Staggeringly Problematic: A Primer on Staggered DiD for Accounting Researchers*

John M. Barrios

Yale School of Management, Yale University, and NBER

john.m.barrios@yale.edu

This version: August 2026

ABSTRACT

I present the staggered difference-in-differences (DiD) method in accessible language to a broad accounting audience from an applied researcher's perspective. I begin by synthesizing recent advances in the econometrics of DiD designs in which multiple units receive treatment at different points in time. Using the Goodman-Bacon decomposition, I illustrate how heterogeneous treatment effects can bias the treatment effect estimate in a staggered DiD estimated with a two-way fixed effects regression. Using the staggered adoption of the 150-hour Rule as an example, I demonstrate several diagnostics and corrections that the econometrics literature has put forward. I close by reviewing what has changed in this literature since its first wave and translating the guidance into a step-by-step checklist that researchers, reviewers, and editors can use to evaluate staggered DiD designs.

Keywords: difference-in-differences; staggered treatments; treatment effect heterogeneity; research design.

JEL Classifications: C18; C21; C23; M41.

*I thank Thomas Bourveau, Matthias Breuer, Ed deHaan, Yael Hochberg, Christian Leuz, Miguel Minutti-Meza, and workshop participants at FARS and the *Journal of Financial Reporting* Research Methods Mini-Conference for helpful conversations, comments, and suggestions. Jin Deng provided excellent research assistance. An earlier version of this paper circulated as an SSRN working paper. All errors are my own.

Staggeringly Problematic: A Primer on Staggered DiD for Accounting Researchers

ABSTRACT: I present the staggered difference-in-differences (DiD) method in accessible language to a broad accounting audience from an applied researcher's perspective. I begin by synthesizing recent advances in the econometrics of DiD designs in which multiple units receive treatment at different points in time. Using the Goodman-Bacon decomposition, I illustrate how heterogeneous treatment effects can bias the treatment effect estimate in a staggered DiD estimated with a two-way fixed effects regression. Using the staggered adoption of the 150-hour Rule as an example, I demonstrate several diagnostics and corrections that the econometrics literature has put forward. I close by reviewing what has changed in this literature since its first wave and translating the guidance into a step-by-step checklist that researchers, reviewers, and editors can use to evaluate staggered DiD designs.

Keywords: difference-in-differences; staggered treatments; treatment effect heterogeneity; research design.

JEL Classifications: C18; C21; C23; M41.

I. INTRODUCTION

Difference-in-differences (DiD) estimation has become the workhorse quasi-experimental tool for measuring the effects of policies and treatments in applied empirical research. Accounting research is no exception (Armstrong et al. 2022). Between 2011 and 2025, more than 1,100 articles published in five leading accounting journals mention a DiD design (Figure 1). However, a recent wave of econometrics research shows that the most common way researchers estimate these designs, a two-way fixed effects (TWFE) regression with a staggered treatment, can produce estimates that are badly biased, even wrong-signed. In this paper, I explain why in plain language, demonstrating both the problem and its fixes, and reduce the guidance to a checklist for authors, reviewers, and editors.

Throughout the paper, I build the discussion around a single running example from the accounting profession: the staggered, state-by-state adoption of the 150-hour Rule. In 1988, 84% of the AICPA's voting members backed a proposal requiring 150 semester hours of college education, roughly five years of study, for new CPA licensure, effective in the year 2000 (Barrios 2022). States, however, adopted the requirement on their own schedules: one state moved in the 1980s, a trickle of states followed through the 1990s, 16 jurisdictions adopted in 2000 alone, and ten states had still not adopted by the end of the sample period I study (see Table 1). Barrios (2022) exploits this staggered rollout to estimate the Rule's effect on the supply and quality of accountants. The setting has all the features that make staggered DiD designs appealing to accounting scholars, a discrete regulatory shock, heterogeneous adoption dates across states, and a set of never-adopting states that can serve as controls. However, as I demonstrate below, it also contains all the elements that render the conventional TWFE estimator problematic.

To see where the trouble starts, begin with the design the staggered version generalizes: the two-group/two-period (2×2) DiD, in which a treated group's before-after change is compared with the change over the same period in a group that never receives treatment. Section II sets out the mechanics. What matters here is the property the 2×2 has and the staggered design gives up: the

researcher can say exactly which comparisons generate the estimate.[1]

However, many accounting research settings depart from the simple 2×2 design because the treatment arrives at different times for different groups. Standard setters phase in accounting standards, courts hand down rulings in different periods, states adopt licensing rules year by year, and firms receive going-concern opinions, switch to Big N auditors, or go public on their own schedules (the first two switch on and off; Section VI returns to that complication). Approximately 30% of DiD studies in the earlier 2011–2020 period used staggered treatment timing (see the note to Figure 1). This type of variation makes these studies susceptible to all of the issues discussed below. In such settings, researchers typically estimate a TWFE regression that includes fixed effects for units (e.g., firms, states, or countries) and for time periods, and then interpret the coefficient on a single binary treatment indicator as the treatment effect.[2] This regression is usually justified by appealing to the canonical 2×2 DiD setup. However, until recently, we had limited understanding of what the TWFE estimator is actually identifying when treatment starts at different times across groups (Goodman-Bacon 2021). Consider the 150-hour setting as an illustration. When a particular state adopts the Rule in 2000, which other states is OLS using as the comparison group, and are some of those states already treated at that point?

The heavy reliance of researchers on the TWFE estimator in staggered settings and the uncertainty about its interpretation have led econometricians to examine the causal content of the standard TWFE estimate (see, e.g., de Chaisemartin and D'Haultfœuille 2020; Goodman-Bacon 2021; Sun and Abraham 2021; Callaway and Sant'Anna 2021; Borusyak et al. 2024; Baker et al. 2022). I rely on the Goodman-Bacon (2021, hereafter GB) derivation to explain why.[3]

Using the GB derivation, I show that the TWFE estimate in a staggered setting is a weighted average of all possible 2×2 DiD estimates in the data, with weights that depend on group size and on the variance of treatment within each 2×2 pair. Some of those comparisons use units treated at

[1] This is the basic structure most econometrics textbooks and DiD survey articles in econometrics use (for example: Angrist and Krueger (1999), Angrist and Pischke (2009), Heckman et al. (1999), Meyer (1995), Cameron and Trivedi (2005), Wooldridge (2010)).

[2] For an overview of fixed effects models in accounting, see Breuer and deHaan (2024).

[3] Related results appear, with differing terminology, in de Chaisemartin and D'Haultfœuille (2020), Sun and Abraham (2021), Athey and Imbens (2022), and Borusyak et al. (2024).

two different points in time: later-treated groups serve as controls before their treatment begins, and earlier-treated groups serve as controls after theirs begins. When the 2000 cohort adopts the Rule, OLS compares it against already-treated mid-1990s adopters. When treatment effects are constant over time, TWFE recovers a variance-weighted ATT. When effects vary within treated cohorts, the already-treated comparisons contaminate the average and, as the simulations below reinforce, the estimate can turn negative even when every underlying effect is positive by construction.

Given these issues, econometricians have proposed several approaches that recover the ATT in staggered settings. I treat the GB decomposition as a diagnostic and review the three first-wave estimators that rebuild the ATT from clean comparisons: the group-time estimator of Callaway and Sant'Anna (2021), the interaction-weighted event-study estimator of Sun and Abraham (2021), and the stacked regression design in the spirit of Cengiz et al. (2019), implementing them all in the 150-hour setting.

Because the econometrics literature has continued to move quickly, I also take stock of what has changed since that first wave of estimators. Imputation-style estimators (Borusyak et al. 2024; Gardner 2022; Wooldridge 2025) now offer efficient alternatives within the familiar regression framework; doubly robust methods discipline how covariates enter (Sant'Anna and Zhao 2020; Caetano and Callaway 2024); pre-trends testing has been rethought (Roth 2022; Rambachan and Roth 2023); and inference with few treated clusters has its own set of fixes (Cameron et al. 2008; MacKinnon et al. 2023). Section VI reviews these developments through the 150-hour lens.

Baker et al. (2022) cover related territory in the *Journal of Financial Economics*. Their paper and mine are complementary: both explain the Goodman-Bacon decomposition and demonstrate TWFE bias through simulations, and they show that heterogeneity-robust estimators can overturn published finance results. I add three things. First, a Roy / reverse-Roy framing of accounting-specific bias sources that links the problem to established selection theory rather than treating it as a purely statistical artifact. Second, a single accounting setting carried through the whole paper: the staggered adoption of the 150-hour CPA licensing rule. The setting is not new; it re-analyzes Barrios (2022). But following one application from diagnosis through each correction, with code,

shows what every step does to the same estimate. Third, a step-by-step checklist in Appendix A covering design, estimation, inference, robustness, and reporting, for authors before submission and for reviewers and editors evaluating staggered DiD papers. Two econometrics surveys cover the same terrain from the other side: Roth et al. (2023) and the practitioner's guide by Baker et al. (2026), written by the architects of the estimators reviewed here, are the field's reference statements of the methods. Within accounting, Gow et al. (2016) pressed for explicit identification arguments and Armstrong et al. (2022) document the turn toward quasi-experiments. This paper is the accounting-side complement, not a substitute: one setting carried from diagnosis through every correction, framed by the selection stories accounting institutions generate, and reduced to a checklist a reviewer can apply.

The paper proceeds as follows. Sections II–III set up the 2×2 design and the TWFE / Goodman-Bacon problem, including simulations. Sections IV–V review the first-wave corrections and implement them in the 150-hour setting. Section VI surveys what has changed since; Section VII concludes.[4]

II. THE TRADITIONAL 2×2 DID DESIGN

The traditional 2×2 DiD design compares the before-after change in a (non-random) treatment group to the change over the same period in a comparison group that never receives treatment. In the running example, this is what a researcher would use if only one state had adopted the 150-hour Rule: the change in that state's CPA exam candidates against the change in never-adopting states over the same years.

Figure 2 walks through the intuition. A naïve post-period cross-section between treated and control firms confounds the treatment effect A with selection bias ($TF - CF$, the fixed treated-control baseline gap). A within-firm time-series difference eliminates the fixed selection term but leaves the common time trend T. Differencing those two differences removes both confounds and isolates A, provided that the assumption of parallel trends holds. Put differently, DiD attributes

[4]Online Appendix OA collects formal estimator statements, the simulation design, and supplemental figures.

to treatment whatever movement in the treated group's outcome the control group's trend cannot account for. (Online Appendix Figure OA1 spells out the same logic in tabular form.)

The running example makes the assumption concrete. When Barrios (2022) compares CPA exam candidates in adopting and non-adopting states, parallel trends requires that, had the adopting states not passed the 150-hour Rule, their candidate counts would have moved in step with those of the non-adopting states. Nothing guarantees this. States chose when to adopt, and anything state-specific and time-varying, such as a regional recession that depresses accounting enrollments, would violate the assumption. The setting also illustrates a subtler threat, anticipation. Prospective CPAs rushed to sit for the exam before the Rule took effect, inflating candidate counts in adopting states in the year just before adoption. A "pre-period" contaminated by anticipation is no longer a clean baseline, which is why the empirical work in Section V pays close attention to the year before each state's adoption.

Define a treated group k and an untreated group U, with pre- and post-periods indexed to k's treatment timing. The 2×2 DiD estimate is:

$$\hat{\delta}^{2\times2}_{k,U} = \left(\overline{y}^{POST(k)}_{k} - \overline{y}^{PRE(k)}_{k}\right) - \left(\overline{y}^{POST(k)}_{U} - \overline{y}^{PRE(k)}_{U}\right) \tag{1}$$

where $\overline{y}$ is the sample mean of the reference group in the reference period. Figure 2 shows this estimator: the left side arranges the four sample means of equation (1) in the familiar 2×2 table; the right side traces the outcome paths on which identification rests.

The mapping of equation (1) into potential outcomes makes the identifying assumption explicit:[5]

$$\begin{aligned} \operatorname{plim} \hat{\delta}^{2\times2}_{k,U} = {} & E\left[Y^1_k \mid Post\right] - E\left[Y^0_k \mid Post\right] \\ & + \left[E\left[Y^0_k \mid Post\right] - E\left[Y^0_k \mid Pre\right]\right] - \left[E\left[Y^0_U \mid Post\right] - E\left[Y^0_U \mid Pre\right]\right] \end{aligned} \tag{2}$$

Equation (2) isolates the ATT in the first line, but only if the second line zeros out, that is, only

[5]For the full mapping of the 2×2 DiD into the ATT, see Angrist and Pischke (2009), Cameron and Trivedi (2005), or Wooldridge (2010).

if the counterfactual post-minus-pre change for the treated group equals the observed change for the untreated group. Because Y_k^0 in the post-period is unobserved, parallel trends is, by definition, untestable.

III. THE TWO-WAY FIXED EFFECTS ESTIMATOR AND THE STAGGERED DID DESIGN

The TWFE Estimator

Because the DiD estimate is built from means, it translates directly into an OLS regression (Angrist and Pischke 2009). Most researchers therefore estimate the effect of a binary treatment in a DiD setting with a TWFE regression, which takes the form:

$$y_{it} = \alpha_i + \alpha_t + \delta^{DD} D_{it} + u_{it} \tag{3}$$

where α_i is a unit fixed effect, α_t is a time period fixed effect, and D_{it} is an indicator variable for a treated unit in post-treated periods. The unit and time period fixed effects subsume the main effect of being a treated unit and the post-period indicators. This regression gives researchers an estimate of their parameter of interest $\hat{\delta}^{DD}$ and standard errors for the estimate. The regression also allows researchers to add covariates and time trends, making the TWFE regression one of the most popular approaches to estimate DiD in the empirical accounting literature (Breuer and deHaan 2024). In the 150-hour application, equation (3) is a regression of log CPA exam candidates on jurisdiction fixed effects, period fixed effects, and an indicator equal to one once jurisdiction i has adopted the Rule, precisely the specification most published staggered DiD papers in accounting use.

