Journal of Financial Economics 144 (2022) 370–395

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

How much should we trust staggered difference-in-differences estimates?☆

Andrew C. Baker[a], David F. Larcker[b], Charles C.Y. Wang[c,*]

[a] Stanford University, Rock Center for Corporate Governance, Stanford, CA, United States
[b] Stanford GSB, Stanford University, Rock Center for Corporate Governance, Stanford, CA, United States
[c] Harvard Business School, Boston, MA, United States

ARTICLE INFO

Article history:
Received 10 December 2021
Revised 19 January 2022
Accepted 19 January 2022
Available online 22 February 2022

JEL classification:
C13
C18
C21
C22
C23

Keywords:
Difference in differences
Staggered difference-in-differences
Generalized difference-in-differences
Dynamic treatment effects
Treatment effect heterogeneity

ABSTRACT

We explain when and how staggered difference-in-differences regression estimators, commonly applied to assess the impact of policy changes, are biased. These biases are likely to be relevant for a large portion of research settings in finance, accounting, and law that rely on staggered treatment timing, and can result in Type-I and Type-II errors. We summarize three alternative estimators developed in the econometrics and applied literature for addressing these biases, including their differences and tradeoffs. We apply these estimators to re-examine prior published results and show, in many cases, the alternative causal estimates or inferences differ substantially from prior papers.

1. Introduction

The estimation of policy effects—either the average effect or the average effect on the treated—is at the core of empirical finance, accounting, and legal studies. "Difference-in-differences" (DiD) is a workhorse estimation approach for making causal inference in these fields and a centerpiece of the "credibility revolution" over the prior thirty years. It typically leverages the passage of laws or market rules (treatment), impacting one set of firms or market participants (treated) but not others (controls), and compares the differences in the outcomes between treated and controls over time to infer causal effects.

A generalized version of this estimation approach that relies on the staggered adoption of laws or regulations (e.g., across states or countries) has become especially popular over the last two decades. Table 1 shows that, from 2000 to 2019, there were 744 papers published in top-five finance (431 papers) or accounting (313 papers) journals that use DiD designs. Among them, 407 (55%

☆ Toni Whited was the editor for this article. For helpful feedback, we thank Toni Whited, an anonymous referee, David Childers, Andrew Goodman-Bacon, Kirill Borusyak, Pamela Jakiela, Robert S. Kaplan, Pedro Sant'Anna (discussant), Edmund Schuster (discussant), and Holger Spamann as well as seminar participants at Stanford GSB, Harvard Business School, the Florida-Michigan-Virginia virtual law and economics workshop, and the LSE/UCL London Law and Finance Workshop for helpful comments and suggestions. We also thank the authors from Fauver et al. (2017) for graciously sharing their data and code, and Beck et al. (2010) for posting their data and code online.

* Corresponding author.
E-mail addresses: charles.cy.wang@hbs.edu, cwang@hbs.edu (C.C.Y. Wang).

https://doi.org/10.1016/j.jfineco.2022.01.004

Table 1
Use of DiD and Staggered DiD in Finance and Accounting: 2000–2019.

	(1) DiD	(2) Staggered DiD	(3) Staggered / All (%)
Journal of Finance	52	30	57.6%
Journal of Financial Economics	163	85	52.1%
Review of Financial Studies	138	75	54.3%
Review of Finance	27	14	51.8%
Journal of Financial and Quantitative Analysis	51	32	62.7%
Finance	**431**	**236**	**54.7%**
Journal of Accounting Research	52	24	46.1%
Journal of Accounting and Economics	63	38	60.3%
The Accounting Review	110	63	57.2%
Review of Accounting Studies	47	28	59.5%
Contemporary Accounting Research	41	18	43.9%
Accounting	**313**	**171**	**54.6%**
Finance and Accounting	**744**	**407**	**54.7%**

Note: Table 1 summarizes the number of papers published in five finance (*Journal of Finance, Journal of Financial Economics, Review of Financial Studies, Review of Finance*, and *Journal of Financial and Quantitative Analysis*) and five accounting (*Journal of Accounting Research, Journal of Accounting and Economics, The Accounting Review, Review of Accounting Studies*, and *Contemporary Accounting Research*) journals in the two decades between 2000 and 2019 that uses DiD or staggered DiD designs in its main analyses. We included those papers that, as of the end of 2019, were accepted for publication in one of these journals. Using Google Scholar's advanced keyword search, we identified the pool of potential papers as those published (or accepted for publication) in the ten journals during the 2000–2019 period in which the term "difference-in-differences" appears anywhere in the article. (We also considered variants without hyphens, which yields identical results. However, searching for abbreviations such as "DID" returned almost every published paper.) We read through each paper to verify which ones employed DiD or staggered DiD designs in their main analyses. This table summarizes the results of our manually collected data. Columns 1 and 2 report the total number of DiD and staggered DiD papers, respectively, published in each journal and for finance, accounting, and all ten journals during the 2000–2019 period. Column 3 reports the percentage of DiD papers that employ staggered DiD designs in each category.

overall and in each of the two fields) use a staggered DiD design, with 394 of the 407 (97%) published since 2010.

The prevalent use of staggered DiD reflects a common belief among researchers that such designs are more robust, and mitigate concerns that contemporaneous trends could confound the treatment effect of interest. However, recent advances in econometric theory (e.g., Borusyak and Jaravel, 2018; Athey and Imbens, 2018; Strezhnev, 2018; de Chaisemartin and D'Haultfœuille, 2020; Borusyak et al., 2021; Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021; Imai and Kim, 2021; Jakiela, 2021; Sun and Abraham, 2021) suggest that standard DiD regression estimates with staggered treatment timing often do not provide valid estimates of the causal estimands of interest to researchers—such as the average treatment effect on the treated (ATT)—even under random assignment of treatment.

This paper explains the intuition behind these theoretical problems, when and how they arise, and how they can lead to incorrect inferences. This paper also summarizes three solutions suggested by the econometrics or applied literature that empirical researchers in finance can apply for circumventing the problems. Importantly, we show that these theoretical problems are likely to matter in actual data and settings that researchers in finance, accounting, and law analyze.

We begin by providing an overview of the recent work in econometrics that explains why *static* treatment effect estimates from staggered DiD designs are not easily interpretable estimates for the ATT (Goodman-Bacon, 2021). In general, these estimates, obtained through two-way fixed effects (TWFE) DiD regressions, are variance-weighted averages of many different "2 × 2" DiDs, each involving the comparison between a treated and an effective control group in a window before and after the treated group receives treatment. In some of the 2 × 2s, already-treated units can act as effective comparison units, whose outcome changes may reflect treatment effects that are subtracted from the changes of later-treated units. Put differently, these regressions introduce a "bad comparisons" problem that differs from a violation of the parallel-trends assumption but is similarly problematic. When treatment effects can change over time ("dynamic treatment effects"), staggered DiD treatment effect estimates can actually obtain the opposite sign of the true ATT, *even if the researcher were able to randomize treatment assignment* (thus where the parallel-trends assumption holds). These theoretical results have far-reaching implications for applied researchers.

To demonstrate the situations under which these problems can arise, we simulate synthetic datasets from Compustat to mimic a standard staggered DiD design in applied corporate finance research: here exploiting staggered changes in state-level laws using a panel of firms whose returns on assets (ROAs) are measured over many years (e.g., Karpoff and Wittry, 2018). Our simulations produce three main insights. First, DiD estimates are unbiased in settings with a single treatment period, even when there are dynamic treatment effects. Second, DiD estimates are also unbiased in settings with staggered timing of treatment assignment and homogeneous treatment effect across firms and over time. Finally, when research settings combine staggered timing of treatment effects and treatment effect heterogeneity, staggered DiD estimates are likely biased. In particular, the combination of staggered treatment timing and dynamic treatment effects accentuates the presence and role of the "bad comparisons" problem in TWFE DiD static effect estimates, which can result in significant estimates with the wrong sign.

Moreover, the biases that arise with *static* staggered DiD estimates are not resolved by implementing event-study estimators. Researchers commonly estimate generalized

TWFE DiD regressions that allow for *dynamic* treatment effects. However, recent work suggests that dynamic effect estimates from such event-study estimators are also problematic. Sun and Abraham (2021) shows that, in the presence of staggered treatment timing and treatment effect heterogeneity, TWFE dynamic effect estimates for one relative-time period is contaminated by the causal effects of other relative-time periods in the estimation sample.

These biases are likely to apply in a large portion of research settings involving staggered treatment assignments and TWFE DiD regressions, because we believe that dynamic treatment effects are the most reasonable default assumption in many economic settings. We also demonstrate why these biases can result in both Type-I and Type-II errors. That is, researchers may conclude that treatment effects exist and that pre-treatment trends in treatment-control outcome differences are not present (consistent with the parallel-trends assumption) when the opposite is true. Researchers may also conclude that treatment effects do not exist, or pre-treatment trends are present, when the opposite is true. Remedying these biases is therefore critical for applied research.

Next, we summarize three alternative estimators developed in the econometrics or applied literature that researchers can apply in settings with staggered treatment timing (e.g., Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Gormley and Matsa, 2011). While the literature has not settled on a standard, the proposed solutions all deal with the biases arising from the "bad comparisons" problem inherent in TWFE DiD regressions by modifying the set of effective comparison units in the treatment effect estimation process. For example, each alternative estimator ensures that firms receiving treatment are not compared to those that previously received it. However, the methods differ in which observations are used as effective comparison units and how covariates are incorporated. We show that these alternative estimators help recover the actual treatment effects using our simulated data. Moreover, we explain the tradeoffs that researchers face when choosing among the three alternatives.

Finally, we demonstrate how these problems affect applied research by examining papers published in the top finance journals over the last decade. We replicate and extend the findings of two papers that apply staggered DiD designs in different settings: from bank deregulation (Beck et al., 2010) to global board governance reform (Fauver et al., 2017). In each paper, we find that the published staggered DiD estimates are susceptible to the biases from treatment effect heterogeneity. For example, treatment effect estimates from the alternative estimators often do not support the papers' original claims. In replicating these papers, we also demonstrate the impact of common specification choices in implementing staggered TWFE DiD regressions. For example, we show how binning relative-time periods in event-study specifications can influence dynamic treatment effect estimates, consistent with the analytical results of Sun and Abraham (2021).

Our paper contributes to the literature by highlighting an important methodological problem that we argue likely applies to a significant subset of applied research in finance and accounting. Contemporaneous papers in these fields have also highlighted the biases with TWFE staggered DiD treatment effect estimates (Barrios, 2021; Zdrojewski and Butler, 2021). We show how these issues raise concerns about spurious effects in empirical work and influence which types of papers empirical researchers pursue and journals publish. We suggest finance and accounting researchers should interpret standard TWFE staggered DiD regression estimates with caution, particularly in cases where treatment effect heterogeneity is the most likely and where the research setting contains relatively few never-treated units. We also suggest opportunities for re-examining critical prior research findings established based on staggered DiD designs or previously rejected research ideas (e.g., due to an absence of estimated treatment effects) relying on such designs. Finally, we offer empirical researchers guidelines for conducting DiD studies in settings with staggered treatment timing and suggestions for mitigating potential pitfalls.

2. A review of the DiD method

2.1. Basic 2×2 design and validity of DiD as causal estimate

The DiD design is one of the most commonly used methods for identifying causal effects in applied economics research. In its simplest form, DiD design involves a single treatment, two discrete periods (pre- and post-treatment), and two groups: units that receive ("treated") and do not receive ("control") treatment. In this "2×2" design, the treatment effect on the outcome of interest can be estimated empirically by comparing the change in the average outcome in the treated units to the change in the average outcome in the control units.

The potential outcomes framework (e.g., Rubin, 2005) formalizes why and when this empirical estimate is valid. Denote $Y_{i,t}(1)$ as the value of the outcome of interest for unit i at time t if the unit receives treatment, and $Y_{i,t}(0)$ as the outcome for unit i at time t if it does not receive treatment. The average treatment effect on the treated (ATT) is typically the causal estimand—the quantities to be estimated—of interest to researchers. It is defined as the difference $Y_{i,t}(1) - Y_{i,t}(0)$ averaged across the units receiving treatment.

The challenge in identifying the ATT stems from a fundamental missing data problem: for any given unit, we only observe one (not both) of the potential outcomes. DiD designs resolve this challenge by implicitly imputing the counterfactual outcomes of treatment units using outcomes for the control units. The validity of this approach rests on the central assumption that the observed *trend* in control units' outcomes mimic the trend in treatment units' outcomes had they not received treatment (i.e., the "parallel-trends" assumption). Letting ATT $= \delta$ and denoting D as an indicator variable evaluating to 1 when unit i is treated and 0 otherwise, we have

$$\begin{aligned}\delta &\equiv \mathbb{E}[Y_{i,1}(1) - Y_{i,1}(0)|D_i = 1] \\ &= \mathbb{E}[(Y_{i,1}(1) - Y_{i,0}(1))|D_i = 1] - \mathbb{E}[(Y_{i,1}(0) - Y_{i,0}(0))|D_i = 1] \\ &= \mathbb{E}[(Y_{i,1}(1) - Y_{i,0}(1))|D_i = 1] - \mathbb{E}[(Y_{i,1}(0) - Y_{i,0}(0)|D_i = 0)].\end{aligned}$$

The first equality defines the estimand of interest but cannot be directly estimated in the data. The second equality

follows from adding and subtracting $Y_{i,0}(0)$ and assuming no anticipation of treatment, so that $Y_{i,0}(0) = Y_{i,0}(1)$. The second equality, particularly the second term, also cannot be directly estimated in the data because $Y_{i,1}(0) - Y_{i,0}(0)$ is unobservable for a unit that receives treatment. The last equality follows from the parallel-trends assumption—$\mathbb{E}[(Y_{i,1}(0) - Y_{i,0}(0))|D_i = 1] = \mathbb{E}\left[\left(Y_{i,1}(0) - Y_{i,0}(0)|D_i = 0\right)\right]$—and can be estimated in the data.[1] To the extent control units' outcome trends do not capture the counterfactual outcome trends for treatment firms, the DiD estimate will be biased.

2.2. Use of regressions in implementing DiD

Researchers commonly obtain DiD estimates through ordinary linear regression (OLS). For example, the ATT from the simple 2×2 case can be obtained as the slope coefficient on the interaction term (β_3) from the following regression:

$$y_{it} = \alpha + \beta_1 D_i + \beta_2 POST_t + \beta_3 (\underbrace{D_i \times POST_t}_{D_{it}}) + \epsilon_{it}, \quad (1)$$

where D_i is an indicator variable for the treated unit, $POST_t$ is an indicator variable for observations in periods $t = 1$, and D_{it} denotes the interaction term.

An advantage of regression-based DiD is that it provides both the point estimate for δ and its standard errors. Another perceived advantage of the regression framework is that it can accommodate more generalized DiD settings because it is "easy to add additional states or periods to the regression setup ... [and] it's easy to add additional covariates" (Angrist and Pischke, 2009).

In settings with more than two units and two time periods, the regression DiD model usually takes the following two-way fixed effect (TWFE) form:

$$y_{it} = \alpha_i + \lambda_t + \delta^{DD} D_{it} + \epsilon_{it}, \quad (2)$$

where α_i and λ_t are unit and time period fixed effects, which subsume the main effects for D_i and $POST_t$. Researchers commonly modify this TWFE model to include covariates, time trends, and dynamic treatment effect estimation (e.g., by separately including indicators for the number of periods before or after the treatment).[2]

Notably, researchers apply the TWFE model to estimate δ in settings with staggered treatment timing. The perceived flexibility of regression DiD models likely contributed to their increasing popularity in applied research over the past two decades.

3. TWFE under staggered treatment timing: The problems

In a DiD with a single treatment period, a typical concern is that contemporaneous trends driven by factors other than the treatment of interest could confound the treatment effect—a violation of the parallel-trends assumption. Staggered DiD designs have been generally viewed as more credible and robust based on the intuition that including multiple treatments plausibly alleviates concerns that contemporaneous trends drive the observed treatment effects.

However, recent work in econometric theory casts doubt on the validity of the TWFE DiD estimator when it is applied to settings with variations in treatment timing. Significant biases may arise when such staggered DiD estimators are used for producing static or dynamic treatment effect estimates. This section summarizes the main issues and provides an intuition for when and why biases arise. We then demonstrate these problems by simulating data commonly encountered by finance researchers.

3.1. Static staggered DiD estimates

Goodman-Bacon (2021) shows that the "static" staggered DiD TWFE treatment effect estimate (δ^{DD} of Eq. (2)) is a "weighted average of all possible two-group/two-period DiD estimators in the data." For example, the TWFE estimate constitutes four possible 2×2s when there are three groups over the sample period (from the earliest period—t_0—to the last period in the data—T): a never-treated group (denoted U), an earlier-treated group (denoted k) that is treated at time t_k^*, and a later-treated group (denoted l) that is treated at t_l^*.[3]

The first two of the possible 2×2 DiD comparisons involve one treatment group (either the earlier- or the later-treated firms) and the untreated group (as control) over the whole sample window (from t_0 to T). The other two possible 2×2s involve comparisons between the different treatment groups. One of these "timing-only" 2×2s compares the earlier-treated firms to the later-treated firms (serving as controls) over the window from t_0 to t_l^* (i.e., in

[1] We note that two additional assumptions underlie the above justification of DiD as a valid estimate for ATT: the first is the assumption that all the expectations exist and are finite, and the second is the stable unit treatment value assumption (SUTVA). SUTVA, also known as the non-interference assumption, says that potential outcomes for a unit depends only on its treatment assignment (i.e., not the treatment assignment of another unit). It implies that only one of the potential outcomes is observed for every member of the population and there are no relevant interactions between members of the population: that is, the observed outcomes are fully specified $y_{i,t} = Y_{i,t}(1)D_i + Y_{i,t}(0)(1 - D_i)$. If SUTVA is violated, then we may observe neither of the potential outcomes, invalidating the DID estimate.

