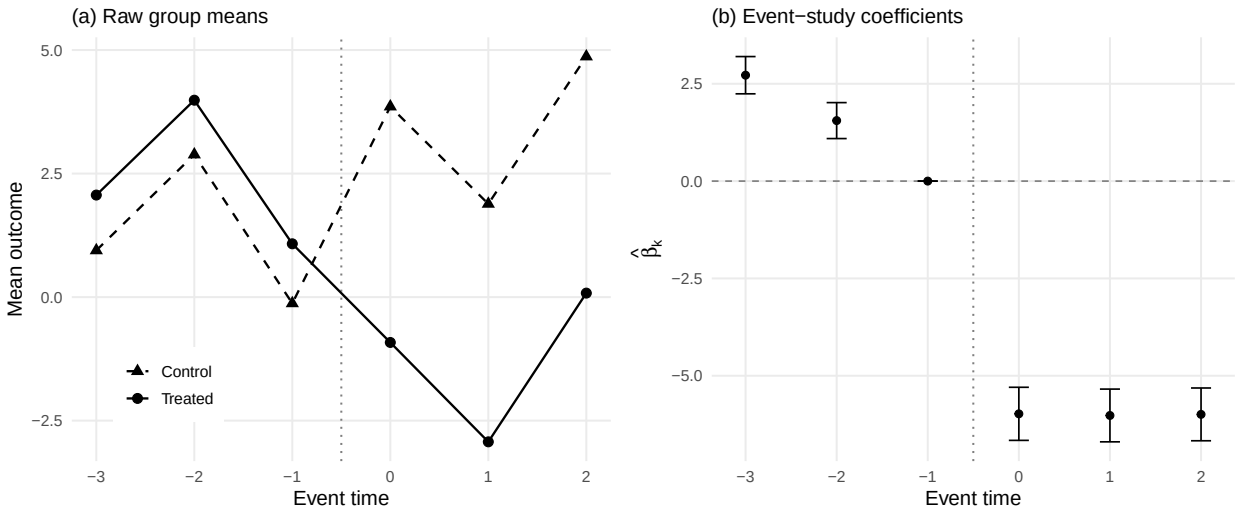

Figure 3: Arbitrary common time trend: the phantom pre-trend is unchanged.

Notes: DGP (9) with a non-monotone common time effect γ_t shared by treated and control units. Panel (a) plots raw group means by event time and panel (b) event-study coefficients $\hat{\beta}_k$, from the same specification as Figure 2. The period fixed effects absorb γ_t, so the pre-treatment coefficients are those of Figure 2. The dotted line marks treatment onset, and event time -1 is the omitted reference.

to the post-period. All units now enter at a common period — so every unit is observed for the same number of pre-treatment periods T_{pre} — but they *exit* at different times. A cohort c is now defined by its exit period and observed for a cohort-specific number of post-treatment periods $T_{\text{post}}(c)$, with $T_c = T_{\text{pre}} + T_{\text{post}}(c)$. The outcome process is otherwise Equation (2), with static, cohort-specific effects τ_c, so both PT-W and PT-P again hold exactly and *no cohort's effect changes over time.*

We simulate three exit cohorts observed for $T_{\text{post}}(c) = 1, 2, 3$ post-treatment periods (common $T_{\text{pre}} = 3$) with heterogeneous effects $\tau = (-2, -6, -10)$, keeping the same unit-level variance as before. Figure 4 shows the result. The pre-treatment coefficients are now flat — with a common entry period, the pre-period composition never changes — but the *post*-treatment coefficients trend steeply (from ≈ -6 at onset to ≈ -8.7 two periods later), even though every cohort's effect is static. The composition change has simply moved to the post-period, where cohorts drop out one by one, and the estimator manufactures dynamic effects exactly as it manufactured pre-trends under differential entry.

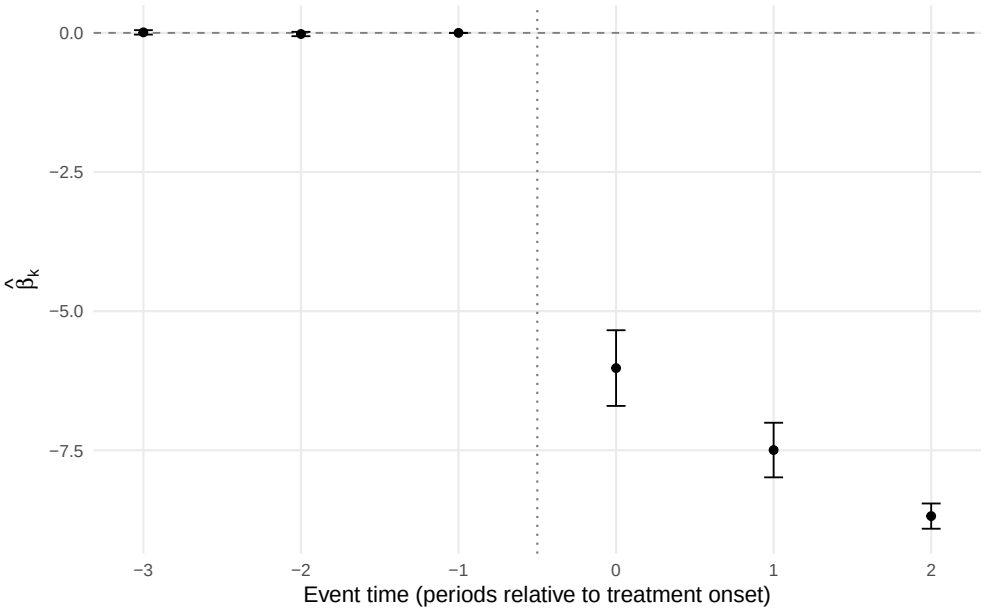

Figure 4: Same entry, differential exit: a phantom in the dynamic effects.

Notes: Event-study coefficients $\hat{\beta}_k$ from the canonical unit-fixed-effects specification. All units enter at a common period (common $T_{\text{pre}} = 3$) but exit after $T_{\text{post}}(c) = 1, 2, 3$ post-treatment periods, with static heterogeneous effects $\tau = (-2, -6, -10)$; both PT-W and PT-P hold. The pre-treatment coefficients are flat, but the post-treatment coefficients trend despite no cohort's effect changing over time, because the composition of observed cohorts changes across the post-period. Bars are 95% confidence intervals; the dotted line marks treatment onset; event time -1 is the omitted reference.

3.2.3 Baseline Differences Across Cohorts

In the settings so far, both notions of parallel trends held. We now let cohorts sit at *different baseline levels*, so that pooled parallel trends (PT-P) fails while within-cohort parallel trends (PT-W) still holds and no unit's outcome changes in the pre-period. We return to the differential-entry design of Section 3.1 but give each entry cohort a different treated–control gap, holding the heterogeneous effects $\tau = (-10, -6, -2)$ fixed. Because the composition of observed cohorts changes across the pre-period, the pooled treated–control difference drifts even though each cohort's gap is constant (Figure 5(a)): PT-P is violated by composition alone.

This is the setting in which unit fixed effects appear most defensible — a researcher would include them not only for efficiency but to absorb the cross-cohort baseline differences that

make the pooled comparison non-parallel. Yet the unit-fixed-effects event study still returns a phantom pre-trend (Figure 5(b)), of essentially the same magnitude as the flat-baseline case ($\approx 2.7 \to 1.5 \to 0$). The fixed effects difference the baselines out, so the reported pre-trend is neither the pooled PT-P violation (which they remove) nor a within-unit trend (there is none): it is the effect-heterogeneity phantom, invariant to the baselines (Appendix A).

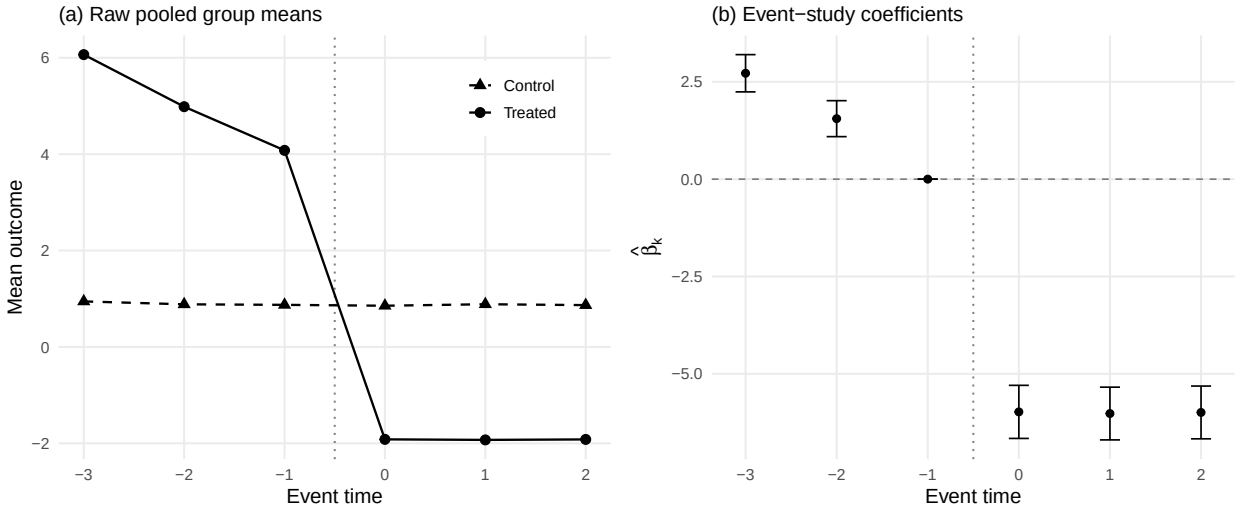

Figure 5: Baseline differences across cohorts: PT-P fails, yet the phantom is unchanged.

Notes: The differential-entry design of Figure 2 with cohort-specific baseline levels (a treated–control gap that varies across entry cohorts) and the same heterogeneous effects $\tau = (-10, -6, -2)$. Panel (a): raw pooled group means. The pooled treated–control gap changes across the pre-period purely because the composition of observed cohorts changes — pooled parallel trends (PT-P) fails — while within each cohort the gap is constant and no unit trends (PT-W holds). Panel (b): event-study coefficients from the unit-fixed-effects specification; the pre-treatment coefficients are the same phantom as Figure 2, because the unit fixed effects difference the baselines out. The dotted line marks treatment onset; event time -1 is the omitted reference.

4 The "Estimand" Problem

Even absent the bias under panel imbalance and heterogeneous effects, the event-study estimator's coefficients are not directly comparable: they target an estimand that is uninformative about the identification assumption.

Consider the common treatment-onset time and differential-entry setting. A pre-treatment coefficient for a period before all cohorts have entered compares the units observed in that

period against the larger set of units observed in the reference period. It therefore does not describe how the *full* set of treated and control units that identify the ATT were trending relative to one another before onset, and therefore does not provide evidence of whether those treated units would, absent treatment, have tracked the controls on average. What the coefficient targets instead is pooled parallel trends (PT-P) — the treated–control mean difference among whoever is observed, period by period — which under compositional change is neither necessary nor sufficient for identification and can be actively misleading. A zero pre-treatment coefficient, for example, may mean that treated and control units genuinely track one another, or it may merely reflect composition change: treated and control units' paths are diverging, but as new units enter, the difference in means among those present in each period stays constant relative to the reference. The coefficient alone cannot distinguish the two. Figure 6 provides an example visualization of this.

The analogous problem applies to the dynamic-effect coefficients under differential exit. A post-treatment coefficient for a period after some cohorts have exited is identified only from the survivors — a smaller, different set of units than the contemporaneous effect $\hat{\beta}_{t^*}$, which is identified from all treated units. Successive lags therefore describe different populations, and a researcher may read the changing composition of surviving cohorts as an effect that grows or shrinks over time when no unit's effect is in fact changing (Section 3.2.2).

This comparability problem is not unique to imbalanced panels. Liu (2025) shows an analogous issue in the leads-and-lags event study under *staggered* adoption with a *balanced* panel: because treatment cohorts are observed over different numbers of pre- and post-treatment periods in event time, the aggregated lead and lag coefficients are estimated from changing subsets of cohorts and are likewise not directly comparable. Here we have argued a related problem arises from compositional change even in the non-staggered adoption setting typically regarded as safe.

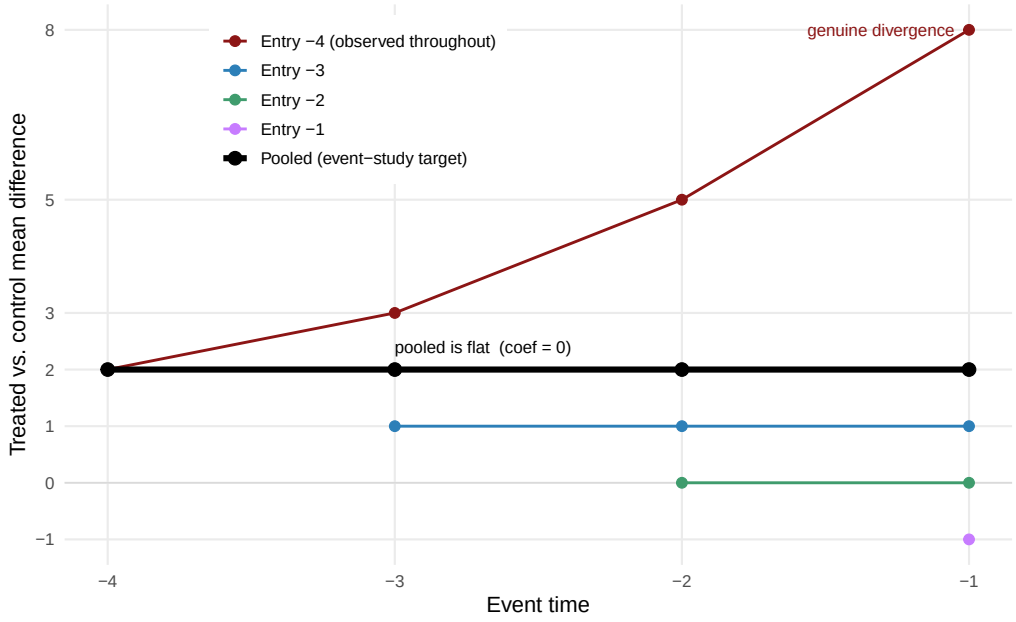

Figure 6: Entering cohorts can mask a genuine pre-trend, leaving a zero coefficient.