Sources of Bias in Staggered DiD Settings

The TWFE estimator breaks down in two distinct ways, worth separating before the math. First, treatment effects vary across units: some firms or states benefit more from a regulation than others. Second, they change over time within a unit, so the effect in year one may differ from the effect

in year five. Cross-unit heterogeneity alone, with time-constant effects, still yields a positively weighted average of true ATTs—the weights are distorted, but no forbidden comparison poisons the average, because a constant effect differences out of the already-treated control. Time-varying effects are the sharper threat, because they contaminate the comparisons that use already-treated units as controls, and together the two can flip the sign of the estimate.

Two selection stories map onto the 150-hour setting. If states where rents from restricting entry were largest lobbied hardest and adopted first, early adopters are high-gain units—the Roy (1951) case—and the variance weights, which penalize cohorts treated for most of the panel, underweight exactly those high-gain units and pull the estimate below the average effect. If instead the states where the Rule was expected to bind least faced the least opposition and moved earliest, early adopters are low-gain units—the reverse-Roy—and the same weighting pushes the estimate up. Which pattern shows up in the data is an empirical question; Section V reads Table 1's cohort-level estimates through this lens.

The second breakdown operates along the time dimension. When treatment effects accumulate, already-treated units show progressively different outcomes in each successive period. The 150-hour Rule is exactly this kind of treatment: a licensing requirement reshapes the pipeline of accounting students over several years, so the effect deepens with time since adoption. TWFE then uses earlier-adopting states as controls for newly adopting states at a moment when the early adopters' candidate counts are still falling from their own treatment. The resulting bias can be severe: a uniformly positive true effect can generate a negative TWFE coefficient. The simulations later in this section demonstrate this with data-generating processes calibrated to realistic accounting panel settings. Goodman-Bacon (2021) formalizes these intuitions. I turn to his decomposition next.

The Goodman-Bacon Decomposition

Staggered treatment timing effectively splits DiD designs into two categories. The first is the 2×2 setup from Section II, where a single firm or a set of firms is treated at the same time—for instance,

a country adopting IFRS in a particular year. The second is a DiD with staggered adoption, where different groups are treated in different periods—such as the phased implementation of the 150-hour Rule across states, with adoption years spanning from the 1980s to 2003 (Table 1). This latter case can be viewed as generating a collection of separate 2×2 DiD comparisons. To clarify how such timing affects the TWFE estimator, I concentrate on GB's decomposition framework.[6]

The GB decomposition theorem states that the TWFE estimator is a weighted average of all potential 2×2 DiD estimates where the weights are based on both the group size and the variance in treatment. Assuming variance-weighted common trends (VWCT) and time-invariant treatment effects, the decomposition shows that the variance-weighted ATT is a weighted average of all possible ATTs. Under more restrictive assumptions, the estimate matches the ATT exactly. However, this is not the case when there are time-varying treatment effects. In settings with time-varying treatment effects and differential timing in treatments, the obtained TWFE estimate is generally biased. This possibility of time-varying treatment effects should be of first-order concern for accounting researchers, as most treatments studied in accounting take effect gradually rather than all at once.

A simple example shows why. Consider three groups: an early treatment group (k), a later treated group (l), and a never treated group (U). In the case of the 150-hour rollout, think of k as the states that adopted the Rule in the mid-1990s, l as the large cohort that adopted in 2000, and U as the ten states that never adopted during the sample period. Groups k and l are both treated, but they differ in timing: k is treated earlier than l (Online Appendix Figure OA2 plots the three outcome paths). They are treated at time periods t_k and t_l, respectively.[7] The treatment exposure matters as the length of time a group spends in treatment determines its treatment variance, which in turn affects the weight that the specific 2×2 comparison receives in the TWFE estimate.

Suppose that these groups are observed for five periods. Call the earliest period, before any group is treated, the "pre" period; the period between the treatment of groups k and l the "mid"

[6]de Chaisemartin and D'Haultfœuille (2020) also derive a similar result.

[7]For example, if k is treated in period two while group l is treated in period four, then group k spends 80% of its time under treatment, or 0.8, while group l spends 40% of its time treated, or 0.4. Define the time spent in treatment for a group as $\overline{D_k} = 0.8$ and $\overline{D_l} = 0.4$.

period; and the period after group l is treated the "post" period. Thus there are four 2×2 DiDs, which is the number of single 2×2 comparisons that make up the TWFE estimator in the current case of two timing groups and a never-treated group.[8]

For notational simplicity, denote each 2×2 DiD as $\hat{\delta}^{2\times2,j}_{a,b}$, where a and b are the comparison groups and j indexes the treatment group. The 2×2 for the case of k and U is then $\hat{\delta}^{2\times2,k}_{k,U}$, equation (1) with the groups relabeled. To see what these comparisons look like, Figure 3 displays the four 2×2 comparisons. In Panels A and B, with only one treatment group and the untreated group, each component is the standard 2×2 DiD:

$$\hat{\delta}^{2\times2,j}_{j,U} = \left(\overline{y}^{POST(j)}_{j} - \overline{y}^{PRE(j)}_{j}\right) - \left(\overline{y}^{POST(j)}_{U} - \overline{y}^{PRE(j)}_{U}\right) \tag{4}$$

where $j = k,l$. But the TWFE estimator also makes 2×2 comparisons in which no untreated units are used at all. In these cases, the estimate relies only on the difference in the timing of the treatment groups. Panel C shows this: before t_l, the early group k acts as the treatment group, and the later treated group l acts as the control. The comparison uses outcomes over the window when treatment status varies across the groups, $MID(k,l)$ against the early group's pre-period $PRE(k)$:

$$\hat{\delta}^{2\times2,k}_{k,l} = \left(\overline{y}^{MID(k,l)}_{k} - \overline{y}^{PRE(k)}_{k}\right) - \left(\overline{y}^{MID(k,l)}_{l} - \overline{y}^{PRE(k)}_{l}\right) \tag{5}$$

Panel D shows the opposite, where the latter group changes treatment after t_l and the earlier treated group k acts as the control. This 2×2 ($\hat{\delta}^{2\times2,l}_{l,k}$) is comparing average outcomes between two periods $POST(l)$ and $MID(k,l)$:

$$\hat{\delta}^{2\times2,l}_{l,k} = \left(\overline{y}^{POST(l)}_{l} - \overline{y}^{MID(k,l)}_{l}\right) - \left(\overline{y}^{POST(l)}_{k} - \overline{y}^{MID(k,l)}_{k}\right) \tag{6}$$

It is this comparison that many applied researchers do not realize is being made when they estimate a TWFE model in a setting with staggered treatments. The decomposition makes it clear that the already-treated group k acts as a control even though it is already treated. This is a result

[8] If there were three timing groups and one untreated control group, then there would be 9 2×2 DiDs.

of group k's treatment assignment indicator not changing over the relevant comparison period ($MID(k,l)$ to $POST(l)$). In the running example, when the 2000 cohort adopts the 150-hour Rule, the regression treats the mid-1990s adopters as "controls" despite these states being five years into their own treatment. No researcher would deliberately choose that comparison, yet OLS makes it automatically.

Each of the four 2×2 DiD estimates in Figure 3 uses only part of the panel's available data. The DiDs that compare treated and untreated units use the entire time period, but only for the two groups involved, so their sample shares are $(n_k + n_U)$ and $(n_l + n_U)$. The 2×2 DiDs that exploit variation between the timing groups (Panels C and D) use only the two groups' observations, and only over the portion of the available time periods that overlaps for both.[9] These comparisons define the generalized TWFE DiD parameter as[10]:

$$\hat{\delta}^{DD} = s_{k,U}\hat{\delta}^{2\times2,k}_{k,U} + s_{l,U}\hat{\delta}^{2\times2,l}_{l,U} + s^{k}_{k,l}\hat{\delta}^{2\times2,k}_{k,l} + s^{l}_{l,k}\hat{\delta}^{2\times2,l}_{l,k} \tag{7}$$

Equation (7) shows that the TWFE estimator is a weighted average of all potential 2×2 DiD estimates. GB also shows that the weights (i.e. $s_{k,U}$) on each of these 2×2 DiD estimates are a function of the treatment cohort's total size, the relative size of the treatment and control groups in the specific DiD subsample, and the timing of the treatment in the specific comparison subsample. For $\hat{\delta}^{2\times2,k}_{k,U}$, the weight is defined as:

$$s_{k,U} = \frac{n_k n_U \overline{D}_k(1-\overline{D}_k)}{\overline{\text{Var}}(\widetilde{D}_{it})} \tag{8}$$

where n_k and n_U are the shares of the sample that fall into groups k and U, and $\overline{D}_k(1-\overline{D}_k)$ is the variance over time of group k's treatment indicator, with $\overline{D}_k$ being the share of periods group k spends treated. In the denominator, $\widetilde{D}_{it}$ is the treatment indicator after the unit and time fixed effects have been partialled out (i.e., the variation the TWFE regression actually uses) and $\overline{\text{Var}}(\widetilde{D}_{it})$

[9] For example, $\hat{\delta}^{2\times2,k}_{k,l}$ uses only group l's pre-period, so its sample share is $(n_k + n_l)(1-\overline{D}_l)$, while $\hat{\delta}^{2\times2,l}_{l,k}$ uses group k's post-period, with share $(n_k + n_l)\overline{D}_k$.

[10] Goodman-Bacon (2021) derives this decomposition by applying the Frisch-Waugh theorem to the TWFE estimator.

is its sample variance. Divided by that total, each comparison's variance contribution becomes a weight; the weights are nonnegative and sum to one. The formula shows that cohort-level variation, not the unit-level variation applied work usually focuses on, drives the estimate: the more states that adopt a law simultaneously, the greater their influence on the aggregate.

The weights also depend on within-group treatment variance, maximized at $\overline{D} = 0.5$, so being treated mid-panel raises a cohort's weight. Lengthening or shortening the sample therefore changes the aggregate estimate even when every underlying 2×2 DiD stays the same. In the 150-hour rollout, the 2000 cohort dominates the TWFE estimate through size alone: it is the largest (16 jurisdictions), while its treatment share of 0.21 sits well below the variance-maximizing 0.5, as the decomposition in Section V makes visible.

The GB decomposition and that of de Chaisemartin and D'Haultfœuille (2020) answer different questions, and the difference matters in practice. GB weights the 2×2 *comparisons*; they are nonnegative and sum to one, so TWFE is a convex combination of the 2×2 estimates. de Chaisemartin and D'Haultfœuille (2020) instead weight the underlying group-time treatment *effects*, and those weights can be negative. The two are complementary: GB identifies which timing comparisons drive the estimate, while de Chaisemartin and D'Haultfœuille (2020) diagnoses whether already-treated controls weight the causal effects perversely. Positive Goodman-Bacon weights are therefore not evidence that the de Chaisemartin and D'Haultfœuille (2020) concern is absent.

The Decomposition in a Potential Outcomes Framework

The Bacon decomposition expresses the TWFE coefficient in terms of sample means, which makes it easy to rewrite in potential outcomes.[11] Define the ATT for timing cohort k in year t as $\text{ATT}_k(t) = E[Y^1 - Y^0 \mid k, t]$, and the window ATT as the average of these over the post-treatment periods for cohort k. The Bacon decomposition writes the probability limit of the TWFE estimator as:

[11] See Callaway and Sant'Anna (2021) and Goodman-Bacon (2021) for the full cohort-ATT derivation.

$$\operatorname*{plim}_{n\to\infty} \hat{\delta}^{DD} = \text{VWATT} + \text{VWCT} - \Delta\text{ATT} \tag{9}$$

Equation (9) is the one-line statement of the problem. The variance-weighted ATT (VWATT) is a positively weighted average of the cohort ATTs, with weights from equation (7). VWCT extends parallel trends to the multi-cohort setting; nonzero VWCT means differential counterfactual trends bias the estimate. The ΔATT term captures within-cohort treatment-effect dynamics: even identical linear trend-breaks across units generate cross-group heterogeneity once earlier-treated groups serve as controls for later ones (Online Appendix Figure OA3 gives the stylized timing-only picture while Figure 4, Panel D, shows the same failure in Monte Carlo).

Two implications follow. First, claiming that $\hat{\delta}^{DD}$ is an ATT implicitly assumes VWCT = 0. Second, bias arises either because the 2×2 weights do not equal sample shares (even with time-constant effects) or, more dangerously, because time-varying effects make the early-as-control comparisons invalid. Figure 4, Panel D, is the empirical counterpart of this second case.

Simulations

To make these concerns concrete, I run simulations in the spirit of Baker (2019). The panels resemble the 150-hour rollout in Table 1: 50 states from 1980 to 2010, adoption cohorts spread across the panel, and ten never-adopting states available as controls in the first three designs. Each state's outcome—think of log CPA exam candidates, or any state-level accounting outcome—combines a time-invariant state component, a common time trend, and an idiosyncratic error, with a treatment effect added once the state's cohort year arrives. For each of four data-generating processes, corresponding to the four panels of Figure 4, I draw the data 1,000 times and recover $\hat{\delta}^{DD}$ from a TWFE regression with state and year fixed effects, clustered by state. Throughout, I benchmark the estimate against the average effect on treated observations, $E[Y^1 - Y^0 \mid D = 1]$, the conventional ATT.[12]

[12]The error standard deviation governs how wide each distribution in Figure 4 is, not where it is centered; Online Appendix OA.5 shows why, and reports the exact probability limits, which I compute analytically rather than by simulation.

The first is the unbiased benchmark: four equally sized cohorts (1983, 1990, 1997, 2002), each state receiving a constant effect drawn around a mean of 0.30, the same for every cohort and constant over time. The second and third keep that structure but let the average effect differ across cohorts, in the spirit of the Roy and reverse-Roy stories from Section III. Under Roy, early adopters gain the most (cohort means 1.20, 0.70, 0.30, 0.10); under reverse-Roy the ordering flips. The same four numbers deliver different true ATTs—0.742 and 0.412—because early cohorts spend more of the panel treated and so contribute more treated observations. The fourth makes the effect dynamic: five cohorts (1984 through 2000), no never-treated states, and an effect accumulating linearly with time since adoption, $\mu_i \times (\text{year} - t_c + 1)$, where t_c is the cohort's adoption year.[13] Every state's treatment effect is positive in every period by construction.