[2] Recent literature examines the assumptions under which the inclusion of time-varying covariates in TWFE DiD regressions lead to consistent estimates for the ATT. Sant'Anna and Zhao (2020) explains that, even when there is only one treatment period, TWFE DiD regression models with time-varying covariates produce consistent estimates for the ATT only under several (and plausibly more stringent) assumptions in addition to the traditional "parallel-trends" and "no-anticipation" assumptions. The requisite additional assumptions include treatment effect homogeneity (i.e., the ATT does not depend on the values of the covariates) and parallel trends in each of the included covariates between the treatment and control groups. For tractability, our paper generally abstracts away from the estimation issues arising from the inclusion of covariates. Nevertheless, this work motivates our replication analysis approach and our recommendation that researchers should produce a variant of TWFE estimates without time-varying covariates as a benchmark.

[3] The decomposition of Goodman-Bacon (2021) assumes a setting in which treatments are irreversible. Other papers, such as de Chaisemartin and D'Haultfœuille (2020), provide alternative decompositions under general conditions (e.g., when treatment can turn on and off).

which the earlier-treated units receive treatment the later-treated firms have not yet received treatment). The other "timing-only" 2×2 compares the later-treated firms to the earlier-treated firms over the window from t_k^* to T (i.e., the later-treated units receive treatment the earlier-treated firms have already received treatment). In this latter comparison, the earlier-treated units are used as controls against which the later-treated outcomes are compared.

We highlight three main results from this decomposition. First, the TWFE estimate of δ^{DD} is a variance-weighted average of the constituent 2×2 DiD estimates, with each 2×2 receiving positive weight. Second, in a significant subset of the constituent 2×2 DiD estimates, treated units can serve the role of effective comparison units, which may be problematic. Of particular concern are the "timing-only" 2×2s in which earlier-treated units act as effective controls (the "potentially problematic 2×2s") for later-treated units. Because changes in the earlier-treated units' outcomes may reflect changes in their treatment effects over time, the resultant DiD estimates could reflect differences in treatment effects over time between different treatment cohorts. (In a research design with G different treatment-timing groups and one untreated group, $G^2 - G$ of the G^2 total constituent 2×2 DiD estimates involve timing-only 2×2s, thus (1-1/G)/2 of the constituent DiDs are potentially problematic.) Third, the weight on each 2×2 estimate used to construct $\widehat{\delta^{DD}}$ is greater when, all else equal, the size of the subsample is larger, the treatment and effective comparison groups are similar in size, or the treatment variance is higher.

These results have important implications for the robustness of TWFE DiD estimates with staggered treatment timing. First, they may differ from the *sample-average* ATT because OLS applies variance weighting and implicitly applies positive weight to the potentially problematic 2×2s. The latter also implies that the TWFE estimate need not have the same sign as the average ATT. For example, even if the ATTs for all treatment cohorts are positive, it is possible to obtain a negative estimate $\widehat{\delta^{DD}}$. Second, the contribution of each constituent 2×2 DiD to the overall TWFE staggered DiD estimate is sample-dependent. For example, all else equal, constituent 2×2 DiD comparisons in which the treatment groups receive treatment closer to the middle of the comparison window receive greater weight because the treatment variance is larger. Changing the panel length alone can therefore change the weights applied to the constituent 2×2s and the TWFE staggered DiD estimate, even when each 2×2 DiD estimate is held constant. Finally, the issues posed by "potentially problematic 2×2s" are mitigated to the extent that units that never receive treatment account for a more significant portion of the sample.

Goodman-Bacon (2021) also examines what causal estimand the TWFE DiD identifies and under which conditions. Like Callaway and Sant'Anna (2021), this paper defines the ATT for a treatment-timing group g at a point in time as the "group-time average treatment effect":

$$ATT(g, \tau) \equiv \mathbb{E}[Y_{i,\tau}(1) - Y_{i,\tau}(0) | E_i = g], \tag{3}$$

where E_i denotes the time when unit i receives treatment and $E_i = g$ for all firms that receive treatment at time period g. $ATT(g, \tau)$ is simply the expected difference between the observed outcome variable for treated firms at time τ and the outcome had the firms not received treatment. This formulation allows for heterogeneity in ATT across treatment cohorts (g) or over time (τ).

Notably, Goodman-Bacon (2021) shows the probability limit of the TWFE DiD estimator consists of three components:

$$\operatorname*{plim}_{N \to \infty} \widehat{\delta^{DD}} = VWATT + VWCT - \Delta ATT. \tag{4}$$

VWATT is the "variance-weighted average treatment effect on the treated," a positively weighted average of the $ATT(g, \tau)$'s for the treatment groups and post-periods across all 2×2s that constitute $\widehat{\delta^{DD}}$. Absent any biases, $\widehat{\delta^{DD}}$ is consistent for this causal estimand. *VWCT* is the "variance-weighted common trend," which extends the parallel-trends assumption for a 2×2 DiD to a setting with treatment timing variation. VWCT is the weighted average of the difference in counterfactual trends (i.e., absent treatment) in the outcome between all pairs of groups and in the windows across all 2×2s that constitute $\widehat{\delta^{DD}}$. This term captures the possibility that different groups might not have the same underlying trend in the outcome without treatment, which will inherently bias any DiD estimate. On the other hand, $VWCT = 0$ if the parallel-trends assumption holds in each constituent 2×2 comparison.

Finally, the last term of Eq. (4) (ΔATT) is a weighted sum of the *change* in $ATT(g, \tau)$ within a treatment-timing group's post-period and around a later-treated unit's treatment window. This term arises because static TWFE DiD estimates implicitly use already-treated groups as effective comparison units for later-treated groups. It quantifies the extent to which, in such situations, the changes in earlier-treated units' outcome values are contaminated by changes in treatment effects over time (e.g., if the full treatment effect takes more than one period to be incorporated). To the extent this occurs, these outcome trends are inappropriate counterfactuals for the later-treated units.

Equation (4) suggests the staggered DiD TWFE estimate can differ from the *sample-average* ATT due to treatment effect heterogeneity either over time or across groups, even when the parallel-trends assumption is satisfied ($VWCT = 0$). When treatment effects are static (where the outcome is shifted by a constant after treatment) but vary across units, $\Delta ATT = 0$ and $\operatorname*{plim}_{N \to \infty} \widehat{\delta^{DD}} = VWATT$. In this case, *VWATT* may differ from the sample-average ATT when there is treatment effect heterogeneity because OLS applies weights on each cohort's *ATT* estimate that generally differ from the sample shares. As explained in Goodman-Bacon (2021), because TWFE uses OLS to combine the constituent 2×2 DiDs efficiently, the VWATT lies along the bias-variance tradeoff, and the weights deliver efficiency by potentially moving the point estimate away from the sample-average ATT. However, because there is no theoretically "correct" weighting, whether the TWFE estimate is desirable ultimately rests on the setting, the research question of interest, or the researcher's objectives.

Second, and more importantly, the staggered DiD TWFE estimate will differ from the sample-average ATT when the treatment effect is "dynamic." That is, instead of a

constant additive effect, the treatment effect is a function of time elapsed since treatment. Equation (4) suggests that time-varying treatment effects can create a bias in the static TWFE DiD estimate because $\Delta ATT \neq 0$. As emphasized earlier, this bias has arbitrary sign and magnitude, and the resultant treatment effect can be either too large, too small, or even have the wrong sign.

3.1.1. Simulations using compustat data

Having summarized the *theoretical* problems with TWFE DiD estimates in settings with staggered treatment timing, we now turn to analyze how these issues may arise in actual data that finance and accounting researchers commonly encounter via a simulation analysis. Similar to Bertrand et al. (2004), we perform a Monte Carlo where the data generating process stems from the empirical distribution of Compustat data, focusing on return on assets (*ROA*) as the outcome of interest. We introduce various treatment effects to firms in treated states, then examine the properties of the resultant TWFE DiD estimates.

We begin with a sample of all firms in Compustat over the 36-year period from 1980 to 2015 that are U.S. incorporated, non-financial, and contain at least five observations. Using this (unbalanced) panel of 176,670 observations, we compute *ROA* and decompose it into year- and firm-fixed effects and residuals:

$$\{\hat{\alpha}_i\}_{i=1}^{I}, \{\hat{\lambda}_t\}_{t=1}^{T}, \text{ and } \{\hat{\epsilon}_t\}_{it=1}^{N} \text{ from } ROA_{it} = \alpha_i + \lambda_t + \epsilon_{it}.$$

For each year, firm, and observation in the sample, we draw year-fixed effects, firm-fixed effects, and *ROA* residuals, respectively, from the empirical distribution. Similar to Bertrand et al. (2004), we also randomly draw states of incorporation for each firm, putting a 1/50 probability on each state.[4] Finally, we randomly assign states into treatment and control groups (in Simulation 1 and 2 below) or into different treatment-timing groups (in Simulations 3–6) with equal probability.

Next, we introduce six different treatment effects to the data generating process for *ROA*. The first three are either constant over time or based on a single treatment period:

$$\widetilde{ROA}_{it}^{1} = 0.5\sigma_{ROA} \times \mathbb{I}[Treat_i] \times \mathbb{I}[t \geq 1998] + \widetilde{\alpha}_i + \widetilde{\lambda}_t + \widetilde{\epsilon}_{it}, \tag{5}$$

$$\widetilde{ROA}_{it}^{2} = 0.05\sigma_{ROA} \times \mathbb{I}[Treat_i] \times \mathbb{I}[t \geq 1998] \times (t - 1997) + \widetilde{\alpha}_i + \widetilde{\lambda}_t + \widetilde{\epsilon}_{it}, \tag{6}$$

and

$$\widetilde{ROA}_{it}^{3} = 0.5\sigma_{ROA} \times (G_{1989} \times \mathbb{I}[t \geq 1989] + G_{1998} \times \mathbb{I}[t \geq 1998] + G_{2007} \times \mathbb{I}[t \geq 2007]) + \widetilde{\alpha}_i + \widetilde{\lambda}_t + \widetilde{\epsilon}_{it}, \tag{7}$$

where $\widetilde{\alpha}_i$, $\widetilde{\lambda}_t$, and $\widetilde{\epsilon}_{it}$ refer to the simulated fixed effects and residuals, σ_{ROA} refers to the sample standard deviation of *ROA* (30.9%), G_t denotes an indicator for units assigned treatment at time t, and $\widetilde{ROA}$ denotes the simulated *ROA*.

The first simulation assumes a single treatment period, in which a random half of the states initiate treatment at $t = 1998$, and a static treatment effect of half of σ_{ROA}. The second simulation also assumes a single treatment period. However, it differs from the first simulation in that the treatment effect is assumed to be dynamic over time, increasing by 5% of σ_{ROA} each year. Instead of a level shift in the outcome (i.e., Simulation 1), Simulation 2 introduces a trend-break in *ROA*. The third simulation allows for staggered timing of treatment assignment with static treatment effects. States are randomly assigned to one of three treatment groups based on the year in which the treatment initiates—1989 ($G_{1989} = 1$), 1998 ($G_{1998} = 1$), or 2007 ($G_{2007} = 1$)—and there are no never-treated units.

Our analysis of TWFE DiD estimates is based on 500 simulated Compustat samples of *ROA*s under each data generating process. Figure 1i depicts the differences in the data generating processes in Simulations 1–3 by plotting the outcome paths (gray lines) and the mean values of the outcome by treatment cohort (the colored lines). For each of the 500 simulated Compustat samples from each data generating process, we estimate a TWFE DiD regression (Eq. (2)) and plot the distribution of treatment effect estimates in Fig. 1ii. In all three simulations, the TWFE DiD estimate is unbiased for the sample-average ATT (vertical dashed line). These simulations suggest that TWFE DiD estimates are valid in settings with a single treatment period (even with dynamic treatment effects) or with no treatment effect heterogeneity across firms and over time (even with variation in treatment timing).

We note that, in an unbalanced sample, there are several ways to define the "sample-average" ATT. We compute a "firm-average" ATT, which first computes each treatment firm's (equal-weighted) average post-period ATT and then computes an (equal-weighted) average ATT across treatment firms. We also compute an "observation-average" ATT, which computes the (equal-weighted) average ATT across all the post-period treatment observations. This computation effectively places greater weight on those treatment firms with more post-period observations. Neither is conceptually more "correct," and they are identical under a balanced panel or static treatment effects. Simulation 2 shows that, under dynamic treatment effects and uniform treatment timing, TWFE DiD estimates are unbiased for the firm-average ATT.

Next, we illustrate the conditions under which TWFE DiD estimates are biased. We conduct three additional simulations (Simulation 4, 5, and 6), each of which follows the staggered treatment timing design of Simulation 3. However, unlike Simulation 3, Simulations 4–6 allow for different forms of treatment effect heterogeneity:

$$\widetilde{ROA}_{it}^{4} = (0.5\sigma_{ROA} \times G_{1989} \times \mathbb{I}[t \geq 1989] + 0.3\sigma_{ROA} \times G_{1998} \times \mathbb{I}[t \geq 1998] + 0.1\sigma_{ROA} \times G_{2007} \times \mathbb{I}[t \geq 2007]) + \widetilde{\alpha}_i + \widetilde{\lambda}_t + \widetilde{\epsilon}_{it}, \tag{8}$$

[4] We also conducted a variant of the simulation which states of incorporation is drawn from the empirical distribution. All of the conclusions from the simulation analysis are qualitatively unchanged. However, the simulated distributions of TWFE DiD estimates are more complex (i.e., multi-modal) due to 56% of the observations being incorporated in Delaware. Our simulations assign firms to states of incorporation with equal probability for parsimony.

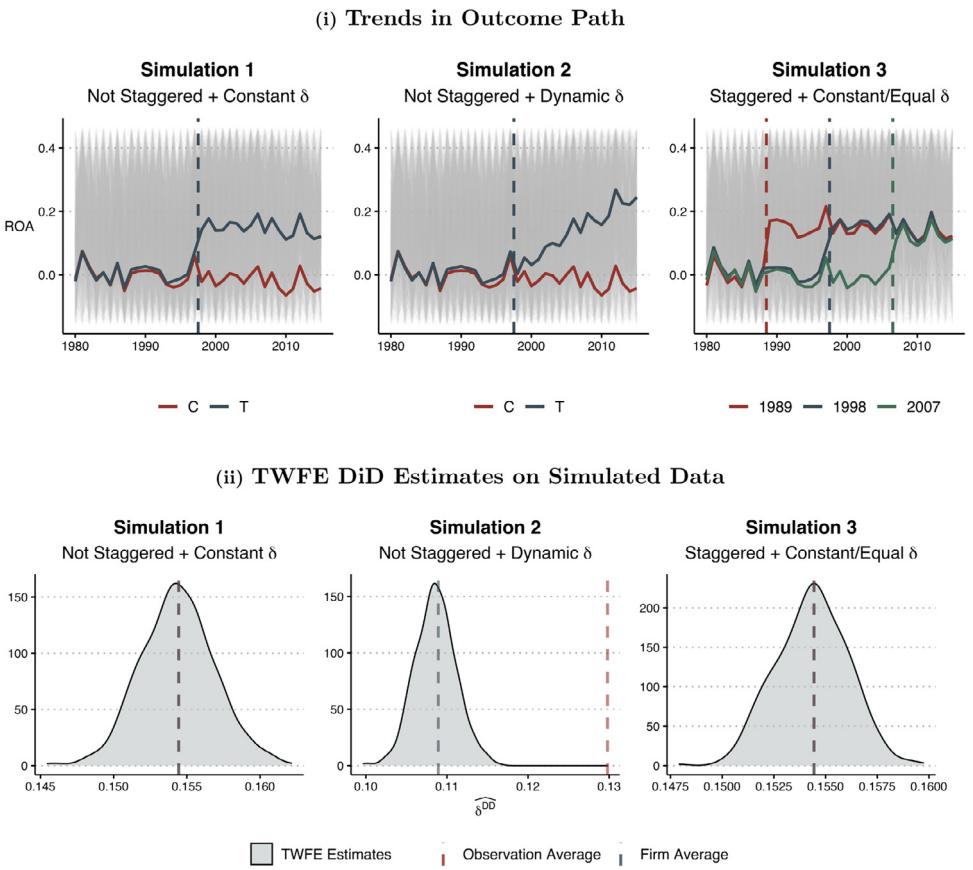


Fig. 1. Simulation: TWFE DiD Estimates Under Uniform Treatment Timing or Treatment Effect Homogeneneity.
Figure 1, panel (i), plots the firm-level outcome path (the gray lines) and the average outcome path by treatment groups (the bold lines) in one of the simulated Compustat datasets for Simulations 1, 2, and 3. To construct a simulated panel dataset, for each year, firm, and observation in the sample, we draw year-fixed effects, firm-fixed effects, and ROA residuals, respectively, from the empirical distribution. We then randomly draw states of incorporation for each firm and randomly assign states into treatment (T) and control groups (C) (i.e., in Simulations 1 and 2) or different treatment timing groups (i.e., in Simulation 3). Finally, we introduce treatment effects to the firms incorporated in treated states. Simulation 1 introduces a single treatment with a static effect (Eq. (5)). Simulation 2 introduces a single treatment with a dynamic effect (Eq. (6)). Simulation 3 introduces three treatments—to firms assigned to the 1989, 1998, or 2007 treatment-timing groups—each with static effects of the same magnitude (Eq. (7)). Panel (ii) plots the distribution of the static TWFE DiD treatment effect estimate ($\widehat{\delta^{DD}}$ from Eq. (2)) from 500 Monte Carlo simulations of our three different data generating processes. The curve represents the distribution of the TWFE estimates, while the dashed vertical lines represent the observation-level or firm-level average ATT.

$$\begin{aligned}\widetilde{ROA}_{it}^{5} = &(0.03\sigma_{ROA} \times G_{1989} \times \mathbb{I}[t \geq 1989] \\ &+ 0.03\sigma_{ROA} \times G_{1998} \times \mathbb{I}[t \geq 1998] \\ &+ 0.03\sigma_{ROA} \times G_{2007} \times \mathbb{I}[t \geq 2007]) \\ &\times [t - 1988G_{1989} - 1997G_{1998} - 2006G_{2007}] \\ &+ \widetilde{\alpha}_i + \widetilde{\lambda}_t + \widetilde{\epsilon}_{it}. \end{aligned} \tag{9}$$

$$\begin{aligned}\widetilde{ROA}_{it}^{6} = &(0.05\sigma_{ROA} \times G_{1989} \times \mathbb{I}[t \geq 1989] \\ &+ 0.03\sigma_{ROA} \times G_{1998} \times \mathbb{I}[t \geq 1998] \\ &+ 0.01\sigma_{ROA} \times G_{2007} \times \mathbb{I}[t \geq 2007]) \\ &\times [t - 1988G_{1989} - 1997G_{1998} - 2006G_{2007}] \\ &+ \widetilde{\alpha}_i + \widetilde{\lambda}_t + \widetilde{\epsilon}_{it}. \end{aligned} \tag{10}$$

Simulation 4 considers static ATTs, like Simulation 3, but allows them to differ across treatment-timing groups. Simulation 5 considers dynamic treatment effects, like Simulation 2, and assumes that the dynamic effects are the same across treatment-timing groups. Simulation 6 considers dynamic treatment effects and allows the trend-breaks to differ across treatment-timing groups.