Notes: A stylized differential-entry design in which controls are flat and cohorts are evenly sized, so each line is one cohort's treated–control mean difference (labeled by entry period). Only the cohort observed throughout (entry −4) is trending — a genuine differential pre-trend — while each later-entering cohort is flat and enters at a progressively lower gap. Because those entrants pull the cross-cohort average down by exactly the right amount, the pooled difference that the event study targets (black) is flat across the entire pre-period, so every pre-treatment coefficient is zero. The entering cohorts mask the one cohort's genuine trend. Event time −1 is the reference period.

5 How do Existing HTE-Robust Estimators Perform?

Recent work on two-way fixed effect estimators has focused extensively on the problems which emerge under staggered adoption. In this setting, vanilla two-way fixed effect estimators may identify effects for each treatment-cohort using already-treated and later-treated units as controls (often referred to as "forbidden comparisons"). When treatment effects are heterogeneous across treatment-cohorts, this results in bias (Goodman-Bacon, 2021; Sun and Abraham, 2021). In addition, the weights that aggregate heterogeneous effects into a point estimate may be negative (often referred to as the "negative-weighting problem"). The model may therefore report an overall negative (positive) effect, even when every unit experiences a positive (negative) effect when entering treatment (de Chaisemartin and D'Haultfœuille, 2020).

Importantly, the mechanism driving the "phantom" bias and "estimand issue" in this paper is distinct from these problems under staggered adoption, as they extend to the non-staggered treatment adoption setting where the forbidden-comparisons and negative weighting problems do not exist. However, a class of new "HTE-robust" (heterogeneous-treatment-effect robust) estimators has emerged to resolve these problems under staggered adoption. It is still possible that the machinery rendering these new estimators robust to effect-heterogeneity incidentally protects them against the new set of problems we show emerge under panel imbalance.

In this section, we analyze the interaction-weighted (IW) (Sun and Abraham, 2021), stacked (Baker et al., 2022), CS-DID (Callaway and Sant'Anna, 2021), and fixed-effect counterfactual imputation ("FEct") (Liu et al., 2024) estimators, and show by simulation that this is not the case. All four estimators fail to sufficiently address the biases stemming from imbalance, because they partition the data by adoption timing, not by entry or exit timing. The IW and stacked estimators suffer from the phantom bias in Proposition 2, because they include unit fixed effects which demean treated units over their entire history. Although the CS-DID and FEct estimators avoid the phantom bias by not demeaning treated units'

outcomes, they still suffer from the composition (estimand) problem shown in Section 4. They implicitly target a pooled estimand, which, under imbalance, may produce misleading evidence for or against pre-trends or dynamic effects. Furthermore, beyond their potential for spurious estimates, all estimators also suffer from the comparability problem discussed in Section 4, because coefficients are identified on different subsets of units.

We focus on the staggered-treatment-adoption setting, because these new estimators were designed to address problems emerging only in this setting. Researchers would have little reason to reach for them in the non-staggered setting. Regardless, the estimators have the same faults in the non-staggered setting.

Design. We simulate a staggered panel in which units may enter treatment in one of two periods, $g \in \{g_1, g_2\}$, alongside a never-treated control group. We refer to the set of units first treated at a common date g as an "adoption cohort". This is separate from an "entry cohort", which denotes when a unit enters the panel.

Each adoption cohort is itself an unbalanced panel of exactly the form studied above: within an adoption cohort, units belong to several entry cohorts (differential entry, common exit) or several exit cohorts (common entry, differential exit), and effects may be heterogeneous across those sub-cohorts. Parallel trends holds throughout in both the within-cohort (PT-W) and pooled (PT-P) senses, treatment effects are static, and outcomes are flat in the pre-period, so any non-zero pre-trend or post-treatment dynamic in the estimated coefficients is spurious. We add modest idiosyncratic noise and use large cohorts, so the results are algebraic properties of the design and estimators, not sampling artifacts.

IW and stacked DiD. The IW (Sun and Abraham, 2021) and stacked DiD estimators (Baker et al., 2022) both effectively estimate a separate event-study with unit fixed effects for each adoption cohort and then aggregate them. But because the adoption cohorts themselves are imbalanced with heterogeneous treatment effects, the phantom bias of Proposition 2 emerges inside each adoption cohort's event-study, biasing the aggregate coefficients.

Panel A of Tables 1 and 2 demonstrates this. Under differential-entry, both estimators report a downward pre-trend even though parallel trends holds exactly, as no unit's outcome changes prior to treatment. Under differential-exit, both estimators report post-treatment coefficients that grow over time, although every unit's effect is static. A falsification check confirms the mechanism — eliminating either the within-adoption-cohort entry/exit heterogeneity or the cross-adoption-cohort effect heterogeneity eliminates the phantom bias.

CS-DID and FEct. Neither the CS-DID estimator (Callaway and Sant'Anna, 2021) nor the FEct estimator (Liu et al., 2024) demean treated units' outcomes. The former estimates two-period two-group effects for each adoption-cohort only using never-treated units as controls, and the latter fits a two-way fixed effects model of untreated potential outcomes using control observations only and imputes treated units' counterfactual outcomes in each period. They therefore avoid the phantom pre-trend bias in the differential-entry, heterogeneous-effect design, where the Sun–Abraham and stacked DiD estimators break (see Panel A of Table 1).

However, the phantom pre-trend bias — non-zero pre-trends when PT-P and PT-W hold exactly, due to composition change and effect heterogeneity — is only one way an estimator may report misleading evidence of pre-trends or dynamic effects. Like the IW and stacked estimators, CS-DID and FEct estimators target a misleading, pooled estimand, which may also result in spurious pre-trend or dynamic effect estimates.

Under differential-exit with heterogeneous effects, both estimate the "survivor-average effect", which is the difference in means between treated and control units in each post-treatment period relative to the reference period, using the units present at that horizon. When exit-cohorts sit at different levels post-treatment, this effect drifts mechanically as cohorts leave. Panel A of Table 2 confirms this, as the estimators suggest the ATT grows from -6 to -10, despite no unit's post-treatment outcomes actually changing.

CS-DID is particularly vulnerable to this estimand issue, as it fails even under effect-

homogeneity. When cohorts differ in baseline levels — so that in the differential-entry setting, PT-W holds but PT-P fails, and in the differential-exit setting, exit-cohorts sit at different post-treatment levels — the estimator reports spurious pre-trends under differential-entry (see Panel B of Table 1) and spurious dynamic effects under differential-exit (see Panel B of Table 2), even when treatment effects are homogeneous across cohorts. The reason is that with an unbalanced panel, the estimator compares changing-composition group means across periods rather than within-unit changes. It correctly measures the treated-control differential among the units observed in each period, but as that set of units changes, a fixed level gap across cohorts gets misreported as a trend.

FEct is immune to this baseline-level mechanism, because its counterfactual model includes unit fixed effects, which difference out level differences across cohorts. Crucially, that model is fit on control observations only, so treated units' pre-treatment outcomes are demeaned using their pre-periods alone, never their post-treatment outcomes, which is what invites the phantom bias. FEct is therefore the only estimator that never reports spurious pre-trends, regardless of effect-heterogeneity and even when PT-P fails while PT-W holds. That said, like the other estimators, it still suffers from the comparability problem of Section 4: its coefficients are still estimated on a changing subset of units, so pre-trend coefficients for periods before all the units that identify the ATT have entered remain uninformative about identification. And because its control-only model cannot capture the post-treatment drop-out structure, under differential-exit with heterogeneous effects, it too can report dynamic effects when none exist.

Table 1: Event-study coefficients under staggered adoption with differential entry and no real pre-trends.

	Pre-treatment ($k < t^*$)						Post-treatment			
Relative period	-6	-5	-4	-3	-2	-1	0	1	2	3
Panel A. Heterogeneous effects ($\tau_c = -2, -6, -10$), common baseline										
Sun–Abraham	-2.30	-2.31	-1.31	-1.33	0.00	—	-6.00	-6.00	-6.00	-5.98
Stacked DiD	-2.30	-2.32	-1.31	-1.33	0.00	—	-6.00	-6.00	-6.00	-5.93
Callaway–Sant'Anna	0.00	-0.02	0.00	-0.02	0.00	—	-6.00	-6.00	-6.00	-5.98
Imputation (FEct)	0.00	-0.01	0.00	-0.01	0.00	—	-6.00	-6.00	-6.00	-5.99
Panel B. Homogeneous effect ($\tau = -4$), baselines $0, 4, 8$ across entry cohorts										
Sun–Abraham	0.03	0.00	0.00	0.00	0.01	—	-3.99	-3.99	-4.00	-3.99
Stacked DiD	0.03	0.00	0.00	0.00	0.01	—	-3.99	-3.99	-4.00	-4.00
Callaway–Sant'Anna	-3.97	-4.00	-2.00	-2.00	0.01	—	-3.99	-3.99	-4.00	-3.99
Pooled (no unit FE)	-3.98	-4.01	-2.01	-2.00	0.00	—	-4.00	-4.00	-4.00	-4.00
Imputation (FEct)	0.02	-0.01	0.00	0.00	0.00	—	-4.00	-4.00	-4.00	-4.00

Notes: Panel A isolates the demeaning phantom by allowing for heterogeneous effects across entry cohorts while holding baselines common. Panel B isolates composition bias by holding effects homogeneous at $\tau = -4$ while letting baseline levels differ across cohorts. Neither setting contains a unit-level trend, and PT-W holds throughout, so every true pre-period coefficient is zero.

Table 2: Event-study coefficients under staggered adoption with differential exit and no real dynamic effects.

	Pre			Post-treatment ($k \geq t^*$)						
Relative period	-3	-2	-1	0	1	2	3	4	5	6
Panel A. Heterogeneous effects ($\tau_c = -2, -6, -10$), common baseline										
Sun–Abraham	-0.01	-0.02	—	-6.01	-6.01	-6.02	-7.15	-7.15	-8.03	-8.02
Stacked DiD	-0.01	-0.02	—	-6.01	-6.01	-6.02	-7.15	-7.15	-8.03	-8.02
Callaway–Sant'Anna	-0.01	-0.02	—	-6.01	-6.01	-6.02	-8.02	-8.03	-10.03	-10.01
Imputation (FEct)	0.00	0.00	—	-6.00	-6.00	-6.01	-8.00	-8.01	-10.01	-10.00
Panel B. Homogeneous effect ($\tau = -4$), baselines $0, 4, 8$ across exit cohorts										
Sun–Abraham	-0.01	0.00	—	-4.00	-4.00	-4.01	-4.00	-3.99	-4.02	-4.00
Callaway–Sant'Anna	-0.01	0.00	—	-4.00	-4.00	-4.01	-2.00	-1.99	-0.01	0.01
Pooled (no unit FE)	0.00	0.00	—	-3.99	-4.00	-4.00	-1.99	-1.98	-0.01	0.01
Imputation (FEct)	0.00	0.00	—	-4.00	-4.00	-4.00	-4.00	-3.99	-4.02	-4.00

Notes: Panel A isolates phantom dynamics by allowing for heterogeneous effects across exit cohorts while holding baselines common. Panel B isolates composition bias by holding effects homogeneous at $\tau = -4$ while letting baselines differ across cohorts. No unit's effect is dynamic, so the true post-treatment path is flat.

6 Changing Composition in a Published Application

Unbalanced panels are common in difference-in-differences applications (Liao, 2023; Weschle, 2021; Clarke, 2020; Fouirnaies and Hall, 2022; Latura and Weeks, 2023; Chiu et al., 2025). In this section, we examine how compositional change in published work leads to the estimand problem of Section 4. We focus on Liao (2023), a recent article in a leading political science journal that relies on a non-staggered design with substantial compositional change. After restricting the sample to a fixed set of units and constructing a more comparable control group, we obtain a substantively different estimate. Appendix B reports two additional replications of Weschle (2021) and Clarke (2020), where the pre-treatment diagnostics similarly

describe changing populations.

6.1 The Published Diagnostics

Liao (2023) asks whether firms that lobbied on immigration in 2017 subsequently received more favorable treatment in H-1B visa applications. Many firms recorded an *increase* in H1-B denial rates in 2017 so the paper's negative estimate suggests lobbying led to a comparatively *smaller* increase in denial rates. The panel is severely unbalanced, as only about a tenth of the possible firm–year cells between 1992 and 2017 are observed.

The published raw-means display. Liao plots average H-1B denial rates for firms that did and did not lobby on immigration in 2017, describes the pre-treatment trends as "quite similar," and states that the figure increases confidence in parallel trends (PT-P). Figure 7 reproduces the published display and adds the number of firms contributing to the estimates, in five-year intervals. The number of observed treated firms rises from 20 in 1992 to 74 at treatment onset in 2017, while the number of control firms grows from 3,677 to 38,188 after peaking at 67,330 in 2004. The displayed means therefore average over changing sets of firms rather than following a common sample over time.

The event study. Liao's headline estimate comes from a static two-way fixed-effects specification with a single post-treatment interaction. In Appendix Table D.14, Liao also reports the canonical event-study specification of Equation (1). Column 2 of the table interacts the 2017 lobbying treatment indicator with year indicators, includes firm and year fixed effects, and clusters standard errors by firm. The specification omits 2016 as the reference period. Because 2017 is the panel's only post-treatment year, the model produces 24 pre-treatment coefficients and a single post-treatment coefficient. Figure 8 plots all 25 coefficients.[8]

The paper characterizes the pre-treatment estimates as "either small and imprecisely

[8]We apply the published specification to the author's replication data and reproduce all 25 coefficients and standard errors exactly.

