How Much Should We Trust Staggered Difference-In-Differences Estimates?

Finance Working Paper N° 736/2021
March 2021

Andrew C. Baker
Stanford University

David F. Larcker
Stanford University, Rock Center for Corporate Governance and ECGI

Charles C.Y. Wang
Harvard University

ECGI Working Paper Series in Finance

How Much Should We Trust Staggered Difference-In-Differences Estimates?

Working Paper N° 736/2021

March 2021

Andrew C. Baker
David F. Larcker
Charles C.Y. Wang

We thank David Childers, Andrew Goodman-Bacon, Kirill Borusyak, Pamela Jakiela, Pedro Sant'Ana (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, Hung, Li, and Taboada (2017) and Wang, Yin, and Yu (2021) for graciously sharing their data and code, and Beck, Levine, and Levkov (2010) for posting their data and code online. Comments are welcome

Abstract

Difference-in-differences analysis with staggered treatment timing is frequently used to assess the impact of policy changes on corporate outcomes in academic research. However, recent advances in econometric theory show that such designs are likely to be biased in the presence of treatment effect heterogeneity. Given the pronounced use of staggered treatment designs in applied corporate finance and accounting research, this finding potentially impacts a large swath of prior findings in these fields. We survey the nascent literature and document how and when such bias arises from treatment effect heterogeneity. We apply recently proposed methods to a set of prior published results, and find that correcting for the bias induced by the staggered nature of policy adoption frequently impacts the estimated effect from standard difference-indifference studies. In many cases, the reported effects in prior research become indistinguishable from zero.

Keywords: Difference in differences; staggered difference-in-differences designs; generalized difference-in-differences; dynamic treatment effects

JEL Classifications: C13, C18, C21, C22, C23

Andrew C. Baker
Researcher
Stanford Graduate School of Business
655 Knight Way
Stanford, CA 94305, USA
phone:
e-mail: bakerandrew370@gmail.com

David F. Larcker*
The James Irvin Miller Professor of Accounting
Stanford University, Graduate School of Business
655 Knight Way
Stanford, CA 94305–7298, United States
phone: +1 650 725 6159
e-mail: dlarcker@stanford.edu

Charles C.Y. Wang
Glenn and Mary Jane Creamer Associate Prof. of Business Administration
Harvard University, Harvard Business School
Soldiers Field
Boston, MA 02163, United States
phone: +1 617 496 9633
e-mail: charles.cy.wang@hbs.edu

*Corresponding Author

How Much Should We Trust Staggered Difference-In-Differences Estimates?

Andrew C. Baker
Stanford Graduate School of Business

David F. Larcker
Stanford Graduate School of Business
European Corporate Governance Institute (ECGI)

Charles C.Y. Wang*
Harvard Business School

March 2021

Abstract

Difference-in-differences analysis with staggered treatment timing is frequently used to assess the impact of policy changes on corporate outcomes in academic research. However, recent advances in econometric theory show that such designs are likely to be biased in the presence of treatment effect heterogeneity. Given the pronounced use of staggered treatment designs in applied corporate finance and accounting research, this finding potentially impacts a large swath of prior findings in these fields. We survey the nascent literature and document how and when such bias arises from treatment effect heterogeneity. We apply recently proposed methods to a set of prior published results, and find that correcting for the bias induced by the staggered nature of policy adoption frequently impacts the estimated effect from standard difference-in-difference studies. In many cases, the reported effects in prior research become indistinguishable from zero.

Keywords: Difference in differences; staggered difference-in-differences designs; generalized difference-in-differences; dynamic treatment effects

JEL: C13, C18, C21, C22, C23

*Baker (abaker2@stanford.edu) is a doctoral candidate at Stanford GSB. Larcker (dlarcker@stanford.edu) is the James Irvin Miller Professor of Accounting (emeritus) at Stanford GSB. Wang (charles.cy.wang@hbs.edu) is the Glenn and Mary Jane Creamer Associate Professor of Business Administration at Harvard Business School. We thank David Childers, Andrew Goodman-Bacon, Kirill Borusyak, Pamela Jakiela, Pedro Sant'Ana (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, Hung, Li, and Taboada (2017) and Wang, Yin, and Yu (2021) for graciously sharing their data and code, and Beck, Levine, and Levkov (2010) for posting their data and code online. Comments are welcome.

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. A workhorse methodological approach in this literature uses the passage of laws or market rules impacting one set of firms or market participants (treated) but not others (controls). This is typically done by comparing the differences in the outcomes between treated and control units after the implementation of a law with the differences in the outcomes between treatment and control units before the law. This methodological approach, called difference-in-differences (DiD), is common in applied microeconomic research, and has been used across policy domains to test the impact of broadly applied policies.

A generalized version of this estimation approach that relies on the staggered adoption of laws or regulations (e.g., across states or across countries) has become especially popular over the last two decades. For example, Table 1 shows that, from 2000 to 2019, there were 751 papers published in (or accepted for publication by) top finance (439 papers) and accounting (312 papers) journals that use DiD designs. Among them, 366 (or 49%) employ a staggered DiD design (50% for finance journals and 47% for accounting journals).[1]

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 suggest that staggered DiD designs often do not provide valid estimates of the causal estimands of interest to researchers—the average treatment effect (ATE) or the average treatment effect on the treated (ATT)—even under random assignment of treatment (e.g., Sun and Abraham, 2020; Borusyak and Jaravel, 2017; Callaway and Sant'Anna, 2020; Goodman-Bacon, 2019; Imai and Kim, 2020; Strezhnev, 2018; Athey and Imbens, 2018; de Chaisemartin and D'Haultfœuille, 2020).

[1] We collected a comprehensive sample of studies that were published in five finance journals (*Journal of Finance*, *Journal of Financial Economics*, *Review of Financial Studies*, *Journal of Financial and Quantitative Analysis*, and *Review of Finance*) and five accounting journals (*Journal of Accounting Research*, *Journal of Accounting and Economics*, *The Accounting Review*, *Review of Accounting Studies*, and *Contemporary Accounting Research*) between 2000 and 2019. We included those papers that, as of the end of 2019, were accepted for publication in one of these journals. We focus on the top five journals in both finance and accounting, as they publish the most influential empirical studies in corporate finance or corporate governance. The vast majority of these DiD and staggered DiD studies were published after 2010.

This paper provides an overview of the recent econometrics literature on staggered DiD: we explain the pitfalls with these designs and the suggested solutions that applied researchers in finance can utilize for circumventing these problems. Importantly, we show that the use of staggered DiD designs often can, and have, resulted in misleading inferences in the finance literature: we show that applying robust DiD alternatives can significantly alter inferences in important papers.

We begin by providing an overview of the recent work in econometrics that explain why treatment effects estimated from a staggered DiD design are not easily interpretable estimates for the ATE or the ATT. In general, such designs produce estimates of variance-weighted averages of many different treatment effects. Moreover, under some conditions—when treatment effects can evolve over time (when there are "dynamic treatment effects")—staggered DiD estimates can obtain the opposite sign compared to the true ATE or ATT, even when the researcher is able to randomize treatment assignment. The intuition is that in the standard staggered DiD approach, already-treated units can act as effective comparison units, and changes in their outcomes over time are subtracted from the changes of later-treated units (the treated). These theoretical results have far reaching implications for applied researchers in finance, accounting, and law.

To demonstrate the situations under which these problems can arise, we simulate and analyze a synthetic dataset that mimics the structure of a standard staggered DiD design in applied corporate governance settings, exploiting changes in state-level laws using a panel of firms whose attributes are measured over many years (e.g., Karpoff and Wittry, 2018). Our simulations produce three main insights. First, DiD estimates are unbiased in settings where there is a single treatment period, even when there are dynamic treatment effects. Second, staggered DiD estimates are also unbiased in settings with staggered timing of treatment assignment and no treatment effect heterogeneity across firms or over time. Finally, when research settings combine staggered timing of treatment effects and treatment effect heterogeneity across firms or over time, staggered DiD estimates are likely to be biased. In fact, these estimates can produce the wrong sign altogether compared to the true average treatment effects.

We then describe three alternative estimators for modifying the standard staggered DiD designs (e.g., Goodman-Bacon, 2019; Callaway and Sant'Anna, 2020; Sun and Abraham, 2020; Cengiz,

Dube, Lindner, and Zipperer, 2019). While the econometrics literature has not settled on a standard alternative approach, the proposed solutions all deal with the bias issues inherent in these design by estimating event-study DiD specifications, and modifying the set of effective comparison units in the treatment effect estimation process. In each case, the alternative estimation strategy ensures that firms receiving treatment are not compared to firms that already received treatment in recent past. However, the methods differ in terms of which observations are used as effective comparison units and how covariates are incorporated in the analysis. Using our simulated data, we show that each of these alternative estimators help to recover the true treatment effects.

Finally, we assess the extent to which these problems likely matter in applied research by applying the alternative DiD estimators to important results published in the top finance and accounting journals over the last decade. We replicate and extend the findings of three important papers that apply staggered DiD designs in a diverse range of settings: from bank deregulation (Beck et al., 2010) and global board governance reform (Fauver et al., 2017) to the deregulation of open market share repurchases (Wang et al., 2021). In each paper, we find that the published DiD estimates are susceptible to the biases created by treatment effect heterogeneity. Once correcting for the use of prior treated units as effective comparison units, the evidence often no longer supports the original paper's findings.

For example, we analyze the findings of Beck et al. (2010), which leverages the staggered bank deregulation across U.S. states between the 1970s and the 1990s to study the impact of bank deregulation on income inequality. Applying a standard staggered DiD design to a panel data of state-level outcomes, including a state-level Gini index that captures each state's income inequality at a point in time, Beck et al. (2010) provides evidence that bank deregulation leads to lower income inequality. We replicate these main findings and show that the staggered DiD estimates are largely driven by comparisons in which earlier-treated firms are used as effective comparison units for later-treated firms (as treated units), suggesting that the main estimates could be susceptible to the biases that arise when there is heterogeneity in treatment effects. In fact, by applying various alternative DiD estimators, which clean up the effective comparison units used to identify the treatment effects, we fail to find compelling evidence of a negative effect of banking deregulation

on income inequality over time. Our analysis suggests either a weakly positive effect or more likely no effect at all.

Similarly, we examine the findings of Fauver et al. (2017), which leverages the staggered implementation of country-level board reforms between 1990 and 2012 to study the effects of board governance on firm value. It applies a staggered DiD design to a panel data of firm-level outcomes, using Tobin's Q as the main outcome variable of interest, and provides evidence that the board reforms increase firm value. Again, after applying various alternative DiD estimators, we fail to find consistent and compelling evidence that the board reforms had a positive effect on Tobin's Q.

Finally, we analyze the findings of Wang et al. (2021), which uses the staggered legalization of stock repurchases across countries to study the effects of such repurchases on firm outcomes. Perhaps the most central result in the paper is the finding that stock repurchases led to significant declines in firm investments, in terms of both capital expenditures (CAPEX) and research and development (R&D). We show that, after applying various alternative DiD estimators that correct for the use of prior treated firms as comparison units in staggered DiD designs, the empirical evidence does not support the conclusion that the legalization of open market repurchases significantly lowered (or had any impact on) repurchasing firms' investing behavior.

Our paper makes several contributions to the applied literature in finance and accounting. DiD designs have become a workhorse tool for causal inference in a large portion of empirical research in these fields. Staggered DiD designs, in particular, are considered the most robust and perhaps desirable variant of such designs. They constitute half of all published DiD papers in top tier finance and accounting journals, and, as noted in Karpoff and Wittry (2018), have been applied to test the effects of a range of policy changes, including banking deregulation on innovation and economic growth, corporate tax changes on firm investment and payout decisions, and the outcome of court decisions on firm value and disclosure policies.

In this context, our paper makes three main contributions. First, we provide an overview of the econometric issues that could impact finance and accounting research. Second, our simulation analysis highlights the circumstances under which staggered DiD designs are most likely to be problematic: when staggered treatment is bundled with dynamic treatment effects. Third, our

empirical analyses suggest that the problems associated with staggered DiD designs are not only theoretical, but they *do* (and are likely to in a significant percentage of cases) impact the inferences of in applied finance and accounting research settings.