Figure 2i shows the simulated outcome paths for Simulation 4–6. As before, for each of the 500 simulated Compustat samples from each data generating process, we estimate a TWFE DiD regression and plot the distribution of treatment effect estimates in Fig. 2ii. In Simulation 4, 5, and 6, TWFE estimates can differ substantially from the sample-average ATT (i.e., they are not centered around either of the vertical dashed lines). Note that Simulation 4 reflects the variance weighting that OLS applies to the constituent 2×2 ATTs, which generally differs from the weighting for a firm-level or observation-level average. Because there is no correct way to weight ATTs across cohorts, Simulation 4 does not suggest that the TWFE estimate is necessarily "wrong." Rather, it reflects a different way of aggregating the overall treatment effect. In our view, when researchers have different ATT estimates

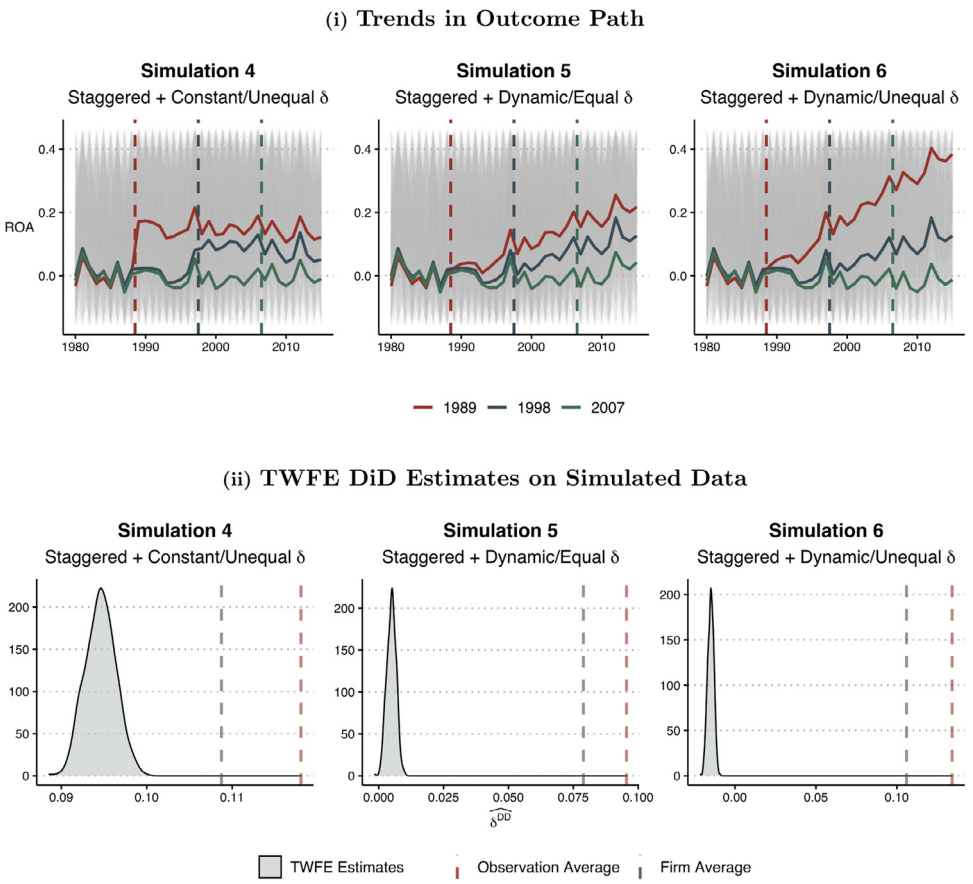


Fig. 2. Simulation: TWFE DiD Estimates Under Uniform Treatment Timing or Treatment Effect Homogeneneity.
Figure 2, panel (i), plots the firm-level outcome path (the gray lines) and the average outcome path by treatment groups (the bold lines) in one of the simulated Compustat datasets for Simulations 4, 5, and 6. To construct a simulated panel dataset, for each year, firm, and observation in the sample, we draw year-fixed effects, firm-fixed effects, and ROA residuals, respectively, from the empirical distribution. We then randomly draw states of incorporation for each firm and randomly assign states into different treatment timing groups: 1989, 1998, or 2007. Finally, we introduce treatment effects to the firms incorporated in treated states. Simulation 4 introduces static treatment effects, where the effect magnitudes differ across treatment-timing groups (Eq. (8)). Simulation 5 introduces dynamic treatment effects, where the dynamics are the same across treatment-timing groups (Eq. (9)). Simulation 6 introduces dynamic treatment effects, where the dynamics differ across treatment-timing groups (Eq. (10)). Panel (ii) plots the distribution of the static TWFE DiD treatment effect estimate ($\widehat{\delta^{DD}}$ from Eq. (2)) from 500 Monte Carlo simulations of the three different data generating processes. The curve represents the distribution of the TWFE estimates, while the dashed vertical lines represent the observation-level or firm-level average ATT.

across cohorts, the ideal weighting across them may depend on the setting, the research questions of interest, or the researcher's objectives.

In contrast, Simulations 5 and 6 generate biased estimates for the sample-average ATT due to past treated units serving as effective comparison units under dynamic treatment effects ($\Delta ATT > 0$). Put differently, these two scenarios differ from Simulation 4 in the sense that, by applying positive weight to the potentially problematic 2×2s, TWFE DiD estimates are clearly wrong.

These simulations show that the combination of staggered treatment timing and treatment effect heterogeneity, either across groups or over time, leads to biased TWFE DiD estimates for the sample-average ATT. This bias can be so severe as to change the researcher's inferences about the direction of the treatment effect. For example, although Simulations 5 leads to biased TWFE DiD estimates of the average ATT, it preserves the correct treatment effect *sign* on average. In contrast, Simulation 6 leads to an average estimated treatment effect that is *negative* and statistically significant, even though the ATT for every treated group is positive.

3.1.2. Intuition via Goodman-Bacon (2021) Diagnostic

To further understand these biases, such as why Simulation 6 produces treatment effects of the wrong sign, we apply a diagnostic test to analyze TWFE estimates' robustness. Specifically, Goodman-Bacon (2021) applies (in Fig. 6 of the paper) its decomposition to analyze the contribution of the constituent 2×2's by plotting the constituent DiD estimates against their implicit weights in the TWFE estimate. Similarly, researchers can analyze the total weights and the weighted-average DiD estimate for each *type* of constituent 2×2s: those involving comparisons of treatment-timing groups vs. never treated groups, those involving comparisons of earlier- vs. later-treated groups (as effective controls), and those involving later- vs. earlier-

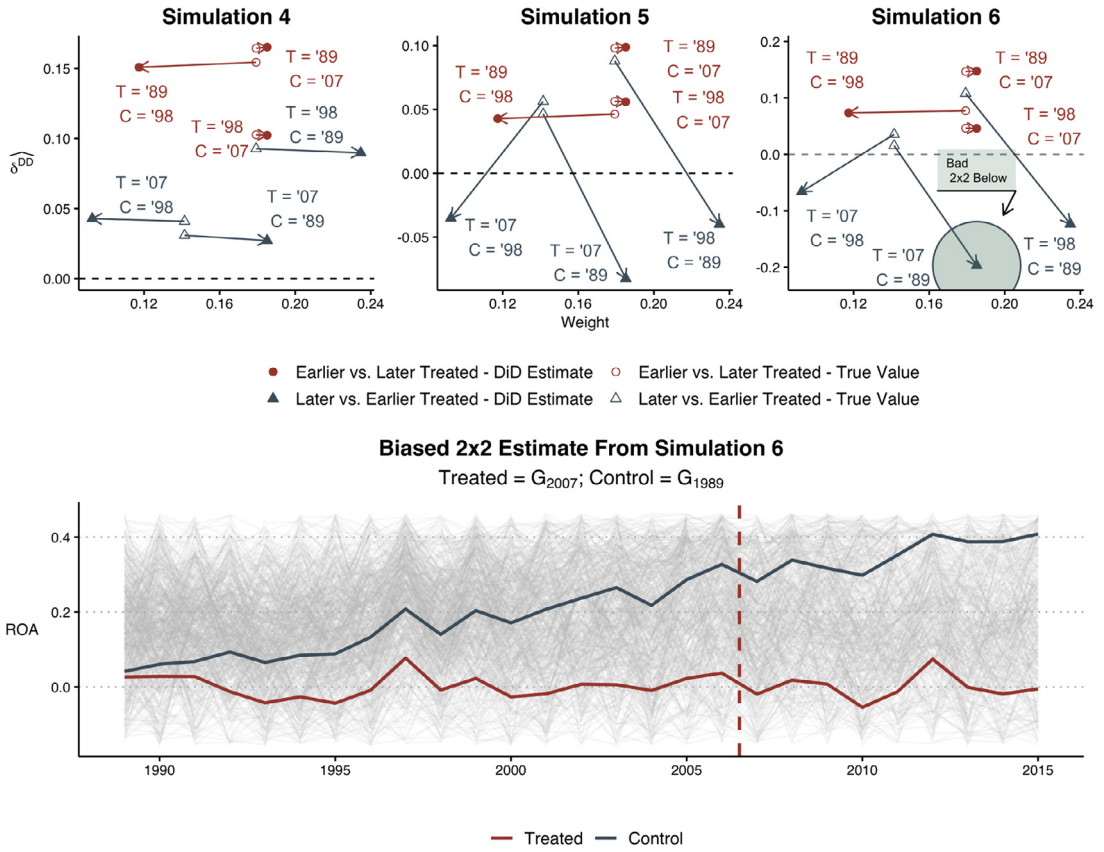


Fig. 3. Simulation: Diagnostics.
Figure 3, upper panel, plots the implicit weight given to each constituent 2×2 in the static TWFE DiD estimate and the effect estimate for that 2×2 in one of the simulated Compustat datasets for Simulation 4, 5, and 6. The solid red circles represent the empirical estimates and TWFE weights for 2×2s using later-treated firms as effective comparisons. The blue triangles represent the empirical estimates and TWFE weights for 2×2s using prior-treated firms as effective comparisons. The hollow circle or triangle represents the firm-average ATT for the treated firms in the corresponding 2×2s. The bottom panel depicts one constituent ("potentially problematic") 2×2 comparison—a comparison of firms treated in 2007 to firms treated in 1989 in Simulation 6. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

treated groups (as effective controls).[5] Of particular concern are situations where later- vs. earlier-treated 2×2s have DiD estimates of a different sign or when they carry substantial total weight in the static TWFE DiD estimate.

The top panel of Fig. 3 illustrates the diagnostic test for Simulations 4, 5, and 6. Because the diagnostic test only applies to balanced panels, in constructing this figure our simulation is modified to artificially induce a balanced panel of firm-year observations from Compustat before drawing fixed effects and residuals from the empirical distribution.

For each of the six constituent 2×2 comparisons, we plot the 2×2 DiD estimate and its overall weight on $\widehat{\delta^{DD}}$. (For the specific formula for computing these weights, see Eqs. (10e), (10f), and (10g) of Goodman-Bacon (2021).) We distinguish the three types of 2×2 comparisons by the marker symbol. Circle markers represent the constituent groups where earlier-treated firms are compared (as treatment) to the later-treated (as effective controls), and triangle markers represent the constituent groups where later-treated firms are compared to the earlier-treated (as effective controls). We also compare each of these constituent DiD estimates to the firm-level average ATT for the treated firms in each 2×2, represented by empty symbol markers and connected to the relevant constituent 2×2 DiD estimate by an arrow. Because the firm-level average ATT is the same across the 2×2s that share the same effective treatment group, we add small perturbations to avoid overlapping marker symbols in Fig. 3 to facilitate the graphical depiction. In addition, because we compute the firm-level average ATT in the sample, 2×2s that share the same effective treatment group also share the same weights for the ATT.

The figure shows that, with heterogeneous static treatment effects under staggered treatment timing (Simulation 4), the constituent DiDs are unbiased for the ATT within each 2×2. However, OLS applies different weights, resulting in an overall weighted average treatment effect that differs from the firm-average ATT. With dynamic treatment effects and staggered treatment timing (Simulations 5 and 6), all the later- vs. earlier-treated comparisons

[5] A Stata package (bacondecomp) written by the Andrew Goodman-Bacon, Thomas Goldring, and Austin Nichols performs the diagnostics discussed in this subsection. An R package (bacondecomp) written by Evan Flack is also available.

yield *negative* estimated treatment effects (i.e., all the blue triangular points lie below zero) that are biased for the ATTs. In contrast, all the earlier- vs. later-treated 2×2s yield *positive* DiD estimates that are unbiased for the ATTs.

The bottom panel of Fig. 3 provides graphical intuition for why constituent 2×2s can produce negative effects despite all ATTs being positive. In particular, we examine a particular 2×2 in Simulation 6 that compares firms treated in 2007 (as treated) to firms treated in 1989 (as controls) in the 1989 to 2015 subsample. This 2×2 illustrates the idea of the ΔATT bias: it yields a negative DiD because the large changes in the outcome for earlier-treated firms, the effective controls, are *subtracted* from the relatively smaller changes in the outcome for later-treated firms, the effective treatment firms in this subsample. Clearly, this comparison is invalid because the control firms' outcome changes are contaminated by changes in treatment effects over time. This example also highlights a critical insight of the Goodman-Bacon (2021) decomposition: under dynamic treatment effects, biases from bad controls can arise even when the parallel-trends assumption holds, as is the case here (i.e., the 2007 and 1989 cohorts have the same expected counterfactual outcomes had they not received treatment).

Finally, the diagnostic test also shows that TWFE down-weights some of the earlier- vs. later-treated comparisons and up-weights some of the later- vs. earlier-treated comparisons, thereby increasing the influence of the potentially problematic 2×2s. Thus, a combination of negative effects estimated from and the significant weights applied to potentially problematic 2×2s results in a TWFE DiD estimate that can significantly deviate from the sample-average ATT.

We note that the decomposition and diagnostic offered by Goodman-Bacon (2021) can at present only be used with balanced panels and do not incorporate covariates. These are atypical features of corporate finance or accounting applications. Nevertheless, we recommend that researchers should always analyze covariate-free variants of DiD analyses as starting points. To the extent possible, we believe this diagnostic test should be applied to analyze the potential biases in TWFE DiD estimators in settings with staggered treatment timing.

3.1.3. Type-I and Type-II errors

Our simulation analyses suggest that the biases in staggered TWFE DiD estimators may result in Type-II errors. For example, in Simulation 5 of Fig. 2i, the TWFE DiD estimates a very small effect that is close to 0, even though the true ATTs for each treatment cohort is positive and large in magnitude. A researcher, expecting editors and referees to be more likely to publish statistically significant results (Andrews and Kasy, 2019; Kim and Ji, 2015), may very well reject such "good" projects, where economically and statistically significant effects exist, or the findings could be important for informing policy.