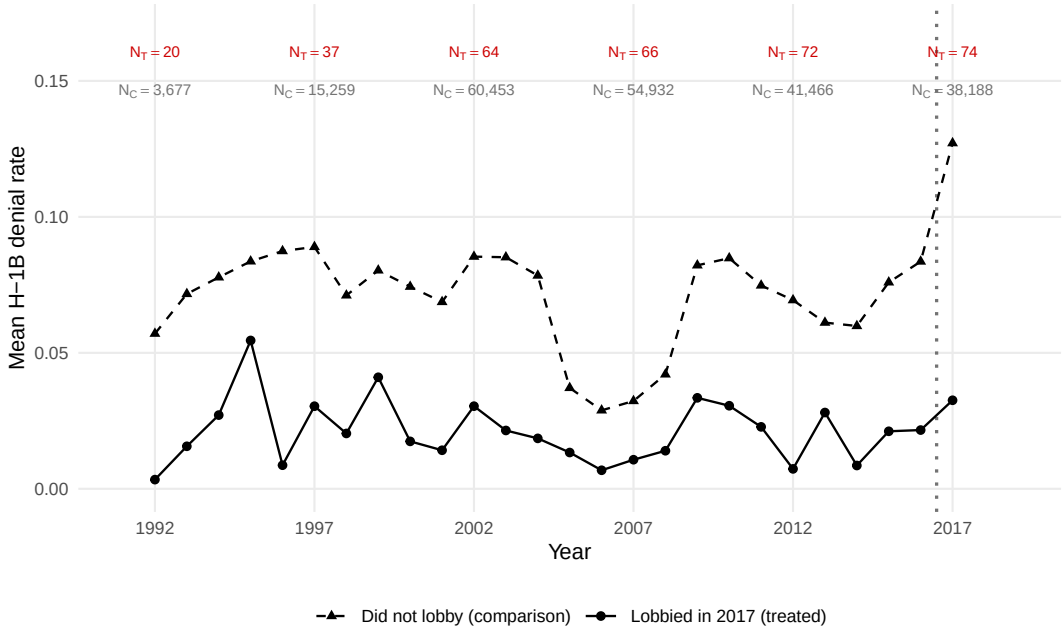

Figure 7: Liao (2023)'s published parallel-trends diagnostic features substantial composition change.

Notes: Mean H-1B denial rate by calendar year for firms that did and did not lobby on immigration in 2017, reproducing the author's estimates exactly. N_T and N_C report the distinct treated and control firms observed in the labeled year. The dotted line marks treatment onset.

estimated or in the opposite direction" of the negative 2017 estimate (Liao, 2023, p. 1427). Of the 24 pre-treatment coefficients, 23 are positive and the only negative one is -0.001. Ten, however, are individually significant. The sign of any individual coefficient also provides little information about pooled parallel trends, which concerns the *path* of pre-treatment estimates. A systematic pattern of pre-treatment movement is concerning, even if every coefficient is of the opposite sign.

More importantly, the coefficients do not describe a common set of units. The firms observed change substantially across the pre-treatment periods, creating the estimand problem described in Section 4. Figure 8 measures the extent of this change by reporting, for selected years, the share of firms observed at treatment onset that are also observed in that year. Among control firms, the overlap is only 3% in 1997 and 13% in 2002. Even in 2016, the omitted reference period and final pre-treatment year, it is just 48%. Among treated firms, the corresponding shares are 49%, 76%, and 95%. The estimand problem therefore persists

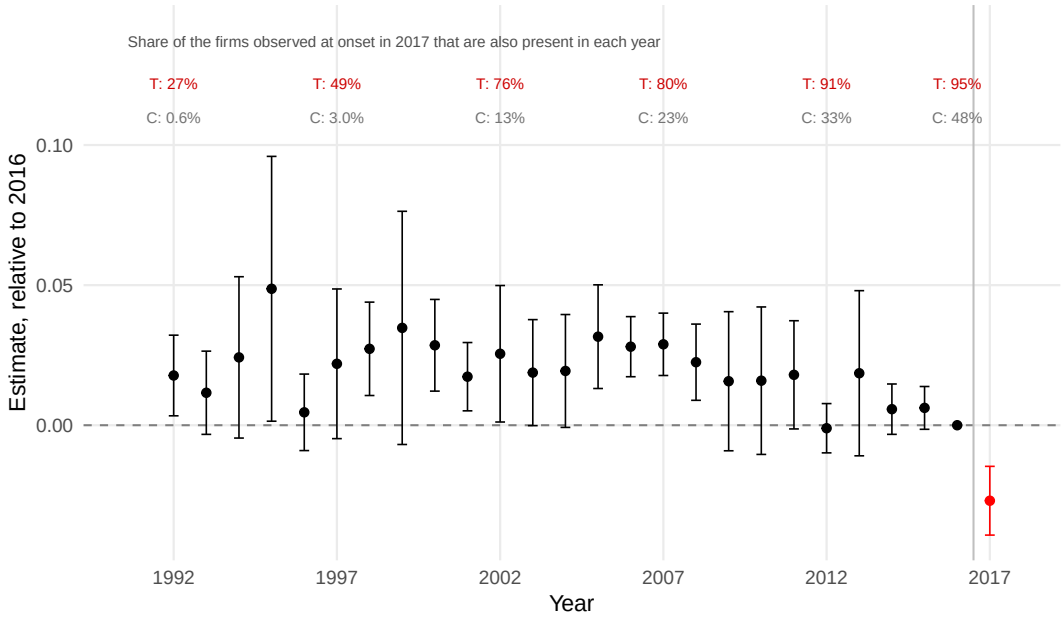

Figure 8: Liao (2023)'s event study and overlap with firms observed at treatment onset.

Notes: Reproduction of Liao (2023) Appendix Table D.14, column 2. The specification includes firm and year fixed effects, standard errors clustered by firm, and omits 2016 as the reference period. The dotted line marks treatment onset. The red point is the 2017 post-treatment estimate. Percentages report the share of treated (**T**) and control (**C**) firms observed in 2017 that are also observed in the labeled year.

through the period immediately preceding treatment.

6.2 Fixing Composition

To address these changes in composition, we first drop the earliest pre-periods and then retain only firms observed throughout, producing a balanced panel. Table 3 reports how each restriction changes the sample and the 2017 estimate. We keep the paper's outcome and treatment definition fixed throughout.

Holding composition fixed over the published 1992–2017 span would retain only 15 of the 74 treated firms observed at onset.[9] The resulting event study would describe only a small fraction of the units that undergo treatment. We instead drop the earliest pre-periods before restricting to a fixed set of units. We begin in 2011, a compromise that provides six pre-periods while retaining 65 of the 74 treated firms. Appendix C reports the tradeoff

[9]Liao's treatment indicator identifies 89 treated firms in Table 3 row A, but only 74 have an observed denial rate at treatment onset.

between pre-periods and treated firm retention for various starting years.

Focusing on 2011–2017 leaves the estimated effect essentially unchanged: −2.73 percentage points over 2011–2017, compared with −2.69 over the published period (Table 3, rows A and B).[10] Restricting the sample further to firms observed for the entirety of the 2011–2017 window attenuates the estimate somewhat to −2.11 percentage points (row C). The corresponding event study follows the same set of firms across all six pre-periods.

Fixing composition resolves the estimand problem, but the pre-treatment path remains concerning. In the balanced panel, the 2011 coefficient is 2.6 percentage points and the subsequent estimates decline toward the omitted 2016 reference period (Figure 9, left panel). A joint test that all five coefficients are zero yields $p = 0.066$. We also test for systematic movement in pre-treatment coefficients, replacing the five year-specific interactions with an interaction between the treatment indicator and a linear time trend. The linear specification pools the five coefficients into a single slope rather than testing each separately, allowing it to detect a consistent decline[consistent decline is weird, I think this sentence should be sign-neutral.] that individual coefficients cannot establish (Roth, 2022). The linear test reveals that the treated-control gap declines by 2.1 percentage points from 2011 to 2016 and we reject a zero slope ($p = 0.006$). With composition fixed, both tests describe the same firms throughout. The remaining problem is thus the comparability of the control and treated firms, given the difference in their pre-treatment paths, rather than panel composition. We turn next to constructing a more comparable control group.

6.3 Constructing a Comparable Control Group

Balancing the panel holds the sample constant across years, but treated and control firms still differ in observable ways. A credible comparison requires control firms that plausibly stand in for how the treated firms would have behaved absent treatment. On the balanced panel, we examine how the two groups differ across a set of firm-level covariates, which we

[10] Liao's headline specification replaces the event study with a single post-treatment indicator and uses all available pre-treatment observations. Those results are reported in Table 3.

take from Orbis in line with Liao (2023). The standardized differences reach 1.5 on firm size and 1.6 on whether the firm is publicly traded. The distribution of firms across entry cohorts also differs (Table 6). Matching therefore serves two purposes: it absorbs unobserved differences associated with entry cohort, and it balances covariates that Liao (2023) considers important.

We therefore match each treated firm to controls drawn from its own entry cohort. A firm first observed in 1992 has been filing H-1B petitions for two decades before our estimation window opens and is likely to differ substantially from one first observed in 2011. Each treated firm is matched to two controls, without replacement, on a propensity score that includes the 2011–2015 denial rates, petition volume, and the Orbis industry, size and listing indicators. The score is estimated across all cohorts.[11] We develop this matching procedure more generally in Section 7.[12]

Of the 65 treated firms in the balanced panel, 64 successfully merge with the Orbis covariates and are then matched to a total of 128 control firms. After matching, the differences in observable characteristics shrink considerably: cohort composition is identical by construction, the standardized difference on firm size falls from 1.5 to 0.04, and the total absolute difference in industry shares falls from 1.0 to 0.22. Outcome balance improves as well, but the 2011–2015 denial rates are included in the propensity score so the fall in their standardized difference from 0.515 to 0.019 is largely mechanical. The 2016 denial rate is the informative case, because it is withheld from the score. Its standardized difference falls from 0.378 to 0.134, which reflects genuine improvement rather than fit.

Matching also attenuates the apparent pre-treatment trend. When the full control pool serves as the comparison group on the balanced panel, the fitted trend implies that the treated–control gap narrows by 2.10 percentage points from 2011 to 2016. When the matched

[11] Entry cohort also enters the pooled score and is the only exact-matching constraint, ensuring that no match crosses cohorts.

[12] Liao (2023) matches on these covariates in a two-period specification (Imai et al., 2023) comparing 2016 with 2017 alone. Because it uses no pre-treatment outcomes, it cannot speak to whether parallel trends holds.

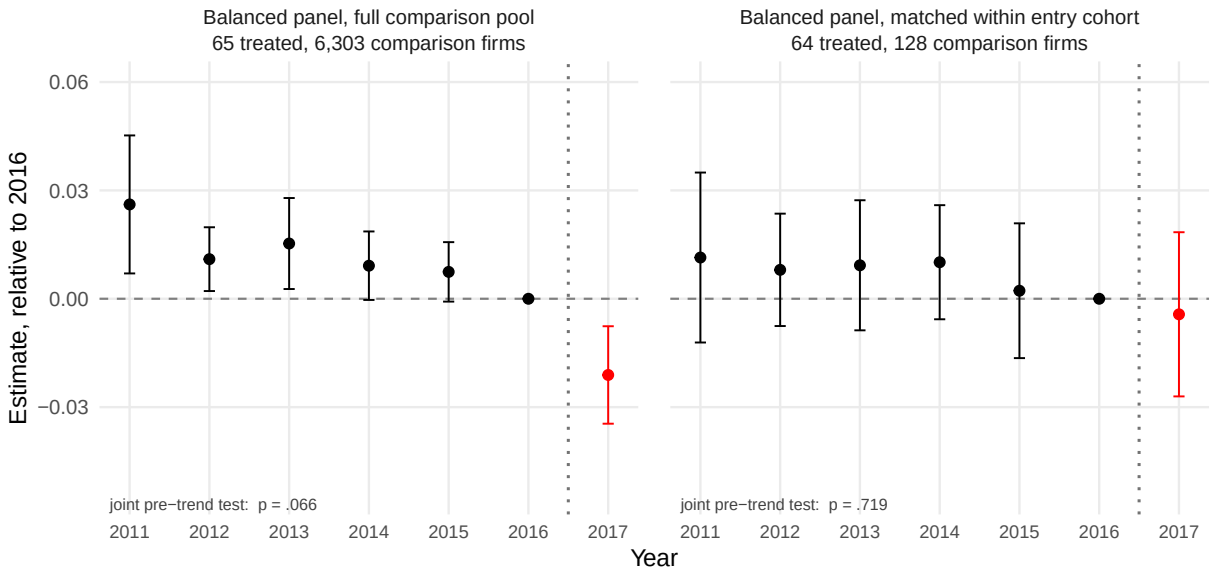

Figure 9: Liao (2023)'s event study with a balanced panel, before and after within-cohort matching.

Notes: Event-study interactions in the paper's own specification, with firm and year fixed effects, standard errors clustered by firm and 2016 omitted, unchanged between panels. The dotted line marks treatment onset. The left panel is row C of Table 3 and the right is row D, both on firms observed continuously from 2011 through 2017. The annotation gives the joint test of the five pre-treatment coefficients.

controls serve as the comparison group, the gap narrows by only 1.05 percentage points. The joint-test p-value rises from 0.066 to 0.719, and the linear-trend p-value rises from 0.006 to 0.258.[13] The matched coefficients nevertheless remain roughly one percentage point above the reference (Figure 9). We therefore interpret the matching as improving observed comparability, not as establishing that either PT-P or PT-W holds.