An implication of our work is that finance and accounting researchers should interpret the treatment effects estimated using staggered DiD designs with caution, particularly in contexts where treatment effect heterogeneity is most plausible. Our analyses also suggest opportunities for re-examining critical prior research findings established based on staggered DiD designs. Ensuring robust inference in these research designs requires adjusting for the bias induced by staggered treatment timing. We conclude by discussing features of the data structure used in empirical finance and accounting studies that make the use of staggered DiD designs particularly problematic, and propose a framework for conducting generalized DiD studies in a robust and structured manner to mitigate the potential problems.

The remainder of the paper proceeds as follows. Section 2 provides a review of a DiD methodology. Section 3 explains the rationale for, and the econometric problems with, staggered DiD designs. Section 4 summarizes diagnostic tests and alternative estimators proposed in recent econometric papers for circumventing the issues with staggered DiD designs. Section 5 replicates the results of three important papers in the corporate finance literature and illustrates how inferences are altered when we apply the alternative estimators suggested in recent work. We conclude the paper by providing some additional considerations about inference in modified DiD designs (Section 6) and a set of recommendations that will help applied researchers in finance improve the credibility of their DiD designs (Section 7).

2 A Review of the DiD Method

The DiD design is one of the most commonly used methods for identifying causal effects in applied microeconomics research. The intuition behind the method can be easily understood by considering a simple variant of the DiD design involving a single treatment, two discrete time periods—pre- and post-treatment—and two treatment groups—units receiving treatment ("treated") and

units never receiving treatment ("control"). In this 2x2 design, the effect of the treatment 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.

To formalize this intuition, denote $Y_{i,t}(1)$ as the value of the outcome of interest for unit i in period t when the unit receives treatment, and $Y_{i,t}(0)$ as the outcome for unit i in period t when it does *not* receive treatment. The average treatment effect (δ) is defined as the average difference $Y_{i,t}(1) - Y_{i,t}(0)$ across the population.[2]

The challenge in identifying δ, however, stems from a fundamental missing data problem: $(Y_{i,t}(0), Y_{i,t}(1))$ refer to *potential* outcomes, and for a given unit at time t we observe only one of the two but not both: i.e., for any unit i at time t, we cannot observe $Y_{i,t}(1) - Y_{i,t}(0)$. The basic idea behind DiD designs is to impute the counterfactual outcomes using the observed outcomes of treatment and control units.

In particular, DiD assumes the observed *trend* in the outcome variable from period $t = 0$ to $t = 1$ in the control units is the same as the trend in the treatment units if they had not received treatment. Under this "parallel-trends assumption," the treatment effect (on the treated) can be estimated using the observed treatment-control unit difference in the pre- and post-treatment differences in the outcome:

$$\begin{aligned}\delta \equiv & \mathbb{E}[Y_{T,1}(1) - Y_{T,1}(0)] = \\ & \mathbb{E}[(Y_{T,1}(1) - Y_{T,0}(1)) - (Y_{T,1}(0) - Y_{T,0}(0))] = \mathbb{E}\left[(Y_{T,1}(1) - Y_{T,0}(1)) - (Y_{C,1}(0) - Y_{C,0}(0))\right],\end{aligned}$$

where T denotes treatment units and C denotes control units. The first equality defines the estimand of interest but cannot be directly estimated in the data; the second equality follows from adding and subtracting $Y_{T,0}(0)$ (which equals $Y_{T,0}(1)$ under the assumption of no treatment anticipation), but again cannot be directly estimated in the data because we do not observe $Y_{T,1}(0) - Y_{T,0}(0)$ for a unit that receives treatment; and the final equality follows from the parallel trends assumption and can be estimated in the data. To the extent treatment and control units have

[2]The average treatment effect on the treated (ATT) is the difference between $Y_{i,t}(1) - Y_{i,t}(0)$ averaged across the units receiving treatment.

different outcome trends, the DiD estimate will be biased (i.e., the last equality will not hold).

In practice, DiD estimates are obtained through a linear regression. As noted in Angrist and Pischke (2009, p.228), the 2x2 DiD can be thought of as a fixed effects estimator. In particular, assuming that the conditional mean of outcomes for treatment and control units follow additive linear structures with group- and period-fixed effects, we obtain:

$$\mathbb{E}[Y_{C,1}(0)] = \alpha_C + \lambda_1, \qquad \mathbb{E}[Y_{C,0}(0)] = \alpha_C + \lambda_0, \ \mathbb{E}[Y_{C,1}(0)] - \mathbb{E}[Y_{C,0}(0)] = \lambda_1 - \lambda_0; \text{ and}$$
$$\mathbb{E}[Y_{T,1}(1)] = \alpha_T + \lambda_1 + \delta, \ \mathbb{E}[Y_{T,0}(1)] = \alpha_T + \lambda_0, \ \mathbb{E}[Y_{T,1}(1)] - \mathbb{E}[Y_{T,0}(1)] = \lambda_1 - \lambda_0 + \delta.$$

Here, the treatment effect parameter of interest is δ,[3] which can be obtained as the slope coefficient on the interaction term (β_3) from the following regression:

$$y_{it} = \alpha + \beta_1 TREAT_i + \beta_2 POST_t + \beta_3 (TREAT_i \cdot POST_t) + \epsilon_{it}, \tag{1}$$

where $TREAT_i$ is an indicator variable for the treated unit, and $POST_t$ is an indicator variable for observations in periods $t = 1$.[4] This "double differencing" is depicted in Figure 1.

An advantage of regression-based DiD is that it provides a point estimate for δ and standard errors for the estimate. Another perceived advantage of the regression framework is that it can accommodate more generalized DiD settings. As mentioned in Angrist and Pischke (2009), it has often been claimed that it is "easy to add additional states or periods to the regression setup ... [and] it's easy to add additional covariates." Most notably, in settings where there are more than two units and two time periods, the regression DiD model usually takes the following two-way fixed effect (TWFE) form:

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

where α_i and λ_t are unit and time period fixed effects, $D_{it} = TREAT_i \cdot Post_t$ is an indicator for a treated unit in treated time periods, and the main effects for $TREAT_i$ and $POST_t$ are subsumed by the unit and time fixed effects. This TWFE regression model can be further modified to include

[3] Note that the parallel trends assumption is built into the additive structure, since the counterfactual change in outcome for treated units, $\lambda_1 - \lambda_0$, is the same as the realized change in the outcome for control units.

[4] Under the additive linear structure, it can be shown that: $\alpha = \lambda_0 + \alpha_C$, $\beta_1 = \alpha_T - \alpha_C$, $\beta_2 = \lambda_1 - \lambda_0$, and $\beta_3 = \delta$.

covariates, time trends, and dynamic treatment effect estimation (e.g., by separately including indicators for the number of periods before or after the treatment), and this flexibility has made regression DiD an increasingly popular model in empirical applied microeconomics over the past two decades.

3 Staggered DiD Designs: The Problems

In theory, a staggered design offer some desirable properties over a DiD with only one treatment period. 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 because including multiple treatment periods plausibly alleviates concerns that the observed treatment effects are driven by contemporaneous trends.

However, recent work in econometric theory casts doubt on the validity and the robustness of the TWFE DiD estimator when there are more than two treatment groups and periods, or when there is variation in treatment timing. In particular, the main coefficient of interest (δ^{DD} of Eq., (2)) is not easily interpretable, and is not consistent for the usual estimands of interest, such as the ATT or ATE. Numerous studies have now shown that this coefficient is in fact a weighted average of many different treatment effects, and can yield estimates with the opposite sign compared to the true ATE or ATT.[5]

3.1 Simulation Example

To illustrate the problems with staggered DiD designs, we begin with a simple simulation example. We generate a dataset with a similar structure to those frequently used in corporate finance and accounting research settings, containing a panel of firms and years corresponding to annual financial reporting data. We assume a particular data generating process that is likely to be found in real-world situations, where a treatment occurs across units at different points in time,

[5]The body of work examining these issues include: Athey and Imbens (2018), Borusyak and Jaravel (2017),Callaway and Sant'Anna (2020), Goodman-Bacon (2019), Imai and Kim (2020), Strezhnev (2018), and Sun and Abraham (2020).

and show that the TWFE DiD estimation produces estimates that can differ dramatically from the true treatment effects.

Assume we are modeling an outcome variable y_{it} on a balanced panel dataset with $T = 36$ years from $t = 1980$ to 2015, and $1,000$ firms (indexed by i). There are both time-invariant unit effects and time-varying year effects in the outcome variable, which in the data are independently drawn from $\mathcal{N}(0, 0.5^2)$. Firms are incorporated in one of 50 randomly drawn states, which differ in terms of whether and when treatment was initiated.

Our first three simulations illustrate the conditions under which TWFE DiD provide unbiased treatment effect estimates. In the first simulation (Simulation 1), we assume that half of the states initiate treatment at $t = 1998$, and each firm's outcome variable of interest is the sum of the unit and year fixed effects, a stochastic error term, and the treatment effect which is drawn from a normal distribution with mean $2 \times \mathbb{I}[Treat] \times \mathbb{I}[t > 1998]$ and variance of 0.2^2, where $\mathbb{I}[Treat]$ is an indicator for the treated firms and $\mathbb{I}[t > 1998]$ is an indicator for the post-treatment period. That is, the average treatment effect is, in expectation, a positive 2 unit increase in the outcome variable. Thus, Simulation 1 represents the standard 2x2 DiD estimate with only two relative-to-treatment time periods (i.e., pre and post periods), one set of treated units that experiences a level shift in the outcome values, and one set of control units whose outcome values are unaffected.

In our second simulation (Simulation 2), we again assume that half of the states initiate treatment at $t = 1998$. However, we allow for the treatment effect to vary over time (what the literature calls "dynamic treatment effects"). Specifically, we draw each firm's treatment effect parameter on the outcome variable δ_i from a normal distribution with mean $.3 \times \mathbb{I}[Treat]$ and variance equal to 0.2^2. Here the treatment effects are additive, so that for any treated firm i the treatment effect in year t equals $\delta_i \times max(0, T - 1998)$.[6] Instead of a level shift in the outcome values, Simulation 2 models treatment effects that accumulate over time.

In our third simulation (Simulation 3), we allow for staggered timing of treatment assignment. Firms are again incorporated in one of 50 randomly drawn states, but now the states are randomly

[6]Note we could model the treatment effects in a number of different ways; all that is needed is for the treatment effect to be dynamic, or not wholly incorporated within one period. This is analogous to the "trend-break" treatment effect described in Goodman-Bacon (2019), but we could also limit the treatment effect accumulation to a set number of years after treatment.

assigned into one of three treatment groups G_g based on the year in which the treatment was initiated: in 1989, 1998, or 2007. All treatment groups are approximately equal-sized (17 states are in G_{1989}, 17 are in G_{1998}, and 16 are in G_{2007}). In this simulation, there are no never-treated units. Rather, each unit is randomly placed into one of fifty states which are sorted into one of three treatment timing groups: $\{G_{1989}, G_{1998}$, or $G_{2007}\}$. Like Simulation 1, we assume again a level shift of 2 units.

Figure 2i shows the outcome paths (gray lines) for the $N = 1,000$ firms in Simulations 1-3, with the colored lines corresponding to the average value of the outcome variable by treatment cohorts. For each simulated dataset, we estimate the standard TWFE DiD regression. In each simulation, we generate a synthetic dataset (each containing 1,000 firms) 500 times, and compare the distribution of the estimated treatment effects to examine whether the resulting $\widehat{\delta^{DD}}$ estimate provides an unbiased estimate of the ATT.

Figure 2ii shows suggests that the TWFE DiD estimate is unbiased for the true ATT (vertical red line) simulated from the data generating process in Simulation 1-3. These simulation results show that TWFE DiD estimates are unbiased in settings where there is a single treatment period, even when there are dynamic treatment effects. They also suggest that TWFE DiD estimates are unbiased in settings with *staggered* timing of treatment assignment and no treatment effect heterogeneity across firms or over time.

Next, we illustrate the conditions under which TWFE DiD produces biased estimates. 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.