Figure 4 plots the distribution of the 1,000 TWFE estimates for each simulation. In Panel A, with staggered timing but a constant, homogeneous effect, the standard TWFE model recovers the true treatment effect: the mass of the distribution is centered on the true ATT of 0.30 (the red vertical line). Panels B and C show the effect of cross-cohort heterogeneity. The variance weighting pulls the mean TWFE estimate below the true ATT in the Roy case (0.458 against 0.742) and above it in the reverse-Roy case (0.659 against 0.412). The mechanism is visible in the weights themselves: the 1983 cohort supplies a quarter of the treated states but receives only 11% of the TWFE weight, because a cohort treated for almost the whole panel has little within-unit treatment variance. When that under-weighted cohort is the high-gain one, the estimate is dragged down; when it is the low-gain one, the estimate is pushed up. The bias is substantial in both directions, roughly two-fifths of the true effect in Panel B and three-fifths in Panel C, but the sign survives, because the weights distort while no forbidden comparison poisons the average.

Panel D shows what happens when effects grow over time: every one of the 1,000 estimates is negative even though every underlying treatment effect is positive by construction, with a mean of -0.236 against a true ATT of 0.570. The estimator is systematically wrong-signed because

[13] Slopes are drawn lognormally around declining cohort means, so earlier adopters accumulate larger cumulative effects and every state's slope is positive. Online Appendix OA.5 records the full simulation design and the seeded Monte Carlo results behind Figure 4. Stata code is available from the author's website.

already-treated cohorts, whose outcomes are still rising from their own treatment, serve as controls for later cohorts; a researcher would conclude a beneficial treatment is harmful, with tight standard errors. The reversal requires two design features: no never-treated states, so every comparison is a timing comparison, and a panel that runs a decade past the last adoption, so in more than a third of the sample years no untreated state exists at all. Replacing ten of the treated states with never-adopting ones, holding the panel at 50 states, moves the probability limit from −0.235 to +0.037: clean controls restore the sign, though the estimate then recovers only about 6.5 percent of the true effect (OA.5 reports the full sensitivity). That contrast is why the first question in Section IV is whether never-treated units exist.

IV. FIRST-WAVE ADVANCES IN STAGGERED DID IMPLEMENTATION

I now turn to the first-wave approaches that obtain interpretable ATTs in staggered treatment settings. The literature focuses on correcting the weighting issues, selecting an appropriate control in each 2×2 comparison, or both.[14]

As Section III showed, TWFE can weight the individual 2×2 DiDs in unintuitive ways—groups in the middle of the panel receive more weight than those at the end—so the aggregate estimate can change when years are added or dropped even if the underlying 2×2s do not. The Bacon decomposition is the natural first diagnostic: plot each 2×2 DiD against its weight, compute average estimates and total weights by comparison type (treated vs. untreated; late vs. early; early vs. late), and ask how much of the TWFE estimate comes from timing variation.[15]

Once the Bacon plot has shown where the TWFE weight sits, three first-wave estimators rebuild the ATT from clean comparisons. Callaway and Sant'Anna (2021) estimate a group-time $ATT(g,t)$ for each adoption cohort and calendar year, then let the researcher aggregate, into an event-study profile or a single overall ATT, with weights that are chosen and reported rather than imposed by

[14] I focus on approaches of immediate use to accounting researchers. For related work, see Arkhangelsky and Imbens (2021), Ben-Michael et al. (2022), and Egami and Yamauchi (2023). Formal statements of the estimators, including the Callaway–Sant'Anna group-time ATT formula, appear in Online Appendix OA.

[15] The decomposition in Goodman-Bacon (2021) requires a balanced panel and does not incorporate covariates, features that are atypical of many accounting applications. The 150-hour panel is nearly balanced (1,901 of 1,908 cells have outcomes), and the decomposition's implied TWFE matches the pooled estimate to the third decimal.

OLS. Controls are never previously treated: either never-treated units or, when those are scarce, not-yet-treated units. Covariates enter through propensity-score reweighting (or the doubly robust extension). In the implementation reported in Section V, an $ATT(g,t)$ for the 2000 adopters two years after adoption is benchmarked against the jurisdictions not yet treated in that year, never against an already-treated mid-1990s adopter.[16]

Sun and Abraham (2021) stay inside the event-study regression but saturate it with cohort × relative-time interactions—their interaction-weighted (IW) estimator—so each cohort gets its own dynamic profile before averaging by sample share. Already-treated units never enter as controls while never-treated or last-treated units do. Applied to the 150-hour Rule example, the large 2000 cohort (16 jurisdictions) contributes in proportion to its size, not its treatment variance.

The stacked approach of Cengiz et al. (2019) builds one clean event dataset per adoption cohort, treated states plus units not treated inside a fixed window, stacks them, and runs a single event-study regression saturated with stack indicators. Previously treated states never appear on the control side. The 150-hour version is one stack for the 1997 adopters plus never-adopters, one for 1998, and so on. The cost is OLS variance-weighting across stacks; Section VI discusses the sample-weighted fix in Wing et al. (2024). Table 3 summarizes these estimators, including the imputation / extended-TWFE estimators taken up in Section VI.

Implementation Guidance: Choosing Among Correction Strategies

Which estimator should a researcher use? It depends on the data, and researchers should usually report more than one. Table 3 and Appendix A Panel B organize the choice around four questions. The first is whether never-treated units exist. If they do—as in the 150-hour setting—all of the estimators in this paper are available, and never-treated controls are the natural default. If they do not, Callaway–Sant'Anna and the imputation estimators remain available with not-yet-treated controls; Sun–Abraham and stacked designs can likewise use the last-treated cohort as a comparison group. The second is whether covariates are needed to make parallel trends plausible: Callaway–

[16]The accompanying R package `did` implements both `control_group = "nevertreated"` and `control_group = "notyettreated"`.

Sant'Anna adjusts via propensity scores, the others through the regression. The third is whether the target is an event-study profile or a single aggregate ATT. The fourth is whether the panel is balanced; if so, begin with the Goodman-Bacon diagnostic, and if not, go directly to a correction method. A large timing weight is a reason to look harder rather than evidence of bias: what matters is whether the timing-only average lies apart from the treated/never-treated average. Two further families sit just beyond these three corrections, the imputation estimators and the saturated "extended" TWFE, which Section VI takes up along with the sensitivity analysis that should follow whichever estimator a researcher settles on.

V. BEST PRACTICES IN IMPLEMENTING A STAGGERED DID DESIGN: THE 150-HOUR RULE

Having used the 150-hour Rule to motivate the problem, I now use it to demonstrate the solutions. Barrios (2022) exploits the staggered timing of state adoption (see Table 1) to evaluate the Rule's effect on the supply and quality of accountants; in the supply tests, which end in 2003, he uses the ten non-implementing jurisdictions as controls.[17] The panel covers the 53 jurisdictions that license CPAs—the 50 states plus the District of Columbia, Guam, and Puerto Rico—and is semi-annual. The Uniform CPA Examination was given twice a year, in May and November. The sample runs from 1985 to 2003 with 1989 unavailable, giving 36 sittings and $53 \times 36 = 1{,}908$ jurisdiction-sittings; seven have a missing outcome, so the estimation sample is 1,901. The setting is well suited to illustrating best practice: it has adoption cohorts of widely different sizes, a long panel, one 1980s adopter, and a pool of jurisdictions untreated throughout the sample (the earliest adopts in 2004), all features described in Sections III and IV.

I begin by replicating the event-study result on state-level data with a log specification; despite the aggregation, it matches Barrios (2022). Figure 5 plots the coefficients. Test-takers show no clear pre-trend except in the year before adoption, when they rise sharply relative to controls, and

[17] I focus on state-level supply data rather than the micro-level university enrollment data used in Barrios (2022), because the robustness tests demonstrated here operate on the aggregate panel.

they fall steadily thereafter.[18] The TWFE DiD coefficient is −1.08, a proportional reduction of roughly two-thirds ($\exp(-1.08) - 1 = -0.66$), as shown in Table 2, Column 1. The plot is where the anticipation in this setting first becomes visible.

Next, I use the Bacon decomposition to examine where that estimate comes from. Figure 6 plots each 2×2 DiD against its weight, with the average effect and total weight for the three comparison types: treated/untreated, early/late, late/early.[19] The TWFE estimate, −1.08, is an average of the y-axis values weighted by their x-axis values. The timing terms ($\hat{\delta}^{2\times2,k}_{k,l}$ and $\hat{\delta}^{2\times2,l}_{l,k}$) carry 53% of the weight, the never-adopting jurisdictions 43%, and the pre-1990 adopter only 4%. Figure 6 also identifies the influential comparisons: the 1997, 1998, 2000, and 2001 cohorts against the never-adopters account for 33% of the weight on their own, and the ten highest-weighted 2×2s for 46%.

Does the 53% timing weight imply that heterogeneity bias is a first-order concern here? The decomposition itself says no, or at least, not obviously. The average of the post-treatment event-study estimates in Figure 5 is −1.02, and the single-coefficient version of that same specification, −1.07, is only 0.05 log points larger. The timing-only average (−1.04) is close to the treated/never-treated average (−1.15), while the comparison against the pre-1990 adopter is somewhat smaller in magnitude (−0.87). When the contaminated comparisons deliver roughly the same answer as the clean ones, high timing weight raises the stakes without changing the conclusion.

Table 1's cohort DiD column speaks to the selection question Section III posed. The earliest estimable cohorts show the smallest declines (−0.55 in 1993, −0.52 in 1994) and the middle cohorts the largest (−1.45 in 1998, −1.61 in 1999), a reverse-Roy pattern: the states that moved first look least affected. By Section III's logic that pattern predicts a TWFE estimate pulled toward zero, and the heterogeneity-robust estimates in Figure 7 do sit at or below −1.08. The reading is suggestive rather than sharp: the earliest cohorts are single jurisdictions whose wild-bootstrap p-values are uninformative, so the pattern rests on point estimates.

[18] This anticipation is documented in Barrios (2022): individuals sat the exam before the new requirement became binding. Replication code is available from the author's website.

[19] The Stata package bacondecomp can generate the individual 2×2 DiD estimates and their weights as well as their graphical representation. Code implementing this for the 150-hour Rule is available from the author's website.

Table 2 estimates several alternative specifications. Every column absorbs jurisdiction and period fixed effects, where a period is one of the 36 semi-annual sittings. Column 2 adds a control for the sitting month, which is collinear with the period fixed effects. In column 3, I drop the never-adopting jurisdictions, while in column 4, I drop the never-adopting jurisdictions and the pre-1990 adopter from the estimation. In column 5, I add jurisdiction-specific linear trends in the sitting index, while in column 6, I control for cohort-specific pre-trends using a two-stage procedure.[20] Consistent with Figure 6, the coefficient remains stable across specifications, ranging from -1.03 to -1.11; column 3's stability without the never-adopters signals limited dependence on clean controls here, not generic robustness (Figure 4, Panel D).[21]

A brief note on inference. The standard errors in Tables 1 and 2 are clustered at the jurisdiction level, where treatment is assigned (Bertrand et al. 2004). With 53 clusters, Table 2's cluster-robust errors are defensible; Table 1's cohort regressions have as few as 11 jurisdictions, most cohorts have three or fewer treated ones, and there the wild-cluster bootstrap p-values reported in its note are the more reliable guide—they weaken several cohorts, exactly the pattern Section VI's few-cluster discussion predicts. Many accounting applications sit in that second case: a handful of adopting states, or a single regulatory event.

I end the replication by implementing the three correction strategies from Section IV. Each answers one question, whether the TWFE weighting distorts the estimate; none answers the anticipation question raised above. Figure 7 graphs the results, and the three agree less closely than the decomposition suggested they would. Callaway–Sant'Anna (Panel A) gives an average post-treatment effect of -1.40 (SE 0.24), Sun–Abraham (Panel B) gives -1.12 (SE 0.22), and the stacked regression (Panel C) gives -1.63 (SE 0.21), against the TWFE estimate of -1.08 (SE 0.15). The estimates span half a log point, though with these standard errors no two are statistically distinguishable.

Much of that spread is not disagreement about weights; it comes from the reference period each estimator uses. The stacked regression sets $t = -1$ to zero; Callaway–Sant'Anna measures

[20] Online Appendix OA.6 sets out GB's two-step pre-trend adjustment as implemented here.

[21] Online Appendix OA.6 describes GB's Oaxaca-Blinder-Kitagawa comparison of weights across specifications.

every post-treatment effect from the year before adoption, though its varying base period reports the pre-treatment placebos as period-to-period changes rather than a normalized zero, so $t = -1$ appears as an estimate (+0.204); and the Sun–Abraham specification here is anchored at $t = -2$, as Figure 5 is, with the endpoints binned at ±5 and the pre-1990 adopter excluded, since a cohort with no in-sample pre-period has no identified effect (Sun and Abraham 2021). Because $t = -1$ is the anticipation year, a specification that measures the decline from it starts from an elevated baseline, and the averages line up accordingly: the two estimators anchored at $t = -1$ give the largest declines (−1.63 and −1.40), while Sun–Abraham, anchored at $t = -2$ and reporting the anticipation directly (+0.166, in line with Figure 5's +0.188), gives −1.12, close to TWFE. The ordering is suggestive rather than a decomposition, because the estimators also differ in control group, averaging window, and time aggregation. The point stands that a reader comparing the three must know which pre-period each is measured from, and that is a question about which baseline is credible rather than about the 2×2 weights.

Panel C is also worth reading for its pre-period. Its pre-treatment coefficients climb steadily toward the omitted year—−0.66, −0.51, −0.49, and −0.39 at $t = -5$ through $t = -2$—which the full-panel event study in Figure 5 does not show. The stacked design differs from the pooled one in more than one way, trimming to ±5-year windows (truncated where the panel ends, so the later event times average over fewer, earlier-adopting cohorts) and restricting controls to never-adopters, so I do not attribute the difference to any single choice. The pattern is still a parallel-trends warning that the pooled specification does not raise, and it argues for a sensitivity analysis of the kind discussed in Section VI.