Matching changes not only the pre-treatment comparison but also the estimated effect. The 2017 coefficient moves further toward zero, to −0.43 percentage points with a standard error of 1.16 (Table 3, row D). The effect is no longer statistically distinguishable from zero. That is a substantively different conclusion from Liao (2023)'s reported estimates, −4.46 percentage points in the static specification and −2.69 in the event study. To be clear, our estimate is noisier than the original result and concerns a narrower sample: the 64 treated firms observed in every year from 2011 through 2017 that merge to the Orbis covariates.

[13]For the linear pre-trend test, a wild cluster bootstrap gives $p = 0.270$, compared with an analytic $p = 0.258$ (Cameron et al., 2008). Both are conditional on the selected match.

Table 3: Sample composition and the estimated effect at each step.

		Firms		Observed	Estimates	
	Sample	Treated	Control	throughout	2017 vs. 2016	Static TWFE
A	1992–2017 (published)	89	384,373	17% / <0.1%	**−2.69** (0.63)	**−4.46** (0.63)
B	2011–2017 window	86	126,786	76% / 5.0%	−2.73 (0.63)	−4.07 (0.75)
C	2011–2017 (balanced)	65	6,303	100% / 100%	−2.11 (0.69)	−3.26 (0.68)
D	+ matched within cohort	64	128	100% / 100%	−0.43 (1.16)	−1.11 (1.08)

Notes: Estimates in percentage points, with standard errors clustered by firm. Liao (2023)'s reported estimates are marked in bold. "Observed throughout" is the share of a sample's firms present in every year of its window, treated share first. The event-study column measures 2017 against 2016, as in Liao's Appendix Table D.14, column 2. The static column replaces the year-specific interactions with a single post-treatment indicator, measuring 2017 against all earlier years pooled, as in Liao's Figure 7 and Appendix Table D.6, column 1.

Nonetheless, the quantity it targets is better identified, because the pre-treatment diagnostic now rests on controls that track the treated firms before onset and on the same firms that identify the effect.

In Liao's panel, restricting the analysis to 2011–2017 preserves six pre-periods while making a balanced panel feasible. Matching within entry cohort then supplies a more comparable control group. Together, these changes produce a substantially different result. In other settings, fixed composition may be more costly. When balancing the full panel would discard most units, composition must instead be held fixed *within* cohorts. Section 7 develops that approach.

7 Within-Cohort Estimation

Requiring fixed composition throughout a panel can discard an excessive number of units. In both the application we explore below, and in a review of published work, we find this to be a common constraint (Clarke, 2020; Weschle, 2021). We instead propose a within-cohort estimation strategy that holds composition fixed *within* entry cohorts. This preserves units that enter in different periods while ensuring that each cohort's pre-treatment path and treatment effect concern the same set of units. This section develops the approach via an application to the 2015 Speaker election in the U.S. House of Representatives.

7.1 The Empirical Setting

The U.S. House elects its Speaker by floor vote at the opening of each Congress. In January 2015, 29 Republicans declined to support the sitting Speaker, John Boehner, a striking defection in a vote that is almost always along party lines. Party discipline is sustained in part by leaders' ability to reward loyalty and punish defection. One plausible lever of punishment runs through fundraising. Leadership is closely tied to a network of corporate donors, and defectors may find that access to those donors is subsequently curtailed (Canes-Wrone and Miller, 2022; Heberlig et al., 2006; Kates and Thieme, 2023; Issue One, 2017; Martin, 2021). Whether such punishment materializes has not, to our knowledge, been established empirically. We estimate the effect of defection on subsequent fundraising from leadership-connected corporate donors.

Our setting combines common treatment onset with differential entry and exit. All defectors are treated in the 114th Congress ($t^* = 114$). The defectors entered the House across five Congresses, from the 110th through the 114th, and therefore form five entry cohorts. We retain 23 of the 29 defectors, excluding six who did not run for reelection following the revolt and therefore have no observed post-treatment fundraising. The control pool consists of Republican House members who did not defect.

Our outcome, Y_{it}, is the log dollar amount that legislator i raises in Congress t from corporate donors connected to party leadership. We construct our outcome from the contribution records available in the Database on Ideology, Money in Politics, and Elections (Bonica, 2024). We classify a corporate donor as leadership-connected if it also contributes to the legislator's own-party leadership in the same Congress. We use $\log(\text{dollars} + 1)$ to accommodate zeros.

Entry timing determines the amount of pre-treatment history each legislator contributes. A legislator entering in the 110th Congress is observed as an incumbent for four periods before treatment, from the 110th through the 113th, whereas a legislator entering in the 114th has no incumbent pre-period. As a result, $T_{\text{pre}}(c)$ declines across entry cohorts while

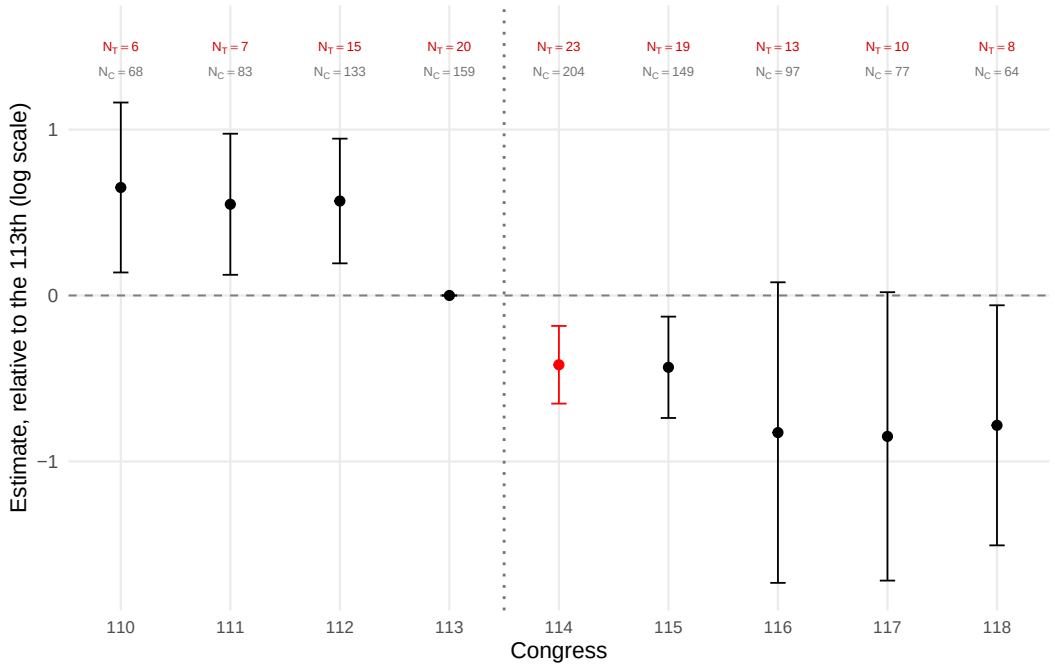

Figure 10: The pooled event study and panel composition by Congress.

Notes: Pooled event-study coefficients from Equation (1), estimated on the observed incumbent panel from the 110th through the 118th Congresses. The 113th Congress is omitted and the dotted line marks treatment onset in the 114th. N_T and N_C report the distinct treated and control legislators observed in the labeled Congress. Bars are 95% confidence intervals with standard errors clustered by legislator.

the post-period is shared, producing the differential-entry structure studied in Section 2.2.

7.2 The Pooled Event Study

We begin by applying the canonical event-study estimator of Equation (1) to the observed panel, pooling legislators across entry cohorts. The sample includes all 23 retained defectors and the full pool of non-defecting Republicans. We impose no restriction on entry or exit timing and thus include all available incumbent observations from the 110th through the 118th Congresses. The specification includes legislator and Congress fixed effects and omits the 113th Congress as the reference period. Figure 10 reports the estimates and the number of treated and control legislators observed in each Congress.

The three pre-treatment coefficients are large and statistically significant, at 0.65, 0.55 and 0.57 log points. They imply that the treated–control gap was substantially larger from

the 110th through the 112th Congresses than in the 113th, a change that appears inconsistent with PT-P.

The pooled estimates, however, do not necessarily indicate a violation of parallel trends. If treatment effects vary across entry cohorts, unit demeaning adds the bias derived in Section 2.2 to any genuine pre-treatment difference. The coefficients also describe changing sets of defectors. Only six of the 23 defectors are observed in the 110th Congress, with the remaining cohorts entering later. The coefficients therefore do not show how the defectors observed at onset were trending relative to a fixed group of controls before treatment.

The full pool of non-defecting Republicans creates a separate comparison problem. Defectors and controls differ substantially in their pre-treatment fundraising. Holding composition fixed does not ensure that these controls track the defectors before treatment onset.

7.3 The Within-Cohort Approach

We develop the within-cohort approach for the common-onset, differential-entry setting of Section 2.1.[14] We fix a common terminal period and retain units observed continuously from entry through that period. Entry is therefore sufficient to define cohorts with fixed composition.

We address the remaining comparison problem by matching treated units to controls within entry cohort. Our approach resembles the logic of PanelMatch (Imai et al., 2023), but constructs the eligible control pool based on entry cohort rather than lagged treatment histories. Within that pool, we match units on pre-treatment outcomes and covariates. We then estimate cohort-specific ATTs without unit fixed effects, thereby avoiding the demeaning that generates bias in pre-treatment coefficients.[15] The procedure has three steps.

1. **Define entry cohorts.** Assign units to cohorts according to their first observed

[14] Although the same demeaning mechanism at play in differential entry can also generate phantom dynamics under differential exit, we focus here on differential entry and return to dynamic effects below.

[15] Because each cohort is a balanced subpanel, the canonical event-study estimator could also be applied within cohort without generating the demeaning bias outlined above. However, the pre-treatment differences between defectors and control legislators would remain.

period. Retain units observed continuously from entry through the common terminal period. Each unit in cohort c is then observed over the same T_c periods. Only cohorts containing both treated and control units contribute to estimation.

2. **Match within cohort.** Match each treated unit to controls from the same cohort using pre-treatment outcomes and covariates. Exclude treated units without comparable controls. The resulting estimand applies only to the retained treated units, not to all treated units in the original sample.

3. **Estimate and aggregate.** Estimate the ATT separately within each matched cohort. Aggregate the cohort estimates so that every retained treated unit receives equal weight.

Let n_c^{ret} denote the number of retained treated units in cohort c and $\widehat{\tau}_c$ their estimated ATT. Giving each retained treated unit equal weight yields

$$\widehat{\tau}_{\text{ret}} = \frac{\sum_c n_c^{\text{ret}} \widehat{\tau}_c}{\sum_c n_c^{\text{ret}}}.$$

Under overlap and PT-W within each matched cohort, this aggregate identifies the ATT for the retained treated units. PT-W requires that, absent treatment, treated units and their matched controls within each cohort would have experienced the same average change in outcomes from the pre- to the post-treatment period. Matching on pre-treatment outcomes and covariates makes this assumption more credible, but is not sufficient to establish that it holds.

Inference depends on how the within-cohort comparisons are constructed. Standard errors from the final outcome regression condition on the realized matches or weights, omitting uncertainty based on the construction of the comparison groups. Analytic methods should account for this uncertainty directly, while resampling methods should repeat the matching or weighting procedure in each sample. If the treated units remain fixed under resampling, the resulting interval is conditional on the retained treated sample.

7.4 Implementation and Results

Implementation. The 23 defectors entered the House across five Congresses and are observed continuously from entry through the revolt. We end the panel in the 114th Congress and exclude control units with gaps in observation, allowing entry to uniquely define five cohorts with fixed composition. For legislators entering in the 113th and 114th Congresses, challenger-campaign fundraising adds one pre-treatment observation. It supplies a 112th-Congress outcome for the 113th cohort and the 113th-Congress reference for the 114th, which has no incumbent pre-period.[16] One defector in the 113th cohort lacks this challenger observation and is excluded, leaving 22 defectors.

We first estimate a separate event study within each entry cohort using all eligible non-defecting Republicans as controls. Each specification includes Congress fixed effects and treatment-by-Congress interactions but omits legislator fixed effects, as discussed above. The top panel of Figure 11 reports the estimates.

We then construct a matched control group within each entry cohort. We estimate a single logistic propensity-score model across the five cohorts using the available pre-treatment outcomes, ideology in the immediately preceding period (a dynamic CF score), and district vote share.[17] The model pools coefficients across cohorts, while exact matching on cohort restricts the eligible controls for each defector. To improve comparability, we exclude two defectors in the 112th cohort who, before treatment onset, raised less than every eligible control in their cohort. We match each of the remaining 20 defectors, without replacement, to the two controls with the closest estimated propensity scores. The bottom panel of Figure 11 reports the cohort-specific event studies estimated on these matched samples.

[16] The pooled specification in Section 7.2 excludes these observations because challenger outcomes from later cohorts would enter in the same period as incumbent outcomes from earlier cohorts. A common Congress fixed effect cannot absorb this cross-cohort difference in incumbency status. Within cohort, treated and control legislators have the same status in each period.

[17] For each pre-treatment Congress, the model includes an indicator for whether fundraising is observed and its interaction with log fundraising, with unobserved outcomes coded as zero. Challenger-campaign periods enter in the same form.