In Simulation 4, we allow the ATT to be constant but differ across treatment-timing groups. Specifically, we draw each firm's treatment effect on the outcome variable from a normal distribution with mean $\delta_g \times \mathbb{I}[Treat] \times \mathbb{I}[t > g]$ and variance of (0.2^2), where $\delta_{1989} = 5$, $\delta_{1998} = 3$, and $\delta_{2007} = 1$. In Simulation 5, we allow for dynamic treatment effects and assume that the dynamic effects are the same across treatment timing groups. We draw each firm's treatment effect parameter on the outcome variable δ_i from a normal distribution with mean 0.3 and variance $(0.2)^2$, and the firm-year

treatment effect is equal to $\delta_i \times max(0, T-g)$ with a variance of $(0.2^2) \times max(0, T-g)$. Here, the simulation assumes an average *yearly* increase in the outcome variable in each year after firms receive treatment, and that these annual increases are the same (in expectation) across treatment timing groups. Finally, in Simulation 6, we again allow for dynamic treatment effects, but now allow the expected annual increases in the outcome variable to differ by treatment timing group. That is, we draw each firm's treatment effect parameter on the outcome variable δ_i from a normal distribution with mean δ_g and variance $(0.2)^2$, where $\delta_{1989} = 0.5$, $\delta_{1998} = 0.3$, and $\delta_{2007} = 0.1$. In this simulation, the firm-year treatment effect for firm i in treatment group G_g is equal to $\delta_i \times max(0, T-g)$ with a variance of $(0.2^2) \times max(0, T-g)$.

Figure 3i shows the outcome paths (gray lines) for the $N = 1,000$ firms in Simulation 4-6, with the colored lines corresponding to the average value of the outcome variable by treatment cohorts. As before, for each of our 500 simulated datasets we estimate and save the standard TWFE DiD regression. In Figure 3ii, the distribution of estimated treatment effects are plotted against the true ATT assumed in the data generating process (the red dotted line).

In each of the three simulations (Simulation 4-6), TWFE DiD treatment estimates differ from the true ATT (i.e., they are not centered around the vertical red line).[7] These simulations suggest that the combination of staggered treatment timing and treatment effect heterogeneity, either across groups or over time (i.e., dynamic treatment effects), leads to biased TWFE DiD estimates for the ATT. In fact, this bias can be so severe as to change the researcher's inferences about the direction of the true treatment effect. For example, although Simulations 4 and 5 lead to biased TWFE DiD estimates of the ATT, they preserve the correct treatment effect *sign* on average. However, Simulation 6 shows that, with heterogeneity in dynamic treatment effects across treated groups, the average estimated treatment effect is *negative* and statistically significant, even though the true effect on every treated group is positive in expectation.

The intuition behind these biases stems from the insight of Goodman-Bacon (2019): the stag-

[7] Note that in Simulation 4 the estimates differ from the "true effect" because of the variance weighting in OLS, whereas we calculate the true effect as being the sample weighted average. This is not necessarily a bias in the estimate, but a different way of aggregating the overall effect. Simulations 5 and 6 however differ from the true value from using past treated units as effective comparison units with dynamic treatment effects, and is a bias under any definition.

gered DiD TWFE approach is actually a "weighted average of all possible two-group/two-period DiD estimators in the data," and treatment-effect estimates are skewed by comparisons between earlier-treated to later-treated when there can be heterogeneity in the ATT. To expand on this intuition, we briefly provide the key insights of the derivation below. Readers seeking more insight on the econometric theory behind these results should refer to Goodman-Bacon (2019) for full details.

3.2 Staggered DiD Estimates and Constituent 2x2 DiD Comparisons

Goodman-Bacon (2019) decomposes $\widehat{\delta^{DD}}$ in a stylized setting where there are just three treatment groups: a never-treated group (denoted U), an early-treatment group (denoted k) that is treated at time t_k^*, and a late-treatment group (denoted l) that is treated at t_l^*. There are three sub-periods in this set up: the pre-period for group k (denoted $T1 = [0, t_k^* - 1]$), the middle period when group k is treated but group l is not (denoted $T2 = [t_k^*, t_l^* - 1]$), and the post-period for group l (denoted $T3 = [t_l^*, T]$). Assume without loss of generality that the true treatment effect is equal to 10 for group k and 15 for group l. Figure 4i depicts each group's dependent variable path over time.

The key question is how δ^{DD} from the TWFE estimation of Eq., (2) maps to the groups and times depicted in Figure 4i. Goodman-Bacon (2019) shows that in this three-group case, δ^{DD} is a weighted average of four possible 2x2 DiD comparisons, depicted in Figure 4ii, each of which can be estimated by Eq., (1) on the subsamples of groups and times.

The first two of the possible 2x2 DiD comparisons involve one treatment group (either the early or the later treated firms) and the untreated group, depicted in Panels A and B. In these cases, the TWFE estimate of δ^{DD} reduces to the standard 2x2 DiD shown earlier:

$$\hat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{T2+T3} - \overline{y}_k^{T1}\right) - \left(\overline{y}_U^{T2+T3} - \overline{y}_U^{T1}\right) \text{ and } \hat{\delta}_{lU}^{2x2} = \left(\overline{y}_l^{T3} - \overline{y}_l^{T1+T2}\right) - \left(\overline{y}_U^{T3} - \overline{y}_U^{T1+T2}\right).$$

The other two possible 2x2 DiD comparisons involve comparisons of the different treatment groups to each other, and do *not* include untreated units in the sample (Panels C and D). In these cases, the TWFE estimate of δ^{DD} is identified from the difference in the *timing* of the treatments

between treatment groups.[8] Panel C depicts one of these scenarios, in which we compare the early-treated firms to later-treated firms in a window (before t_l^*) in which later-treated firms do not receive treatment and the treatment status varies in the early-treated firms. Thus, the early-treated units (k) act as the treatment group and the later-treated units (l) effectively serve the role of a control group:

$$\hat{\delta}_{kl}^{2x2,k} = \left(\overline{y}_k^{T2} - \overline{y}_k^{T1}\right) - \left(\overline{y}_l^{T2} - \overline{y}_l^{T1}\right).$$

Panel D depicts the second of these scenarios, in which we compare the early-treated firms to later-treated firms in a window (after t_k^*) in which early-treated firms already received treatment and the treatment status varies in the later-treated firms. Now, the later-treated units acts as the treatment group and the early-treated units effectively serve the role of a control group:

$$\hat{\delta}_{k,l}^{2x2,l} = \left(\overline{y}_l^{T3} - \overline{y}_l^{T2}\right) - \left(\overline{y}_k^{T3} - \overline{y}_k^{T2}\right).$$

Generalizing from the above, in a research design with K different treatment timing groups, there are a total of K^2 constituent 2x2 DiD estimates. There are K^2-K timing-only constituent 2x2 DiD estimates comparing earlier and later-treated groups, and K constituent 2x2 DiDs involving treated and untreated groups. The weights on each of these 2x2 estimates used to construct $\widehat{\delta^{DD}}$ are functions of the absolute size of the subsample, the relative size of the treatment and effective comparison groups in the subsample, the timing of the treatment in the subsample, and the magnitude of the treatment variance in the subsample.

We highlight three main insights that follow from the Goodman-Bacon (2019) derivation. First, the TWFE estimate of $\widehat{\delta^{DD}}$ in a staggered DiD design is simply the weighted average of the constituent 2x2 DiD estimates. Second, in a significant subset of the constituent 2x2 DiD estimates, treatment group units can serve the role of effective comparison units (i.e., in (1-1/K)/2% of the constituent 2x2 comparisons), because their treatment assignment does not change over the relevant window. Third, the contribution of each constituent 2x2 DiD to the overall TWFE staggered DiD

[8] Note that whereas the constituent 2x2 DiDs involving treated and untreated units use the entire time period (Panels A and B), the other constituent 2x2 DiDs (the "timing-only" DiDs in Panels C and D) only use a portion of the available time periods: $\hat{\delta}_{k,l}^{2x2,k}$ only uses group l's pre-period, while $\hat{\delta}_{k,l}^{2x2,l}$ uses group k's post-period.

estimate is sample dependent. For example, because the weights applied to a constituent 2x2 DiD is greater when the size of the subsample or the magnitude of the treatment variance is greater, changing the panel length alone can change the staggered DiD estimate, even when each 2x2 DiD estimate $\widehat{\delta^{DD}}$ is held constant. Similarly, all else equal, constituent 2x2 DiD comparisons in which the treatment groups receive treatment closer to the middle of the panel receive greater weight, because the variance is the treatment indicator variable is larger. These seem to us to be normatively undesirable properties.

3.3 Staggered DiD Estimates and Average Treatment Effect on the Treated

It is also possible to analyze how the staggered DiD TWFE estimate relates to the average treatment effect on the treated (ATT), the usual estimand of interest in DiD studies. Goodman-Bacon (2019) and Callaway and Sant'Anna (2020) define the ATT for a timing group g (i.e., all firms that receive treatment during a certain period) at a point-in-time τ (called the "group-time average treatment effect") as

$$ATT_g(\tau) \equiv \mathbb{E}[Y_{i,\tau}(1) - Y_{i,\tau}(0)|g].$$

$ATT_g(\tau)$ 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 generalized formulation allows for heterogeneity in the ATT, either across groups (g) or over time (τ).

The TWFE DiD averages outcomes in pre- and post-periods, so we can re-define the average $ATT_g(\tau)$ in a date range W with T_W periods as:

$$ATT_g(W) \equiv \frac{1}{T_W} \sum_{t \in W} \mathbb{E}[Y_{i,t}(1) - Y_{i,t}(0)|g]$$

Goodman-Bacon (2019) derives the probability limit of the TWFE DiD estimator $\widehat{\delta^{DD}}$ (assuming T is fixed and N grows) using the constituent 2x2 DiD decomposition, as a simple combination of

three components:

$$\operatorname*{plim}_{N\to\infty} \hat{\delta}^{DD} = VWATT + VWCT - \Delta ATT. \tag{3}$$

VWATT is the "variance-weighted average treatment effect on the treated", which is just the positively weighted average of the ATTs for the units and periods that act as treatment groups across the 2x2 estimates that make up $\widehat{\delta^{DD}}$ (e.g., the ATT for the early- and later-treated groups in the 3-group example above). VWCT is the "variance-weighted common trend", which extends the parallel trend assumption for DiD to a setting with timing variation. VWCT is the average of the difference in counterfactual trends between pairs of groups and different time periods using the weights from the previous decomposition, and captures how differential trends map to bias in the $\widehat{\delta^{DD}}$ estimate. This term captures the possibility that different groups might not have the same underlying trend in outcome dynamics, which will inherently bias any DiD estimate.[9]

Finally, the last term ΔATT is a weighted sum of the *change* in treatment effects within each unit's post-period with respect to another unit's treatment timing. This term enters the coefficient estimate because already-treated groups act as effective comparison units for later-treated groups, and thus the 2x2 estimators (which subtract changes in the control units from changes in the treated units) will subtract *both* the average change in untreated outcomes *and* the treatment effect from earlier periods, assuming that the treatment effect takes more than one period to be incorporated in the outcome variable.

The decomposition of Eq., (3) suggest that biases can arise in the staggered DiD TWFE estimate in two ways, due to treatment effect heterogeneity over time or groups, even when the parallel trends assumption is satisfied ($VWCT = 0$). For example, if treatment effects vary across units, but not over time, then

$$\Delta ATT = 0 \text{ and } ATT_g(W) = VWATT = \sum_{g \neq U} ATT_g \times w_g^T,$$

where w_g^T is a function of the decomposition weights, and is a combination of sample shares and treatment variance. In general these weights are not equal to the sample shares, so $\widehat{\delta^{DD}}$ will not

[9] Although the importance of the parallel trends assumption has been long acknowledged, it is inherently untestable. Under the assumption that the parallel trends assumption holds, the VWCT term is 0.

equal the sample ATT.[10] The VWATT will give more weight to units treated towards the middle of the panel, so if the treatment effects during that period differ materially from other treatment effects, the coefficient could be biased for the sample-share-weighted ATT.