The replication illustrates how an event study (Figure 5), a Bacon decomposition (Figure 6), and heterogeneity-robust estimators (Figure 7) fit together. It also illustrates their limits, which is the more useful lesson. Every correction implemented above targets the same object—the weight OLS places on each underlying comparison—and all three return a large, persistent decline, as TWFE does. None of them touches on the anticipation visible in the year before adoption, when candidates sat the exam early to beat the deadline. That is a mistimed treatment date, not a weighting problem,

and no reweighting can repair it.

That leaves the question of what the anticipation actually is. The year-before coefficient in Figure 5 is +0.188, or roughly a 21% increase in candidates relative to trend. Candidates close to eligible under the old requirement had an obvious incentive to sit before the deadline, and the spike is the size of that response. That makes it a finding in its own right, and it bears on how the main estimate should be read. Pull-forward on its own cannot produce the post-period decline: borrowing candidates from later years should generate a dip that fades once the borrowed cohort is exhausted, and Figure 5 shows the opposite, -0.632 at $t = 0$, -1.119 at $t = 1$, and no recovery through $t = 5$. A decline that deepens and stays points to a change in the flow of new candidates, not a shift in the timing of a fixed stock. The spike plausibly changes the composition of the post-Rule pool as well, since those who rushed were the candidates closest to the old requirement, which is one place where the supply estimates here may connect to the quality results in Barrios (2022).

So, the honest reading of this section is narrower than the number of estimators in it suggests. A flat pre-period and a sharp post-period break are useful diagnostics, but they do not certify the design; Roth (2022) shows that conventional pre-trends tests are often underpowered. A researcher who thought anticipation was the binding threat would move the event date or drop the anticipation window; a researcher who thought differential trends were binding would report a sensitivity analysis. Here, though, the anticipation is part of the economic story and not just a nuisance to be differenced away. The rush to sit the exam before the Rule bound shows how candidates responded to it. A paper that takes either remedy still owes the reader an account of that response.

VI. RECENT DEVELOPMENTS: WHAT HAS CHANGED SINCE THE FIRST WAVE

The corrections implemented in Section V are what I call the first wave: they diagnosed the problem and provided workable fixes. The literature has since consolidated. Roth et al. (2023) and de Chaisemartin and D'Haultfœuille (2023) synthesize the econometrics, and Baker et al. (2026) distill it into a practitioner's guide. The framework has also outgrown the binary, absorbing treatment: de Chaisemartin and D'Haultfœuille (2026) develop estimators for treatments that switch

on and off or vary over time, and Callaway et al. (2023) extend the design to continuous treatment intensity. The frontier has moved from "is my TWFE estimate biased?" to which estimand I want, how covariates enter, what the pre-trends actually tell me, and whether the inference is honest; I take these up in turn, staying with the 150-hour Rule as the running example.

Imputation Estimators and the Extended TWFE

A second family, developed largely after the first wave, approaches the problem through imputation. Borusyak et al. (2024) estimate the unit and time fixed effects from untreated observations only—here, the never-adopting states plus the pre-adoption years of the adopters—and impute each treated observation's counterfactual from them. The effect for each treated state-year is the actual outcome minus that counterfactual, and these can be averaged however the researcher wants. Gardner (2022) reaches the same place through a two-stage regression, and Wooldridge (2025) shows that a TWFE regression saturated with cohort-by-calendar-period interactions—the "extended" TWFE—recovers the same cohort-level ATTs within OLS; his two-way Mundlak equivalence, by contrast, describes the *uncorrected* estimator. Under parallel trends and no anticipation these are more efficient than the first-wave alternatives, because they use the whole untreated sample rather than discarding observations. The efficiency has a price: leaning on the full pre-period makes imputation more sensitive to parallel-trend violations in distant pre-treatment years. Because the 150-hour anticipation spike falls in the year just before adoption, a researcher using imputation there should say which pre-periods identify the counterfactual.

Wooldridge's approach has a second payoff for accounting researchers. In companion work, Wooldridge (2023) carries the saturated specification to nonlinear models (i.e., logit, probit, and Poisson) so that staggered DiD with a binary outcome (a restatement, a going-concern opinion, an auditor switch) or a count outcome (comment letters, misstatements, CPA exam candidates) no longer forces a choice between a misspecified linear model and an uncorrected nonlinear one. Given how many accounting outcomes are binary or counts, this is the development most likely to matter in practice.

Covariates: From Afterthought to Design Choice

In the TWFE tradition, covariates are something a researcher "adds" to the regression, often at a referee's request. The new literature treats them as a design choice with their own issues. Sant'Anna and Zhao (2020) develop doubly robust DiD estimators combining an outcome regression with propensity-score weighting; the estimator is consistent if either model is right, and it underlies the covariate-adjusted version of Callaway and Sant'Anna (2021). Caetano and Callaway (2024) show that controlling for contemporaneous covariates can introduce bias when those covariates themselves respond to treatment—"bad controls" transported to DiD. The 150-hour setting makes the danger concrete: controlling for contemporaneous state accounting-sector employment would partial out part of the effect itself, because the Rule also reshapes that sector. Here the guidance is straightforward. Condition on covariates measured before treatment, say why parallel trends is more plausible conditionally, and if covariates matter use a doubly robust implementation rather than tossing controls into a TWFE regression. The design sensitivity that Shipman et al. (2017) document for propensity-score matching in accounting applies with equal force here.

Pre-Trends: Testing Less, Bounding More

The first wave treated a flat pre-period as the design's seal of approval. Two papers by Jonathan Roth and coauthors complicate that. Roth (2022) shows that conventional pre-trends tests are often underpowered—a violation large enough to matter can survive the test—and that conditioning on having "passed" one distorts inference, because the samples that pass are selected. Rambachan and Roth (2023) replace pass/fail with a sensitivity analysis: instead of asking whether pre-trends are exactly zero, ask how large a post-treatment violation—disciplined by the observed pre-trends—would have to be to overturn the result. Their HonestDiD framework provides confidence sets under those assumptions. The 150-hour application also shows the framework's limit: the bounds are disciplined by the observed pre-trends, and here the anticipation spike contaminates $t = -1$, so the pre-trends are not measuring the counterfactual. The event date must be moved to the spike year before the sensitivity analysis is meaningful. Reporting that analysis, rather than asserting the

pre-period "looks flat," is where the literature has landed.

Inference with Few Treated Clusters

Bertrand et al. (2004) taught applied researchers to cluster at the level of treatment assignment. The recent literature refines the message for the situations accounting researchers face. With 53 clusters, as here, conventional cluster-robust standard errors perform reasonably well. But many accounting settings have few clusters, or few *treated* clusters: a regulation adopted by six states, a rule change affecting one exchange, a court ruling covering a single circuit. Then cluster-robust standard errors are biased downward, sometimes severely, and t-statistics of three or four can be artifacts of the variance estimator rather than evidence of an effect. The wild cluster bootstrap (Cameron et al. 2008) restores approximately correct inference when clusters are few overall, but MacKinnon and Webb (2018) show that the restricted version severely under-rejects when the *treated* clusters are few; there, randomization inference in the spirit of Conley and Taber (2011) is the sturdier route. Canay et al. (2021) characterize when the bootstrap works with few large clusters, and MacKinnon et al. (2023) give the practice guide. Table 1 makes this concrete: several adoption cohorts contain a single jurisdiction, their wild-bootstrap p-values are uninformative (above 0.40), and only the pooled estimands deliver credible standard errors here.

Stacked Regression, Refined

Section V reported a stacked estimate of -1.63, the largest of the three corrections in magnitude, and attributed much of its distance from the others to the reference period rather than to weighting. The issue of weighting remains an open question, and the literature has now addressed it in depth. Wing et al. (2024) formalize the stacked design: they show that the naive stacked OLS estimator converges to a variance-weighted average that generally differs from the sample-weighted ATT, and they derive corrective sample weights that restore a clean estimand while preserving the design's appeal (i.e., trimmed, balanced event windows and controls that are never previously treated). Researchers who favor stacking for its transparency should use the weighted version; the

unweighted stack in Figure 7, Panel C, is best viewed as a diagnostic companion to the corrected estimator rather than the state of the art.

What Should Accounting Researchers Do Now?

So what can we take away from the prior review? Three practical conclusions follow. First, the choice of estimator has narrowed where it matters: when never-treated units exist, treatment is absorbing, parallel trends holds, and anticipation is absent, Callaway–Sant'Anna, Sun–Abraham, imputation, and properly weighted stacking all recover interpretable averages of cohort-time ATTs. When they disagree materially, that is itself a diagnostic, usually pointing to pre-trend violations, anticipation, or overweighted cohorts. Second, the reporting standard has risen: state the estimand, name the comparison group, show a heterogeneity-robust event study, and report a parallel-trends sensitivity analysis (Baker et al. 2022). Third, none of this substitutes for design: no estimator rescues a setting without credible comparison units or with badly measured treatment dates. Appendix A organizes these obligations in the order a researcher, or a reviewer, should think about them.

The guidance above is for the canonical design: a binary, absorbing treatment, effects measured in one outcome metric, and no interference across units. Three departures deserve flags, and Online Appendix OA.8 develops each. Triple-difference designs rest on a single parallel-trends assumption for the relative outcome (Olden and Møen 2022), so a staggered triple difference is a staggered DiD in that outcome and inherits every weighting problem above. The estimators also assume that one unit's treatment leaves other units' outcomes alone: in the 150-hour setting, a candidate who responds to the requirement by sitting the examination in a non-adopting jurisdiction raises the control count while lowering the treated one, which would push every estimate in Section V away from zero, and no reweighting corrects it. And parallel trends is specific to the chosen functional form; it can hold in levels and fail in logs, or the reverse (Roth and Sant'Anna 2023), so Section V's estimates carry the assumption in logs.

VII. CONCLUSION

Difference-in-differences is now a workhorse design in archival accounting research, and roughly 30 percent of those applications exploit staggered timing. Staggering has real advantages—more distinct treatment events make it harder for one confound to explain the results—but the TWFE coefficient is then not an easily interpretable ATT, and when effects vary across cohorts or over time it is generally biased.

Using the Goodman-Bacon framework and the 150-hour Rule, I show where the bias comes from—forces that in simulation turn uniformly positive effects into a negative estimate—and how to diagnose and correct it. The first-wave corrections and the additions since (imputation, HonestDiD, few-cluster inference, corrected stacking) are now standard, and Appendix A lays out the resulting checklist. One habit is worth resisting: running every estimator, reporting agreement, and letting that stand in for a robustness section. The estimators reviewed here correct the same defect—OLS's weights on the underlying 2×2 comparisons—and every one still assumes parallel trends, so five tests answer one threat five times.

The question worth asking is which threat the design actually faces and whether the chosen method speaks to it. If the concern is that early adopters gain more and their gains accumulate, the decomposition and any heterogeneity-robust estimator address it. If the concern is anticipation, as here, no reweighting helps; move the event date or drop the window. If adopting and non-adopting states were on different trajectories, the remedy is a sensitivity analysis, not another estimator. The method should follow from a stated theory of what could go wrong; reviewers might ask which threat a paper defends against before counting estimators. A study that names its threats and confronts them is worth more than one that reports every estimate with no argument for why.

Staggered DiD forms the basis of many empirical claims in accounting about standards, regulation, and enforcement; rigorous methods to implement it already exist, and both researchers and reviewers should employ them. The default practices used today will eventually be superseded; what endures is clearly specifying which comparisons underpin each estimate.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics 138*(1), 1–35.

Angrist, J. D. and A. B. Krueger (1999). Empirical strategies in labor economics. In *Handbook of Labor Economics*, Volume 3, pp. 1277–1366. Elsevier.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arkhangelsky, D. and G. W. Imbens (2021). Double-robust identification for causal panel data models. NBER Working Paper w28364, National Bureau of Economic Research.

Armstrong, C. S., J. D. Kepler, D. Samuels, and D. J. Taylor (2022). Causality redux: The evolution of empirical methods in accounting research and the growth of quasi-experiments. *Journal of Accounting and Economics 74*(2–3), 101521.

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

Baker, A. (2019). Difference-in-differences methodology. Blog post, archived at `https://web.archive.org/web/20191208152654/https://andrewcbaker.netlify.com/2019/09/25/difference-in-differences-methodology/`.

Baker, A., B. Callaway, S. Cunningham, A. Goodman-Bacon, and P. H. C. Sant'Anna (2026). Difference-in-differences designs: A practitioner's guide. *Journal of Economic Literature 64*(2), 498–557.

Baker, A. C., D. F. Larcker, and C. C. Y. Wang (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics 144*(2), 370–395.

Barrios, J. M. (2022). Occupational licensing and accountant quality: Evidence from the 150-hour rule. *Journal of Accounting Research 60*(1), 3–43.

Ben-Michael, E., A. Feller, and J. Rothstein (2022). Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society, Series B 84*(2), 351–381.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Borusyak, K., X. Jaravel, and J. Spiess (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies 91*(6), 3253–3285.

Breuer, M. and E. deHaan (2024). Using and interpreting fixed effects models. *Journal of Accounting Research 62*(4), 1183–1226.

Caetano, C. and B. Callaway (2024). Difference-in-differences when parallel trends holds conditional on covariates. Working paper, arXiv:2406.15288.

Callaway, B., A. Goodman-Bacon, and P. H. C. Sant'Anna (2023). Difference-in-differences with a continuous treatment. Working paper, arXiv:2107.02637.

Callaway, B. and P. H. C. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics 90*(3), 414–427.

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Canay, I. A., A. Santos, and A. M. Shaikh (2021). The wild bootstrap with a "small" number of "large" clusters. *The Review of Economics and Statistics 103*(2), 346–363.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics 134*(3), 1405–1454.