TWFE biases may also result in Type-I errors, where true treatment effects are zero on average, but the estimated effects are not. This is because ΔATT bias in Eq. (4) can be non-zero even when $VWATT$ is zero. For example, the treatment effect could be heterogeneous across cohorts but is on-average zero. However, differences in dynamic effects across groups can lead to large constituent 2×2 DiD estimates (i.e., the potentially problematic 2×2s) that TWFE up-weights, leading to a significant aggregate TWFE DiD estimate.

To illustrate this idea, we make the following modification to Simulation 6:

$$\widetilde{ROA}_{it}^{6'} = .03\sigma_{ROA}\Phi \times \mathbb{I}[t \geq g] \times [t-g] + \widetilde{\alpha}_i + \widetilde{\lambda}_t + \widetilde{\epsilon}_{it} \quad (11)$$

for an observation i assigned to treatment group $g \in \{1989, 1998, 2007\}$, where $.03\sigma_{ROA}\Phi$ is a normal distribution centered at zero with a standard deviation of $0.03 \cdot \sigma_{ROA}$. Instead of assigning pre-determined trend-breaks to each of the three treatment groups, as in Simulation 6, this modified simulation now draws the trend-breaks from a distribution centered around zero. Thus, we allow for heterogeneity in dynamic treatment effects across firms, but where the ATT is zero in expectation.

As before, we run this simulation 500 times. We estimate the TWFE DiD regression for each simulated Compustat panel and compute the t-Statistic, using standard errors clustered at the state level. Figure 4i plots the distribution of t-Statistics across the simulations and shows that the TWFE regression produces significant treatment effect estimates at the 5% level (or t-Statistics larger than 1.96 in absolute value) in 79% of the cases. In untabulated results, we find very similar results when restricting the average ATT within each simulated panel, at the observation- or the firm-level, to be exactly zero in the sample: we continue to find that TWFE estimates are significant at the 5% level in about 80% of the simulated samples. At this level of treatment effect heterogeneity, the biases associated with TWFE staggered DiD regressions lead to a large degree of over-rejection (excess Type-I errors).

We also analyze how much treatment effect heterogeneity is required to create spurious inferences in these regressions. We repeat the above exercise for different levels of treatment effect heterogeneity (different percentages of σ_{ROA}), from zero to 10 percent of the empirical ROA distribution. At each level of heterogeneity, we run the simulation 500 times as above and compute the percent of the simulations that yielded a t-Statistics larger than 1.96 in absolute value.

The results, plotted in Fig. 4ii, show that just a little bit of treatment effect heterogeneity can have a significant impact on the degree of over-rejection. When there is no heterogeneity, 95% of the simulations (represented by the horizontal dashed line) produce insignificant t-Statistics (a 5% Type-I error rate), as expected. However, Type-I error rates increase quickly as we introduce a small degree of treatment effect variation. For example, when the standard deviation of the trend-break is one percent of σ_{ROA}, more than half of the simulations produced significant t-Statistics, even though the average ATT is zero. As we further increase treatment effect heterogeneity, the percent of simulations that produce insignificant t-Statistics stabilize to around 20% (Type-I error rates stabilize to 80%). In summary, these simulations show that the TWFE DiD estimator's biases could easily result in spurious inferences.

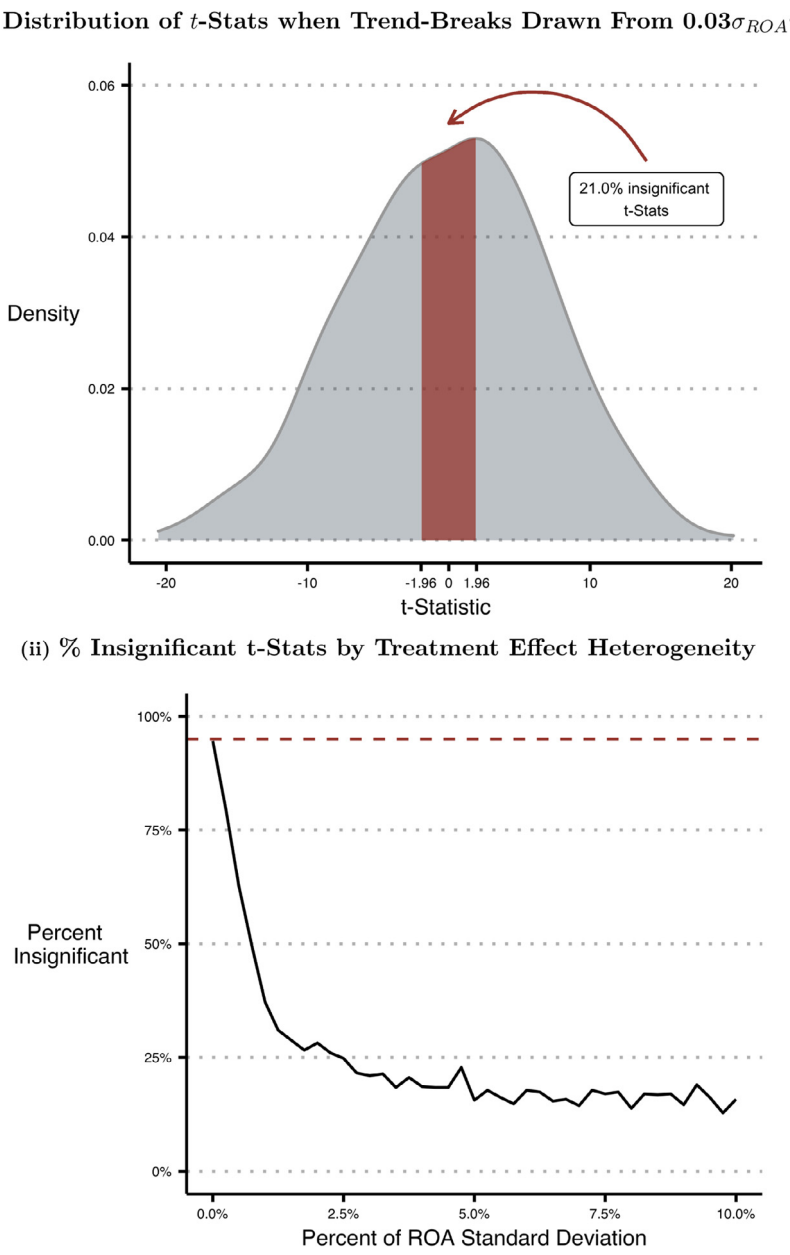


Fig. 4. Simulation: TWFE DiD Estimates When Expected ATT = 0.
Figure 4, panel (i), plots the distribution of *t*-Statistics across 500 iterations of Simulation 6' (Eq. (11)). All aspects of this simulation are the same as Simulation 6, except that trend breaks are drawn from a normal distribution with mean zero and standard deviation of 3% of the empirical *ROA* standard deviation in Compustat. The red shaded region in the distribution represents the insignificant (at the 5% level) *t*-Statistics from the 500 simulations. Panel (ii) plots the percent insignificant *t*-Statistics as a function of the treatment effect heterogeneity (i.e., the variation in the trend-break distribution in terms of the percent of ROA standard deviation). At each level of treatment effect heterogeneity, we repeat the exercise illustrated in panel (i) and record the percent of insignificant *t*-Statistics across 500 simulations. The horizontal dashed line represents 95%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Dynamic staggered DiD estimates

We have thus far focused on the biases associated with static TWFE DiD estimators, in which there is a single aggregate treatment effect parameter of interest. However, the combination of treatment effect heterogeneity and staggered treatment timing also biases dynamic TWFE DiD specifications (or "event study" specifications).

Researchers often estimate dynamic treatment effects using a generalized variant of Eq. (2):

$$y_{it} = \alpha_i + \lambda_t + \sum_{l=-K}^{-2} \mu_l D_{it}^l + \sum_{l=0}^{L} \mu_l D_{it}^l + \epsilon_{it}, \tag{12}$$

where $D_{it}^l = \mathbb{I}[t - E_i = k]$ is an indicator for a treatment unit i in cohort E_i (the period of treatment) being k

periods from the start of treatment. Instead of using a single binary treatment indicator (D_{it} in Eq. (2)), the event-study specification utilizes a set of relative-time indicators: the first summation in Eq. (12) captures the time periods leading up to the treatment ("leads") and the second summation captures the time periods following treatment ("lags").[6] Equation (12) follows the standard practice of excluding the relative-time indicator for the period before treatment to avoid multicollinearity; in settings with no never-treated units, *two* excluded (pre-treatment) relative-time indicators are necessary (Borusyak et al., 2021). Thus, the main parameters of interest, the μ_l's, are interpreted as the difference between the outcome differences between treated and untreated observations l periods from treatment relative to the outcome differences between treated and untreated observations in the excluded periods.

Researchers implement event-study designs to analyze post-treatment effect dynamics and pre-treatment trends in outcome differences between treated and control units. The absence of observable "pre-trends" is often provided as evidence consistent with the parallel-trends assumption, which is not directly testable.

Sun and Abraham (2021) ("SA") shows that TWFE dynamic treatment estimates from Eq. (12) are also biased when there is variation in treatment timing and treatment effect heterogeneity. Like Callaway and Sant'Anna (2021) ("CS"), SA defines the estimand of interest as the ATT for a particular treatment cohort at a particular time:

$$CATT_{g,l} = \mathbb{E}[Y_{i,g+l} - Y^{\infty}_{i,g+l} | E_i = g]. \tag{13}$$

CATT, or the "cohort-specific ATT," is defined in relative treatment time terms and compares a treatment unit's potential outcome at a point in time if it received treatment in time period g ($Y_{i,g+l}$) to the counterfactual outcome if the unit *never* receives treatment ($Y^{\infty}_{i,g+l}$).

One of SA's main results (see Proposition 3 in the paper) shows that, even when the parallel-trends and no-anticipation assumptions hold, the population μ_b is a linear and non-convex combination of post-treatment CATTs from both its own relative period l and other relative periods.

$$\begin{aligned}\mu_b = &\sum_{l' \in b, l' \geq 0} \sum_{g} w^b_{g,l'} CATT_{g,l'} \\ &+ \sum_{b' \neq b, b' \in \mathcal{B}} \sum_{l' \in b', l' \geq 0} \sum_{g} w^b_{g,l'} CATT_{g,l'} \\ &+ \sum_{l' \in b^{excl}, l' \geq 0} \sum_{g} w^b_{g,l'} CATT_{g,l'}\end{aligned} \tag{14}$$

The first term of this decomposition is what researchers would like to identify because it represents the weighted average of the CATTs across treatment cohorts in post-treatment periods ($l' \geq 0$) within the relative-time bin b of interest.[7] The latter two terms represent linear combinations of post-treatment CATTs across cohorts in other relative-time periods that are included in (the second term) or omitted (the third term) from the dynamic specification but belong to the sample. The bias stemming from the last two terms shows that the TWFE dynamic effect estimate for one relative-time period is contaminated by causal effects of other periods. The results of SA are thus an extension to Goodman-Bacon (2021) to dynamic effect estimates: the biases associated with TWFE DiD regressions under treatment effect potentially invalidates *every* event-study coefficient.

In addition, SA shows that even with treatment effect homogeneity, which resolves the biases in static TWFE DiD estimates, dynamic treatment effect estimates can remain contaminated by CATTs from excluded periods (e.g., the last term of Eq. (14)). In such a case, a combination of treatment effect homogeneity and ensuring that only pre-treatment periods (or generally those periods where CATT=0) are excluded prevents the contamination. (Under no anticipation, the pre-period CATTs are zero.) Thus, the choice of excluded relative-time periods can lead to biases in TWFE dynamic specifications.

SA's analyses also show that the common practice of trimming (i.e., dropping from the sample) or binning (i.e., grouping) distant relative-time indicators does not resolve the contamination problems.[8] In fact, effect estimates from event-study specifications that bin relative-time periods continue to be contaminated by CATTs from periods in other relative-time bins, even under all of the assumptions above (i.e., homogeneity and CATT=0 for excluded relative-time periods). In these cases, a sufficient condition to avoid contamination is to group relative-time periods into bins only when their treatment effects are the same. Thus, the choice of relative-time bins *per se* can also lead to biases in TWFE dynamic specifications.

Finally, a key implication of SA's results is that the common practice of testing pre-trends using the coefficients on the leads (e.g., the in the first summation term of Eq. (12)) is not generally valid. The reason is that a given pre-treatment coefficient does not identify the relevant pre-period CATT but is contaminated by CATTs from *all* relative-time periods and across treatment cohorts. This contamination can result in significant pre-period estimates when pre-trends in CATTs do not exist or insignificant pre-period estimates when pre-trends exist.

[6] The standard static specification in Eq. (2) can be expressed in terms of the post-treatment D^l_{it}s:

$$y_{it} = \alpha_i + \lambda_t + \delta^{DD}\left(\sum_{l \geq 0} D^l_{it}\right) + \epsilon_{it}.$$

[7] The decomposition of is done for a more general version of Eq. (12) that allows for the grouping (or "binning") and the exclusion of relative-time periods:

$$y_{it} = \alpha_i + \lambda_t + \sum_{b \in \mathcal{B}} \mathbb{I}\{t - E_i \in b\} + \epsilon_{it},$$

where E_i is the treatment period for unit i, the set $\mathcal{B}$ collects disjoint sets b of relative periods, and some relative periods could be omitted from the specification. The excluded set is denoted $b^{excl} = \{l : l \notin \bigcup_{b \in \mathcal{B}} b\}$. SA shows that the weights on the first, second, and third terms of Eq. (14) sum to 1, 0, and -1, respectively.

[8] One difference noted in SA is that, for a given coefficient in the dynamic specification, trimming mechanically removes the contamination stemming from CATTs in relative-time periods trimmed from the specification.

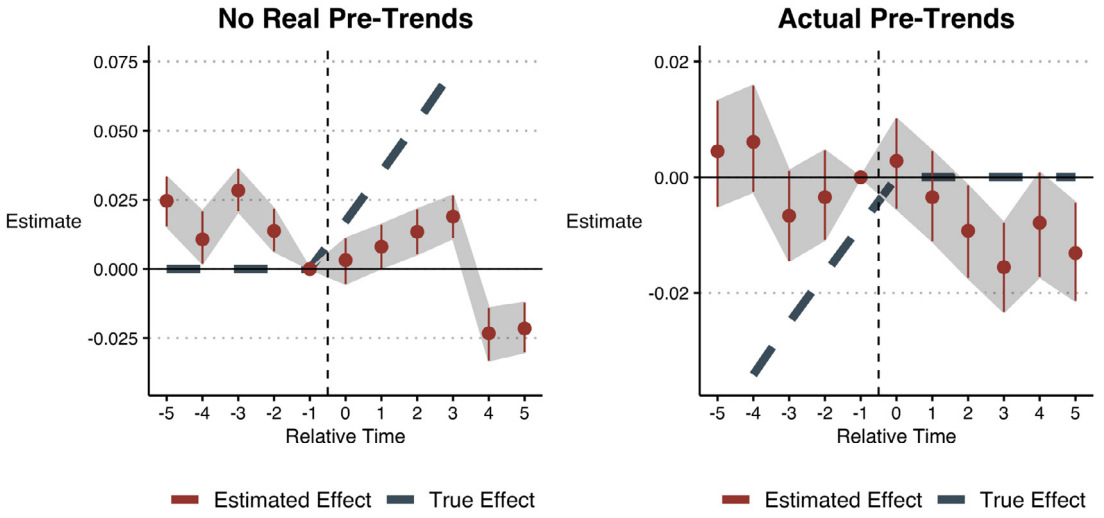


Fig. 5. Simulation: TWFE Dynamic Treatment Effect Estimates with No Real Pre-Trends and Actual Pre-Trends.
Figure 5, left-hand panel, plots the distribution of event-study estimates based on a variant of Simulation 6 (Eq. (10)) in which there are no pre-period trends and post-treatment trend-breaks of the three different cohorts are $\delta_{1989} = 0.10\sigma_{ROA}$, $\delta_{1998} = 0.05\sigma_{ROA}$, and $\delta_{1998} = 0.01\sigma_{ROA}$, where σ_{ROA} is the empirical *ROA* standard deviation in Compustat. For each of the 500 simulated Compustat panel datasets, we estimate a TWFE event-study specification (Eq. (12)) that includes relative-time indicators for the five years before and after the year of treatment (Relative Time = 0). We exclude the relative-time indicator for the year prior to treatment (Relative Time = -1). Moreover, we combine relative-time periods more than five years before treatment into one bin and relative-time periods more than five years after treatment into another bin. For each relative-time period from -5 to 5, we plot the point estimate (the solid circle), the 95% confidence interval (the vertical lines intersecting the solid circles), and the observation-average ("true") ATT for each relative-time period (the dashed line). The right-hand panel plots the distribution of event-study estimates based on Simulation 7 (Eq. (15)), which has pre-treatment trends and but no treatment effects. Aside from the data generating function, all other aspects of this simulation are the same as the left-hand-side panel.

(For a helpful simplified example, we refer readers to Section 3.7 of their paper.)

3.2.1. Simulation analysis

To illustrate that TWFE event-study estimates lead to misleading inferences, consider a variant of Simulation 6 (Eq. (10)) in which the trend-breaks of the three different cohorts are $\delta_{1989} = 0.10\sigma_{ROA}$, $\delta_{1998} = 0.05\sigma_{ROA}$, and $\delta_{1998} = 0.01\sigma_{ROA}$. Like the previous simulations: each treatment cohort has a positive average treatment effect, the parallel-trends assumption holds, and treatment-control ROA differences are zero in expectation in each pre-treatment period (i.e., no pre-trends).