Second, and more importantly, the coefficient will be biased for the sample ATT when the treatment effect for treated units vary over time. That is, instead of a constant additive effect (where the outcome is shifted by a constant δ after treatment), there are dynamics to the treatment effect so that δ is a function of time elapsed since treatment.[11] In this case, time-varying treatment effects generate heterogeneity across the 2x2 DiDs and bias the estimates away from the VWATT because $\Delta ATT \neq 0$. With time-varying treatment effects, $\widehat{\delta^{DD}}$ uses already-treated units as effective comparison units and will yield estimates that are too small, or even wrong-signed if the early treated units have larger (in absolute value) treatment effects than later treated groups.

4 Diagnostics and Alternative Estimators

While the econometric literature has settled on the methodological challenge posed by TWFE estimation of DiD with staggered treatment timing, a number of alternative DiD estimation techniques have been proposed to circumvent the problem. In essence, each alternative estimator relies on event study DiD (rather than pooled DiD), which accommodates the possibility of dynamic treatment effects, but modifies the set of units that can act as effective comparison units in the estimation process. However, the proposed methods differ in terms of how the effective comparison units are selected and how covariates are dealt with. Below we describe a diagnostic to help researchers identify potential biases in staggered DiD designs. Then, we describe three alternative estimators that circumvent the biases that can arise in such designs. We illustrate each method using data from Simulation 6 of Figure 3, in which we found the largest biases in the staggered DiD treatment effect estimate ($\widehat{\delta^{DD}}$).

[10] As explained in Goodman-Bacon (2019), because TWFE uses OLS to combine 2x2 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 ATT.

[11] Dynamic treatment effects are likely in many research settings: most "event study" DiD estimates document post-treatment trends in the estimated effects. This might be less of a problem when using excess returns (although it would impact studies of post-earnings announcement drift), but is almost certainly a problem for non-stationary outcomes commonly used in the literature, such as valuation (i.e., Tobin's Q) or performance (ROA) variables.

4.1 Goodman-Bacon (2019) Diagnostic

In addition to the decomposition results above, Goodman-Bacon (2019) proposes a series of diagnostic tests to examine the robustness of the TWFE DiD estimate. In particular, Goodman-Bacon (2019) (see, e.g., Figure 6 in the paper) demonstrates that a useful diagnostic test for identifying potential biases in staggered TWFE DiD estimates is to plot the constituent 2x2 DiD estimates by each constituent comparison's implicit assigned weight (which is a function of treatment timing and group size) and constituent comparison's type (e.g., earlier vs. later treated states or later vs. earlier treated states).

Take for example, Simulation 6 from the prior section. Recall that, because every unit in Simulation 6 is incorporated in a state that eventually receives treatment, the identification of $\widehat{\delta^{DD}}$ is based on the variation in the timing of treatment: there are no never-treated units, since the variation in treatment status depends not on *whether* treatment was ever assigned but *when* treatment was assigned. Under dynamic treatment effects in such a setting, the potential for bias in $\widehat{\delta^{DD}}$ is greatest in the constituent 2x2 involving comparisons of later-treated firms (as treatment firms) to early-treated firms (as effective comparison units). This is because, around the period when the later-treated firms receive treatment, the changes in the later-treated firms' outcomes are relatively small compared to the contemporaneous changes in the outcomes of the early-treated firms.

Figure 5 display the diagnostic test suggested by Goodman-Bacon (2019) for Simulations 4, 5, and 6. For each of the six subgroups (recall that for 3 timing groups there are $3^2 - 3 = 6$ constituent 2x2 comparison groups), we plot the constituent 2x2 DiD estimate and its overall weight on $\widehat{\delta^{DD}}$.[12] Each of these points is represented by a marker symbol: the circle markers represent the constituent groups where earlier treated (as treatment firms) are compared to later treated (as effective comparison) units; the triangle markers represent the constituent groups where later treated (as treatment firms) are compared to earlier treated (as effective comparison) units. We also compare each of these constituent 2x2 DiD estimates to the true (simulated) ATTs of the underlying observations, which are represented by empty symbol markers and connected to the

[12] For the specific formula for computing these weights, see Eq. (10e), (10f), and (10g) of Goodman-Bacon (2019).

relevant constituent 2x2 DiD estimates by an arrow.

The graph shows that all the later vs. earlier treated comparisons yield *negative* estimated treatment effects (i.e., all the blue triangular points lie below zero), while all the earlier vs. later treated comparisons yield *positive* estimated treatment effects. This pattern is consistent with treatment effect heterogeneity producing biased TWFE DiD estimates in staggered treatment timing settings. The bottom panel of Figure 5 provides graphical intuition using a specific constituent 2x2 DiD grouping that compares firms treated in 2007 (as treatment firms) to firms that were treated in 1989 (as control firms) from 1989 to 2015. Under dynamic treatment effects, later vs. earlier treated comparisons yield negative effects because the large changes in the outcome for earlier treated firms, which are used as effective comparison units, are *subtracted* from the relatively smaller changes in the outcome for later treated firms, which are the effective treatment firms.

We note that the decomposition and diagnostic offered by Goodman-Bacon (2019) can at present only be used with balanced panels and do not incorporate covariates, which are atypical features of many corporate finance applications. Nevertheless, we believe that researchers should always analyze covariate-free variants of DiD analyses as starting points. Thus, we believe these tools should be broadly used by applied researchers interested in implementing DiD designs with staggered timing.

4.2 Alternative DiD Estimators

We now present three types of alternative estimators that have been suggested in the econometrics literature.[13] Although the field has not yet settled on an established standard, we believe that applied researchers leveraging settings in which TWFE DiD could be biased should implement at least one such method to test the robustness of inferences. To ensure that the estimation process is not contaminated by comparisons of late versus earlier treated firms, all three of these remedies suggest comparing the treated firms to a "clean" set of control firms. However, each remedy differs in terms of how these control firms are identified and used.

[13] An additional method not discussed in detail here is that from de Chaisemartin and D'Haultfœuille (2020). Their paper primarily discusses the one-period instantaneous effects, although they have implemented a multi-period version in Stata called **did_multiplegt**.

In general, these alternative methods rely on event study DiD designs, which can accommodate the possibility of dynamic treatment effects, and are implemented by including leads and lags of the treatment variable instead of a single binary indicator variable. That is, the TWFE regression with event-time indicators takes the following form:

$$y_{it} = \alpha_i + \lambda_t + \sum_k \delta_k \mathbb{I}[t - E_i = k] + \epsilon_{it}, \tag{4}$$

where y_{it} is the outcome variable of interest, α_i and λ_t are unit and time fixed effects, E_i is the time period when treatment begins for unit i, and $\mathbb{I}[t - E_i = k]$ is an indicator for being k years from the treatment starting.

We argue that in implementing event study analyses like Eq., (4), researchers should include the *full* set of relative-time indicator variables, excluding only the necessary number of relative time indicators to avoid multicollinearity. In general, we follow standard practice by excluding the relative time indicator for the period before treatment, so that the coefficients for the relative time indicators can be viewed as the mean differences from the average value of the outcome in the period before treatment. As noted in Borusyak and Jaravel (2017), when there are no never-treated units in the sample, two relative time indicators need to be omitted to avoid multicollinearity. In these cases, we also drop the most negative relative time indicator (i.e., the minimum k that appears in the dataset), so that the coefficients for the relative time indicators can be viewed as the mean differences from the an average value of the outcomes in two specific relative periods prior to treatment.

There are a number of advantages to the event study DiD estimator, and in general we suggest that it should be a part of any DiD analysis.[14] The results from the standard DiD capture the mean difference between groups of units before and after policy adoption. Because most studies involve settings with multiple time periods, raising the possibility of dynamic treatment effects that can

[14]We do note that event study DiD models will, in general, be less powered than binary indicator regressions, due to fitting an increased number of parameters. However, we believe that the advantages of tracing the timing of differences between treated and untreated units more than makes up for the loss of precision. It is always possible to combine groups of event study estimates (i.e. all relative indicators post-treatment) after fitting the model to increase power.

complicate the aggregated differences between post- and pre-treatment adoption outcomes. One important advantage of an event study DiD is elucidating the timing of treatment effects: it allows researchers to break down the average difference captured in $\widehat{\delta^{DD}}$ into the differences between treated and comparison units at each time period relative to the treatment adoption. Because each relative-time indicator is only turned on once for each unit, event study DiD designs help to resolve some of the variance-weighted averaging concerns described above. In addition, these designs also help researchers evaluate the credibility of the parallel trends assumption (i.e., by observing trends in the coefficients on pre-period relative time indicators).

4.2.1 Callaway and Sant'Anna (2018) Estimator

Callaway and Sant'Anna (2020) considers the identification and estimation of treatment effect parameters using DiD with multiple time periods, variation in treatment timing, and where the parallel trends assumption may only hold after conditioning on observables. When the treatment effect can differ by treatment groups and over time, there are numerous causal parameters of interest: the ATT is a function of treatment group g, where a group is defined by when units are first treated (e.g., firms in 2006 and firms treated in 2009 are in separate groups), and time period t. Callaway and Sant'Anna (2020) calls these causal parameters, denoted $ATT(g,t)$, "group-time average treatment effects," and proposes a two-step estimation strategy with a bootstrap procedure to conduct asymptotically valid inference that adjusts for autocorrelation and clustering. The methodology also allows for the estimation of aggregate treatment effects by either relative time (i.e., the event study approach) or by calendar time.

Following the notation in Callaway and Sant'Anna (2020), the inference problem is set up as follows. Assume there are T periods where $t = 1, \ldots, T$, with D_{it} a binary variable equal to 1 if a unit is treated and 0 otherwise. Define G_g to be a binary variable that is equal to 1 when a unit is first treated in period g, and C as a binary variable equal to 1 for never-treated units. For each unit, exactly one of $\{G_1, ..., G_T\}$ or C is equal to 1. Denote the generalized propensity score as $p_g(X) = P(G_g = 1 | X, G_g + C = 1)$, which is the probability that an individual is treated conditional on having covariates X and conditional on being a member of a group g or a control

group C.

Callaway and Sant'Anna (2020) show that, under these assumptions, the group-time average treatment effect can be semi-parametrically identified as:

$$ATT(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1})\right] \tag{5}$$

This is just a weighted average of the "long difference" in the outcome variable, with the weights depending on the propensity score which is normalized to sum to one. The intuition is to take observations from the control group and group g, omitting other groups, and then up-weight observations from the control group that have characteristics similar to those frequently found in group g and down-weight observations from the control group that are rarely in group g. Note that a control unit can either be one that never receives treatment, or one which has not yet received treatment by period t.[15] This re-weighting ensures that the covariates of the treatment and control group are balanced. The authors provide an open-source package that allows for inverse probability weighting (IPW) or doubly-robust methods in the estimator in addition to the standard regression approach that we focus on in Section 5 of the paper.[16] In the Appendix, we provide a stylized illustration of how the CS estimator works in the standard regression approach.

4.3 Sun and Abraham (2020) Estimator

Sun and Abraham (2020) focuses exclusively on the event-study context, which includes leads and lags of the treatment variable instead of a single binary indicator variable. Sun and Abraham (2020) proves that in the event study context, where the timing of treatment varies across units, lead/lag regressions can also produce causally uninterpretable results for similar reasons to those discussed above in the context of a binary indicator variable. Their proposed method estimates the dynamic effect for each treatment cohort (equivalent to group G_g from Callaway and Sant'Anna (2020)), and then calculates the weighted average of these cohort-specific estimates, with weights

[15] In both cases, the CS estimator is asympotically unbiased. However, using not-yet-treated control firms drops fewer observations and presumably has higher power to detect treatment effects.

[16] The packages is called **did** and is on CRAN. The "notyettreated" option in their program implements the version that uses not-yet-treated firms as effective comparison units.

equal to each cohort's respective sample share. Sun and Abraham (2020) focuses on the "cohort-specific average treatment effects on the treated" ($CATT$), k periods from initial treatment, which is conceptually similar to $ATT(g,t)$ of Callaway and Sant'Anna (2020).