Conley, T. G. and C. R. Taber (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics 93*(1), 113–125.

de Chaisemartin, C. and X. D'Haultfœuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–2996.

de Chaisemartin, C. and X. D'Haultfœuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal 26*(3), C1–C30.

de Chaisemartin, C. and X. D'Haultfœuille (2026). Difference-in-differences estimators of intertemporal treatment effects. *The Review of Economics and Statistics 108*(4), 863–880.

Egami, N. and S. Yamauchi (2023). Using multiple pretreatment periods to improve difference-in-differences and staggered adoption designs. *Political Analysis 31*(2), 195–212.

Gardner, J. (2022). Two-stage differences in differences. Working paper, arXiv:2207.05943.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

Gow, I. D., D. F. Larcker, and P. C. Reiss (2016). Causal inference in accounting research. *Journal of Accounting Research 54*(2), 477–523.

Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999). The economics and econometrics of active labor market programs. In *Handbook of Labor Economics*, Volume 3, pp. 1865–2097. Elsevier.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics 232*(2), 272–299.

MacKinnon, J. G. and M. D. Webb (2018). The wild bootstrap for few (treated) clusters. *The Econometrics Journal 21*(2), 114–135.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics 13*(2), 151–161.

Olden, A. and J. Møen (2022). The triple difference estimator. *The Econometrics Journal 25*(3), 531–553.

Rambachan, A. and J. Roth (2023). A more credible approach to parallel trends. *Review of Economic Studies 90*(5), 2555–2591.

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

Roth, J. and P. H. C. Sant'Anna (2023). When is parallel trends sensitive to functional form? *Econometrica 91*(2), 737–747.

Roth, J., P. H. C. Sant'Anna, A. Bilinski, and J. Poe (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics 235*(2), 2218–2244.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), 135–146.

Sant'Anna, P. H. C. and J. Zhao (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics 219*(1), 101–122.

Shipman, J. E., Q. T. Swanquist, and R. L. Whited (2017). Propensity score matching in accounting research. *The Accounting Review 92*(1), 213–244.

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

Wing, C., S. M. Freedman, and A. Hollingsworth (2024). Stacked difference-in-differences. NBER Working Paper 32054, National Bureau of Economic Research.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Wooldridge, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal 26*(3), C31–C66.

Wooldridge, J. M. (2025). Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators. *Empirical Economics 69*(5), 2545–2587.

APPENDIX A

A Researcher's Checklist for Staggered DiD

This appendix collects the paper's guidance into a checklist ordered the way a project unfolds: design, estimation, inference, robustness, and reporting. Authors can walk through it before submission; reviewers and editors can use it to structure their evaluation of a staggered DiD study. The final column applies each item to the paper's running example, the staggered adoption of the 150-hour Rule.

Panel A: Design Stage

#	Question	What to do	In the 150-hour application
D1	When does each unit get treated?	Tabulate adoption dates and cohort sizes before running anything. Note very early adopters, singleton cohorts, and cohorts near the middle of the panel, which OLS overweights.	Table 1: one pre-1990 adopter, cohorts of 1–16 jurisdictions from 1993–2003; the 16-jurisdiction 2000 cohort dominates the TWFE weights through size, despite a treatment share of only 0.21.
D2	Do clean comparison units exist?	Identify never-treated units or, failing that, not-yet-treated units, including the last-treated cohort (Sun and Abraham 2021). If treatment is universal and simultaneous, no staggered-DiD estimator applies.	Ten jurisdictions never adopt during the 1985–2003 sample; they anchor the Sun–Abraham and stacked estimates, while the Callaway–Sant'Anna implementation uses not-yet-treated controls.
D3	Is treatment absorbing, binary, and dated correctly?	Verify treatment does not switch off and that adoption dates are measured at the level and timing at which the rule binds. Non-absorbing or continuous treatments need dedicated estimators (de Chaisemartin and D'Haultfœuille 2026; Callaway et al. 2023).	No state repealed the Rule; effective dates come from state statutes.
D4	Can units anticipate treatment?	Ask whether agents can shift behavior before the official date; if so, redefine the event date or exclude the anticipation window.	Candidates rushed to sit the exam before the Rule bound, producing a spike in the year before adoption (Barrios 2022).

#	Question	What to do	In the 150-hour application
D5	Are covariates needed for parallel trends?	If early and late adopters differ systematically, condition on covariates measured *before* treatment; avoid contemporaneous controls that may respond to treatment (Sant'Anna and Zhao 2020; Caetano and Callaway 2024).	Adoption timing varies with state political economy; pre-treatment state characteristics, not contemporaneous accounting-sector employment, are the admissible controls.

Panel B: Estimation Stage

#	Question	What to do	In the 150-hour application
E1	What is the TWFE estimate made of?	Estimate the TWFE baseline, then run the Goodman-Bacon (2021) decomposition (derived for a balanced panel): plot each 2×2 against its weight and sum the weight on timing-based comparisons.	Figure 6: 53% of the TWFE estimate of −1.08 comes from timing comparisons; timing-only estimates average −1.04, close to the never-adopter average of −1.15—though see E2: the robust estimators still move the estimate.
E2	Which heterogeneity-robust estimator fits the data?	Match the estimator to the setting: Callaway–Sant'Anna (2021) with never/not-yet-treated controls and covariates; Sun–Abraham (2021) for event studies; stacking with corrective weights (Cengiz et al. 2019; Wing et al. 2024); imputation for efficiency (Borusyak et al. 2024; Gardner 2022; Wooldridge 2025); Wooldridge (2023) for nonlinear outcomes.	Section V implements Callaway–Sant'Anna, Sun–Abraham, and stacking; estimates range from −1.12 to −1.63 around the TWFE value of −1.08, and much of that range reflects differing reference periods (Figure 7).
E3	What do the dynamics look like?	Plot the event study from a heterogeneity-robust estimator—not only from TWFE—with a clear reference period and confidence bands.	Figures 5 and 7: flat pre-period except the anticipation year; effects deepen after adoption.
E4	What is the estimand?	State the target parameter (overall ATT, cohort ATTs, event-time profile) and the aggregation weights before comparing estimates across methods.	This paper's Callaway–Sant'Anna estimate averages event-time effects over the post-window; OLS's variance weights answer a different question, rarely the one posed.

Panel C: Inference Stage

#	Question	What to do	In the 150-hour application
I1	At what level is treatment assigned?	Justify the clustering level from the assignment and sampling design (Bertrand et al. 2004; Abadie et al. 2023), rather than by convention.	Treatment varies at the jurisdiction level, so cluster by jurisdiction even though outcomes are semi-annual.
I2	How many clusters—and how many treated clusters?	With few clusters, or few treated clusters, use the wild cluster bootstrap or design-based alternatives and say so (Cameron et al. 2008; Canay et al. 2021; MacKinnon et al. 2023).	The 53 jurisdiction clusters support conventional inference for the aggregate estimate, but singleton adoption cohorts in Table 1 cannot support cohort-level clustered inference.

Panel D: Robustness Stage

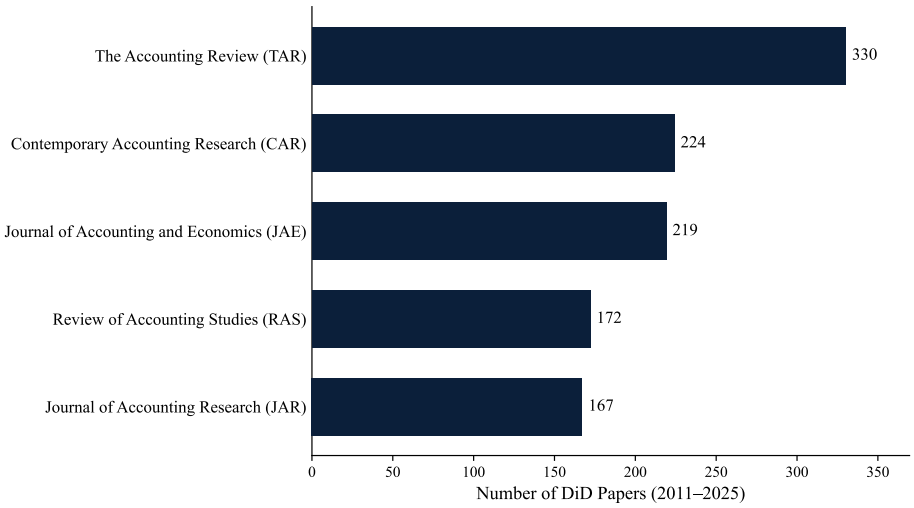
#	Question	What to do	In the 150-hour application
R1	Do alternative estimators agree?	Report at least one heterogeneity-robust estimator beside TWFE, and say what agreement establishes: these differ in weighting, not in requiring parallel trends.	All three give a large negative estimate but range from -1.12 to -1.63; much of that range tracks the reference period each uses rather than its weights. None addresses anticipation.
R2	How fragile is parallel trends?	Do not rely on a passed pre-trends test (Roth 2022). Report HonestDiD bounds (Rambachan and Roth 2023): how large a violation would overturn the result?	The anticipation spike contaminates $t = -1$, so the bounds apply only after the event date is moved to the spike year.
R3	Is the estimate stable across samples?	Re-estimate without never-treated units to see how much identification rests on them—a large move is informative and stability is not reassurance (Figure 4, Panel D). Sample-window changes move the weights, so some movement is expected.	Table 2: the coefficient moves little without the never-adopters or the pre-1990 adopter, and with trends and pre-trend adjustments—limited dependence on clean controls here, not a validated design.

Panel E: Reporting Stage

#	Question	What to do	In the 150-hour application
P1	Can the reader reconstruct the comparison?	State the estimator, the comparison group, the estimand, and the aggregation weights in the text—not only in a table note.	"Group-time ATTs relative to not-yet-treated jurisdictions, averaged over post-treatment event times" is one line.

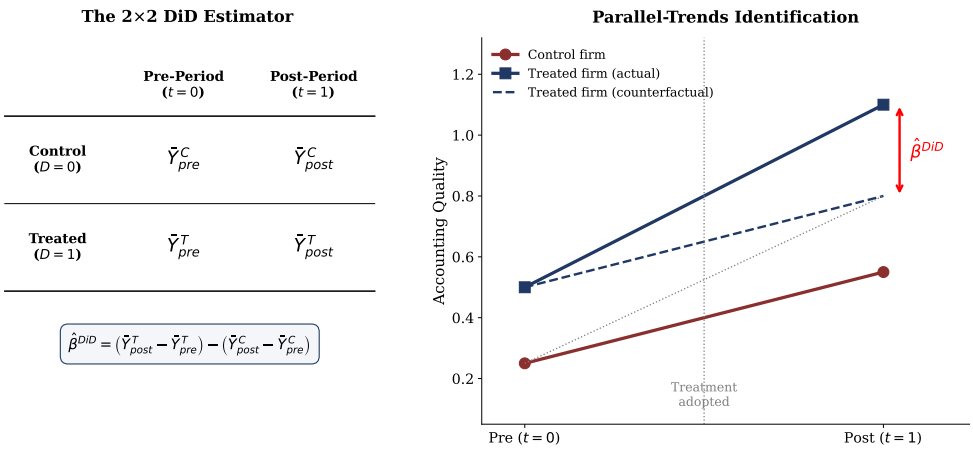
#	Question	What to do	In the 150-hour application
P2	Are the diagnostics visible?	Show the decomposition plot and the robust event study in the paper or appendix; report the weight on timing comparisons.	Figures 6 and 7 serve this role.
P3	Can others replicate it?	Name the packages and versions, post code, and state data availability, consistent with journal policy.	Simulation and replication code are available from the author's website.

Figure 1: DiD Papers in Top Accounting Journals



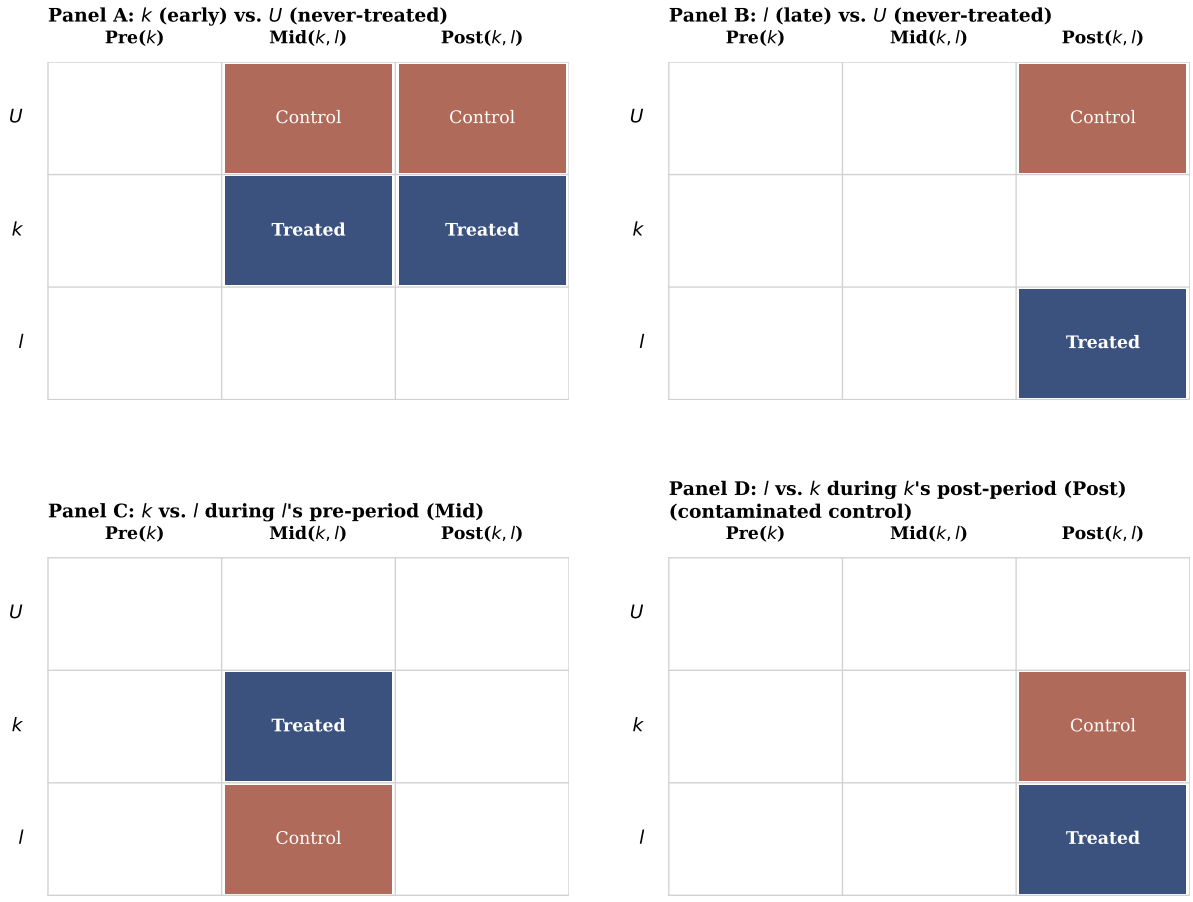
Note: This figure reports the number of papers in TAR, JAR, JAE, CAR, and RAS that contain a difference-in-differences keyword, 2011–2025 (N = 1,112). Keywords: "difference-in-difference," "difference-in-differences," "diff-in-diff," "difference in difference," and "difference in differences." Roughly 30% of DiD papers in 2011–2020 used staggered treatment timing, from the author's full-text classification of that sample (staggered or generalized DiD). Produced by `01_figure1_did_prevalence.py`.