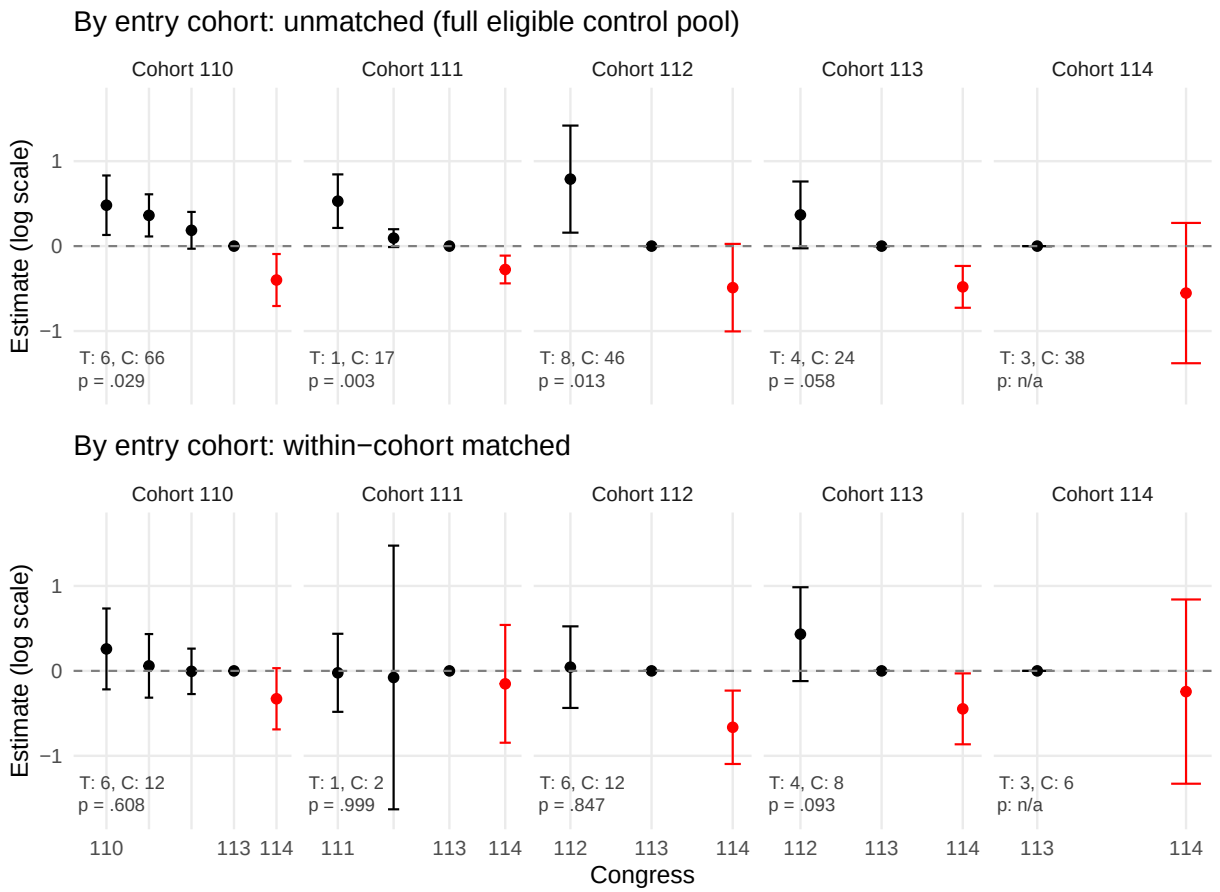

Figure 11: Cohort-specific event studies before and after matching.

Notes: Cohort-specific event studies estimated without legislator fixed effects. The top panel uses all eligible non-defecting Republicans within each entry cohort, the bottom uses 1:2 propensity-score matches from the same cohort. Each cohort reports treated and control counts, **T** and **C**, and the p-value from a joint test of its pre-treatment coefficients. Counts remain fixed across periods within each panel. The bottom panel excludes two defectors in the 112th cohort without comparable controls. The 113th Congress is omitted. The 114th cohort has no testable coefficient before onset. Bars are 95% confidence intervals with standard errors clustered by legislator.

Pre-treatment comparisons. The top panel holds composition fixed within entry cohort and compares the 22 retained defectors with all 191 eligible controls. The pre-treatment coefficients remain pronounced. Joint tests reject that all pre-treatment coefficients are zero in three of the four testable cohorts, while the 113th cohort yields a p-value of 0.058. Even after holding composition fixed and removing unit demeaning, pre-treatment differences remain between defectors and the full control pool.

The bottom panel reports cohort-specific event studies for the 20 retained defectors and

their matched controls. Across the four testable cohorts, the pre-treatment coefficients move toward zero and no joint test rejects the null. The smallest p-value is 0.09 in the 113th cohort. With only the reference-period observation before onset, the 114th cohort provides no pre-treatment test. The flatter pre-treatment paths make PT-W more plausible, although the improvement is partly by construction because pre-treatment outcomes enter the propensity score. In the Appendix, we provide additional evidence that the matched controls are similar on observables. Ideology and district vote share are similar in most cohorts, although the 114th cohort is somewhat less balanced (Appendix Figure 17).[18] The raw outcome trajectories also generally track before onset (Appendix Figure 15). Together, these diagnostics increase confidence that the matched controls approximate the fundraising path the retained defectors would have followed, absent defection.

Estimation and inference. Within each matched cohort, we estimate $\widehat{\tau}_c$ as a two-period difference-in-differences between defectors and matched controls across the 113th and 114th Congresses. We aggregate the five estimates using weights proportional to the number of retained defectors, n_c^{ret}. For inference, we run 1,000 bootstrap replications, resampling the eligible controls with replacement within entry cohort and rerunning the matching and estimation in each. The 20 retained defectors remain fixed across replicates. The resulting inference incorporates uncertainty from selecting the comparison units but remains conditional on the treated sample that defines the aggregate estimand.

Results. Table 4 reports the cohort-specific and aggregate estimates. Among the 20 retained defectors, we estimate a 0.43 log-point reduction in leadership-connected corporate fundraising (bootstrap SE 0.06, 95% percentile CI $[-0.60, -0.37]$), a roughly 35% decrease. Every cohort estimate is negative, ranging from -0.15 to -0.66 log points.

[18] The aggregate estimate is nearly unchanged when the 114th cohort is excluded.

Table 4: Cohort-specific and aggregate effects of defection on corporate fundraising.

	Defectors		
Entry cohort (Congress)	All (n_c)	Retained (n_c^{ret})	ATT (log scale)
110	6	6	−0.33
111	1	1	−0.15
112	8	6	−0.66
113	5	4	−0.45
114	3	3	−0.24
Total	**23**	**20**	**−0.43**

Notes: Rows are entry cohorts, defined by the Congress in which a defector first entered the House. The first count column lists all defectors in each cohort. The second lists those retained in the matched analysis. Cohort ATTs come from within-cohort 1:2 propensity-score matching. The aggregate estimate weights cohorts by n_c^{ret}.

Dynamic effects under differential exit. The main analysis avoids differential exit by ending the panel in the 114th Congress, the last period observed for every retained legislator. This holds composition fixed within cohort but limits the estimate to the treatment onset period. Extending the cohort event studies would change which defectors each coefficient describes, leaving the later coefficients not directly comparable to the treatment-onset estimate.

Later outcomes are nonetheless informative about whether the fundraising decrease persists among the legislators who remain in office. Appendix Figure 16 provides an illustrative extension to later periods. For each legislator with at least one post-114 observation, we average the available outcomes into a single observation so each legislator receives equal weight regardless of how many subsequent Congresses they are observed. Because the outcome averages across post-treatment periods, the resulting coefficient does not identify the effect for any particular Congress. It also rests on a smaller set of units, the 16 of the 20 retained defectors observed after the 114th.[19]

[19] If defection affects whether a legislator remains in office, conditioning on post-114 observation introduces selection. A causal interpretation therefore requires additional assumptions about the observation process.

Reporting. Within-cohort estimates are interpretable only if readers can see which units each estimate describes. We recommend reporting, for each cohort, the periods over which it is observed and its treated and control counts, along with any treated units dropped for missing outcomes or lack of comparable controls. Cohorts without enough pre-treatment history to assess PT-W should be identified explicitly. Finally, when cohort estimates are aggregated, cohort-specific paths should be reported alongside the aggregate, together with the weights used to construct it.

Staggered adoption. With staggered adoption, cohorts must be defined jointly by entry and adoption timing. Differential exit adds an additional problem, because later coefficients then describe only the units still present. Estimators robust to heterogeneous treatment effects are not a substitute. Section 5 shows that imbalance survives within adoption cohorts. Defining cohorts on more dimensions shrinks the number of units sharing a cell, so overlap is the practical constraint on this extension.

8 Discussion

In recent years, numerous papers have demonstrated issues with two-way fixed effect designs under effect-heterogeneity. However, this work has primarily focused on settings with staggered treatment-adoption, treating the common-onset setting (where ever-treated units enter treatment in the same period) as safe, and typically assumes a balanced panel in their theoretical setup, despite settings where units are observed for different sets of periods being quite common.

This paper shows that even in the common treatment-onset setting, panel imbalance creates two distinct problems for the canonical event-study estimator. First, we prove that the estimator is biased under effect-heterogeneity. When units enter (exit) the sample in different periods, the estimator may fabricate non-existent, or conceal genuine, pre-trends (dynamic effects). Second, even setting the bias aside, the estimator targets the wrong

estimand for assessing identification and dynamic effects. Coefficients may be estimated on a different set of units than those used to estimate the treatment effect. Therefore, pre-treatment coefficients for periods before all units have entered do not provide evidence of how the units identifying the ATT were trending prior to treatment. And analogously, post-treatment coefficients after some units have exited are not directly comparable to the treatment effect. Researchers may conflate the changing composition of units with evidence against or for pre-trends or dynamic effects.

Existing HTE-robust estimators do not generally resolve both problems. Some avoid the demeaning bias, but differential entry or exit can still make their coefficients describe changing populations. We instead define effects within cohorts of units observed in the same periods. This holds composition fixed, permits comparisons against cohort-specific controls, and avoids the unit demeaning that generates phantom pre-trends and dynamics.

We find the estimand problem in three published applications (Liao, 2023; Weschle, 2021; Clarke, 2020). In each, the published parallel-trends diagnostic compares changing sets of units across the pre-period. We examine Liao (2023) in detail. In this setting, restricting the analysis to 2011–2017 makes a balanced panel feasible while retaining most treated firms. The canonical event study estimates display a declining pre-treatment path, even after balancing the panel. Matching within entry cohort flattens the pre-treatment trajectory and yields a substantively different result. The estimate attenuates toward zero and is estimated imprecisely, though the balanced panel describes a narrower set of firms.

In the Boehner application, the canonical event study is estimated on an imbalanced panel and displays three large, positive pre-treatment coefficients. Here, dropping units to create a balanced panel would retain only six of the 23 defectors. We therefore estimate the event study separately within entry cohort. This accounts for differential entry and avoids unit demeaning, but pre-treatment differences remain between treated units and the full control pool. We then match within cohort to construct a more comparable control group, which moves the pre-treatment coefficients toward zero. Aggregating the matched cohort estimates

yields an estimated 0.46 log-point reduction in leadership-connected corporate fundraising among the 20 retained defectors, or approximately 37%.

Event-study coefficients are informative about parallel trends and treatment dynamics only to the extent they are unbiased and describe a common set of units. In an unbalanced panel, neither condition is guaranteed. Researchers must therefore account for both unit demeaning and changing composition before interpreting the estimated path.

References

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo, and James A. Robinson. 2019. "Democracy Does Cause Growth." *Journal of Political Economy*, 127(1): 47–100.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press. (§5.2, regression DiD with leads and lags.)

Autor, David H. 2003. "Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing." *Journal of Labor Economics*, 21(1): 1–42.

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

Bonica, Adam. 2024. Database on Ideology, Money in Politics, and Elections: Public version 4.0 [Computer file]. Stanford, CA: Stanford University Libraries. https://data.stanford.edu/dime.

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 H. C. 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.

Canes-Wrone, Brandice, and Kenneth M. Miller. 2022. "Out-of-District Donors and Representation in the US House." *Legislative Studies Quarterly*, 47(2): 361–395.

Chiu, Albert, Xingchen Lan, Ziyi Liu, and Yiqing Xu. 2025. "Causal Panel Analysis under Parallel Trends: Lessons from a Large Reanalysis Study." *American Political Science Review*. Forthcoming.

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

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

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.

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

Heberlig, Eric, Marc Hetherington, and Bruce Larson. 2006. "The Price of Leadership: Campaign Money and the Polarization of Congressional Parties." *The Journal of Politics*, 68(4): 992–1005.

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.

Issue One. 2017. "The Price of Power: How Political Parties Squeeze Influential Lawmakers to Boost their Campaign Coffers." Technical Report, Issue One.

Jacobson, Louis S., Robert J. LaLonde, and Daniel G. Sullivan. 1993. "Earnings Losses of Displaced Workers." *American Economic Review*, 83(4): 685–709.

Kates, Sean, and Sebastian Thieme. 2023. "Fundraising Events and Non-Ideological Donation Motivations." IAST Working Paper No. 159, Institute for Advanced Study in Toulouse. https://publications.ut-capitole.fr/id/eprint/48740/.

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.

Liu, Ziyi. 2025. "Cohort-Anchored Robust Inference for Event-Study with Staggered Adoption." Working paper, arXiv:2509.01829.

Martin, Jonathan. 2021. "Liz Cheney's Consultants Are Given an Ultimatum: Drop Her, or Be Dropped." *The New York Times*, October 21. https://www.nytimes.com/2021/10/21/us/politics/liz-cheney-kevin-mccarthy.html.

Paglayan, Agustina S. 2021. "The Non-Democratic Roots of Mass Education: Evidence from 200 Years." *American Political Science Review*, 115(1): 179–198.

Roth, Jonathan. 2022. "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends." *American Economic Review: Insights*, 4(3): 305–322.

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

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.

A Proof of Proposition 2: the Exact Phantom Coefficients

We derive the closed form for the pre-treatment coefficients $\hat{\beta}_k$, $k < t^*$, from Equation (1) under the data-generating process of Equation (2). Throughout, $r \equiv t^* - 1$ is the omitted reference period, $N \equiv \sum_c n_c$ the number of treated units, and $\bar{\tau} \equiv \frac{1}{N}\sum_c n_c \tau_c$ their count-weighted average effect.