The key theoretical result in this paper is that, even when using an event-study estimation technique rather than a single binary indicator variable, the coefficients on the TWFE lead/lag indicators may be biased, because the weights assigned to the different $CATT$s need not be positive without assuming treatment effect homogeneity. Specifically, the fixed effects estimands for k periods relative to treatment can be written as non-convex averages of not only the $CATT$ from that period, but also $CATTs$ from other periods. This is similar to the result in Goodman-Bacon (2019) that $\Delta ATT \neq 0$ with dynamic treatment effects, although the event study framework does solve the OLS variance-weighted issues brought up by Goodman-Bacon (2019) in the binary indicator context.

The proposed alternative estimation technique in Sun and Abraham (2020) uses an interacted specification that is saturated in relative time indicators D_{it}^k and cohort indicators $1\{G_g = g\}$ to estimate each $CATT_{g,k}$, which they call an "interaction-weighted" (IW) estimator. The DiD using the IW estimator is estimated simply by:

$$y_{it} = \alpha_i + \lambda_t + \sum_e \sum_{k \neq -1} \delta_{g,k}(1\{G_g = g\} \cdot D_{it}^k) + \epsilon_{it}. \tag{6}$$

Of note, when using the Sun and Abraham (2020) IW method, the only units used as effective comparison units are those that are never-treated (or the last treated group, which is then never used as an effective treated unit). The standard event-study DiD plots can be re-created by taking the weighted average over cohorts for relative time period k, with the weights equal to the share of each cohort in the relevant periods.

4.4 Stacked Regression Estimator

Another approach to estimating DiD with time varying treatments and treatment effect heterogeneity is "stacked regression". A published example of stacked regression is Cengiz et al. (2019),

which estimates the impact of minimum wage changes on low-wage jobs across a series of 138 prominent state-level minimum wage changes between 1979 and 2016 in the United States using a difference-in-differences approach. In Online Appendix D, Cengiz et al. (2019) notes that there are issues in aggregating discrete DiD estimates through OLS, and as a robustness check uses stacked-data regressions.

To do this, the authors create event-specific datasets, including the outcome variable and controls for the treated state and all other "clean controls" that don't have a material change to the state minimum wage within the eight year estimation window ($t = -3$ to $t = 4$). They then stack these event-specific data sets in relative time to calculate an average effect across all 138 events using a single set of treatment indicators. These stacked regressions are of the form:

$$y_{itg} = \alpha_{ig} + \lambda_{tg} + \sum_{k} \delta_k \mathbb{I}[t - E_i = k] + \epsilon_{itg}.$$

The only difference between this functional form and the standard event-study DiD estimand is that you need to saturate the unit and time fixed effects with indicators for the specific stacked dataset.[17]

As the authors note, this is an alternative to a baseline TWFE DiD estimate, but "uses a more stringent criteria for admissable control groups, and is more robust to possible problems with a staggered treatment design in the presence of heterogeneous treatment effects." By stacking and aligning events in event-time, this approach is equivalent to a setting where the events happen contemporaneously, and it prevents using past treated units as effective comparison units, which may occur with a staggered design. Moreover, by dropping all control states with any state-level minimum wage increases within a defined event window, this method guards against bias due to heterogeneous treatment effects that show up in the ΔATT term from Goodman-Bacon (2019).

[17] Unlike the Callaway and Sant'Anna (2020) estimator, stacked regression still uses OLS to weight treatment effects, with the attendant differences from sample-weighted averages. This could potentially be problematic when different stacked samples don't have coverage for the full treatment effect range set in the stacking process. As a result, we recommend that authors use, whenever possible, the estimator from Callaway and Sant'Anna (2020), which does not suffer from the same concerns. In addition, given that researchers control the length of the stacking window, many of these weighting issues can be amelioriated by changing the estimation range or dropping treatment groups without full observations over the specified range. Finally, in our experience these weighting issues are much less pronounced than those from standard TWFE models.

4.5 Simulation Example: Alternative Estimators

Figures 3i and 3ii demonstrates that the primary risk of bias from using TWFE DiD estimators comes from settings with staggered adoption and treatment effect heterogeneity. While the remedies discussed above in Goodman-Bacon (2019), Sun and Abraham (2020), Callaway and Sant'Anna (2020), and Cengiz et al. (2019) have been theoretically or conceptually justified, we use the data from Simulations 4-6 to demonstrate their effectiveness.

The results are presented in Figure 6, and show that each of the new proposed methods is able to approximately recover the true treatment path with staggered treatment timing and dynamic and heterogeneous treatment effects. The stacked regression approach generates estimates that are slightly biased above the true treatment effect average, which is likely a result of using OLS variance-weighting rather than weighting explicitly by the sample share as in Callaway and Sant'Anna (2020) and Sun and Abraham (2020).[18]

The results of this stylized simulation show how in settings with panel data and heterogeneous treatment effects that vary through time, the simple TWFE DiD estimate will be biased. This is of practical concern for most applied empirical work; treatment assignment is often staggered and bunched, and there is normally little reason to believe that treatment effects are homogeneous across time or units. In addition, the normal event study plot in applied work shows a post-treatment path characterized by gradual incorporation of treatment effects, and not a single discontinuous jump at the period of treatment. In these instances, the simple two-way differencing inherent to the TWFE DiD model will create a bias from using prior-treated units as effective comparison units, that will either shrink or even flip the sign of the treatment coefficient. However, all of the proposed models by Callaway and Sant'Anna (2020), Sun and Abraham (2020), and Cengiz et al. (2019) are sufficient to recover the true treatment path in the data.

[18] In addition, the estimates as presented in Figure 6 show tighter confidence intervals for the Sun and Abraham (2020) estimates. This is simply because my implementation of the Sun and Abraham (2020) estimator does not adjust for timing-group uncertainty, while the Callaway and Sant'Anna (2020) estimator does.

5 Applications

In this section, we examine three papers, covering a broad sample of empirical research topics in finance and accounting that use DiD analysis with staggered treatment assignment. We stress that each of these are methodologically strong papers, published in top journals with credible claims to identification and strong theoretical justifications and rationalizations. They were all published before the advent of the new literature on flaws in TWFE estimation, and our replications are not meant as any claim to flaws in the paper. Rather, we mean to show here that the limitations to TWFE DiD can create genuine issues of inference even for papers with strong designs and claims to causality.

For each paper, we first replicate a portion of the published results. We then provide diagnostic tests demonstrating the distribution of treatment timing, and if possible[19] use the decomposition method from Goodman-Bacon (2019) to test whether the aggregate treatment effect is driven by potentially biased or unbiased samples. Finally, we apply some of the remedial methods reviewed in Section 4.2 to test whether the published results are robust to DiD methods that correct for the biases induced by time-vary treatment assignment and treatment effect heterogeneity. In this section we focus exclusively on the Callaway and Sant'Anna (2020) estimator and the stacked regression approach, both to keep the number of analyses manageable, and because the Sun and Abraham (2020) and Callaway and Sant'Anna (2020) are conceptually very similar, with the exception that the latter allows for the inclusion of covariates.

5.1 Beck, Levine, and Levkov (2010)

Beck, Levine, and Levkov (2010) ("BLL") is one of many papers to analyze the effect of bank branching deregulation that occurred wholesale across the United States, staggered across time and occurring mostly between the 1970s and the 1990s.[20] Over this period, most states removed restrictions on interstate banking, and by the end of the period none continued to outlaw the practice. In this paper, the authors exploit the cross-state and intertemporal variation in deregulation to ana-

[19] The Goodman-Bacon (2019) decomposition is only possible as of now for balanced samples

[20] Dozens of papers have used the staggered rollout of bank deregulation as an identifying shock, although BLL is arguably the most influential such paper with over 500 Google Scholar citations at the time of writing.

lyze the implications of financial regulation on income inequality. Prior research had shown that national technological innovations, including the invention of the ATM, triggered branch deregulation at the state level, which reduced the monopoly power of local banks and weakened their ability and desire to fight against deregulation. However, anecdotal evidence and economic theory provide conflicting predictions for the distributional effects of bank deregulation.

Our replication and extension will focus on Table II and Figure III of the paper, which suggest that bank deregulation reduced income inequality using a DiD analysis with binary treatment indicator and event study specifications, respectively.[21] They create state-level Gini index measures using the March Supplement of the Current Population Survey from 1977 to 2007. Their sample includes prime age individuals (25-54) that have 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. While the authors use multiple measures of state-level inequality in their paper, for parsimony we focus on only one: the log of the Gini index. The dataset includes observations for 31 years and 48 states plus the District of Columbia, for a total of 1,519 observations.

5.1.1 Binary Indicator Variable

Table 2 presents the results of the DiD analysis using a binary indicator variable for state-year observations following the passage of bank deregulation.[22] As in BLL, we estimate the impact of bank deregulation as the coefficient estimate on δ^{DD} from the regression:

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

where the outcome variable is the natural logarithm of the Gini coefficient, measured at the state-year level, α_i and λ_t are firm and year fixed effects, and D_{it} is an indicator set to 0 before a state allows interstate bank branching, and 1 afterwards. The estimated coefficient $\widehat{\delta^{DD}}$ is the causal effect of deregulation on state level inequality, assuming the parallel trends assumption holds. In

[21] The data and code used to replicate these results are publicly available at https://dataverse.nl/dataset.xhtml?persistentId=hdl:10411/15996.

[22] The results of Table 2 are similar with and without time-varying covariates, and are broadly consistent across measures of inequality. As a result we focus on the results using the Gini index and without covariates for parsimony.

Table 2 we replicate the main finding of BLL: the point estimate and inference is identical to BLL, that bank deregulation reduced income inequality.

Next, we use the Goodman-Bacon (2019) diagnostic to decompose the aggregate estimate into its constituent components: i.e., the portion driven by comparisons in which early treated states are treated firms and later treated states are effective comparison units, and the portion driven by comparisons in which later treated states are treated firms and earlier treated states are effective comparison units (where the potential for bias is the greatest).[23] The overall ATT is a weighted average of the ATT for each of these two groupings (the total ATT estimate $\widehat{\delta^{DD}}$ is equal to the weighted average for early vs. late treated states times its total weight (0.005 $\times$ 0.143) and the weighted average for late vs. early treated states times its total weight (-0.027 $\times$ 0.857)).

The decomposition, reported in Figure 7, indicates reason for concern. BLL's documented negative effects of bank deregulation on the income distribution are driven by comparisons in which later treated states are the treatment states and earlier treated states are effective comparison units. In those constituent DiD groups comparing earlier treated firms (as the treatment states) to later treated states (as effective comparison units), with less concern about potential bias, we find that the DiD estimates are on average positive and close to zero in magnitude. Notably, Figure 7 suggests that the negative effect documented in BLL are driven by a small number of constituent 2x2 comparisons that produce negative ATT estimates and carry a large weights in the pooled OLS regression.

5.1.2 Event Study DiD

In addition to the single binary indicator approach to DiD, BLL also explores the "dynamics of the relation between deregulation and inequality" by including a series of dummy variables to "trace out" the yearly effects of deregulation on inequality. This approach, commonly referred to as an event study DiD, is reported in Figure III of BLL, which plots the coefficients and the standard errors of the event time indicators from the following regression:

[23] Figure 7ii graphically compares each 2x2 constituent DiD and its weight in the pooled OLS estimate across the two types of comparisons. The top panel of Figure 7 summarizes the data points in each panel by taking their weighted averages. These weighted averages are represented as horizontal red lines in Figure 7ii.

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

Instead of a single binary indicator (i.e., D_{it} in the previous specification), this specification uses 25 separate indicator variables for the years *relative* to the year of adoption: from $t-10$ to $t+15$.[24] The year of treatment, or relative year $t = 0$, is not included and is the implicit reference year for the other indicator coefficients. In addition, BLL present their results by subtracting the average of the pre-adoption coefficients from all of the plotted relative-time means, forcing the pre-adoption coefficients to be centered at zero. It is unclear what the justification for this procedure is, but we nevertheless replicate the results from the event study DiD in Figure 8, Panel A.