Figure 2: The 2×2 DiD Estimator



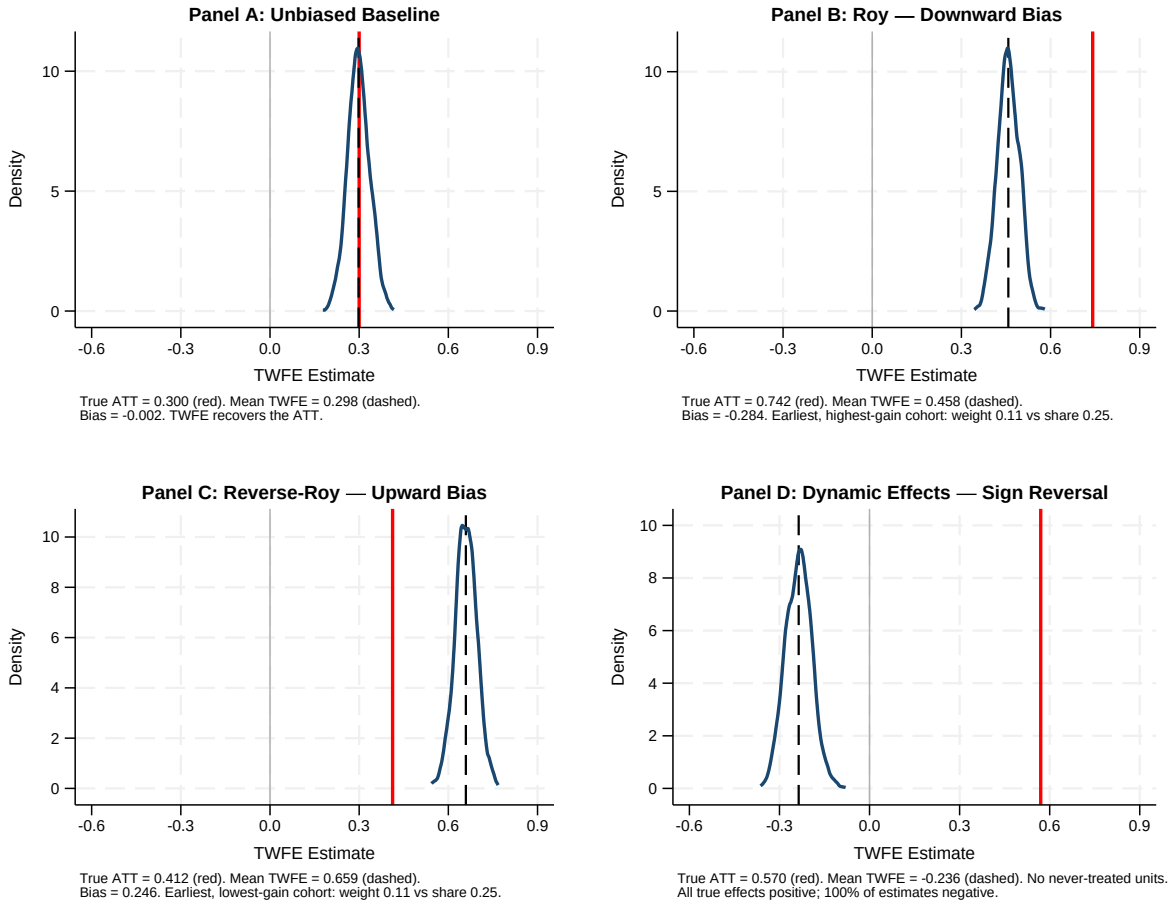
Note: This figure is a graphical representation of equation (1). Left: the four sample means in the standard 2×2 table. Right: treated and control outcome paths; the DiD estimate is the gap between the treated unit's actual and counterfactual post-period outcomes under parallel trends. Online Appendix Figure OA1 provides the companion tabular walkthrough.

Figure 3: The Four 2×2 Comparisons in a Three-Group Staggered Design



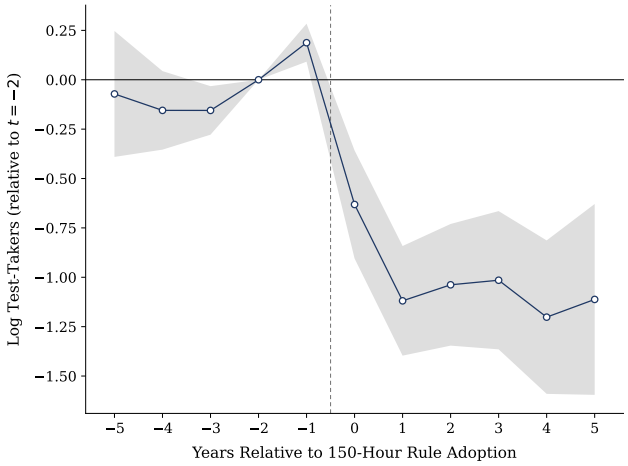
Note: This figure shows the groups and periods that generate the four 2×2 DiDs with early-treated k, late-treated l, and untreated U. Panels A–B: treated vs. untreated. Panel C: early vs. late in the late group's pre-period. Panel D: the forbidden comparison—late vs. already-treated early in the late group's post-period. Online Appendix Figure OA2 shows the underlying three-group timeline.

Figure 4: Staggered TWFE DiD Estimates: Monte Carlo Simulations



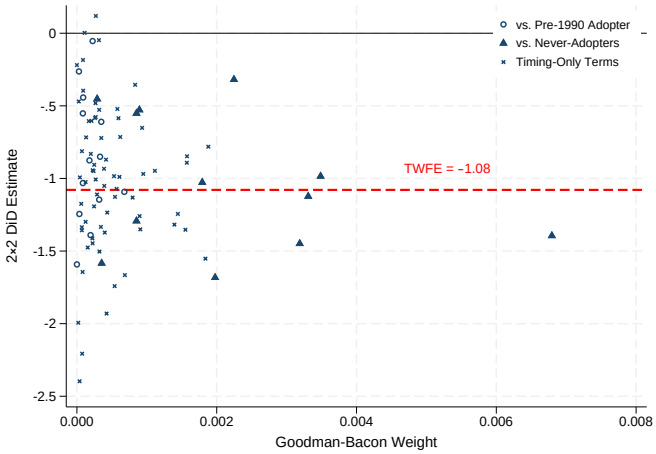
Note: This figure plots distributions of TWFE estimates from 1,000 simulations. Panel A: constant homogeneous effects—TWFE recovers the true ATT (0.300). Panels B–C: Roy / reverse-Roy cross-cohort heterogeneity—the weights distort the estimate in both directions but do not flip its sign. Panel D: accumulating effects and no never-treated states—100% of estimates are negative despite every true effect being positive by construction. Red line: the true ATT, defined throughout as the average effect on treated observations; dashed black: mean TWFE; thin grey: zero. The error standard deviation governs the width of each distribution but not its center, so the bias shown is invariant to it. Full DGP details, exact probability limits, and the sensitivity of Panel D to the never-treated share are in Online Appendix OA.5. Produced by `02_figure4_montecarlo.do`.

Figure 5: Event Study: Effect of the 150-Hour Rule on CPA Exam Test-Takers



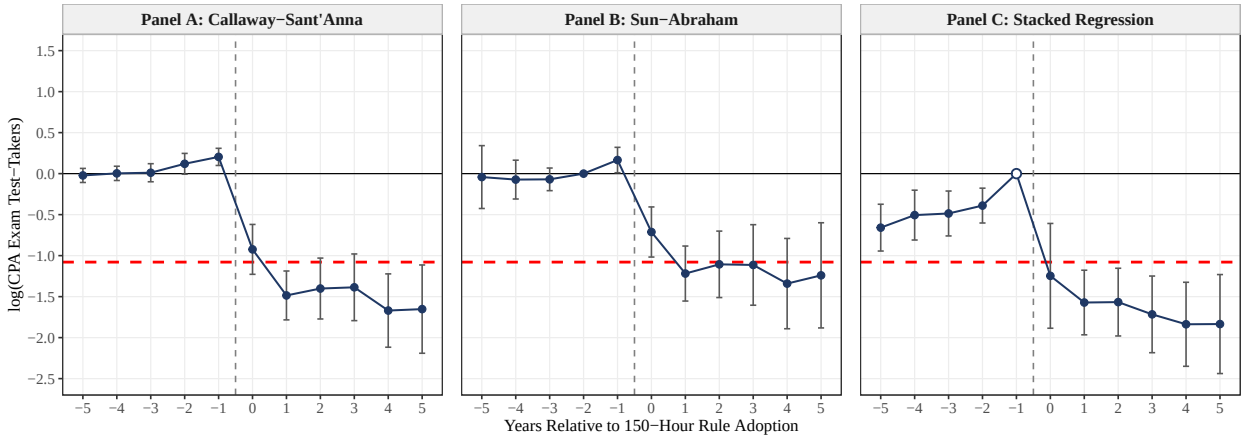
Note: This figure plots TWFE event-study estimates with jurisdiction and year fixed effects and a sitting-month control; 95% CI shaded; SEs clustered by jurisdiction. The single-coefficient version of this specification gives $\hat{\beta} = -1.07$, against -1.08 for the period fixed-effects specification of Table 2, column 1; the pooled estimate quoted in the text is the latter. Either implies roughly a two-thirds reduction in test-takers ($\exp(-1.08) - 1 \approx -0.66$). The outcome is the log of one plus the number of Uniform CPA Examination candidates per jurisdiction-sitting. The panel covers 53 licensing jurisdictions over 36 semi-annual sittings (May and November, 1985–2003, excluding 1989); $N = 1{,}901$. Event time is measured in years relative to adoption, and $t = -2$ is the omitted baseline. Leads and lags beyond ±5 are accumulated into the endpoint bins. Produced by `03_figure5_event_study.do` and `03b_figure5_plot.py`.

Figure 6: Bacon Decomposition: 150-Hour Rule and CPA Exam Test-Takers



Note: This figure plots each 2×2 DiD against its weight. Open circles: vs. the pre-1990 (always-treated) adopter. Triangles: vs. never-adopters (none adopts before 2004). Crosses: timing-only. Red line: TWFE −1.08. Timing comparisons carry 53% of the weight; never-adopters 43%; the pre-1990 adopter 4%. The timing-only average (−1.04) lies close to the TWFE. Weights are plotted as returned by `bacondecomp` and are not normalized to sum to one; the percentages above are shares of their total, the normalization of Section III, and the weighted mean of the 2×2s reproduces the TWFE estimate. The components pool the two directions of each timing pair; re-run with `ddetail` on the balanced 50-jurisdiction sub-panel, the timing weight splits into 39% clean earlier-versus-later and 11% forbidden later-versus-earlier, with weighted means −0.97 and −0.92. Sitting-level treatment dates give 12 timing groups against Table 1's 11 cohorts. Produced by `04_figure6_bacon_decomp.do`.

Figure 7: Heterogeneity-Robust Estimators: 150-Hour Rule



Note: This figure plots event-study profiles from Callaway–Sant'Anna (Panel A), Sun–Abraham (Panel B), and stacked regression (Panel C). Dashed red line: baseline TWFE (−1.08). Average post-treatment effects, with jurisdiction-clustered standard errors, are −1.40 (0.24), −1.12 (0.22), and −1.63 (0.21), respectively; the Callaway–Sant'Anna figure is that estimator's aggregate over all post-treatment periods ($t = 0$ to 10), while Panels B and C average over the plotted window. Vertical bars are pointwise 95% confidence intervals, plotted as the estimate plus or minus 1.96 standard errors; the hollow marker in Panel C is the omitted reference period. The panels use different reference periods: Panel A reports $t = -1$ as an estimate under its varying base period, Panel B omits $t = -2$ (as Figure 5 does) with endpoints binned at ±5 and the pre-1990 adopter excluded, and Panel C omits $t = -1$. Because $t = -1$ is the anticipation year, the choice of baseline accounts for part of the difference between the three averages; the panels also differ in control group, fixed effects, averaging window, and time aggregation. Panel A uses not-yet-treated controls and is estimated on jurisdiction-year means, comparing cohort-by-period cells directly rather than through fixed effects; Panels B and C use never-adopters (none adopts before 2004) on the semi-annual panel, Panel B with jurisdiction and sitting fixed effects, Panel C with jurisdiction and calendar-year fixed effects interacted with each stacked window. Produced by `05_figure7_correction_methods.do`, `05_figure7_stacked.R`, and `06_figure7_combine.R`.