Step 1: define the least-squares problem. As in the main text, we let $\mathrm{ES}_k(w)$ denote the kth-period coefficient that the canonical event-study estimator returns when run on outcome w. By the decomposition in Corollary 1,

$$\hat{\beta}_k = \beta_k + b_k$$

where $\beta_k = \mathrm{ES}_k[Y(0)]$ and $b_k = \mathrm{ES}_k[W]$ with $W_{it} \equiv D_i \, \mathbf{1}(t \geq t^*) \, \tau_{c(i)}$.

Under (2), the untreated potential outcome is μ_1 in each period for every treated unit and μ_0 in each period for every control unit. It is constant within each unit and thus absorbed entirely by the unit fixed effects. Consequently,

$$\beta_k = 0 \implies \hat{\beta}_k = b_k.$$

Therefore, our goal is to solve $\hat{\beta}_k = b_k = \mathrm{ES}_k[W]$, where

$$W_{it} = \begin{cases} 0 & D_i = 0, \\ 0 & D_i = 1, \ t < t^*, \\ \tau_{c(i)} & D_i = 1, \ t \geq t^*, \end{cases} \tag{10}$$

and $\mathrm{ES}_k[W]$ are the $\hat{\beta}_k$ interaction coefficients returned from estimating the following event

study on W,

$$W_{it} = \alpha_i + \lambda_t + \sum_{k \neq t^*-1} \beta_k D_i \, \mathbf{1}(t = k) + \varepsilon_{it}. \tag{11}$$

The period effects λ_t are identified from the control units alone. Controls have $W_{it} = 0$ throughout, so they are fit exactly by $\hat{\alpha}_i = \hat{\lambda}_t = 0$.

Treated units in a cohort are identical, so we collapse their unit fixed effects into a common cohort-level effect a_c. Given $\hat{\lambda}_t = 0$, we can define the fitted value for a treated unit in cohort c at period t as $\hat{y}_{ct} = \hat{a}_c + \hat{\phi}_t$. With cohort effects $\{a_c\}$ and period effects $\{\phi_t\}$, the regression in (11) is thus solved via the following least squares problem fit on treated units only[20]

$$\min_{\{\hat{a}_c\},\{\hat{\phi}_t\}} \sum_c n_c \left[\sum_{t=e_c}^{t^*-1} \left(0 - \hat{a}_c - \hat{\phi}_t\right)^2 + \sum_{t=t^*}^{\mathcal{T}} \left(\tau_c - \hat{a}_c - \hat{\phi}_t\right)^2 \right]. \tag{12}$$

Given the controls are fixed at 0, the estimated event-study coefficient $\hat{\beta}_k$ is the change in the treated fitted value from the reference period r to period k. The cohort effect a_c cancels, leaving

$$\hat{\beta}_k = \hat{\phi}_k - \hat{\phi}_r. \tag{13}$$

Note that this is consistent with the reference period being omitted, since $\hat{\beta}_r = \hat{\phi}_r - \hat{\phi}_r = 0$ mechanically.

Step 2: solving (12). We solve (12) through its first-order conditions, differentiating the objective J with respect to each parameter and setting it to zero. We use the following notation:

- $N \equiv \sum_c n_c$ and $\bar{\tau} \equiv \frac{1}{N} \sum_c n_c \tau_c$ (as above): the total number of treated units and the count-weighted average effect.
- $\bar{a} \equiv \frac{1}{N} \sum_c n_c a_c$: the count-weighted average cohort effect.

[20] because $\hat{\lambda}_t$ and α_i for control units are pinned to 0, rendering control units' residuals equal to 0, they need not factor into the objective.

- $S_k \equiv \{c : e_c \leq k\}$: the cohorts that have entered by period k (i.e. those observed at k).
- $N_k \equiv \sum_{c \in S_k} n_c$ and $\bar{a}_{S_k} \equiv \frac{1}{N_k} \sum_{c \in S_k} n_c a_c$: the count and average cohort effect over just those cohorts.

Notice the objective (12) splits into a pre-period sum $\sum_c n_c \sum_{t=e_c}^{t^*-1} (-\hat{a}_c - \hat{\phi}_t)^2$, and a post-period sum $\sum_c n_c \sum_{t=t^*}^{\mathcal{T}} (\tau_c - \hat{a}_c - \hat{\phi}_t)^2$, which we differentiate separately.

Post-period effect p. All cohorts are present in every post-period and have constant outcomes across the post-periods, so the post-period effects ϕ_t ($t \geq t^*$) share a single common value, which we denote p. Only the post-block depends on p, and it collapses to $\sum_c n_c T_{\text{post}} (\tau_c - a_c - p)^2$; differentiating with respect to p:

$$\begin{aligned}
\frac{\partial J}{\partial p} &= \sum_c n_c T_{\text{post}} \frac{\partial}{\partial p} \big(\tau_c - a_c - p\big)^2 \\
&= -2\, T_{\text{post}} \sum_c n_c \big(\tau_c - a_c - p\big) \overset{\text{set}}{=} 0 \\
\implies &\quad \sum_c n_c \tau_c - \sum_c n_c a_c - p\, N = 0 \\
\implies &\quad p = \bar{\tau} - \bar{a}.
\end{aligned} \tag{14}$$

Pre-period effects ϕ_t ($t < t^*$). For a pre-period $t < t^*$, the effect ϕ_t is estimating using only the cohorts observed at t, namely S_t. Cohorts that have not yet entered contribute nothing. We can therefore write:

$$\begin{aligned}
\frac{\partial J}{\partial \phi_t} &= \sum_{c \in S_t} n_c \frac{\partial}{\partial \phi_t} \big(0 - a_c - \phi_t\big)^2 \\
&= 2 \sum_{c \in S_t} n_c \big(a_c + \phi_t\big) \overset{\text{set}}{=} 0 \\
\implies &\quad \phi_t = -\,\bar{a}_{S_t}.
\end{aligned} \tag{15}$$

Each pre-period effect is thus minus the average cohort effect among the cohorts present

that period. Because S_t grows as t increases (later cohorts enter), $\bar{a}_{S_t}$ — and hence ϕ_t — shifts across pre-periods; this movement is the phantom.

Cohort effect a_c. Differentiating with respect to a_c:

$$\begin{aligned}
\frac{\partial J}{\partial a_c} = n_c \left[\sum_{t=e_c}^{t^*-1} \big[-2(0 - a_c - \phi_t) \big] + \sum_{t=t^*}^{\mathcal{T}} \big[-2(\tau_c - a_c - p) \big] \right] &\overset{\text{set}}{=} 0 \\
\implies \quad \sum_{t=e_c}^{t^*-1} (0 - a_c - \phi_t) + \sum_{t=t^*}^{\mathcal{T}} (\tau_c - a_c - p) &= 0 \\
\implies \quad -T_{\text{pre}}(c)\, a_c - \sum_{t=e_c}^{t^*-1} \phi_t + T_{\text{post}}\, \tau_c - T_{\text{post}}\, a_c - T_{\text{post}}\, p &= 0 \\
\implies \quad T_c\, a_c = T_{\text{post}}\, \tau_c - \sum_{t=e_c}^{t^*-1} \phi_t - T_{\text{post}}\, p, & \qquad (16)
\end{aligned}$$

using $T_{\text{pre}}(c) + T_{\text{post}} = T_c$.

Every cohort is present at the reference period ($e_c < t^*$, so $e_c \leq r$), hence $\hat{\phi}_r = -\bar{a}$. Combining with (15),

$$\hat{\beta}_k \;=\; \hat{\phi}_k - \hat{\phi}_r \;=\; \bar{a} - \bar{a}_{S_k}, \qquad k < t^*. \qquad (17)$$

The phantom coefficient at pre-period k is thus the gap between the average treated cohort effect over the full sample and its average over only the cohorts observed at k. Substituting (14) and (15) into (16) gives the following for $\{a_c\}$:

$$a_c \;=\; \frac{1}{T_c} \left(T_{\text{post}}\, (\tau_c - \bar{\tau}) \;+\; T_{\text{post}}\, \bar{a} \;+\; \sum_{t=e_c}^{t^*-1} \bar{a}_{S_t} \right). \qquad (18)$$

Step 3: proof by contradiction of Proposition 2. By assumption effects are heterogeneous: $\tau_c \neq \tau_{c'}$ for some $c \neq c'$. Now suppose, for contradiction, that $\hat{\beta}_k = 0$ for all $k < t^*$. By (17) this means $\bar{a}_{S_k} = \bar{a}$ for every pre-period k. Order cohorts by entry, $e_1 < \cdots < e_C$. At $k = e_1$, $S_{e_1} = \{1\}$, so $a_1 = \bar{a}$. For $j = 2, \ldots, C-1$, the set $S_{e_j} = S_{e_{j-1}} \cup \{j\}$ averages $\bar{a}$ having added cohort j to a set already averaging $\bar{a}$, which forces $a_j = \bar{a}$; then $a_C = \bar{a}$ follows from $\sum_c n_c a_c = N\bar{a}$. With all $a_c = \bar{a}$, the sum in (18) equals $T_{\text{pre}}(c)\, \bar{a}$, so (18) becomes

$T_c\bar{a} = T_{\text{post}}(\tau_c - \bar{\tau}) + T_{\text{post}}\bar{a} + T_{\text{pre}}(c)\bar{a} = T_{\text{post}}(\tau_c - \bar{\tau}) + T_c\bar{a}$. Hence $T_{\text{post}}(\tau_c - \bar{\tau}) = 0$, and since $T_{\text{post}} \geq 1$, $\tau_c = \bar{\tau}$ for all c, which is a contradiction. Therefore, $\hat{\beta}_k \neq 0$ for some $k < t^*$. □

A.1 Illustrative Examples

Closed form with two cohorts. This example illustrates via a simple case with two cohorts, what factors affect the magnitude of the bias.

When $C = 2$, solving (18) gives $a_2 - a_1 = \frac{T_{\text{post}}}{T_2}(\tau_2 - \tau_1)$, so by (17) the coefficient at any pre-period observing only the earlier-entering cohort is

$$\hat{\beta}_k \;=\; \frac{n_2}{N}\,\frac{T_{\text{post}}}{T_2}\,(\tau_2 - \tau_1), \tag{19}$$

while $\hat{\beta}_k = 0$ at pre-periods observing both. The phantom is nonzero precisely when $\tau_1 \neq \tau_2$, and grows with the effect gap, the later cohort's share n_2/N, and the fraction T_{post}/T_2 of that cohort's panel lying after treatment.

Verifying closed-form solution against simulation (Figure 2). In Figure 2 of the main text, we simulated a panel using the proof's DGP. Here we verify that the proof's closed-form solution for $\hat{\beta}_k$ matches the simulation results.

We have three equal-sized cohorts observed for $T_{\text{pre}} = 3, 2, 1$ pre-periods (so $T_c = 6, 5, 4$, $T_{\text{post}} = 3$) with $\tau = (-10, -6, -2)$, hence $\bar{\tau} = -6$ and $\tau_c - \bar{\tau} = (-4, 0, 4)$. Fixing $\bar{a} = 0$ without loss[21], and then solving (18) gives $a = \{-2.7,\ -0.3,\ 3.0\}$. Then using (17), we get

$$\hat{\beta}_{-3} = \bar{a} - a_1 = 2.7, \qquad \hat{\beta}_{-2} = \bar{a} - \tfrac{a_1 + a_2}{2} = 1.5, \qquad \hat{\beta}_{-1} = 0,$$

matching the simulated coefficients of Figure 2 exactly.

[21] Only the a_c relative to one another matter: adding the same constant to every a_c and subtracting it from every ϕ_t leaves all fitted values, and hence every $\hat{\beta}_k$, unchanged. We use this freedom to set their average to zero.

A.2 Further Remarks

Adding common time trends $\lambda_t \neq 0$. DGP (2) assumed there was no common time trend between treated and control units (i.e. $\lambda_t = 0$). We note here that adding non-zero period effects to (2) does not affect the result in Proposition 2, because they are absorbed exactly by λ_t and thus, as before, drop out of the treated-only objective in (12).

Adding cross-cohort baseline differences. The DGP (2) assumed every cohort enters with the same pre-treatment baseline outcome. We now allow cohort-specific baselines, such that PT-P fails but PT-W holds, and show the phantom result is unchanged.

Let a treated unit in cohort c have constant outcome $\mu_1 + \gamma_c$ before treatment (controls at μ_0), with the level γ_c varying across cohorts. Each unit is still constant over its pre-period, so within any cohort, treated and control remain parallel (PT-W holds). But because the treated-control gap now differs across cohorts, the pooled gap drifts as the observed composition of cohorts changes (PT-P fails).

No unit trends in the pre-period, and PT-W — the condition that holds composition fixed, and is thus, in this paper's view, the relevant identifying assumption — is satisfied, so the pre-trend of interest is genuinely zero. The estimator nonetheless reports the exact same phantom coefficients. By Corollary 1, $\hat{\beta}_k = \mathrm{ES}_k[Y(0)] + \mathrm{ES}_k[W]$. The cohort-specific baselines enter only $Y(0)$, and since each unit's $Y(0)$ is constant over time (just at a cohort-specific level), it is absorbed entirely by that unit's fixed effect, so $\mathrm{ES}_k[Y(0)] = 0$. The phantom again comes only from W, which contains no baselines, leaving $\hat{\beta}_k = \bar{a} - \bar{a}_{S_k}$ exactly as in (17).

B Further Published Designs with Panel Imbalance

This appendix reports two additional published designs in which the evidence available to assess parallel trends rests on a sample different than the one used to estimate the main

Table 5: MPs behind the seven entry panels of Weschle (2021) Figure 3.

	MPs observed			Share of year taken	
Position	−2	−1	0	−2	−1
Minister	36	37	65	55%	57%
Minister of state	40	40	59	68%	68%
Parliamentary secretary	71	71	97	73%	73%
Frontbench team	52	56	108	48%	52%
Committee chair	47	49	72	65%	68%
Shadow cabinet	**40**	**81**	**218**	**18%**	**37%**
Committee membership	**27**	**57**	**415**	**7%**	**14%**

Notes: Event time is measured in years relative to taking up the position, with 0 denoting the year of entry. Each share divides the pre-position count by the count at event time 0.

effect. The main results from both papers were reproduced from the authors' own replication packages before any of the counts below were computed.