As before, we generate 500 simulated Compustat samples of *ROA*s. With each sample, we estimate a TWFE event-study specification (Eq. (12)) that includes relative-time indicators for the five years before and after the year of treatment (Relative Time = 0). To avoid collinearity, we exclude the relative-time indicator for the year prior to treatment (Relative Time = -1). Following standard practice in the literature, we bin relative-time periods further out in the event study window (i.e., more than five years before or after treatment).

Figure 5, left-hand panel, plots the distribution of estimated coefficients on the relative-time indicators. We also plot the observation-level average ATT for comparison. This figure confirms the theoretical result established in SA: in the presence of heterogeneous treatment effects, TWFE event-study estimates are biased. The post-period effect estimates are negatively biased relative to the ATT. The effect estimates for years four and five after treatment are both negative and statistically significant even though the actual effects are positive in both cases.

Notably, the pre-treatment event-study estimates are also biased. Despite no real pre-trends in the data generating process, the TWFE dynamic specification produced positive and statistically significant coefficients on pre-treatment relative-time indicators. A researcher may infer from the observed pre-trends that the parallel-trends assumption is violated, and that any post-treatment effect estimates are likely spurious, despite economically significant true effects and the parallel-trends and no-anticipation assumptions being valid.

Biases associated with TWFE event-study estimates could also lead researchers to infer a lack of pre-trends when the parallel-trends assumption does not hold. Consider the following data generating process:

$$\begin{aligned}\widetilde{ROA}_{it}^{7} = {} & (0.01\sigma_{ROA} \times G_{1989} \times \mathbb{I}[t < 1989] \\ & + 0.035\sigma_{ROA} \times G_{1998} \times \mathbb{I}[t < 1998] \\ & + 0.035\sigma_{ROA} \times G_{2007} \times \mathbb{I}[t < 2007]) \\ & \times [t - 1988G_{1989} + 1997G_{1998} + 2006G_{2007}] + \widetilde{\alpha}_i \\ & + \widetilde{\lambda}_t + \widetilde{\epsilon}_{it}. \end{aligned} \quad (15)$$

Each cohort has a treatment effect of zero in expectation; however, cohorts differ in terms of the pre-treatment trends in the outcome. Unlike Simulation 6, where group-specific trend-breaks in *ROA* apply after treatment, in this simulation, group-specific trend-breaks apply *prior* to treatment. Thus, this data generating process violates the parallel-trends assumption.

However, estimating the TWFE dynamic specification described above produces an intriguing and comparatively

persuasive event-study plot, shown in the right-hand panel of Fig. 5. All pre-period coefficients are statistically indistinguishable from zero, consistent with the parallel-trends assumption; moreover, the post-period coefficients suggest a negative treatment effect over a longer horizon (e.g., three to five years after treatment). However, this spurious event-study plot is driven by the biases associated with TWFE event-study estimates and highlighted in SA.

Together, the biases associated with static and dynamic TWFE DiD estimates can lead to both Type-I and Type-II errors. They may also influence researchers' choice of projects. Remedying these biases, we believe, is paramount for applied research.

4. Alternative estimators

While the econometric literature has settled on the theoretical problems with TWFE staggered DiD estimators, it has proposed several alternative DiD estimation techniques to circumvent them. We highlight three estimators applied researchers should consider, either formally developed in the econometrics literature or adopted as a remedy in the applied literature. In essence, each estimator modifies the units that can act as effective comparison units to avoid comparing treatment units to inappropriate controls. However, the remedies differ in terms of which observations may serve as effective control units and their complexity and flexibility.

4.1. Callaway and Sant'Anna (2021) and Sun and Abraham (2021)

The first two estimators, developed by CS and SA, are closely related. Each relies on first estimating the individual cohort-time-specific treatment effects (e.g., Eqs. (3) or (13)), allowing for treatment effect heterogeneity, then aggregating them to produce measures of overall treatment effects. However, CS and SA differ methodologically regarding flexibility, accommodation of covariates, choice of control groups, and inference.

The simplest variant of the CS estimator boils down to estimating cohort-time-specific treatment effects through simple 2×2s with clean controls. For example, the treatment effect of a particular treatment group (i.e., treated at time g) can be estimated via the following regression

$$y_{it} = \alpha_1^{g,\tau} + \alpha_2^{g,\tau} \cdot \mathbb{I}\{E_i = g\} + \alpha_3^{g,\tau} \cdot \mathbb{I}\{t = \tau\} + \beta^{g,\tau} \cdot (\mathbb{I}\{E_i = g\} \times \mathbb{I}\{t = \tau\}), \quad (16)$$

using observations at time τ and $g-1$ from treated units i with $\mathbb{I}\{E_i = g\} = 1$, or from a set of clean control units.[9] CS allows for not-yet-treated, last-treated, or never-treated as clean controls, and shows that $\beta^{g,\tau}$ is a valid estimator for $ATT(g,\tau)$ under no anticipation and unconditional parallel trends. CS also derives estimators that are consistent for $ATT(g,\tau)$ under more general conditions, such as when the parallel-trends assumption holds conditionally on covariates, including an outcome-regression-based estimator (Heckman et al., 1997), an inverse-probability-weighted estimator (Abadie, 2005), and a doubly robust estimator (Sant'Anna and Zhao, 2020).[10]

SA proposes a fully parametric regression-based estimator that estimates the full set of cohort-specific relative-time treatment effects (i.e., each $CATT_{g,l}$ in the sample) jointly using an interacted specification that is saturated in relative-time indicators D_{it}^k and cohort indicators $\mathbb{I}\{E_i = g\}$:

$$y_{it} = \alpha_i + \lambda_t + \sum_{g \notin C} \sum_{l \neq -1} \mu_{g,l} (\mathbb{I}\{E_i = g\} \cdot D_{it}^l) + \epsilon_{it}. \quad (17)$$

SA shows that, by including the full set of cohort-specific relative-time indicators, $\mu_{g,l}$, are consistent for the CATTs under unconditional parallel trends and no anticipation. We note that in implementing Eq. (17), always-treated firms are dropped, and the only units that can be used as effective controls are those that are never-treated or last-treated.[11] (When the last-treated are used as controls, they are never used as treated units.)

There are two main differences between CS and SA's methodologies for estimating group-time *ATTs*. First, CS allows for greater flexibility in selecting control groups: whereas SA allows only for never-treated or last-treated comparison units, CS additionally accommodates not-yet-treated units as controls. Second, CS allows for (pre-treatment and static) covariates (i.e., when conditional parallel-trends assumptions are more appropriate), while SA does not. When there are no covariates, and never-treated firms are used as effective controls, CS and SA provide numerically equivalent estimates.

SA and CS also differ in terms of inference, which is not the focus of our paper. SA uses pointwise inference of average ATTs, whereas CS develops and argues for simultaneous confidence intervals, which can be estimated with a simple multiplier bootstrap procedure. SA directly estimates the asymptotic standard errors of its interaction-weighted estimator and does not use bootstrapping.

Finally, both CS and SA provide solutions for aggregating the group-time ATTs. SA's interaction-weighted three-step estimator focuses on event study type aggregation: the average CATT for a particular relative-time period τ uses the weighted average of the $CATT(g,\tau)$ over treatment cohorts using the sample shares of each cohort in the relevant periods. CS considers a variety of possible aggregations of group-time *ATTs*. It is, of course, possible to apply a weighting scheme like SA to create event-study plots. However, for researchers interested in a single overall effect estimate, CS recommends first computing the average *ATTs* for each treatment cohort (across all post-treatment periods) then reporting the weighted average ATTs across cohorts (e.g., weighting by each cohort's sample share). This type of aggregation

[9] Researchers may specify different baseline periods. For estimating pre-treatment effects (e.g., $\tau = g-4$), CS uses the prior period (e.g., $\tau = g-5$) as the baseline.

[10] The authors provide an open-source R package (did) that implements all three types of estimators and allows for different types of clean controls. A Stata package (csdid) written by Fernando Rios-Avila is also available.

[11] The authors provide a Stata package (eventstudyinteract) that implements their interaction-weighted estimator. An R package (fixest) written by Laurent Berge is also available.

produces an estimate of the average effect of participating in the treatment experienced by all the units that ever participated, similar in spirit to the interpretation of the 2×2 static DiD estimate. It is also possible to apply these alternative weightings to SA's CATT estimates.

Overall, we view SA as being perhaps simpler (and quicker) to execute because it simultaneously estimates all the group-time treatment effects in one regression and does not use bootstrapping for inference. However, the CS approach is more flexible (e.g., allowing for covariates and the use of not-yet-treated controls). In addition, it offers more robust modeling options (e.g., outcome-regression, inverse-probability-weighted, and doubly-robust estimators, as well as simultaneous confidence intervals that account for multiple-testing of relative-time indicators). For these reasons, our replications of prior results in Section 5 focuses on the application of the CS estimator.[12]

4.2. Stacked regression estimator

An alternative approach developed by applied researchers for circumventing the issues with TWFE DiD estimators is a "stacked regression" (see, e.g., Gormley and Matsa, 2011; Cengiz et al., 2019; Deshpande and Li, 2019). We describe here one implementation of this approach, used in Cengiz et al. (2019). The idea is to create event-specific "clean 2×2" datasets, including the outcome variable and controls for the treated cohort and all other observations that are "clean" controls within the treatment window (e.g., not-yet-, last-, or never-treated units). For each clean 2×2 dataset, the researcher generates a dataset-specific identifying variable. These event-specific data sets are then stacked together, and a TWFE DiD regression is estimated on the stacked dataset, with dataset-specific unit- and time-fixed effects. This approach can be applied using either a static or a dynamic specification (Eqs. (2) or (12)). The only difference in the estimation equation between the standard TWFE approach and a stacked regression alternative is defining the main variables within each event-specific dataset, so that unit- and time-fixed effects are saturated with indicators for dataset identifiers (e.g., α_{ig} and λ_{tg}).

In essence, the stacked regression estimates the DiD from each of the clean 2×2 datasets, then applies variance weighting to combine the treatment effects across cohorts efficiently. This approach is likely the most easily implementable solution for researchers interested in producing aggregated treatment effect estimates via OLS while circumventing the problems introduced by staggered treatment timing and treatment effect heterogeneity. In addition, this estimator is efficient: it relies on OLS to determine the weights on the clean 2×2 DiDs, trading off bias for efficiency. However, relative to the CS or SA approaches, the stacked regression estimator provides less flexibility for aggregation and may be inconsistent for the sample-average ATT.

[12] de Chaisemartin and D'Haultfœuille (2020) also develop an approach for estimating treatment effects under treatment timing variation and treatment effect heterogeneity under more general settings, where treatments may be reversible. (Both CS and SA assume staggered adoption of irreversible treatments.) However, de Chaisemartin and D'Haultfœuille (2020) focuses on recovering simultaneous treatment effects rather than the estimation of dynamic effects, and it does not allow for covariates. For these reasons, we do not emphasize this approach.

4.3. Simulation: Alternative estimators

Figure 6 compares the three alternative estimators under Simulations 1–6 (examined in Section 3.1.1). Here, we focus on a static estimator for the overall treatment effect from participating in treatment. CS and SA are unbiased for the sample ATT in the data in each case. (Note the sample ATT in this figure is different from that of Fig. 2, because we only calculate the treatment effects for the cohorts with valid available comparison units, and only for the five years post treatment assignment.) On the other hand, stacked regressions can differ from the sample-average ATT, particularly when there is heterogeneity in treatment effects across cohorts or time. These differences reflect the alternative weighting of the constituent clean 2×2s implicit in the stacked regression approach compared to CS or SA; they are not the result of potentially problematic 2×2 comparisons under dynamic treatment effects. Because OLS determines these weights by trading off bias for efficiency, stacked regression estimators also exhibit greater efficiency (i.e., a tighter distribution) in Fig. 6 relative to CS or SA. Notably, none of these alternatives exhibit the sign-flip problem of TWFE DiD estimators (i.e., Simulation 6 of Fig. 2).

Figure 7 compares event-study estimates using each of the alternative approaches. For parsimony, we focus on Simulation 6, for which TWFE's biases are most severe. Each of the alternative estimators is able to recover the true treatment path. The stacked regression approach generates slightly larger estimates relative to the sample-average ATT for each relative-time period, again resulting from the use of OLS variance-weighting rather than weighting by sample shares.

Figures 6 and 7 illustrate that each of the alternative estimators is effective for estimating treatment effects in settings with staggered treatment timing and heterogeneous treatment effects. Although the field has not yet settled on an established standard, we believe that applied researchers leveraging settings with staggered treatment timing should implement at least one of these alternatives.

5. Applications

We examine two papers published in top finance journals that rely on TWFE staggered DiD regressions to evaluate the effects of policies. Each was published before the advent of the econometrics literature on the flaws of TWFE estimation, applied the methodological tools available at the time, and had credible claims to causal identification. We replicate a portion of the main results, provide diagnostic tests demonstrating the distribution of treatment timing and the Goodman-Bacon (2021) decomposition when possible, and evaluate the extent to which the published results are robust to DiD methods that correct for the biases induced by treatment timing variation and treatment effect heterogeneity. We focus on the CS and stacked regression estimators for parsimony.

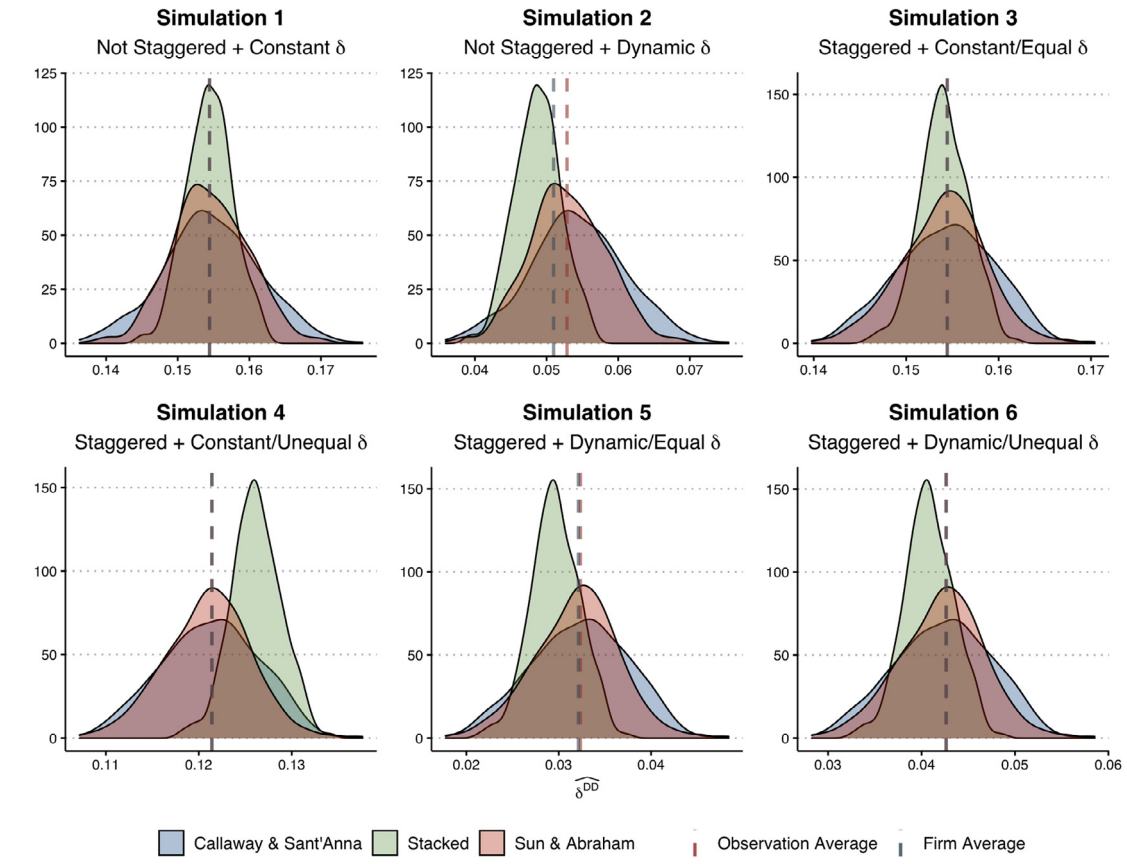


Fig. 6. Simulation: Distribution of Static Effect Estimates of Alternative Estimators.
Figure 6 plots the distribution of static treatment effect estimates for the three alternative estimators explained in Section 4. These distributions are generated based on applying the alternative estimators to each of the 500 simulated Compustat ROA panel datasets under Simulations 1–6. For each data generating process, we overlay the three distributions. The dashed vertical lines represent the observation-level or firm-level average ATT for the five-year period post treatment.