Table 1: The 150-Hour Rule Rollout: Treatment Times, Group Sizes, Treatment Shares, and Cohort DiD Estimates

150-Hour Rule Year (t_k)	Number of Jurisdictions	% Share	Treatment Share	Cohort DiD Estimate (vs. Never-Treated)
Never Adopted in Sample	10	18.87	—	—
Pre-1990 Adopters	1	1.89	—	—
1993	1	1.89	0.61	−0.552†
1994	1	1.89	0.53	−0.516†
1995	2	3.77	0.50	−1.027†
1996	1	1.89	0.39	−1.293†
1997	4	7.55	0.36	−1.094*** (0.293)
1998	4	7.55	0.33	−1.449*** (0.293)
1999	3	5.66	0.25	−1.610*** (0.336)
2000	16	30.19	0.21	−1.149*** (0.343)
2001	6	11.32	0.17	−1.014** (0.353)
2002	1	1.89	0.11	−1.584†
2003	3	5.66	0.03	−0.218 (0.386)

Notes: This table lists the effective years of the 150-hour Rule from Barrios (2022), with the number and share of adopting jurisdictions and the share of sample periods each cohort spends treated. The sample covers the 53 jurisdictions that license CPAs (50 states, the District of Columbia, Guam, and Puerto Rico) over 36 semi-annual examination sittings (May and November, 1985–2003, excluding 1989); shares are of 53. Each cohort DiD compares that cohort with the ten never-adopting jurisdictions only, drawn from the same 1,901-sitting estimation sample and using the same treatment indicator as the rest of the paper, absorbing jurisdiction and year fixed effects (Table 2 uses sitting fixed effects). The dependent variable is the log of one plus the number of Uniform CPA Examination candidates per jurisdiction-sitting. Jurisdiction-clustered standard errors in parentheses, with significance from the t distribution on $G-1$ degrees of freedom, where G is the number of jurisdictions in that regression; *** $p<0.01$, ** $p<0.05$; † fewer than three adopting jurisdictions, so no clustered standard error is reported. Wild-cluster bootstrap p-values (Webb weights, 9,999 draws), reported in the replication package, weaken but do not overturn the starred cohorts (the 1997 and 1999 p-values rise from below 0.01 to 0.018 and 0.027) and are uninformative for single-jurisdiction cohorts ($p>0.40$), illustrating the few-treated-cluster caution of Section VI. Produced by `06_table1_cohort_did.do`.

Table 2: TWFE DiD Estimates of the Effect of the 150-Hour Rule on Supply of CPAs: Alternative Specifications

	(1) Baseline	(2) Sitting-Month Control	(3) No Untreated Jurisdictions	(4) Switching Jurisdictions	(5) Jurisdiction-Specific Trends	(6) Cohort Pre-Trends
150-hour Rule	−1.079***	−1.079***	−1.028***	−1.040***	−1.037***	−1.105***
	(0.149)	(0.149)	(0.122)	(0.130)	(0.110)	(0.165)
Observations	1,901	1,901	1,541	1,505	1,901	1,901
Adjusted R^2	0.828	0.828	0.816	0.816	0.897	0.871
Mean test-takers (level)	425.98	425.98	350.92	351.59	425.98	425.98
Sitting-month control	No	Yes	No	No	No	No
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
Jurisdiction FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports DiD estimates of the effect of the 150-hour Rule on log CPA exam test-takers. All columns absorb jurisdiction and period fixed effects, where a period is one of the 36 semi-annual examination sittings. (1) baseline TWFE; (2) adds a sitting-month control, which is collinear with the period fixed effects (the 0.0002 movement is numerical); (3) drops never-adopters; (4) drops never-adopters and the pre-1990 adopter (switching jurisdictions only); (5) adds jurisdiction-specific linear trends in the sitting index; (6) fits a linear trend by adoption cohort on pre-treatment sittings only, subtracts it from the outcome in all periods, and re-estimates. Jurisdiction-clustered standard errors in parentheses; *** $p<0.01$. Produced by `07_table2_specifications.do`.

Table 3: A Menu of Staggered DiD Estimators

Approach	Estimand	Admissible controls	Covariates	When / package
Goodman-Bacon	TWFE as weighted avg. of 2×2s	— (diagnostic)	No	Balanced panel; first look at weights (`bacondecomp`)
Callaway–Sant'Anna	Cohort-time ATT, then researcher-weighted aggregate	Never- or not-yet-treated	PS / DR	Explicit ATTs + covariates (`did`)
Sun–Abraham	Cohort dynamics, sample-share IW avg.	Never- or last-treated	Regression	Event-study focus (`eventstudyinteract`)
Stacked	Event-window ATT on stacked datasets	Not previously treated in window	Regression	Transparent clean 2×2s; prefer weighted stack
Imputation / ETWFE	Unit-time effects, then average	Untreated obs. only	Regression; nonlinear OK	Efficiency; binary/count outcomes (`did2s`)

Notes: This table summarizes the estimators in Sections IV and VI. Formal statements appear in Online Appendix OA. "Weighted stack" refers to Wing et al. (2024). "DR" = doubly robust. Package names are indicative; equivalents exist in Stata and R.

Online Appendix OA

Staggeringly Problematic: A Primer on Staggered DiD for Accounting Researchers

Formal Statements, Simulation Design, and Supplemental Figures

This online appendix collects (i) the formulas behind the estimators summarized in Section IV and Table 3, (ii) the Monte Carlo design behind main-text Figure 4, and (iii) supplemental figures that support magnitudes already stated in the main text.

OA.1 Callaway and Sant'Anna (2021)

Callaway and Sant'Anna (2021) estimate a group-time average treatment effect on the treated under parallel trends (possibly conditional on covariates) and no anticipation. Let $G_g = 1$ if a unit is first treated in period g, let $C = 1$ for units in the chosen comparison group (never-treated or not-yet-treated), and let $p_g(X) = P(G_g = 1 \mid X, G_g + C = 1)$ denote the propensity score. One IPW representation of the group-time ATT is

$$\mathrm{ATT}(g,t) = E\left[\left\{\frac{G_g}{E\left[G_g\right]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{E\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right\}\left(Y_t - Y_{g-1}\right)\right]. \tag{OA.1}$$

The building blocks $\mathrm{ATT}(g,t)$ are then aggregated into an event-study profile or an overall ATT with researcher-chosen weights. The R package `did` implements both never-treated and not-yet-treated comparison groups, as well as the doubly robust extension of Sant'Anna and Zhao (2020). Previously treated units never enter as controls.

OA.2 Sun and Abraham (2021)

Sun and Abraham (2021) work inside an event-study regression saturated with cohort × relative-time interactions. Let G_g indicate membership in adoption cohort g, as in OA.1, and let D_i^l indicate that calendar time is l periods relative to i's adoption (the superscripted D^l is a relative-time indicator, distinct from the treatment status D_{it}). Their interaction-weighted (IW) estimator first recovers cohort-specific dynamic effects from the saturated regression (using never-treated or last-treated units as controls), then averages those effects by each cohort's sample share:

$$\hat{\mu}_l = \sum_g \hat{\nu}_{g,l}\,\widehat{\Pr}\{G_g = 1 \mid D^l = 1\}, \tag{OA.2}$$

where $\hat{\nu}_{g,l}$ is the estimated effect for cohort g at relative time l. The resulting $\hat{\mu}_l$ is the event-study profile reported in applications. In Stata, `eventstudyinteract` implements the IW estimator.

OA.3 Stacked Regression

Following Cengiz et al. (2019), construct one event dataset s for each adoption cohort: the treated cohort plus units that remain untreated throughout a fixed event window around that cohort's adoption date. Stack the datasets and estimate

$$Y_{ist} = \alpha_{is} + \alpha_{ts} + \sum_{l \neq -1} \beta_l D^l_{ist} + u_{ist}, \tag{OA.3}$$

where α_{is} and α_{ts} are unit-by-stack and time-by-stack fixed effects and D^l_{ist} are relative-time indicators within stack s. Previously treated units never appear on the control side of any stack. Unweighted stacked OLS converges to a variance-weighted average across stacks; Wing et al. (2024) derive sample weights that restore a clean ATT while preserving the design's trimmed windows.

OA.4 Imputation and Extended TWFE

Borusyak et al. (2024) and Gardner (2022) estimate unit and time fixed effects from untreated observations only and impute each treated observation's counterfactual; the treatment effect is the residual, which can then be averaged arbitrarily. Wooldridge (2025) reaches the same cohort-level ATTs through a TWFE regression fully saturated with cohort × calendar-period interactions (the "extended" TWFE; the same paper's two-way Mundlak result is, by contrast, an equivalence for the uncorrected estimator), and Wooldridge (2023) extends the saturated specification to logit, probit, and Poisson—of particular value when the accounting outcome is binary or a count. Packages include `did_imputation` / `didimputation` and `did2s`.

OA.5 Simulation Design

Main-text Figure 4 reports Monte Carlo distributions of the TWFE estimator under four data-generating processes (DGPs). This section records the design in enough detail to reproduce the figure. The simulations follow the spirit of Baker (2019) but are calibrated to a state-year panel that resembles the 150-hour rollout in main-text Table 1. Stata code (seeded) is available from the author's website.

Panel structure and outcome equation

Each replication draws a balanced panel of $N = 50$ states observed annually from 1980 to 2010 ($T = 31$). The untreated potential outcome is

$$Y_{it}^0 = \alpha_i + \alpha_t + \varepsilon_{it}, \tag{OA.4}$$

where α_i is a time-invariant state effect, α_t is a common time effect (the notation of main-text equation (3)), and ε_{it} is idiosyncratic noise. States are assigned to adoption cohorts in equal blocks (details below); because α_i is drawn independently of the state index, block assignment and random assignment are equivalent here. Once state i's cohort adoption year t_c arrives, a treatment effect is added to the observed outcome. Think of y_{it} as the log number of CPA exam candidates—or any state-level accounting outcome with unit and time structure.

In every replication I estimate the standard TWFE regression

$$y_{it} = \alpha_i + \alpha_t + \delta^{DD} D_{it} + u_{it}, \tag{OA.5}$$

where $D_{it} = 1\{t \geq t_{c(i)}\}$ for adopting states and $D_{it} = 0$ otherwise, with standard errors clustered by state (`reghdfe`). Each DGP is repeated 1,000 times. The seed for the reported run is 20240101.

Four data-generating processes

Two conventions apply throughout. First, the benchmark is the conventional ATT, $E[Y^1 - Y^0 \mid D = 1]$, the average effect over treated state-years. Because early cohorts spend more of the panel treated, this differs from the simple average across cohort effects; I report both below, and the sign of every bias is the same under either. Second, the idiosyncratic error standard deviation is 0.40. Since `reghdfe` is linear in the outcome while the within-transformation of D_{it} does not involve the outcome, $E[\hat{\delta}^{DD}]$ is exactly invariant to this parameter and only the dispersion of each distribution changes. I therefore report exact probability limits, computed by regressing the deterministic treatment effect on D_{it} with two-way fixed effects, alongside the Monte Carlo means.

The exact TWFE weights on the four cohorts in Panels A–C are 0.108, 0.289, 0.326, and 0.276, against an equal sample share of 0.250. The 1983 cohort is heavily under-weighted because a cohort treated for nearly the whole panel has little within-unit treatment variance; the remaining three cohorts are all somewhat over-weighted.

Panel A — Unbiased baseline. States are assigned to one of four equally sized adoption cohorts (1983, 1990, 1997, 2002) of ten states each, with ten never-treated states (20%). Every treated state receives a constant treatment effect $\mu_i \sim N(0.30,\, 0.04^2)$, the same mean for every cohort and constant over time. The true ATT is 0.300 under either aggregation. Under this DGP

the TWFE estimator is unbiased: the exact probability limit is 0.3000 and the mean of the 1,000 estimates is 0.298.

Panel B — Roy (early adopters gain more). Same staggered structure as Panel A, but cohort means differ: 1.20, 0.70, 0.30, and 0.10 for the 1983, 1990, 1997, and 2002 cohorts, respectively. Effects remain constant within cohort over time. The true ATT is 0.742 (0.575 equally weighted across cohorts). Because the under-weighted 1983 cohort is the high-gain one, the estimate is dragged down: exact plim 0.4577, Monte Carlo mean 0.458, bias −0.284, or 7.9 sampling standard deviations. Distorted weights, but no forbidden comparison poisons the average.

Panel C — Reverse-Roy (early adopters gain less). Cohort means flip to 0.10, 0.30, 0.70, and 1.20. The true ATT is 0.412 (again 0.575 equally weighted); it differs from Panel B's because the cohorts contributing the most treated observations now carry the smallest effects. The under-weighted 1983 cohort is now the low-gain one, so the same channel operates in reverse: exact plim 0.6576, Monte Carlo mean 0.659, bias +0.245, or 6.7 sampling standard deviations.

Panel D — Trend-break (effects accumulate). Five adoption cohorts (1984, 1988, 1992, 1996, 2000) of ten states each and no never-treated states. The treatment effect grows linearly with time since adoption,

$$\tau_{it} = \mu_i \times (t - t_c + 1) \quad \text{for } t \geq t_c, \tag{OA.6}$$

with per-period slopes μ_i drawn lognormally around cohort means 0.070, 0.055, 0.040, 0.030, and 0.020, respectively—so earlier adopters accumulate larger cumulative effects. The lognormal draw makes every state's slope positive with probability one, so every state's treatment effect is positive in every post-adoption period by construction. The true ATT is 0.570 (0.480 equally weighted). This is the design that produces the paper's headline result: the exact probability limit is −0.2353, the Monte Carlo mean is −0.236, and 100% of the 1,000 draws are negative. Already-treated cohorts, whose outcomes are still rising from their own treatment, serve as controls for later cohorts.

What the sign reversal requires. Two features of Panel D drive the result. First, there are no never-treated states, so every comparison is a timing comparison and treated-versus-never comparisons carry none of the identifying weight. Second, the panel runs to 2010, a decade after the last adoption, so in eleven of the thirty-one sample years no untreated state exists at all; the two-way within transformation still assigns those years non-zero weight, with the sign flipped for early adopters. Both matter quantitatively. Holding the panel at 50 states and replacing ten of the treated states with never-adopting ones—each of the five cohorts shrinks from ten states to eight—moves the probability limit from −0.235 to +0.037; truncating the panel at 2001 instead of 2010 moves it to +0.019. Clean controls and a shorter post-adoption window restore the sign, though in the first case the estimate still understates the true effect by more than half. The sensitivity is the point: it is the quantitative counterpart of the clean-comparison-units question in the Appendix A checklist (item D2).