The event study results reported in Panel A of Figure 8 suggest a negative effect of deregulation on inequality. In the pre-adoption period, the coefficients on the relative time dummies are all centered around zero, and there is little evidence of differential trends between states that deregulate and those states that they are compared against. Following deregulation, there is an immediate and statistically significant negative effect that ultimately settles to around a 4% decline in the Gini index, which BLL argue represents about 60% of the variation of inequality after controlling for state and year fixed effects.

However, these results use the full panel of observations with surveys stretching from 1977 through 2007, even though all states deregulated by 1999 (see Figure 7i). There are, as a result, *no* effective control states that can identify a DiD for those observations. 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. As a result, the relative time indicators for those states do not have a natural interpretation within a DiD framework. We make three changes to the event study results from Figure III in Beck et al. (2010), which we report in Figure 8. In Panel B we plot the coefficients directly from the regression results, rather than subtracting the mean of the pre-adoption coefficients (this results in a simple shift of the plotted coefficients). Second, rather than "binning" the indicators at $t-10$ and $t+15$ as in Beck et al. (2010), we follow Sun and Abraham (2020) and Borusyak and Jaravel (2017) and include the full set of relative indicator

[24] The most negative and the most positive relative time indicators are also set to 1 for all years earlier than 10 years before adoption, or all years greater than 15 years post adoption, respectively.

variables in the estimating equation in Panel C.[25] Finally, in Panel D, to mimic the assumptions of Callaway and Sant'Anna (2020) and Sun and Abraham (2020), where prior treated units cannot act as comparison units for later treated ones, we remove all observations for states that deregulated before 1977 as well as observations after 1999 when every unit is treated.[26]

While the results in Panel B suggest a similar interpretation with the original event study plot, the plots in Panels C and D show that changing from binning the most negative and positive relative time indicators to fully saturating the time dummies by individual relative year leads to the opposite conclusion. Over the long run, banking deregulation is followed by an on-average increase in income inequality. However, both of the estimates exhibit significant evidence of pre-trends, suggesting that the parallel trends assumption underlying this DiD is likely not satisfied.

5.1.3 Alternative DiD Estimators

In this section we apply the remedies to correct for potential biases in TWFE DiD estimates. First, we implement the Callaway and Sant'Anna (2020) estimator in two ways: one which uses only the final states to deregulate as effective comparison units (Panel A), and one which uses future treated states as comparison units as well (Panel B).[27] The results are reported in Figure 9i, and generally indicate that there is no robust linkage between the deregulation of bank branching and inequality in the state-level income distribution.

In addition to the dynamic event study point estimates and confidence intervals provided by the estimator, we also report a rough measure of overall effect within different portions of the event window in the figures. Here we use the simple aggregation procedure from Equation 3.12 in

[25] As noted in both papers, you must omit *two* relative time indicators to avoid perfect collinearity, at least in staggered adoption designs with no never-treated units. As a result we drop the relative time indicators for the most negative relative time period and the period prior to treatment.

[26] In all subsequent extensions, we follow Callaway and Sant'Anna (2020) and Sun and Abraham (2020) by not allowing prior treated units to act as controls. In effect, all of the alternative estimators presented in this paper restrict the sample of potential control units in some manner to ensure valid comparisons in the constituent DiD analyses. 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 of the question being asked.

[27] The Callaway and Sant'Anna (2020) estimator can be used as either a regression, IPW, or doubly robust estimator. For purposes of general comparison across methods we focus on the standard regression-based approach in this paper, but IPW and doubly robust methods are appealing across a range of applications and should be considered for use by researchers.

Callaway and Sant'Anna (2020), which first takes the average of the individual $ATT(g,t)$ estimates across treatment groups and relative time periods e, and is a corollary to the ATT in the standard 2x2 DiD design. We provide these estimates, and the associated p-values for the pre- and post-treatment periods within our restricted window $[-5,+10]$ around treatment. Consistent with the visual evidence from the event study coefficients, when using future-treated units as comparison units there is no significant post-treatment change in the outcome variable after deregulation. Using only the final states to deregulate as comparison states there is, if anything, marginal evidence of an *increase* in inequality after deregulation, although this is driven by an anomalous increase late in the post-treatment period and the pre-treatment aggregated trends are statistically different from zero.

Finally, we apply stacked regression models to the state-year inequality data. We perform the stacking again in two approaches. In Panel A of Figure 9ii, we stack cohort-specific datasets that include observations from states that deregulate in a certain year (treated states) and all states that do not deregulate within 10 years (matched control states). In Panel B, we stack cohort-specific datasets that include all states that deregulate in that year (treated states) and all other state-year observations that are pre-treatment (matched control states).[28] 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, interacting the state and year fixed effects with stack-specific indicators.

We also provide summary values similar to the aggregated ATT estimates from the Callaway and Sant'Anna (2020) estimator. From each of the stacked datasets, we create pre- and post-treatment datasets that include all of the control units, and either the pre ($[-5,-1]$) or post ($[-1,10]$) treatment observations for the treated units. We then create a simple binary indicator treatment variable that is set to 1 for the treated units in every period except for the reference relative period (-1). We report the associated coefficient and p-value from the stacked regression using the binary treatment variable on both datasets. The stacked regression results in Figure 9ii

[28] For example, in the first stacking approach we would not include any observations from states that deregulated in 1990 for states that deregulate in 1985. In the second approach we would include the observations for states that deregulate in 1990, but only up until 1989.

again show little compelling evidence of significant changes in inequality following from banking deregulation.

In summary, the DiD results in Beck et al. (2010) suggest that the deregulation of bank branching across states in the latter half of the 20th century led to causally-induced decreases in state-level inequality. However, using more robust approaches suggested by recent econometrics literature, we conclude that there is no evidence of any relation between deregulation and inequality.

5.2 Fauver, Hung, Li, and Taboada (2017)

Fauver et al. (2017) (FHLT) analyzes the relation between board reforms and firm value using international data. While a long literature exists on the relationship between board governance practices and firm operating performance or value in the US, 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 that call for the separation of the chairman and CEO positions. The FHLT sample consists of firms in countries with a reform in the 1990-2012 period and with available stock price data, and data availability can vary by country.

FHLT's identification strategy relies on the staggered implementation of country-level board reforms and the variation in firm-level data. The main result is that firm value (as measured by Tobin's Q) increases on average following the reforms, and that the increase in value occurs on or after the board reforms become effective in the country. The paper also documents no differential trends in Q prior to the reform.

5.2.1 Binary Indicator DiD

We begin by replicating the main result of FHLT. Their main regression specification is of the form:

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

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 equaling 1 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 firm-year samples: one that restricts to $[t-5, t+5]$ years around the board reform ($t = 0$), and another that maintains the full sample of available observations between 1990 and 2012. Given that our modified DiD procedures require excluding potential control firms that face a board reform within the estimation window, we focus on the results that use the full data panel.[29] In addition, the paper uses two different effective dates for defining the board reform "treatment": one that uses the timing of the "major" board reforms, as defined by the authors, and another that uses the timing of the first board reforms. The top panel of Figure 10 depicts the timing of country-level board reforms, broken down by both major reforms and first reforms.[30]

Columns 1 and 2, Table 3, replicate the main results of Fauver et al. (2017) (i.e., Table 4B of their paper) that use the full data panel and using both reform definitions. In addition, in columns 3 and 4, we also produce the DiD estimates without covariates.[31] Consistent with FHLT, our replication shows that board reforms increase Tobin's Q. We obtain positive and statistically significant coefficients on *Post*, and the results are similar with (columns 1 and 2) and without (columns 3 and 4) the inclusion of covariate. Effect sizes using the timing of the first reforms are about 20% to 50% larger than those using the major reforms.

5.2.2 Alternative Estimators

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 (2019) diagnostic. However, we are unable to do so here because the Goodman-Bacon (2019) approach only works with balanced panels and FHLT's panel is highly unbalanced. Instead, we

[29] Note that even without doing a modified DiD analysis, restricting the years to the 11-year window around treatment leads to potentially underpowered results. While studies frequently look at windows around treatment for the indicator variables, by dropping all observations outside that window you reduce the set of potential control units with observations unaffected by past or pending board reforms, which is particularly problematic in a setting such as this where the authors have no never-treated units.

[30] Because we are using firm-level data, different countries receive different weights in the DiD (as a result of having different numbers of listed firms). The country weights are represented by the shading of the timing tiles, with darker tiles representing more firm-year observations.

[31] 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. The authors use the **areg** Stata command for fixed effects regressions, which gives different standard error estimates from either **reghdfe** in Stata or **felm** in R (which give identical results).

proceed by examining how FHLT's inferences are affected under alternative estimators, beginning with implementing an event-study DiD design and proceeding to the Callaway and Sant'Anna (2020) and stacked regression approaches. In these analyses, we focus on the specifications without the inclusion of covariates.

For the event study DiD analyses, we modify the panel dataset and exclude observations after the final treatment. When considering the first reforms as "treatment," all countries are treated by 2006. If we consider the major reforms as "treatment," all firms are treated by 2007. Thus, although the data in the study contains observations through 2012, there are no effective control units after those years for the DiD. In light of these timing considerations, we present the results of two event study specifications. In each, we remove observations after every firm is headquartered in a country with board reforms (2006 for the first reforms and 2007 for major reforms), and we set the relative-time indicators to zero for the firms in the last treated countries.[32]

Using the modified datasets, we estimate the following model:

$$Q_{it} = \alpha_i + \lambda_t + \sum_{k=min+1}^{max} \beta_k \mathbb{I}[t - E_i = k] + \epsilon_{it},$$

where again Q_{it} is Tobin's Q, α_i and λ_t are country and quarter-year fixed effects, E_i is the year of the reform for firm i, and $\mathbb{I}[t - E_i = k]$ is an indicator for being k years from the reform. Although FHLT also presents event study results, our specification differs in that we include the *full* set of relative-time indicator variables, excluding the most negative relative time indicator and the indicator for time $t = -1$ to avoid multicolliniearity.

In the bottom panel of Figure 10, we report the β_k coefficients estimates and their standard errors. We report the event study estimates for the major reforms in Panel A and the estimates for the first reforms in Panel B. Figure 10 suggests there is little evidence of an increase in firm value around board reforms.

We next examine the relation between board reforms worldwide and Tobin's Q using the Callaway and Sant'Anna (2020) estimator and a stacked regression approach. For parsimony, we focus

[32] Because we drop observations when all units are treated, these estimates both because of the alternative specification and because of sample differences. We confirm that sample composition does not entirely drive the difference; in unreported results, binary regressions on the modified sample produce similar results to the published estimates.

only on analyses that allow the greatest use of pre-treatment observations. To implement the Callaway and Sant'Anna (2020) estimator, we use future-treated firms as control units in their pre-treatment years. We keep only the $ATT(g,t)$ estimates within the estimation window (from $t-5$ to $t+5$ years here) from the estimator. The results, presented in Figure 11, again show no evidence of any impact of the reform on firm valuation, either when we consider major reforms (Panel A) or first reforms (Panel B) as the treatment definition.

For stacked regressions, we stack cohort-specific datasets that include all firms headquartered in countries that implemented board reforms in a treatment cohort year as well as all other pre-treatment firm-year observations. We keep only observations within the estimation window (from $t-5$ to $t+5$ years here) and fit an event study regression specification on the stacked dataset. To avoid multicollinearity, we omit the indicator for year $t-1$ and interact the firm and year fixed effects with stack-specific indicators. Figure 11 reports the results of the stacked regressions using both major reforms (Panel C) and first reforms (Panel D) as the treatment definition. These results are similar to the findings using the Callaway and Sant'Anna (2020) estimator.[33] In conclusion, using the alternative approaches suggested in the econometrics literature, we fail to find robust evidence that board reforms geared towards independence and CEO duality have a positive impact on firm valuation worldwide.