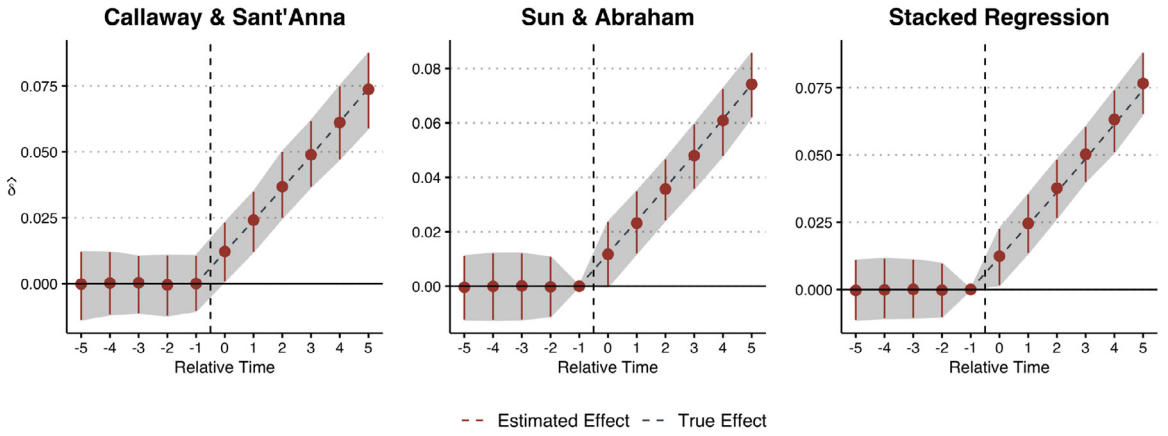


Fig. 7. Robust DiD Methods with Staggered Treatment Assignment and Dynamic Treatment Effects.
Figure 7 plots the distribution of treatment effect estimates by relative-time period for the three alternative estimators explained in Section 4. These distributions are generated based on applying the alternative estimators to each of the 500 simulated Compustat ROA panel datasets under Simulation 6 (Eq. (10)), for which TWFE DiD estimates are highly biased. For each relative-time period from −5 to 5, we plot the point estimate (the solid circle), the 95% confidence interval (the vertical lines intersecting the solid circles), and the observation-level average ATT for each relative-time period (the dashed blue line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
The Impact of Deregulation on Income Inequality.

Panel A: Replication using TWFE	No Controls	With Controls
	(1)	(2)
	Log Gini	Log Gini
Bank deregulation	−0.022***	−0.018***
	(0.008)	(0.006)
Observations	1519	1519
Adj. R2	0.51	0.54
Panel B: Alternative Estimators	Callaway & Sant'Anna	Stacked Regressions
	(1)	(2)
	Log Gini	Log Gini
Bank deregulation	0.001	0.000
	(0.007)	(0.005)

Note: Table 2, panel A, replicates the TWFE estimates of the effects of banking deregulation on inequality (using the natural logarithm of the Gini index as a proxy) from Table II of Beck et al. (2010). The regression includes state- and year-fixed effects, and standard errors are clustered at the state level. We report the results with and without controls found in Table II of their paper. Panel B reports static effect estimates from Callaway and Sant'Anna (2021) and the stacked regression approach, using treatment observations from five years before to ten years after the year of treatment, consistent with the event-study estimates in Fig. 10, and their clean controls (not-yet-treated observations). *,**, and *** denote two-tailed significance tests at the 10%, 5%, and 1% levels, respectively.

5.1. Beck et al. (2010) (BLL)

BLL analyzes the income distribution effects of bank branching deregulation in the United States, which occurred across states and was staggered over time. By exploiting the cross-state intertemporal variation in deregulation, BLL finds that the removal of interstate banking restrictions led to a decline in income inequality.

We begin by replicating the main result from Table II of BLL, which provides static treatment effect estimates. The authors use multiple measures of state-level income inequality and find similar results across them. For parsimony, our replication focuses on the log of the state-level Gini index as the outcome of interest. Table 2 presents the results from the following static DiD regression

$$\text{Log(Gini)}_{st} = \alpha_s + \lambda_t + \delta^{DD} D_{st} + \epsilon_{st},$$

where α_s and λ_t are state and year fixed effects, and D_{st} is an indicator set to 0 before a state allows interstate bank branching and one afterward. We report results without (column 1) and with (column 2) the time-varying covariates used in the paper, as BLL does in Panels A and B of its Table II. Table 2, Panel A, replicates BLL's point estimates.[13]

[13] The data and code used to replicate these results are publicly available at https://dataverse.nl/dataset.xhtml?persistentId=hdl:10411/15996. BLL creates state-level Gini index measures using the March Supplement of the Current Population Survey from 1977 to 2007. The sample includes prime-age individuals (25–54) with non-negative personal income, excluding individuals with missing observations of key variables and those with total personal income below the 1st or above the 99th percentile of the distribution of income, among other restrictions. Overall, the dataset includes 31 years and 48 states plus the District of Columbia, totaling 1519 observations.

Next, we provide diagnostics on the TWFE estimate in column 1, Table 2. Figure 8i plots the treatment timing across states, suggesting that there is significant variation, with most of the deregulation occurring between the 1970s and 1990s. However, the treatment timing variation also suggests that potentially problematic 2×2s could influence the static TWFE estimate. To examine this possibility, we implement the Goodman-Bacon (2021) diagnostic. Figure 8ii graphically compares each 2×2 constituent DiD and its weight in the pooled OLS estimate across the two types of comparisons. Moreover, the bottom panel (iii) summarizes the data points in each panel by taking their weighted averages, represented as horizontal red lines in Fig. 8ii. The decomposition indicates a reason for concern. BLL's documented negative effects on income inequality are driven by a relatively small number of potentially problematic 2×2s. These 2×2s comparing later-treated states to earlier-treated states (as effective controls) produce an average negative effect and receive a weight of 0.86 in the overall TWFE estimate. In contrast, the clean 2×2s that compare earlier-treated to later-treated states (as effective controls) produce an average effect close to zero and receive a relatively low weight of 0.14 in the overall TWFE estimate.

We also replicate BLL's event-study analysis (Figure III of their paper), which plots the event-time coefficients and the standard errors from the following regression:

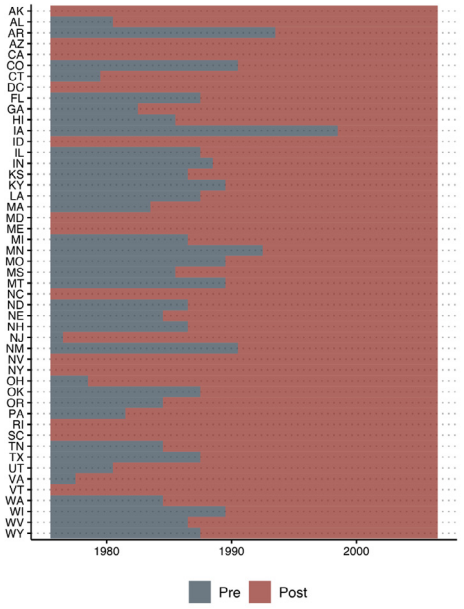
$$\text{log(Gini)}_{st} = \alpha_s + \lambda_t + \beta_1 D_{st}^{-10} + \beta_2 D_{st}^{-9} \ldots \beta_{25} D_{st}^{+15} + \epsilon_{st}.$$

Instead of a single binary indicator (i.e., D_{st} in the previous specification), this specification uses 25 separate indicator variables for the years *relative* to the year of adoption (g), from $g-10$ to $g+15$. In addition, BLL bins the most distant relative-time periods: all years earlier than 10 years before adoption are grouped in the $g-10$ bin and all years greater than 15 years post adoption are grouped in the $g+15$ bin. The year of treatment ($g+0$) is excluded from the specification as the reference year. In addition, BLL implements a normalization that subtracts the average of the pre-adoption coefficients from all of the plotted relative-time coefficients. The theoretical justification for this normalization is unclear; practically speaking, it achieves the effect of forcing the pre-adoption coefficients to be centered at zero.

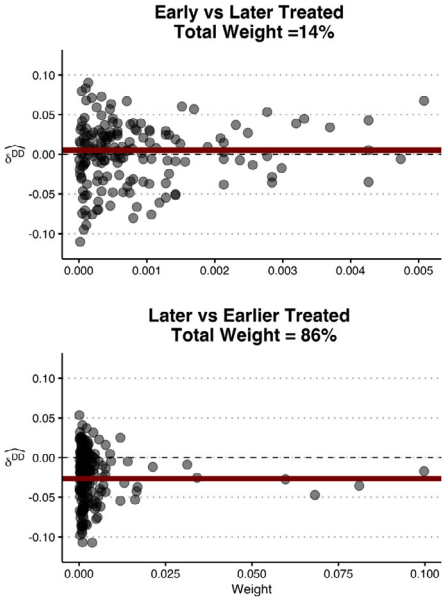
Figure 9, Panel A, replicates BLL's event-study plot, which suggests a negative effect of deregulation on income inequality. In the pre-adoption period, the relative-time dummies' coefficients are centered around zero, suggesting a lack of pre-trends and consistent with the parallel-trends assumption. However, following deregulation, an immediate and statistically significant negative effect appears and settles to a 4% decline in the Gini index over time.

We make three changes to BLL's event-study analysis to analyze the impact of the specification choices it makes. In Panel B, we plot the coefficients directly from the regression results without subtracting the mean of the pre-adoption coefficients. This adjustment shifts the event-study plot upwards without changing the trends; however, most post-period coefficients' confidence intervals now cover zero. In Panel C, we additionally remove binning; we estimate a "fully dynamic" specification, following

(i) Distribution of Treatment Timing



(ii) 2x2 Weights and Estimates



(iii) Overall Weights and Estimates, by Timing Type

Type	Weighted Average	Total Weight
Earlier vs Later Treated	0.005	0.143
Later vs Earlier Treated	-0.027	0.857

Fig. 8. BLL: Treatment-Timing Plots and Goodman-Bacon Decomposition Diagnostic.
Figure 8, panel (i), plots the timing of banking deregulation across states in the Beck et al. (2010) sample. Blue tiles represent pre-deregulation observations, and red tiles represent post-deregulation observations. Panel (ii) plots the TWFE weights and 2×2 DiD estimates for each treatment-timing cohort, broken down by early- (as treatment) vs. later-treated states (as controls) comparisons (in the upper half of the panel) and later- (as treatment) vs. earlier-treated states (as controls) comparisons (in the lower half of the panel). Each dot is a unique comparison between treatment-timing cohorts (e.g., states treated in 1990 compared to states treated in 1985). Panel (iii) reports the weighted average for each comparison type (bold horizontal lines in panel (ii) plots) and the total weight applied by TWFE. The overall TWFE ATT estimate is the weighted sum of each weighted average. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Sun and Abraham (2021) and Borusyak and Jaravel (2018), that includes the complete set of relative-time indicators in the estimation.[14] However, we report only those coefficients from $g-10$ to $g+15$, following BLL's original event-study analysis. Furthermore, in Panel D, we remove all the states that deregulated prior to 1977 (i.e., states that are always-treated in the sample) and all the observations after 1999 (i.e., when all states have deregulated).[15]

Panels C and D show that removing binning significantly changes the event-study plots. These event-study plots now show an upward trend in income inequality, in contrast to Panels A and B. These results are consistent with the theoretical analysis of Sun and Abraham (2021) discussed above: under heterogeneous treatment effects, binning relative-time periods *per se* can bias TWFE staggered DiD dynamic effect estimates.

We stress the possibility that *none* of the event studies in Fig. 9 paint an accurate picture of the effects of banking deregulation on income inequality. Under treatment heterogeneity, all of the dynamic TWFE staggered DiD estimates in each panel could be biased. Therefore, we apply the CS and stacked regression approaches to provide a better benchmark for BLL's dynamic effect estimates.

[14] As noted in both papers, you must omit *two* relative-time indicators to avoid perfect collinearity in staggered adoption designs with no never-treated units. We drop the relative-time indicators for the most negative relative-time period in addition to the year of treatment.

[15] BLL's estimation uses the full panel of observations with surveys stretching from 1977 through 2007 even though all states deregulated by 1999 (see Fig. 8i). As there are no effective control units, post-1999 data cannot be used to identify a treatment effect. In addition, 13 states adopted branch reforms *before* the data started, and thus have no pre-adoption observations from which to calculate the first difference. We note that it is possible for researchers to justify the use of prior-treated units as comparisons units—for example, a number of years after treatment when it can be safely assumed that treatment effects no longer accrue. However, such choices are best justified based on knowledge of the institutional details relating to the research setting.

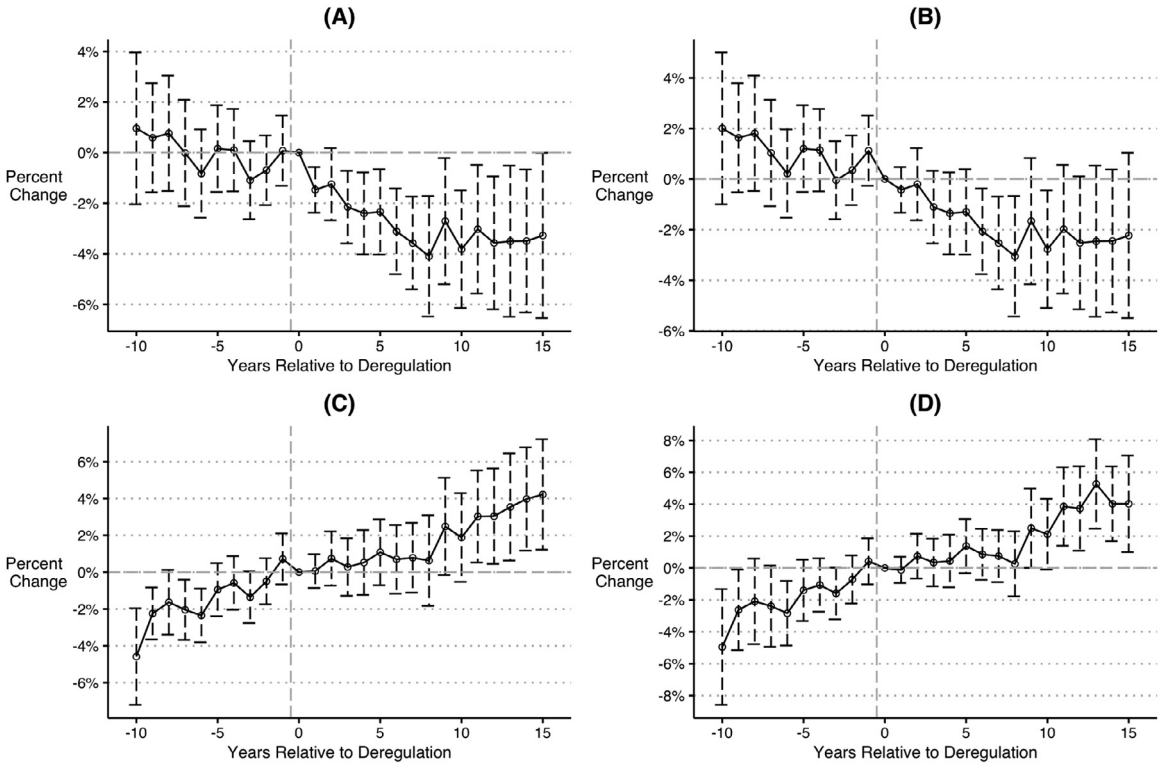


Fig. 9. BLL: TWFE DiD Event-Study Plots.
Figure 9 plots TWFE event-study estimates and 95% confidence intervals for relative-time periods from $l = g - 10$ to $l = g + 15$ around deregulation ($l = g + 0$). Panel (A) reports estimates from our replication of the event-study analysis in Beck et al. (2010). In this specification, the average value pre-adoption coefficients is subtracted from all of the event-study estimates so that the pre-period coefficients are centered at zero. Panel (B) presents estimates from a specification similar to (A) but does not subtract the average of the pre-treatment coefficients. Panel (C) presents estimates from a specification similar to (B). However, it removes the binning of relative-time periods more than ten years prior to deregulation and the binning of relative-time periods more than fifteen years after deregulation. Panel (D) presents estimates from a specification identical to (C) but estimated on a modified sample, which drops all state-year observations for which deregulation occurred before the beginning of the panel dataset and drops all observations after all states have deregulated.

In Fig. 10i, we implement the CS estimator in two ways: one that uses the last-adopting states (Panel A) and one that uses later-adopting states (Panel B) as effective comparison units. We aggregate group-time treatment effects by relative time and report both the point estimates and standard errors for relative-time periods from $g - 5$ to $g + 10$. Across both panels, the CS estimates do not suggest a decline in income inequality after banking deregulation. If anything, both panels suggest marginal evidence of an *increase* in inequality several years after deregulation.

Figure 10ii reports event-study estimates using the stacked regression approach. In Panel A, we stack cohort-specific datasets that include observations from states that deregulate in a certain year (treated) and all states that do not deregulate within 10 years (effective controls). In Panel B, we stack cohort-specific datasets that include all states that deregulate in that year (treated) and all other state-year observations that are not-yet-treated (effective controls). We keep only state-year observations within -5 and 10 years of the given treatment year and estimate the event-study specification on the stacked data, using dataset-specific time- and state-fixed effects. These results are similar to those using CS and show little evidence of a significant decline in income inequality following banking deregulation.

Finally, we provide alternative estimates of the overall inequality effect from banking deregulation. Table 2, Panel B, reports the overall treatment effect estimate using CS (column 1) and stacked regression (column 2). (For parsimony, we implement only the regression-based CS estimator and not the inverse-probability-weighted or doubly-robust variants.) In both cases, we use later-treated states as effective controls and, to stay consistent with the event-study analysis in Fig. 10i and ii, only include the relative-time periods in the $g - 5$ to $g + 10$ window. Because all states are treated by 1999, none of the stacked datasets include observations after 1999.

The CS and stacked regression aggregate effect estimates are similar in magnitude, close to zero, and statistically insignificant at the 10% level. These results differ substantially from the negative and statistically significant static estimate in Panel A. Together, both static and dynamic estimates from the alternative estimators raise doubts about whether banking deregulation impacted income inequality.