Weschle (2021): uneven composition across panels. Weschle (2021) asks whether taking up an influential position in the UK House of Commons increases an MP's private-sector earnings. We reproduce the paper's main estimates in Table 1 exactly. Its parallel-trends evidence comes from Figure 3, which contains fourteen panels covering entry into and departure from seven positions. The seven entry panels compare average logged earnings among MPs who take up each position with those of MPs who never hold it. They cover the two years before entry, the year of entry and the first two years afterward. The paper summarizes all fourteen panels with a single claim: "there are no significantly diverging pre-position earnings trends for any of these positions." We focus on the seven entry panels.

The number of MPs supporting each pre-treatment paths varies substantially across panels and event times. Using the position-timing variables distributed with the paper, Table 5 reports the MPs observed at each displayed point.

At event time −2, coverage ranges from 7% for committee membership to 73% for parliamentary secretary. The committee-membership panel rests on 27 MPs at that point, compared with 415 at entry. The paper's common characterization therefore covers panels whose early pre-position observations retain most of the MPs observed at entry and pan-

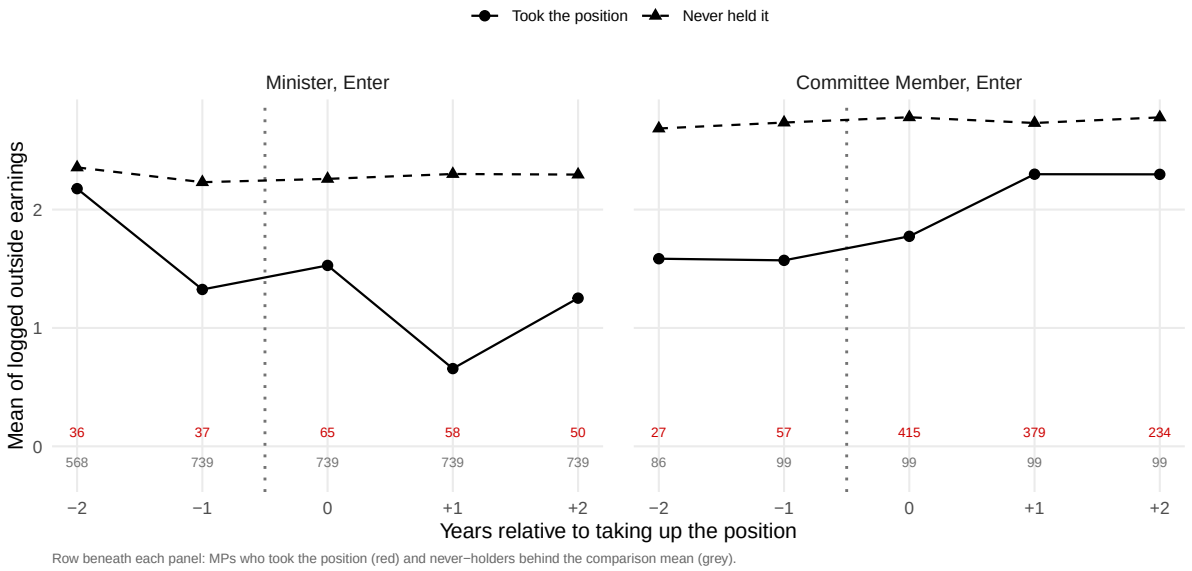

Figure 12: Two panels of Weschle (2021) Figure 3, with the MPs behind each plotted point.

Notes: Mean logged real outside earnings by event time, reproduced from the author's replication files. The comparison series averages earnings among MPs who never hold the position over the calendar years in which each position-taker is observed, then re-indexes those means to that MP's event time. Rows beneath each panel report the position-takers and never-holders contributing to each point.

els where they describe only a small subset. Figure 12 adds the underlying counts to the minister and committee-membership panels.

Clarke (2020): changing composition across the pre-period. Clarke (2020) asks whether joining an ideological faction changes a U.S. House member's donor base. The paper estimates a staggered difference-in-differences among Democrats using candidate and Congress fixed effects. We reproduce all five specifications in Appendix Table A3 to the two decimals reported. The underlying panel is severely unbalanced. Only 3,772 of 12,155 possible candidate–Congress cells are observed, and the median candidate appears in 5 of the 17 Congresses. Candidates enter across all 17 Congresses, while 73% of candidates leave before the final one. Only 3% of units have interior gaps in observation, so the imbalance primarily reflects differential entry and exit rather than intermittent missingness.

The paper reports no event study. We construct one using the author's estimation sample and specification, then add the counts behind each coefficient (Figure 13). Across the four

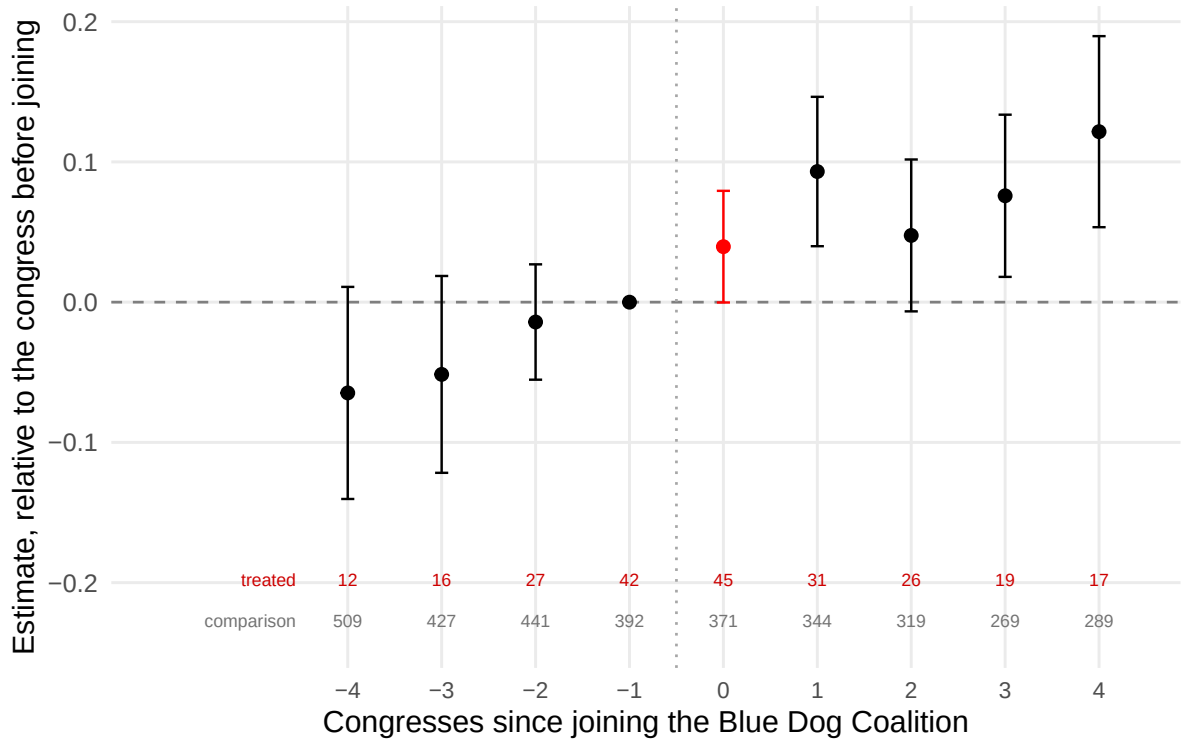

Figure 13: Event study based on the data and specification of Clarke (2020), with the counts behind each coefficient.

Notes: Event-study interactions for Blue Dog Coalition membership, estimated on the paper's sample of Democratic candidates. The specification includes candidate and Congress fixed effects and the paper's four controls, with standard errors clustered by candidate. Event time is measured in Congresses relative to joining, with the preceding Congress omitted and the tails binned at −4 and +4. The 38 candidates who are already members when first observed have no pre-treatment period and are excluded, leaving 45 with an observed onset. Rows beneath the panel report the treated and control candidates behind each coefficient.

pre-treatment periods, the treated count rises from 12 to 42, compared with 45 at treatment onset. The control count remains between 269 and 509. The pre-treatment coefficients therefore describe successively expanding treated populations rather than the full group observed at onset. In this application, the change in support lies primarily on the treated side.

C Target Populations and Inference in the Liao (2023) Diagnostics

This appendix supplements the reanalysis of Liao (2023) in Section 6. It reviews the paper's additional diagnostic evidence, reports the balance achieved by our matching procedure, examines the sensitivity of inference to the cluster-robust approximation, and shows how sample retention changes across fixed-composition windows.

The paper's other robustness checks. Beyond the raw-means figure, Liao (2023) reports an event-study specification in Appendix Table D.14, two placebo specifications in neighboring columns, and a matched difference-in-differences comparing 2016 with 2017 (Imai et al., 2023). None of those analyses are presented as a parallel-trends display. Section 6.1 reproduces and plots the event-study coefficients, showing that the units supporting them change substantially across years. The placebo specifications test whether the estimated effect appears at a false onset date or among firms that lobbied in 2016. Because neither test examines the pre-treatment path, neither speaks directly to the parallel-trends diagnostic.

The matched design is a two-group, two-period comparison. All fourteen specifications compare 2016 with 2017 and exclude observations from prior periods. Matching holds the comparison population fixed, leaving 70 treated firms, but provides only one pre-treatment period. As a result, the reported balance concerns outcome levels rather than trends of treated and control units in the periods preceding onset. The paper's evidence on pre-treatment trends therefore comes from the pooled display and the event-study specification. The tables below provide additional detail on both.

Balance and control-group choice. Table 6 reports balance for the within-cohort matching specification in row D of Table 3. It compares the resulting sample with the unmatched control pool and with an otherwise identical match that does not require treated and control

firms to share an entry cohort.

The choice of firm covariates follows Liao (2023). His matching code exact-matches on two-digit NAICS industry and firm size class, and assesses balance on denial rates, industry, size, public listing, sales and employment. We use industry, size and listing, which are observed for 93% of control firms and 98% of treated firms in the balanced panel. Sales and employment are too sparsely populated to use in matching. We supplement these variables with entry cohort, to account for unobservables likely correlated with entry timing (Section 4), and with pre-treatment denial rates and petition volume.[22]

Because the 2011–2015 denial rates enter the propensity score, improved balance over those years cannot independently validate the matched design. The more informative comparison is the withheld 2016 outcome. Its standardized difference falls from 0.378 to 0.134, even though 2016 was not used to estimate the score. This remains evidence about balance in outcome levels, not trends.

Petition volume shows the largest remaining imbalance, with a standardized difference of about 0.33 under both matching specifications. The alternative that includes entry cohort only in the propensity score achieves somewhat better balance on public listing and the held-out 2016 outcome, but leaves a cohort-composition gap of 0.27. We prefer the within-cohort specification because it ensures that the control group reproduces the treated firms' entry composition. That choice is made on design grounds rather than in response to the resulting estimate.

The choice of estimation window. Table 7 demonstrates how sample retention and the 2017 estimate vary with the window over which fixed composition is imposed. Each row retains only firms observed in every year from the indicated starting year through 2017. At treatment onset, 74 treated and 38,188 control firms are observed

Treated-firm retention falls from roughly 89% to 64% as the window expands to sixteen

[22]The firm characteristics are not included in the replication package. We retrieved them from Orbis on 23–24 August 2026 by matching on Bureau van Dijk identifiers, so they may not correspond to the vintage used in the published analysis.

Table 6: Balance before and after matching on the fixed-composition sample.

		Propensity score matched	
	Full control pool	Plus exact on cohort	Cohort in score only
Control firms	5,854	128	128
Composition: total absolute difference in category shares			
Entry cohort	1.00	**0.00**	0.27
Industry	1.02	0.22	0.23
Absolute standardized difference in means			
Denial rate, 2011–15	0.515	0.019	0.169
Denial rate, 2016 (held out)	**0.378**	**0.134**	**0.125**
Petition volume	1.151	0.325	0.330
Size class	1.532	0.039	0.000
Listed	1.577	0.287	0.144

Notes: All columns use the same 64 treated firms with Orbis covariates. Column 2 corresponds to row D of Table 3. The composition measure is the total absolute difference between the treated and control category shares. It ranges from 0 when the distributions coincide to 2 when they are disjoint. Each matched specification selects two controls per treated firm without replacement using nearest-neighbor matching on the propensity score. Both scores include entry cohort, industry, size class, public listing, 2011–2015 denial rates and petition volume. Column 2 also requires exact matching on entry cohort. The 2016 denial rate is withheld from the score and provides a held-out balance check.

Table 7: Sample retention and estimated effects by window length.