5.3 Wang, Yin, and Yu (2021)

Our last replication is the analysis of Wang et al. (2021) (WYY), which examines how share repurchases impacted firm operations—including investment, profitability, and firm value. The paper's results provide evidence informing the corporate governance debate, both in the U.S. and abroad, about the implications of share repurchases (Fried and Wang, 2019, 2021). In recent years, asset managers (Fink, 2015), leading corporate lawyers (Lipton, 2015), and senior politicians (Biden, 2016) have raised concerns that repurchases deprive firms of the capital needed for long-term investment. These concerns have led to proposals in the U.S. for limiting or banning open market

[33] While the aggregated p-value for the post-treatment observations using the first set of reforms is marginally significant at the 10% level, there is evidence of pre-treatment trends in the outcome variable that call that result into question.

repurchases (e.g., Senator Tammy Baldwin's *Reward Work Act* and Senator Sherrod Brown's *Stock Buyback Reform and Worker Dividend Act*).

To identify the causal effect of share repurchases on firm outcomes, WYY leverages the staggered legalization of open market share repurchases across 17 countries from 1985 to 2010, and studies the long-term outcomes of repurchasing firms: those that repurchased shares within a two-year window of share repurchase legalization. WYY's main results show that repurchasing firms experienced subsequent declines in investment—CAPEX and R&D—as well as declines in firm value, profitability, and innovation. For purposes of our exercise, we will focus on the effects of repurchases on CAPEX and R&D.

5.3.1 Binary Indicator Specification

We begin by replicating the main result of WYY, which is based on the following specification:

$$Invest_{ict} = \alpha_i + \lambda_t + \delta^{DD} Rep_{ct} + \gamma' \mathbf{x}_{ict} + \epsilon_{ict}.$$

The dependent variable is one of the measures of investment, CAPEX or R&D, measured at the firm-year level; α_i and λ_t are firm and year fixed effects; Rep_{ct} is an indicator set to 1 for firm-years in countries c that have legalized open market repurchases in place and 0 otherwise; and $\mathbf{x}_{ict}$ are time-varying covariates measured at the firm level. The sample includes only those firms that repurchased shares within the two-year window following legalization.

Table 4, columns 1, 2, 4, and 5, replicate the main results of Wang et al. (2021) (i.e., columns 5-8 of Table 5 in their paper). All of these DiD specifications include firm and year fixed effects, but differ based on whether a "short" or a "long" set of covariate controls are included. The "short" regression models (columns 1 and 4) control for total assets, net sales, net income, leverage, and return on assets (ROA); the "long" models (columns 2 and 5) include additional covariates for sales growth, net profitability (EBIT / Sales), investment intensity (PPE/Sales), the quick ratio, and market share.

In columns 3 and 6 of Table 4, we report the results of the TWFE DiD specification without the inclusion of covariate controls. For each investment variable, the estimates from this model

are consistent with those obtained from the short and long models.[34] (For brevity, our subsequent analyses focus on specifications that do not include covariates.) In all cases, the negative and significant coefficients suggest that share buybacks that resulted from the legalization of open market repurchases led to a significant decline in firm-level investment (as measured in CAPEX and R&D).

5.3.2 Alternative Estimators

As with FHLT, we are unable to apply the Goodman-Bacon (2019) diagnostic, due to the highly unbalanced panel of the WYY sample. We proceed by examining how WYY's inferences are affected by using alternative estimators, beginning with implementing an event-study DiD design and proceeding to the Callaway and Sant'Anna (2020) and stacked regression approaches.

Figure 12i reports the timing of the legalization of open market share repurchases across countries, where weights are represented by the shading of the timing tiles, with darker tiles representing more firm-year observations. There are 17 nations in the sample, all of which deregulate within the data panel, although many do not have firm-year observations in the beginning of the panel. The most represented countries in the data are Canada and Taiwan. By 2010, all 17 countries had deregulated.

We begin by providing event study DiD analyses of CAPEX and R&D. We modify the dataset in WYY to exclude all observations on or after 2010, the final legalization year: all observations in the sample after this point are treated, and following Callaway and Sant'Anna (2020) and Sun and Abraham (2020) these observations cannot serve as valid comparison units to help identify treatment effects. Using the modified dataset, which sets firms in Kuwait (the last liberalizing country) to be effectively untreated comparison units in the sample, we estimate the following model:

$$Invest_{ict} = \alpha_i + \lambda_t + \sum_{k=min+1}^{max} \beta_k \mathbb{I}[t - E_i = k] + \epsilon_{itc},$$

where $Invest_{ict}$ is the firm-year investment measure (i.e., CAPEX or R&D); α_i and λ_t are again

[34] In untabulated results we also confirm that the estimates reported in Table 4 remain virtually unchanged when estimated on a common set of observations, consistent with the choice of covariates having minimal impact on the treatment effect estimates.

country and year fixed effects; E_i is the year of reform for firm i located in country c; and $\mathbb{I}[t-E_i = k]$ is an indicator for being k years from the legalization of open market share repurchases. Although WYY also presents event study results, our specification differs in that it fully saturates the model in relative-time indicator variables, excluding the most negative indicator and the indicator for relative time $t = -1$ to avoid multicolliniearity.

The event study coefficients for relative time periods $k \in \{-5, +5\}$ are reported in Figure 12ii, with Panel A reporting results for CAPEX and Panel B reporting results for R&D expenditures. Unlike the results for the TWFE binary indicator estimates reported in Table 4, the event study estimates indicate an increasing in trend in investment around legalization, although these plots also suggest pre-legalization differences in investment trends between treated and comparison firms.[35]

Next, we test the relation between repurchases and investment using the Callaway and Sant'Anna (2020) estimator and stacked regression DiD approach. As before, for parsimony we focus on analyses that allow the greatest use of pre-treatment observations. Thus, we use future-treated firms as control units in their pre-treatment years, and aggregate only the $ATT(g,t)$ estimates within the estimation window (from $t-5$ to $t+5$ years here) from the estimator. The aggregated ATT estimates for CAPEX and R&D, reported in Panels A and B, Figure 13, continue to show little evidence of a decline in firm investment following the legalization of open market repurchases.

To implement stacked regressions, we stack cohort-specific datasets that include all firms in countries that legalized repurchases in a treatment cohort year as well as all other pre-treatment firm-year observations as comparison units. We keep only observations within the estimation window (from $t-5$ to $t+5$ years here) and fit an event study regression specification on the stacked dataset. To avoid multicollinearity, we omit the indicator for year $t-1$ and interact the firm and year fixed effects with stack-specific indicators. The stacked regression event study estimates for CAPEX and R&D, reported in Panels C and D, Figure 13, again do not suggest a systematic negative shift in firm investment. Thus, after correcting for the use of prior treated firms as comparison units in staggered DiD designs, the empirical evidence does not support the conclusion that the

[35] Wang et al. (2021) also report event study results, which differ markedly from ours. These differences stem from WYY's omission of pre-deregulation relative time indicators more than two years before treatment as well as the choice to bin all treatment indicators greater than five years following deregulation.

legalization of open market repurchases significantly lowered (or had any impact on) repurchasing firms' investing behavior.

6 Robust Inference

Although generally beyond the scope of the article, in this section we mention some additional considerations regarding inference in modified DiD designs. Most of the alternative estimators considered in this paper use event study designs rather than single parameter estimators. As mentioned in Section 4.2, we believe that the case for event study DiD estimation is strong, and should be a primary study design used by researchers conducting DiD analyses of policy shocks. However, inference on treatment paths and/or aggregate effects is not straightforward, and researchers should be aware of considerations and advancements in this area. In this section we briefly discuss two such concerns: simultaneous confidence intervals and set identification in the presence of parallel trend violations.

In the replications and analyses in this paper we follow standard practice and report pointwise confidence intervals for the relative time indicators in event study designs. However, given the multiple hypothesis testing inherent to studying numerous relative-time treatment effect parameters, such pointwise intervals do not asymptotically cover the whole path of the group-time average treatment effects with a fixed probability. As a result, the confidence intervals for such estimates will typically be too small, and fail to achieve simultaneous coverage. A number of alternative confidence bands have been proposed in the literature, including Bonferroni, Šidák, and projection, although sup-t bands are typically preferred in such settings (Olea and Plagborg-Møller, 2019). Simultaneous confidence intervals for event study designs have been advocated by both Freyaldenhoven, Hansen, and Shapiro (2019) and Callaway and Sant'Anna (2020), and are generated by their estimator.

Another active area of research regarding inference and DiD designs is sensitivity analysis that allows for possible violations of parallel trends. As noted earlier, although the parallel trends assumption is generally un-testable in practice, standard methods for inference with DiD are valid only

when it holds. Building upon Manski and Pepper (2018), Rambachan and Roth (2020) proposes a methodology that provides valid inference under weaker assumptions and sensitivity analyses with respect to those assumptions. With this method a researcher imposes that violations of the parallel trends assumptions are restricted to a set Δ, which typically allows the effect estimates to be set-identified even where point-identification is unwarranted. These restrictions can encompass a range of intuitions already present in DiD practice, such as whether the presence and magnitude of pre-treatment differences between treated and untreated units is probative of post-treatment counterfactual differences in trends. Researchers can report how the range of consistent estimates changes with assumptions on the sign and shape of trend differences motivated (hopefully) by context-specific information. In addition, this bounding approach can be used to test whether and how much DiD estimates depend on the functional form of the outcome (Roth and Sant'Anna, 2020).[36]

7 Conclusion and Recommendations

We have shown that the commonly used TWFE DiD specification is susceptible to biased estimates, both because of the variance-weighting implicit in ordinary least squares, and more importantly due to the embedded use of past treated units as effective comparison units for later-treated units. Using a simulation analysis, we showed how this bias arises, and the ease with which it can contaminate estimates under even a straightforward data generating process. Finally, we show how these concerns are not merely conceptual, but impact estimates in published studies, particularly those without never-treated units to act as comparison units. We conclude by providing a set of practical recommendations for applied researchers interested in exploiting staggered DiD settings for causal inference.

1. Researchers should provide a graphical depiction of the treatment timing of their DiD indicator variable. The distribution of treatment is a key ingredient to any staggered DiD design, and the use of TWFE estimation often masks a lack of variation in treatment timing used to

[36]The authors provide an open-source program to implement such sensitivity analyses. This program is called HonestDiD and is available at https://github.com/asheshrambachan/HonestDiD.

compare treated and control units.

2. Settings in which treatment timing varies across a long period of time are more susceptible to biases that arise from dynamic treatment effects, and therefore requires further diagnostic and robustness tests.

3. DiD estimates should always be presented, at least once, without the inclusion of covariates. Since at least Meyer (1995), it has been shown that including covariates in a linear fashion is inappropriate if the treatment has different effects across subgroups in the population. Even new DiD methods that allow for parallel trends to hold only after conditioning on covariates (e.g., Abadie (2005) and Callaway and Sant'Anna (2020)) do not argue for including post-treatment control variables. It should be made clear whether the results are driven by the inclusion of controls.

4. If possible (i.e., if the data is a balanced panel), the binary indicator DiD estimates should be broken down using the Goodman-Bacon (2019) diagnostic (e.g., similar to Figure 5). This will show the percentage of the estimate that is driven by different types of treatment timing comparisons, and how the weighted average ATT in each group differs. Given what we know about the risks of using past-treated units as effective comparison units, reporting these decomposition results will increase the credibility of TWFE DiD estimates.

5. Binary indicator DiD estimates should be accompanied by event study estimates tracing out the timing of outcome differences between treated and control units. While this has become increasingly common in practice, the manner of conducting event study DiD varies widely across papers. We suggest fitting the full set of possible relative time indicator variables in the event study DiD, even if only reporting a subset of time indicators of interest.

6. With events study estimators (and stacked regression event study estimators in particular), the length of time in the event windows impacts the DiD estimator. This is a design choice that should be defended by the researcher and should be guided by the specific research question and institutional knowledge.

7. The modified DiD methods of Callaway and Sant'Anna (2020), Sun and Abraham (2020), and the stacked regression approaches all focus on generating unbiased estimates of DiD treatment effects by being explicit about the pool of acceptable control units. However, they produce marginally different estimates under marginally different assumptions. We propose that authors present results from a range of methods without time-varying covariates to show the robustness of results.[37]

8. To the extent that time-varying post-treatment covariates do need to be included as controls, based on theoretical justification or for robustness purposes, they can be included readily through stacked regression models, which maintain similar assumptions to the standard TWFE estimate while preventing prior treated units from acting as effective comparison units.