Our replication of BLL highlights the potential severity of the biases in TWFE staggered DiD treatment effect es-

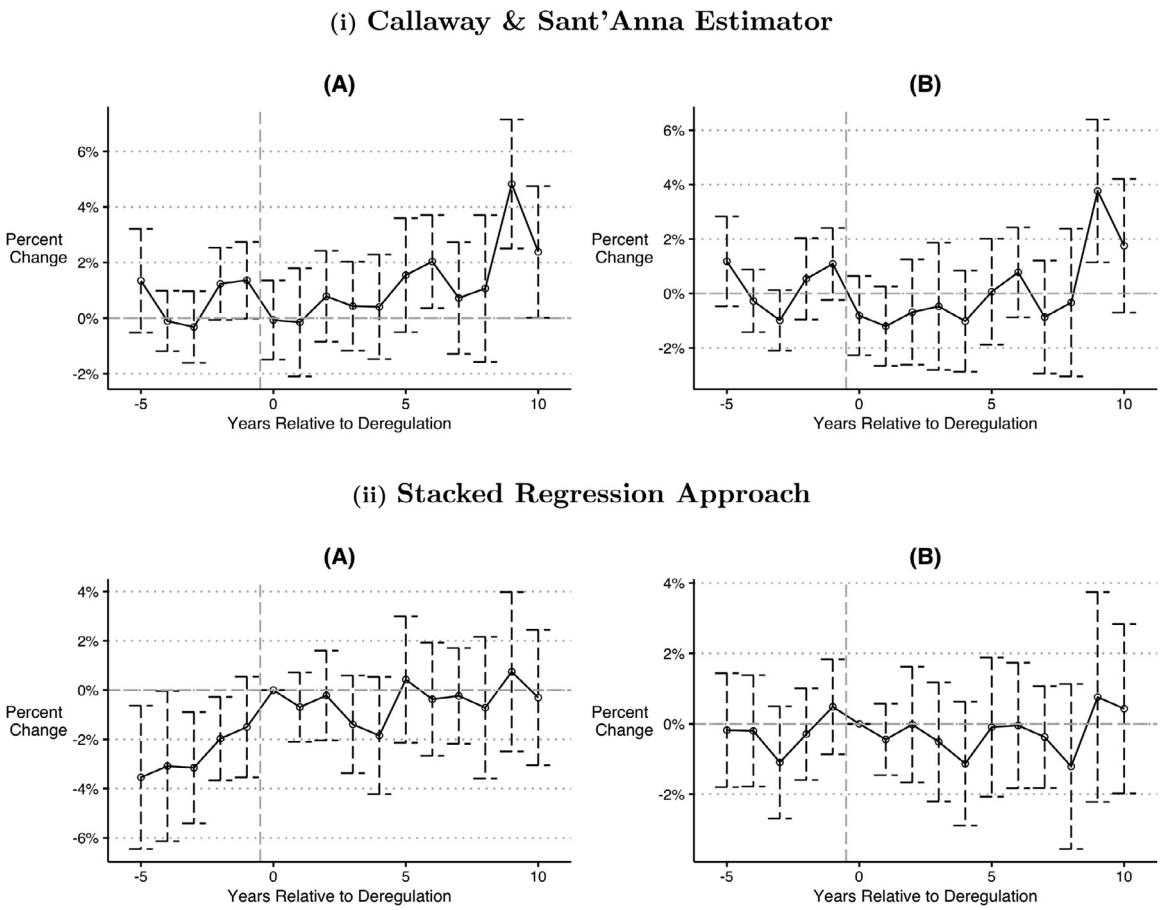


Fig. 10. BLL: CS Estimator and Stacked Regression Event-Study Plots.
Figure 10 plots the event-study estimates and 95% confidence intervals for relative-time periods from $l = g - 5$ to $l = g + 10$ around deregulation ($l = g + 0$), estimated using the Callaway and Sant'Anna (2021) regression estimator and stacked regression estimator described in Section 4. Panel (i) provides two figures of event-study coefficients from the Callaway and Sant'Anna (2021) estimator that differ based on the effective control units used in the estimation. Figure (A) includes only the last-treated states as effective comparison units, while figure (B) uses not-yet-treated states. Panel (ii) provides two figures of event-study coefficients from the stacked regression approach that differ based on the effective control units used in the estimation. Figure (A) uses all states that do not deregulate within ten years of the treated cohort in each stacked dataset, while figure (B) uses all state-year observations that are pre-treatment relative to the treatment cohort's treatment timing in each stacked dataset. Thus, Figure (B) allows more observations to act as effective control units than figure (A).

timates and the substantial differences in inferences from applying the remedies suggested by the econometrics or applied literature. These biases could lead researchers to infer significant effects when they do not exist. We explain in Section 3 and Figs. 4 and 5 why this is possible in both static and event-study staggered DiD specifications.

5.2. Fauver et al. (2017) (FHLT)

A long literature in corporate governance examines the relation between board governance practices and firm performance or value in the US; however, there is scant evidence in other countries. FHLT analyzes data on 41 major board reforms worldwide that either impose or recommend board, audit committee, or auditor independence or call for the separation of the chairman and CEO positions. The paper's identification strategy relies on the staggered implementation of these country-level board reforms from 1990 to 2012. FHLT finds that the reforms increased average firm value, as measured by Tobin's Q.

We begin by replicating FHLT's main regression specification:

$$Q_{it} = \alpha_i + \lambda_t + \delta^{DD} Post_{it} + \gamma' \mathbf{x}_{it} + \epsilon_{it},$$

where Q_{it} is a firm-year measure of Tobin's Q, α_i and λ_t are firm- and year-fixed effects, $Post_{it}$ is an indicator evaluating to one for firm-year observations after a board reform in a firm's headquarter country, and $\mathbf{x}_{it}$ are time-varying firm and country-level controls intended to mitigate confounding events and correlated omitted variables.

The paper uses two different effective dates for defining the board reform "treatment": one based on the timing of the "major" board reforms, as defined by the authors, and another based on the timing of the first board reforms. Figure 11 plots the timing of major and first board reforms across countries. Because we are using firm-level data, different countries receive different weights in the DiD

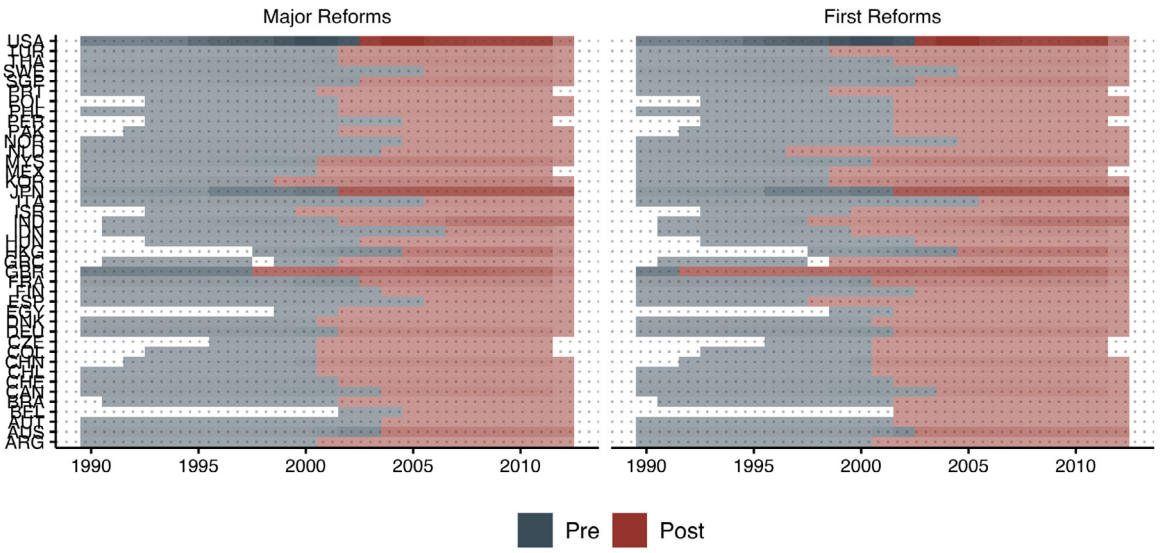


Fig. 11. FHLT: Treatment-Timing Plots.
Figure 11 plots the timing of board reforms (both the major reforms and the first reforms) across countries in the Fauver et al. (2017) sample. Blue tiles represent pre-reform observations, red tiles represent post-reform observations, and empty tiles represent missing data. The shade of the tile indicates the number of firm-year observations for each country: countries with more firm-year observations appear with darker tiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
The Impact of Board Reforms on Firm Value.

Panel A: Replication using TWFE	With Covariates		Without Covariates	
	(1) Major Reform	(2) First Reform	(3) Major Reform	(4) First Reform
Reform	0.096***	0.149***	0.110**	0.136**
	(0.03)	(0.05)	(0.05)	(0.07)
Control variables	Yes	Yes	No	No
Firm fixed effects	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes
Observations	196,016	196,016	196,016	196,016
Adj. R2	0.580	0.581	0.536	0.536
Panel B: Alternative Estimators	Callaway & Sant'Anna		Stacked Regressions	
	(1) Major Reform	(2) First Reform	(3) Major Reform	(4) First Reform
Reform	0.062	0.116	0.063	0.166***
	(0.135)	(0.094)	(0.051)	(0.055)

Note: Table 3, panel A, replicates the TWFE estimates of the effects of board reform on Tobin's Q from Table 4B of Fauver et al. (2017). The first two columns replicate the published values with firm and country covariates. In the third and fourth columns, we present the results without including covariates. All estimates use firm and year fixed effects, and robust standard errors are clustered at the country level. Panel B reports static effect estimates from Callaway and Sant'Anna (2021) and the stacked regression approach, using treatment observations from five years before to five years after the year of treatment, consistent with the event-study estimates in Fig. 13, and their clean controls (not-yet-treated observations). *,**, and *** denote two-tailed significance tests at the 10%, 5%, and 1% levels, respectively.

due to having different numbers of listed firms. In the plot, countries with more firm-year observations appear with darker tiles.

Table 3, Panel A, columns 1 and 2 replicate the results of Fauver et al. (2017) (i.e., Table 4B of their paper) that use the full data panel with both reform definitions. We replicate the point estimates exactly but obtain slightly different standard errors due to how different software packages calculate clustered standard errors in fixed-effects regressions.

We also report TWFE DiD estimates without covariates in columns 3 and 4, which are similar to those of columns 1 and 2 both in terms of economic magnitudes and statistical significance. Consistent with FHLT, our replication shows that board reforms are associated with higher Tobin's Q. Effect sizes using the timing of the first reforms

are about 20% to 50% larger than those using the major reforms.

Ideally, in analyzing the degree to which FHLT's TWFE DiD estimates are susceptible to potential biases due to treatment effect heterogeneity, we would implement the Goodman-Bacon (2021) diagnostic. However, the diagnostic applies to only balanced panels, and FHLT's panel is highly unbalanced.

Our analysis of FHLT's results thus proceeds by examining how effect estimates differ under alternative estimators. Following the structure of the BLL replication above, we start by examining FHLT's event-study TWFE DiD estimates and comparing the results to the dynamic effect estimates under the CS and stacked regression approaches. Finally, we provide an aggregate value effect estimate of board reforms on the adopting countries' firms using the CS and stacked regression approaches and compare them to FHLT's main static effect estimates. We focus on the specifications that do not include covariates in these analyses.

We replicate BLL's event-study analysis (Table 4 Panel C of their paper), which estimates the following regression:

$$Q_{it} = \alpha_i + \lambda_t + \beta_1 D_{it}^{-1} + \beta_2 D_{it}^{0} + \beta_3 D_{it}^{+1} + \beta_4 D_{it}^{+2} + \epsilon_{it}.$$

Instead of a single binary indicator (i.e., $Post_{it}$ in the previous specification), this specification uses four separate indicator variables for the years *relative* to the year of adoption. This estimation equation follows FHLT's convention for denoting relative-time periods, in which Year 1 (D_{it}^{+1}) is the first effective year of board reform (as opposed to the usual convention of Year 0 (D_{it}^{0})).

We highlight several sample and specification choices FHLT makes in estimating this event study regression. First, this specification is estimated on a truncated sample that drops observations outside of five years prior to or five years after the first year of reform adoption. Second, this specification excludes all relative-time periods more than two years prior to the first year of reform ($D_{it}^{-2}, D_{it}^{-3}, D_{it}^{-4}$) as reference periods. Third, FHLT combines all relative-time periods from Year 2 to Year 6 into one bin (D_{it}^{+2}).

Figure 12i, Model 1, replicates FHLT's event-study estimates. Instead of reporting them in table form, we plot the point estimates and 95% confidence intervals. Model 1 suggests a positive value effect from the adoption of major board reforms.

Similar to our analysis of BLL's event-study design, we make three changes to FHLT's event-study analysis to analyze the impact of the specification choices it makes. In Model 2, we include additional pre-period relative-time indicators in the specification that were omitted in Model 1: D_{it}^{-2} and D_{it}^{-3}. Sun and Abraham (2021) shows that the choice of exclusion periods could lead to biases in TWFE dynamic effect estimates, particularly when researchers exclude *post*-treatment relative-time periods from event-study specifications. Our modification in Model 2 shows a similar pattern to Model 1, suggesting that the choice of excluding additional pre-period relative-time indicators did not have a meaningful impact on the overall inference.

In Model 3, we additionally remove binning; we estimate a fully dynamic specification that includes the complete set of relative-time indicators in the estimation. To implement this specification, we omit the relative-time indicators from one and five years prior to the reform to avoid perfect collinearity. The resultant event-study plot provides a dramatically different picture: we no longer observe an apparent positive effect after reform, and all the point estimates are much closer to zero.

Finally, in Model 4, we estimate a fully dynamic specification using the whole sample. We include the complete set of relative-time indicators in the estimation (beyond five years before and after the reform), exclude the indicator for the most negative relative-time period and the indicator for the year prior to the reform, and report only those coefficients from the window between the five years prior to five years after reform. In estimating this model, we also exclude observations after the final treatment, like Panel D of Fig. 9, because these observations have no effective controls thus cannot be used to identify treatment effects. None of the remedies described in Section 4 use these observations.[16] The event-study plot for Model 4 is similar to that for Model 3 and again does not suggest strong evidence of a positive value effect from major board reforms. One difference is that Model 4 identifies an additional coefficient (i.e., for $g-4$), which is possible because it uses observations outside of the ten-year window surrounding the reform.

Figure 12ii analyzes FHLT's event-study estimates of the value effects of the first board reforms, following the same sequence of modifications as Fig. 12i. Again, we show that including additional pre-period relative-time indicator results in similar dynamic effect estimates. However, once we relax the binning of post-treatment relative-time periods or estimate the fully dynamic specification on the whole sample, we no longer find strong evidence of positive value effects stemming from the first board reforms. These results reinforce the important role of binning relative-time periods on event-study estimates, as shown in our BLL replication. To the extent that dynamic effects in Tobin's Q apply after the implementation of board reforms, FHLT's choice to begin binning one year after the reform (e.g., instead of binning the most distant relative-time periods) could accentuate the potential bias in TWFE event-study estimates (Sun and Abraham, 2021).

It is possible that none of the event studies in Fig. 12 paint an accurate picture of the value effects of board reforms. Under treatment heterogeneity, all of the dynamic TWFE staggered DiD estimates in each panel could be biased.

To provide a benchmark for FHLT's dynamic effect estimates, we apply the CS and stacked regression approaches. We focus only on the variants that use the greatest number of pre-treatment observations and choose later-treated firms as clean control units. As with Fig. 12, we report only the dynamic effect estimates within the

[16] Specifically, when considering the first (major) reforms as "treatment," all countries are treated by 2006 (2007). Although the panel data contains observations through 2012, the observations after the final-treatment years are not useful for the DiD without imposing more structure on the length of treatment effect dynamics and are dropped from our two event-study analyses. We set the relative-time indicators to zero for the firms in the last-treated countries.