Window	Pre-periods	N_T	N_C	N_T retained	2017 estimate
2013–2017	4	66	8,467	89%	−2.14 (0.67)
2012–2017	5	65	7,294	88%	−2.22 (0.68)
2011–2017	6	65	6,303	88%	−2.11 (0.69)
2010–2017	7	62	5,386	84%	−1.89 (0.72)
2009–2017	8	60	4,651	81%	−1.59 (0.72)
2007–2017	10	52	3,672	70%	−1.39 (0.78)
2005–2017	12	50	2,768	68%	−1.32 (0.77)
2003–2017	14	49	1,981	66%	−0.68 (0.80)
2001–2017	16	47	1,497	64%	−0.70 (0.84)
1999–2017	18	35	707	47%	−0.49 (1.07)
1997–2017	20	25	332	34%	−0.64 (1.36)
1995–2017	22	21	179	28%	+0.21 (1.32)
1992–2017	25	15	49	20%	−1.46 (1.86)

Notes: Each row retains firms observed in every year of the indicated window. *Treated retained* is the percentage of the 74 treated firms observed at treatment onset that remain after imposing this requirement. The estimate is the 2017 coefficient from the event-study specification used in Table 3, measured in percentage points with firm-clustered standard errors in parentheses.

pre-treatment years, then declines more sharply to 20% over the full published period. The longer windows are constrained primarily by entry timing. A window beginning in 2001 can

retain firms that entered in any earlier year, whereas a window beginning in 1992 necessarily excludes every firm that entered later. Control-firm retention falls more rapidly throughout, reproducing the asymmetry underlying the published display. The estimated effect also moves toward zero, from -2.11 with six pre-treatment years to -0.70 with sixteen. Because each window retains a different population, these estimates are not a conventional robustness sequence. We report the full range to make transparent the tradeoff between pre-treatment history and sample retention

D Additional Empirical-Application Diagnostics

This appendix provides supporting diagnostics for the within-cohort matching estimator in Section 7.3. It reports the candidate-level pre-treatment outcomes used for support trimming (Figure 14), matched raw trajectories within cohort (Figure 15), cohort event studies extended one period beyond the common endpoint (Figure 16), and covariate balance after matching (Figure 17)

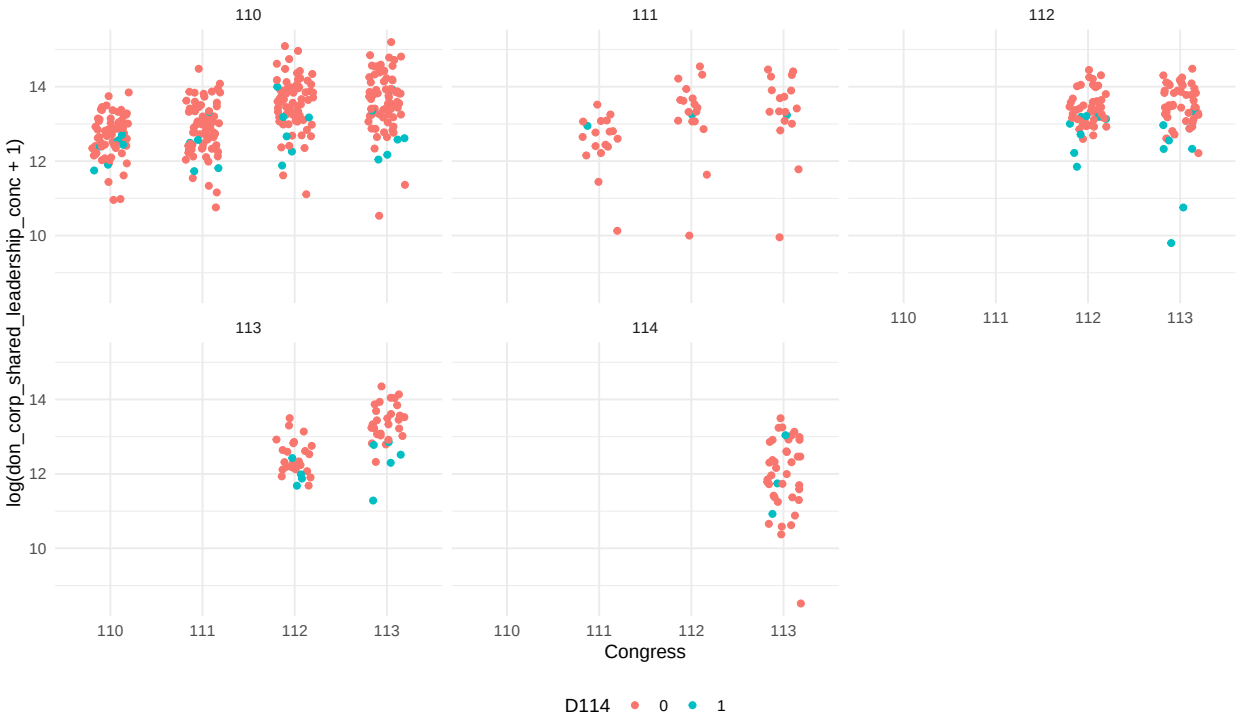

Figure 14: Candidate-level pre-treatment outcomes, by entry cohort.

Notes: Each point is a legislator's pre-treatment log corporate outcome in a given Congress, faceted by entry cohort. Color distinguishes defectors (D114 = 1) from non-defecting controls (D114 = 0). Defectors whose pre-period outcomes fall outside the support of the controls in their cohort are dropped before matching, two in the 112th cohort and one in the 113th. Each marker is a single legislator–Congress observation, so no confidence intervals are shown.

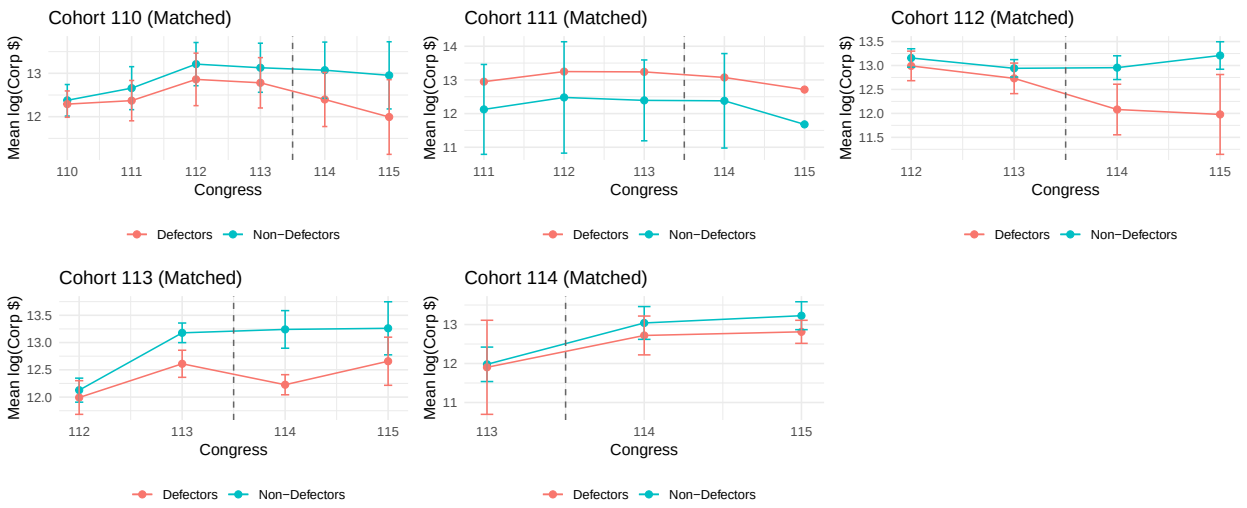

Figure 15: Within-cohort raw trajectories for matched defectors and their controls.

Notes: Mean log corporate fundraising for matched defectors and non-defectors, by Congress, faceted by entry cohort. Bars are ±1.96 times the standard error of each group mean within a Congress. That is the sampling error of the plotted raw mean, unclustered, not the legislator-clustered interval of the event-study figures. The dashed line marks treatment onset at the 114th Congress.

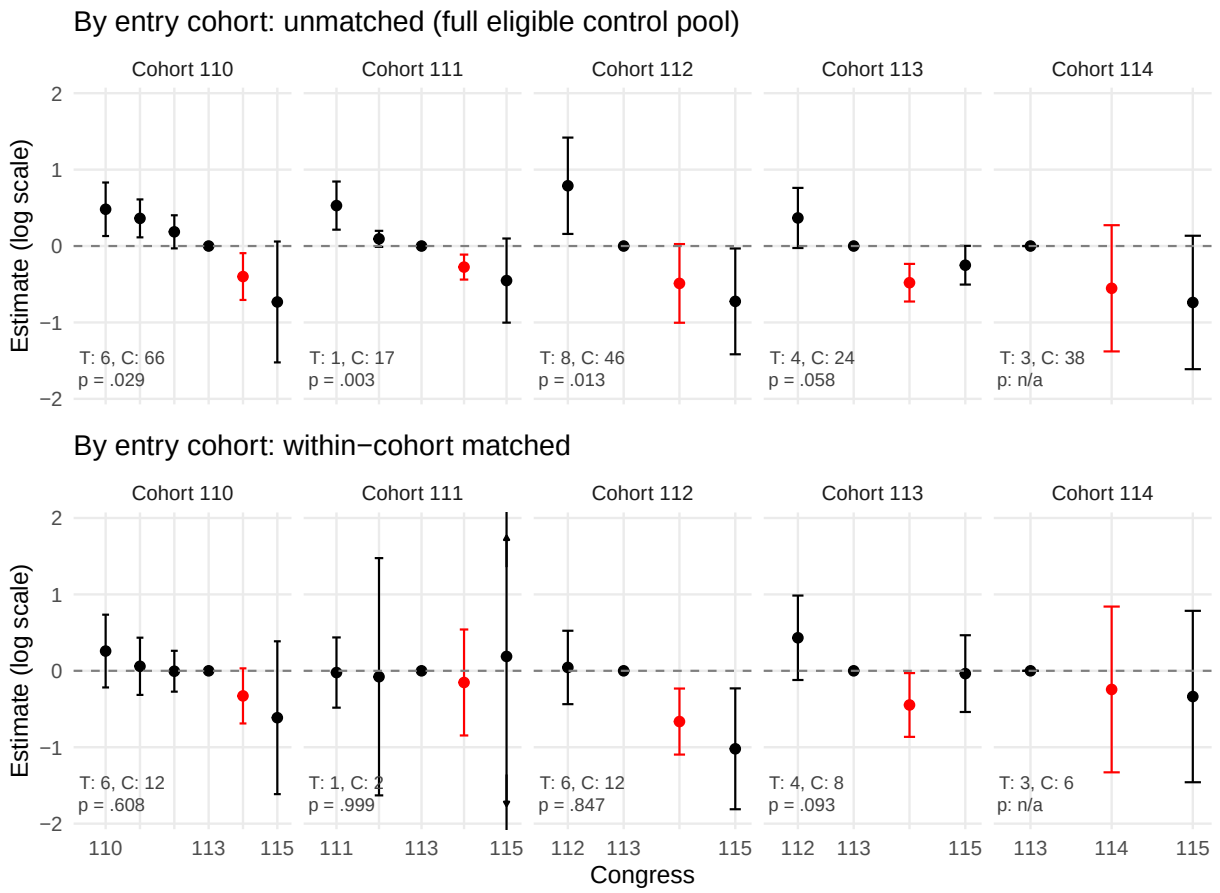

Figure 16: The cohort event studies carried one period past common exit.

Notes: Figure 11 extended by one post-treatment period, with specification, reference period and matching unchanged, so the pre-treatment coefficients are identical. The Congresses after the 114th are averaged into a single observation, so every legislator still serving contributes the same post-treatment length. That equalizes duration but not presence. Four defectors leave the House after the revolt and appear in no post-114 period, so the final point in each panel rests on fewer legislators than the rest and the printed counts describe every displayed period except the last.

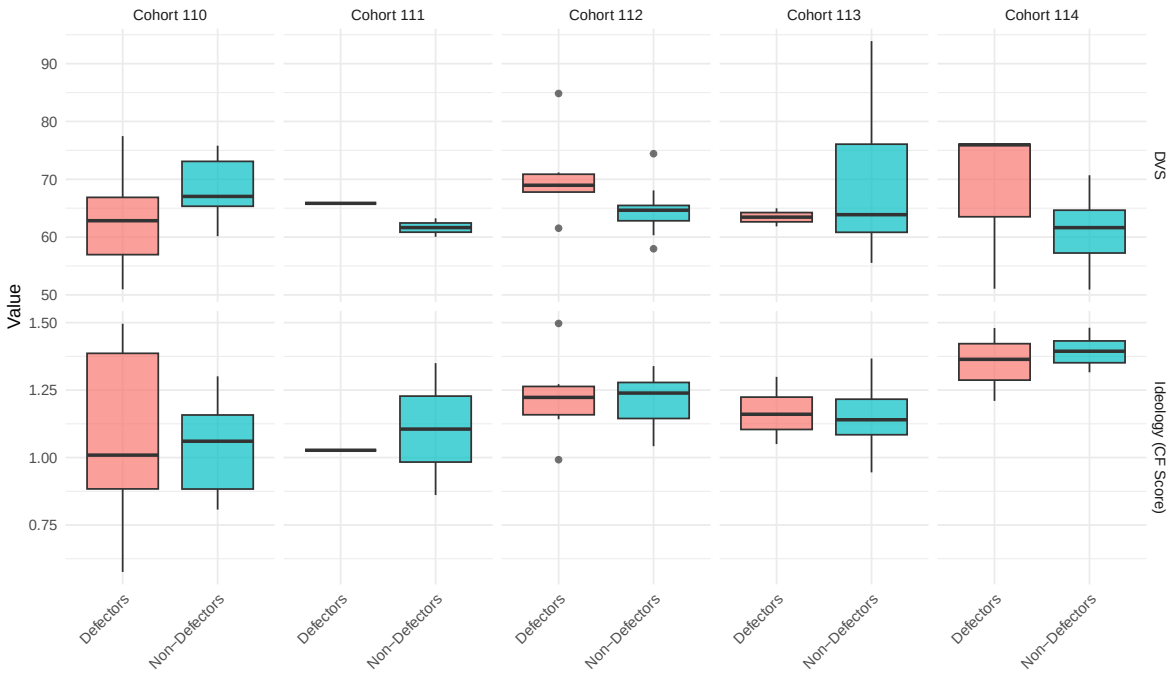

Figure 17: Within-cohort covariate balance after matching.

Notes: Distributions of district vote share (DVS) and ideology (dynamic CF score) for matched defectors and non-defectors, by entry cohort. Balance is close in most cohorts. The 114th cohort is the least well balanced, with three treated units and a single pre-period. Dropping it barely moves the aggregate estimate. Boxes span the interquartile range with the median line, and whiskers extend to 1.5× the interquartile range. No confidence intervals are shown.