What the figure shows

Main-text Figure 4 plots the kernel density of the 1,000 $\hat{\delta}^{DD}$ draws for each DGP. The solid red vertical line marks the true ATT in every panel; the dashed black line marks the mean TWFE estimate; a thin grey line marks zero. Panels A–C illustrate that cross-cohort heterogeneity alone warps the weights but need not flip the sign. Panel D shows that time-varying effects can.

OA.6 Two Goodman-Bacon Diagnostics

The two-step pre-trend adjustment

Column 6 of main-text Table 2 reports GB's two-step pre-trend adjustment. First, I fit a linear trend by adoption cohort using pre-treatment sittings only, and subtract the fitted trend from the outcome in every period. Jurisdictions that do not adopt inside the sample window form one group and are entirely pre-treatment; the single pre-1990 adopter has no in-sample pre-period, so no trend is fitted for it and its outcome passes through unchanged. Second, I estimate the uncontrolled regression on the transformed outcome. The adjusted estimator is unaffected by effect dynamics and does not change the 2×2 weights. It requires enough pre-periods before any unit is treated, which is why the pre-1990 adopter cannot contribute.

Comparing weights across specifications

GB also recommends an Oaxaca-Blinder-Kitagawa comparison: run several specifications of the DiD regression and compare both the 2×2 estimates and the weights they receive. Abstracting from bias, if the aggregate coefficient moves because the weights moved, the specification is delivering a different estimand; if it moves because the underlying 2×2 estimates moved, that is more consistent with bias. Researchers should also watch how adding controls changes the variation the TWFE estimator uses.

OA.7 Supplemental Figures

The figures below support the schematic and tabular arguments of the main text. The printed paper keeps the DiD prevalence chart (Figure 1), the 2×2 identification graph (Figure 2), the four 2×2 comparisons (Figure 3), the Monte Carlo distributions (Figure 4), and the 150-hour application (Figures 5–7).

Figure OA1: 2×2 DiD Framework—Tabular Illustration

Panel A: Cross-sectional Difference

Company	**Accounting Quality**
Treated Firm	Y=TF +A
Control Firm	Y=CF

Panel B: Time Series Difference

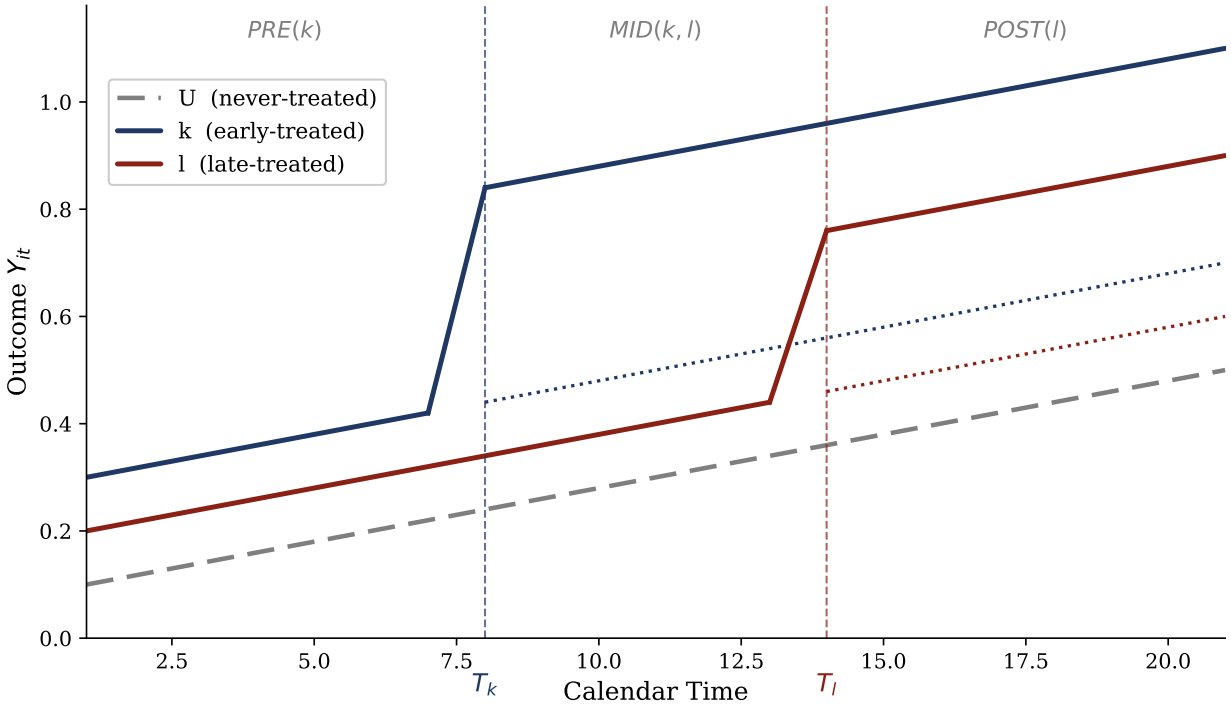
Company	**Time**	**Accounting Quality**
Treated Firm	Pre	Y=TF
	Post	Y=TF+(T+A)

Panel C: Difference in Differences

Company	**Time**	**Accounting Quality**	**Diff$_1$**	**Diff$_2$**
Treated Firm	Pre	Y=TF		
	Post	Y=TF+(T+A)	T+A	
				A
Control Firm	Pre	Y=CF		
	Post	Y=CF+T	T	

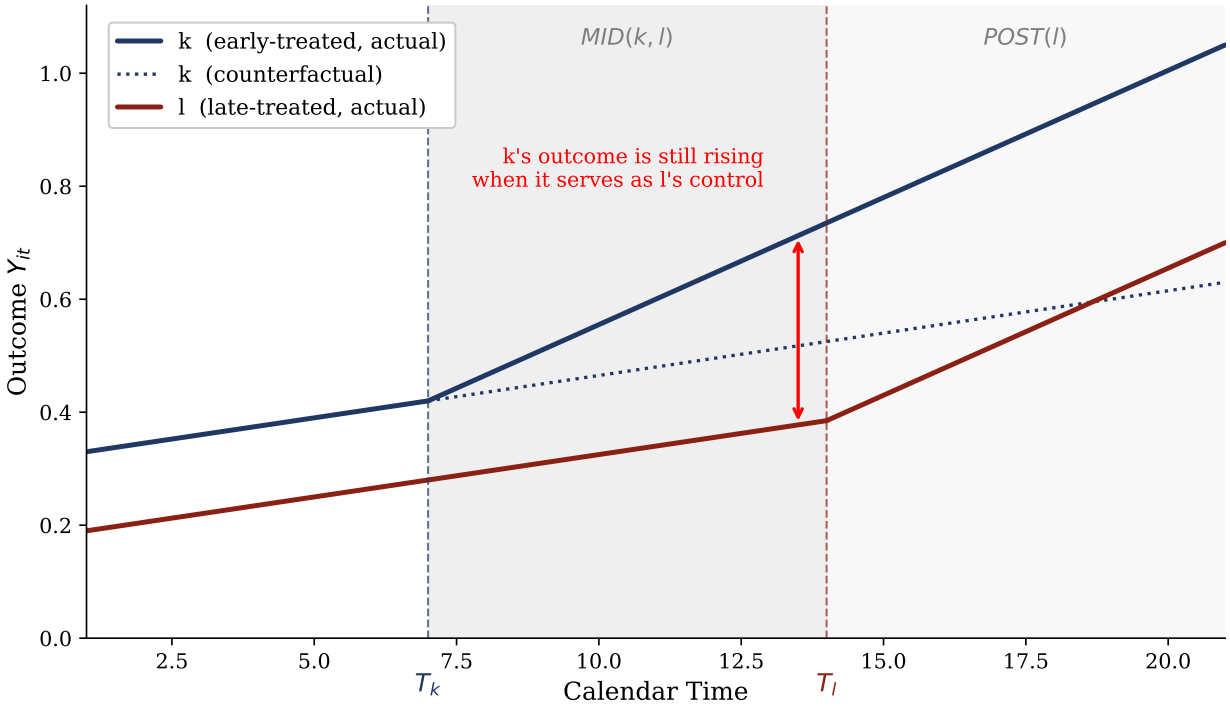
Note: This figure is the tabular companion to main-text Figure 2. Panel A: the cross-sectional difference confounds the treatment effect A with the selection term $TF - CF$. Panel B: the within-firm time-series difference removes selection but confounds A with the common trend T. Panel C: differencing the two differences recovers A under parallel trends.

Figure OA2: Staggered DiD with Three Groups



Note: This figure plots outcome paths for never-treated U, early-treated k (treated at T_k), and later-treated l (treated at T_l). Periods: PRE(k), MID(k, l), and POST(l). Solid lines (dashed for U) are actual paths; dotted lines are counterfactuals. Companion to main-text Figure 3. Presentation follows Goodman-Bacon (2021). Produced by `08_figures_OA2_OA3.py`.

Figure OA3: Bias When Treatment Effects Vary Over Time



Note: This figure shows a stylized timing-only DiD with a linear trend-break treatment effect rather than a level shift. The late-vs-early 2×2 in POST(l) is biased—and can even be negative despite a positive, growing true effect. Main-text Figure 4, Panel D, shows the same failure in Monte Carlo. Presentation follows Goodman-Bacon (2021). Produced by `08_figures_OA2_OA3.py`.

OA.8 Designs and Assumptions Beyond the Primer

Main-text Section VI closes by flagging three departures from the canonical design, which assumes a binary, absorbing treatment, one outcome metric, and no interference across units. This section develops each and points to the literature that handles it.

Triple differences

A triple-difference (DDD) design adds a within-jurisdiction comparison to the DiD: affected versus unaffected groups, before versus after, in adopting versus non-adopting jurisdictions. Olden and Møen (2022) show the design requires one parallel-trends assumption rather than two: trends in the relative outcome, the within-jurisdiction gap between affected and unaffected groups, must be parallel across adopting and non-adopting jurisdictions. That result carries the whole primer over: a staggered triple difference is a staggered DiD estimated on the relative outcome, so the weighting arithmetic of Section III, the Bacon decomposition, and the corrected estimators of Section IV apply to it unchanged, and the Appendix A checklist can be run with the relative outcome in place of y_{it}.

Spillovers and interference

Every estimator in the paper assumes one unit's treatment leaves other units' outcomes alone. The 150-hour setting shows the local version of the threat: the outcome counts candidates sitting in a jurisdiction, and a candidate facing the 150-hour requirement at home who instead sits in a non-adopting jurisdiction is subtracted from a treated count and added to a control count. The DiD contrast then overstates the true decline, and because the contamination sits in the comparison rather than the weights, none of the corrections in Section IV touches it; Callaway–Sant'Anna, Sun–Abraham, and the stack would all inherit the same inflation. Diagnostics are design-specific. A natural first pass here is geographic: if cross-border sitting drives part of the estimate, never-adopters bordering adopting jurisdictions should show larger candidate increases than remote ones, and aggregating adjacent jurisdictions into single units should shrink the estimate.

Non-absorbing and non-binary treatments

Several of the introduction's motivating examples are treatments that can switch off, and Section VI notes that the framework has outgrown the binary absorbing case. de Chaisemartin and D'Haultfœuille (2026) develop estimators built from comparisons of switchers to non-switchers that remain valid when treatment turns on and off; Callaway et al. (2023) study continuous treatments and show that comparing outcomes across doses requires a stronger, dose-specific parallel-trends assumption. For an absorbing binary rule like the 150-hour requirement none of this binds, but a researcher whose treatment reverses, a going-concern opinion withdrawn, an auditor switched back, should start from those papers rather than force the treatment into an absorbing indicator.

Functional form

Roth and Sant'Anna (2023) show that parallel trends is a functional-form-specific assumption: it can hold in levels and fail in logs, or the reverse, and it is invariant to the transformation only under strong conditions on the distribution of untreated outcomes. The choice is therefore substantive, not cosmetic. Section V estimates everything in $\log(1 + \text{candidates})$, so the identifying assumption is parallel trends in that transformed outcome, and the proportional translations $(\exp(\beta) - 1)$ inherit it. The practical advice is to pick the form the economics dictates, state that parallel trends is assumed in that form, and report the untransformed estimate as a check when the theory is agnostic.

References

Baker, A. (2019). Difference-in-differences methodology. Blog post, archived at https://web.archive.org/web/20191208152654/https://andrewcbaker.netlify.com/2019/09/25/difference-in-differences-methodology/.

Borusyak, K., X. Jaravel, and J. Spiess (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies 91*(6), 3253–3285.

Callaway, B., A. Goodman-Bacon, and P. H. C. Sant'Anna (2023). Difference-in-differences with a continuous treatment. Working paper, arXiv:2107.02637.

Callaway, B. and P. H. C. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics 134*(3), 1405–1454.

de Chaisemartin, C. and X. D'Haultfœuille (2026). Difference-in-differences estimators of intertemporal treatment effects. *The Review of Economics and Statistics 108*(4), 863–880.

Gardner, J. (2022). Two-stage differences in differences. Working paper, arXiv:2207.05943.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

Olden, A. and J. Møen (2022). The triple difference estimator. *The Econometrics Journal 25*(3), 531–553.

Roth, J. and P. H. C. Sant'Anna (2023). When is parallel trends sensitive to functional form? *Econometrica 91*(2), 737–747.

Sant'Anna, P. H. C. and J. Zhao (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics 219*(1), 101–122.

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

Wing, C., S. M. Freedman, and A. Hollingsworth (2024). Stacked difference-in-differences. NBER Working Paper 32054, National Bureau of Economic Research.

Wooldridge, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal 26*(3), C31–C66.

Wooldridge, J. M. (2025). Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators. *Empirical Economics 69*(5), 2545–2587.