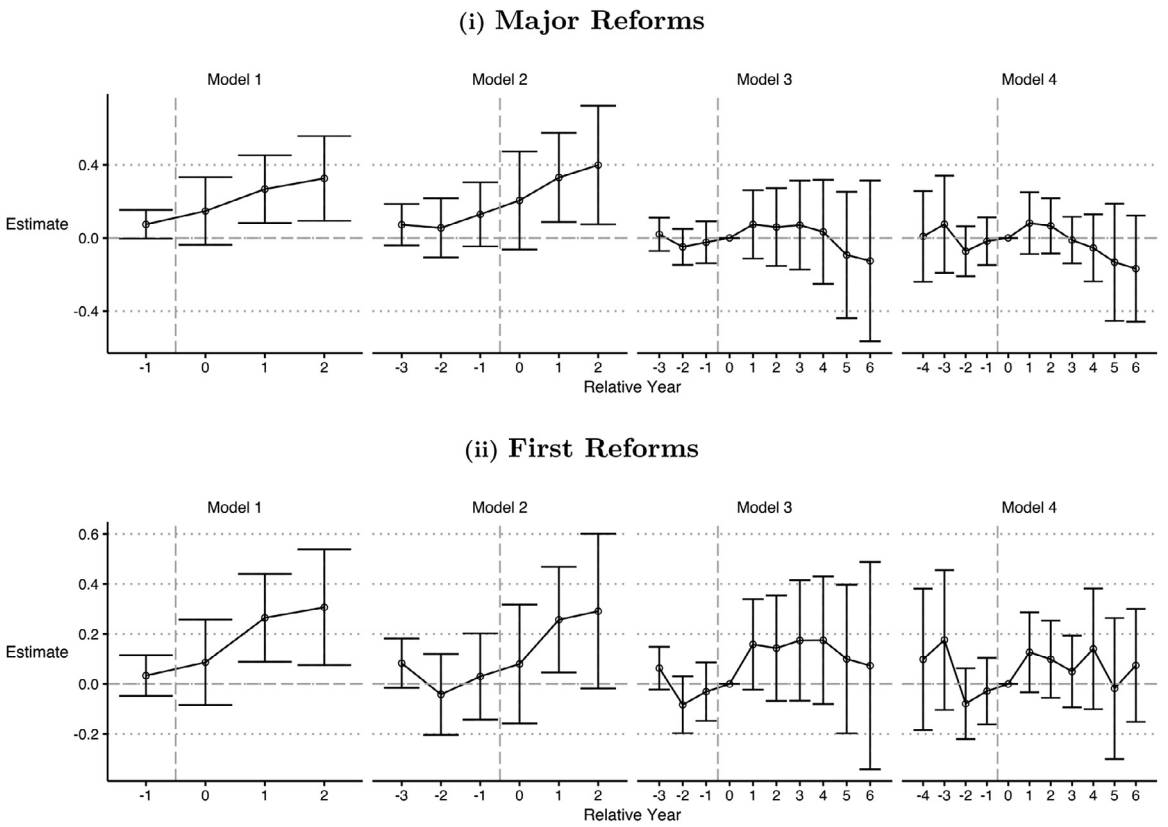


Fig. 12. FHLT: TWFE DiD Event-Study Plots.
Figure 12 plots TWFE event-study estimates and 95% confidence intervals for relative-time periods from $l = g-4$ to $l = g+6$ around board reform ($l = g+1$). Panel (i) plots estimates from specifications that use major reforms as the treatment of interest. Model 1 reports estimates from our replication of the event-study analysis in Fauver et al. (2017). This specification is estimated on a truncated sample that drops observations outside of five years prior to ($l = g-4$) or five years after ($l = g+6$) the first year of reform adoption; it excludes all relative-time periods more than two years prior to the first year of reform ($l \in \{g-4, g-3, g-2\}$) as reference periods; and it combines all relative-time periods from Year 2 to Year 6 into one bin ($l = g+2$). Model 2 presents estimates from a specification similar to Model 1 but adds additional pre-period relative-time indicators that are omitted in Model 1: $l \in \{g-3, g-2\}$. Model 3 presents estimates from a specification similar to Model 2 but removes binning of relative-time periods from Year 2 to Year 6. Finally, Model 4 presents estimates from a fully dynamic specification that includes the complete set of relative-time indicators in the estimation, similar to Model 1, but estimated on the whole sample. Panel (ii) reports the same estimates for the first reforms following the same sequence of modifications as in panel (i).

$g-5$ to $g+5$ window. However, unlike Fig. 12, we revert to the convention of denoting the year of treatment as relative-time period of 0.

Figure 13i and ii present the results from CS and stacked regression estimators. They again do not suggest strong evidence of a statistically significantly positive impact of the reforms on firm valuation, either when we consider major reforms (Panel A) or first reforms (Panel B) as the treatment. Both CS and stacked regression estimates suggest a statistically significant negative effect on firm value five years after a major reform.

Finally, we provide alternative estimates of the overall value effect from adopting board reforms. Table 3, Panel B, reports the overall treatment effect estimate using the regression-based CS estimator (column 1) and stacked regression estimator (column 2). In both cases, we use not-yet-treated firms as effective controls and, to stay consistent with the event-study analysis in Fig. 13, only include the relative-time periods in the $g-5$ to $g+5$ window.

The CS aggregate effect estimate is statistically insignificant at the 10% level, regardless of whether we consider first reforms or major reforms as the treatment. Moreover, compared to the TWFE estimates in columns 3 and 4 of Panel A, these CS estimates are smaller in magnitude and have larger standard errors.

The stacked regression aggregate effect estimates are mixed. For major reforms, stacked regression also produces a statistically insignificant effect at the 10% level, and the point estimate is similar to that of CS. However, stacked regression produces a positive and significant effect associated with the first reforms. This is in part because stacked regression estimates are generally more precise and have lower standard errors compared to CS. Another reason is that the first reform effect estimate from stacked regression is relatively large.

To understand this significant first reform effect estimate, we further scrutinize the corresponding dynamic effect estimates in Fig. 13ii, Panel B. This plot reveals some evidence of pre-trends: the effect estimate three years prior to reform ($g-3$) is negative, statistically significant, and monotonically increasing thereafter until the year of reform. Because the stacked regression aggregate effect

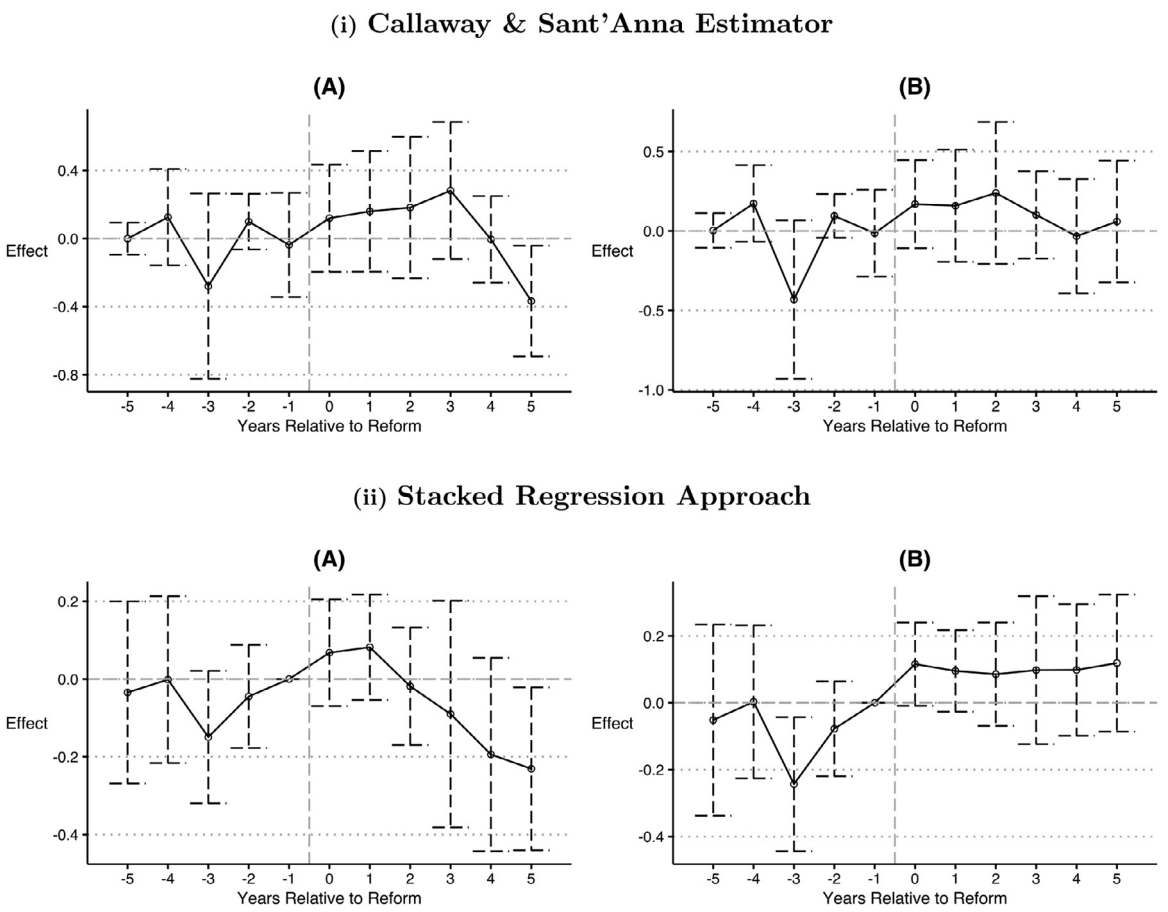


Fig. 13. FHLT: CS Estimator and Stacked Regression Event-Study Plots.
Figure 13 plots the event-study estimates and 95% confidence intervals for relative-time periods from $l = g - 5$ to $l = g + 5$ around board reform ($l = g + 0$) estimated from the Callaway and Sant'Anna (2021) regression estimator and stacked regression estimator described in Section 4. Panel (i) provides two figures showing the event-study coefficients from the Callaway and Sant'Anna (2021) estimator using not-yet-treated units as effective comparison units, but differ based the source of treatment variation. Figure (A) uses major reforms while figure (B) uses first reforms as the treatment of interest. Panel (ii) provides two figures showing the event-study coefficients from the stacked regression approach, each using not-yet-treated units as effective comparison units. Figure (A) uses major reforms, while figure (B) uses first reforms as the treatment of interest.

estimate is, in essence, the difference between the post-period and the pre-period dynamic effect estimates, the significantly positive aggregate effect in part reflects the presence of pre-trends. In contrast, the CS aggregate effect estimates essentially aggregates the post-period dynamic effects shown in Fig. 13i, Panels A and B.

To summarize, while our analysis of FHLT suggests that first reforms could have been associated with a significant increase in firm value, this finding could be confounded by the presence of pre-trends. For example, it is possible that countries adopted initial board reforms after periods of relatively low firm value. While a full exploration of these issues is beyond the scope of our analysis, at a minimum, our analysis suggests that the value effects of global board reforms are not as robust as initially believed.[17]

6. Conclusion and recommendations

Staggered DiD regressions commonly used by applied researchers are susceptible to biases introduced by treatment effect heterogeneity. We argue that these biases apply to a large portion of research settings in finance, accounting, and law involving staggered treatment timing. We conclude by providing a set of practical recommendations for applied researchers interested in exploiting such settings for causal inference.

1. TWFE DiD regressions are appropriate in settings with a single treatment period or where homogeneous treatment effects can be assumed. In the latter case,

[17] Fauver et al. (2021) (FHLT21) argues that the conclusions in FHLT are valid after performing various analyses to address the aforementioned issues. FHLT21 applies the stacked regression approach to a new sample that adds observations from three additional countries that never adopted board reforms between 1990 to 2012: Ivory Coast, Venezuela, and Vietnam. We do not scrutinize these new analyses here. Our main goal is to assess the robustness of FHLT's original findings to the alternative estimators proposed, using the original dataset, a standard set of clean controls, and a minimal amount of sample modifications to facilitate estimation.

researchers should provide theoretical justification for homogeneity.

2. Researchers continuing to report TWFE staggered DiD regressions should provide an assessment of the likelihood of bias. We suggest that it is a good practice to plot the treatment timing across cohorts: significant variation in treatment timing suggests the possibility of biases. We also recommend decomposing the static TWFE DiD estimator (e.g., the Goodman-Bacon, 2021, decomposition) when possible. When such decomposition is not available (e.g., if the panel is unbalanced), and never-treated firms are appropriate effective controls (i.e., the parallel-trends assumption is likely to hold), researchers may report the percent never-treated observations in the sample: the larger the percentage of never-treated units, the less problematic the biases associated with TWFE staggered DiD regressions. In addition, researchers should articulate the expected heterogeneity in treatment effects. For example, the larger the expected long-run effects, the more likely are TWFE biases.
3. Researchers implementing TWFE staggered DiD event-study specifications should avoid binning relative-time periods unless they have reasons to believe homogeneous effects apply in the relative-time periods within a bin. We suggest fitting the complete set of possible relative-time indicators in the event-study DiD, even if reporting coefficients on only a subset of them. We also recommend that researchers manually specify and justify the reference periods, which should generally be pre-treatment periods with no expected treatment anticipation. Specifying regression models with multicollinearity may lead statistical software packages to automatically drop relative-time periods, which can create biases (e.g., if post-treatment relative-time periods are omitted).
4. With differential treatment timing and justifiable concern for bias, researchers should apply at least one of the alternative estimators. Those wishing to stay close to TWFE staggered DiD regressions can implement stacked regressions as a baseline. We suggest that, in doing so, researchers report a variant of the stacked regression without time-varying covariates to understand the robustness of the effect estimates and the degree to which they rely on the inclusion of controls (see footnote 3). For a more flexible estimator, we recommend researchers apply the regression, inverse-probability-weighted, or doubly-robust variants of Callaway and Sant'Anna (2021). Another alternative is to analyze each treatment event separately: e.g., estimate a separate TWFE DiD regression for each event using clean controls. Such an approach does not provide an aggregation of the treatment effects, though the resultant distribution of treatment effects may be helpful to report.
5. In applying the alternative estimators, researchers should justify their choice of "clean" comparison groups—not-yet treated, last treated, or never treated—and articulate why the parallel-trends assumption is likely to apply. When using not-yet- or last-treated units as comparison groups in a given event window, researchers should also validate the assumption of no anticipatory effects for these units.
6. Regardless of the estimators used, static DiD estimates should be accompanied by event-study estimates that trace out the timing of outcome differences between treated and control units. In both cases, the length of time in each treatment cohort's event window included in the analysis can impact the treatment effect estimates. For example, the aggregate treatment effect estimate based on CS depends on how many post-treatment relative-time periods' ATTs are aggregated. This is a design choice that should be defended by the researcher and guided by the specific research question and relevant institutional knowledge.

We believe these practices will significantly increase the credibility of staggered DiD studies.

References

Abadie, A., 2005. Difference-in-differences estimators semiparametric. Rev. Econ. Stud. 72 (1), 1–19.

Andrews, I., Kasy, M., 2019. Identification of and correction for publication bias. Am. Econ. Rev. 109 (8), 2766–2794.

Angrist, J.D., Pischke, J.-S., 2009. Mostly Harmless Econometrics. Princeton University Press.

Athey, S., Imbens, G., 2018. Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption. Working Paper.

Barrios, J.M., 2021. Staggeringly problematic: a primer on staggered DiD for accounting researchers. J. Financ. Rep. Forthcoming.

Beck, T., Levine, R., Levkov, A., 2010. Big bad banks? The winners and losers from bank deregulation in the United States. J. Finance 65 (5), 1637–1667.

Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates? Q. J. Econ. 119 (1), 249–275.

Borusyak, K., Jaravel, X., 2018. Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume. Working Paper.

Borusyak, K., Jaravel, X., Spiess, J., 2021. Revisiting Event Study Designs: Robust and Efficient Estimation. Working Paper.

Callaway, B., Sant'Anna, P.H.C., 2021. Difference-in-differences with multiple time periods. J. Econom. 225 (2), 200–230.

Cengiz, D., Dube, A., Lindner, A., Zipperer, B., 2019. The effect of minimum wages on low-wage jobs. Q. J. Econ. 134 (3), 1405–1454.

de Chaisemartin, C., D'Haultfœuille, X., 2020. Two-way fixed effects estimators with heterogeneous treatment effects. Am. Econ. Rev. 110 (9), 2964–2996.

Deshpande, M., Li, Y., 2019. Who is screened out? Application costs and the targeting of disability programs. Am. Econ. J. 11 (4), 213–248.

Fauver, L., Hung, M., Li, X., Taboada, A.G., 2017. Board reforms and firm value: worldwide evidence. J. Financ. Econ. 125 (1), 120–142.

Fauver, L., Hung, M., Li, X., Taboada, A.G., 2021. Re-examining Board Reforms and Firm Value: Response to "How Much Should We Trust Staggered Differences-in-Differences Estimates?" by Baker, Larcker, and Wang (2021). Working Paper.

Goodman-Bacon, A., 2021. Difference-in-differences with variation in treatment timing. J. Econom. 225 (2), 254–277.

Gormley, T.A., Matsa, D.A., 2011. Growing out of trouble? Corporate responses to liability risk. Rev. Financ. Stud. 24 (8), 2781–2821.

Heckman, J.J., Ichimura, H., Todd, P.E., 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Rev. Econ. Stud. 64 (4), 605–654.

Imai, K., Kim, I.S., 2021. On the use of two-way fixed effects regression models for causal inference with panel data. Polit. Anal. 29 (3), 405–415.

Jakiela, P., 2021. Simple Diagnostics for Two-way Fixed Effects. Working Paper.

Karpoff, J.M., Wittry, M.D., 2018. Institutional and legal context in natural experiments: the case of state antitakeover laws. J. Finance 73 (2), 657–714.

Kim, J.H., Ji, P.I., 2015. Significance testing in empirical finance: a critical review and assessment. J. Empir. Finance 34, 1–14.

Rubin, D.B., 2005. Causal inference using potential outcomes. J. Am. Stat. Assoc. 100 (469), 322–331.

Sant'Anna, P.H.C., Zhao, J., 2020. Doubly robust difference-in-differences estimators. J. Econom. 219 (1), 101–122.
Strezhnev, A., 2018. Semiparametric Weighting Estimators for Multi-period Difference-in-Differences Designs. Working Paper.
Sun, L., Abraham, S., 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. J. Econom. 225 (2), 175–199.
Zdrojewski, A., Butler, A.W., 2021. Are Two-Way Fixed-Effect Difference-in-Differences Estimates Blowing Smoke? A Cautionary Tale from State-Level Bank Branching Deregulation. Critical Finance Review (Forthcoming).