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

[37] Note that this does not have to require a large number of alternative estimates, as the Callaway and Sant'Anna (2020) and Sun and Abraham (2020) estimators are numerically identical without covariates for the post-treatment indicators.

References

Abadie, A. (2005). Semiparametric Difference-in-Difference Estimators. *Review of Economic Studies' 72*, 1–19.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics*. Princeton University Press.

Athey, S. and G. W. Imbens (2018). Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption.

Beck, T., R. Levine, and A. Levkov (2010, 10). Big bad banks? The winners and losers from bank deregulation in the United States. *Journal of Finance 65*(5), 1637–1667.

Biden, J. (2016). How Short-Termism Saps the Economy. *Wall Street Journal, September 27, http://www.wsj.com/articles/how-short-termism-saps-the-economy-1475018087*.

Borusyak, K. and X. Jaravel (2017). Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume *.

Callaway, B. and P. H. Sant'Anna (2020). Difference-in-Differences with multiple time periods. *Journal of Econometrics*.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019, 8). The Effect of Minimum Wages on Low-Wage Jobs*. *The Quarterly Journal of Economics 134*(3), 1405–1454.

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

Fauver, L., M. Hung, X. Li, and A. G. Taboada (2017, 7). Board reforms and firm value: Worldwide evidence. *Journal of Financial Economics 125*(1), 120–142.

Fink, L. (2015). Letter To Ceos. *March 31, http://www.shareholderforum.com/access/Library/20150331_BlackRock*

Freyaldenhoven, S., C. Hansen, and J. M. Shapiro (2019). Pre-Event Trends in the Panel Event-Study Design. *American Economic Review 109*(9), 3307–3338.

Fried, J. M. and C. C. Wang (2021). Short-Termism, Shareholder Payouts, and Investment in the EU. *European Financial Management*, 1–25.

Fried, J. M. and C. C. Y. Wang (2019). Short-Termism and Capital Flows. *Review of Corporate Finance Studies 8*(1), 207–233.

Goodman-Bacon, A. (2019). Difference-in-Differences With Variation in Treatment Timing.

Imai, K. and I. S. Kim (2020). On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data. Technical report.

Karpoff, J. M. and M. D. Wittry (2018). Institutional and Legal Context in Natural Experiments: The Case of State Antitakeover Laws. *Journal of Finance 73*(2), 657–714.

Lipton, M. (2015). Some Thoughts For Boards Of Directors In 2016. *December 9, Harvard Law School Forum on Corporate Governance and Financial Regulation, https://corpgov.law.harvard.edu/2015/12/09/some-thoughts-for-boards-of-directors-in-2016/*.

Manski, C. F. and J. V. Pepper (2018). How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics 100*(2), 232–244.

Meyer, B. D. (1995). Natural and Quasi-Experiments in Economics. *Journal of Business and Economic Statistics 13*(2), 151–161.

Olea, J. L. M. and M. Plagborg-Møller (2019). Simultaneous confidence bands: Theory, implementation, and an application to SVARs. *Journal of Applied Econometrics 34*(1), 1–17.

Rambachan, A. and J. Roth (2020). An Honest Approach to Parallel Trends. *Working paper*.

Roth, J. and P. H. Sant'Anna (2020). When is parallel trends sensitive to functional form?

Strezhnev, A. (2018). Semiparametric weighting estimators for multi-period difference-in-differences designs.

Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* (xxxx), 1–25.

Wang, Z., Q. E. Yin, and L. Yu (2021). Real effects of share repurchases legalization on corporate behaviors. *Journal of Financial Economics 140*(1), 197–219.

Appendix: CS Estimator in Event-Study DiD Setting

We present a stylized example to show how the Callaway and Sant'Anna (2020) (CS) estimator works for regression-based DiD. Although the CS estimator can accommodate the inclusion of covariates, for clarity we focus on the simple setting without the inclusion of covariates.

Assume that a researcher desires to test the causal effect of a treatment using a DiD estimator, and she observes the following data on the outcomes of four units (A, B, C, and D) over six periods.

Time	A	B	C	D
1	2	2	4	3
2	2	2	1	3
3	4	4	**5**	4
4	2	4	**2**	1
5	3	2	**5**	**4**
6	1	4	**2**	**6**

In this example, A and B never receive treatment; C receives treatment in period 3; and D receives treatment in period 5 (treated observations are bolded and underlined and treatment). Here we have two treatment groups: G_3 (unit C) and G_5 (unit D). In what follows, we work in relative event time and, for notational convenience, re-center $ATT(g,t)$ around the treatment period period g to $ATT(g,k)$, where $t = g + k$. Thus, the ATT for a cohort in the first year of treatment is denoted $ATT(g,0)$.

We show how one of the ATT parameters is estimated, focusing on $ATT(G_3,1)$ or the ATT estimate for G_3 (unit C) in year 4. For simplicity, we illustrate the "never-treated" approach in constructing the CS estimator, where the only comparison units are drawn from units that never receive treatment (Units A and B). To estimate $ATT(G_3,1)$, we first calculate the "long-difference" in the outcome for each unit: between relative time period e (here time period 4) and the reference time period. Reference periods are chosen from the pre-period sample: for post-treatment observations (i.e., $ATT(g,k)$ for $k \geq 0$), the CS estimator uses the year before treatment, or year 2 for unit C.[38]

In estimating $ATT(G_3,1)$, the relevant long-differences for the treated and potential control units are: $\Delta A = 2-2 = 0$, $\Delta B = 4-2 = 2$, and $\Delta C = 2-1 = 1$. Without covariates, $ATT(G_3,1)$ is simply the difference in means between these long differences in the treated and

[38] Because of the need for pre-period reference points in estimating $ATT(g,k)$, for G_3, there are two pre-treatment periods but only one pre-treatment causal estimand of interest ($ATT(G_3,-1)$) that can be identified. For pre-treatment observations (i.e., $ATT(g,k)$ for $k < 0$), the CS estimator uses the lagged relative time period as the reference, which is missing in time period 1. On the other hand, four post-treatment ATTs for G_3 ($ATT(G_3,0),...,ATT(G_3,3)$) can be identified. For G_5, on the other hand, there are three pre-treatment ($ATT(G_5,-1), ATT(G_5,-2), ATT(G_5,-3)$) and two post-treatment ($ATT(G_5,0), ATT(G_3,1)$) ATTs that can be identified.

control observations: $\Delta C - \frac{\Delta A + \Delta B}{2} = 1 - \frac{0+2}{2} = 0$.[39] Similarly, without covariates, $ATT(G_3, -1)$ can be obtained as $\Delta C - \frac{\Delta A + \Delta B}{2} = (1-4) - \frac{(2-2)+(2-2)}{2} = -3$, suggesting the presence of a pre-period trend.

With the set of $ATT(g,k)$ estimates, the overall aggregate effect in relative event time is calculated by averaging all of the estimates for each relative time period, weighting by the sample share of each group. In the above example, we have two different ATT estimates for relative time period $k = 1$: $ATT(G_3, 1)$ and $ATT(G_5, 1)$. Thus, an estimate of the ATT for one period after the treatment ($k = 1$) would be the average of $ATT(G_3, 1)$ and $ATT(G_5, 1)$. These can be estimated in regression form: without covariates, the regression simply yields group means, but covariates can also be included in the outcome regression. Standard errors of the relative event period ATTs are calculated using the influence functions from the underlying regressions and a multiplier-type bootstrap procedure.

While this is a highly stylized example of how the Callaway and Sant'Anna (2020) estimator works, it reflects the intuitive properties of the estimator and shows how it avoids the problems identified in the literature. By clearly specifying the potential control units, the estimator prevents prior treated units, and their potentially dynamic treatment effects, from impacting the group-time average treatment effects.

[39] An alternative approach, which can be implemented using the "notyetreated" option from the **did** package on CRAN to compute the CS estimator, is to also include as control units those observations that receive treatment in later years but are not yet treated in relative time period e. In the above example, this would mean that we would include the long difference for Unit D in the estimate for $ATT(G_3, 1)$, since D is not yet treated by time period 4. The new estimate for $ATT(G_3, 1)$ would then be $1 - \frac{0+2-2}{3} = 1$.

Fig. 1. Graphical Illustration of DiD

Figure 1 depicts the steps taken in conducting a standard difference-in-differences analysis in the simple two-period/two-units case. In this stylized example, the treated unit receives treatment between the Pre and Post period, while the control unit does not receive treatment. Panel A presents hypothetical trends in the outcome variable for both groups. The treated unit increases from a value of 2 in the pre-treatment period, to 5 in the post-treatment period, whereas the control unit only increases from 1 to 2. Panels B and C show how the "double-differencing" process works under an assumption of parallel trends. In panel B we compute the trends for each group by subtracting the pre-treatment level from each unit. The first difference for the treated group is equal to 3 (5 - 2), while the corresponding difference is 1 (2 - 1) for the control unit. If we assume that the treated unit and the control unit would have experienced the same trend in the outcome without the treatment, then the difference in these differences is an unbiased estimate of the treatment effect, shown in Panel C.

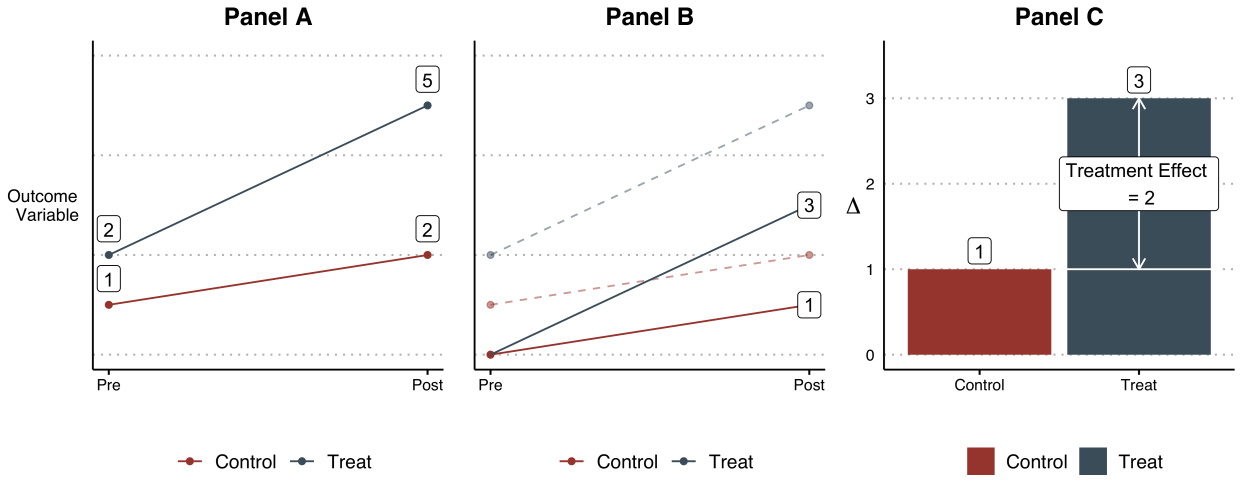


Fig. 2. Simulation: TWFE DiD Estimates Under Uniform Treatment Timing or Treatment Effect Homogeneneity

Panel (i) plots the outcome path by unit in the simulated data. The outcome is generated as the sum of firm and year fixed effects, drawn from $\mathcal{N}(0, 0.5^2)$, a treatment effect, and a random noise term $\epsilon_{it} \sim \mathcal{N}(0, 0.5^2)$. The treatment effects are either constant (Simulations 1 and 3) or time-varying (Simulation 2). In addition, Simulations 1 and 2 only have one treatment period, and a set of firms receiving the treatment (T), and a set not receiving treatment (C). In Simulation 3 all firms receive treatment, and are randomly assigned to one of three treatment groups with treatment beginning in 1989, 1998, or 2007. Panel (ii) plots the distribution of treatment effect estimates $\widehat{\delta^{DD}}$ from 500 Monte Carlo simulations of our three different data generating processes. The curve represents the distribution of the estimates, while the red vertical line is the true treatment effect imputed in the data. The standard TWFE DiD estimator is unbiased for all three simulations.

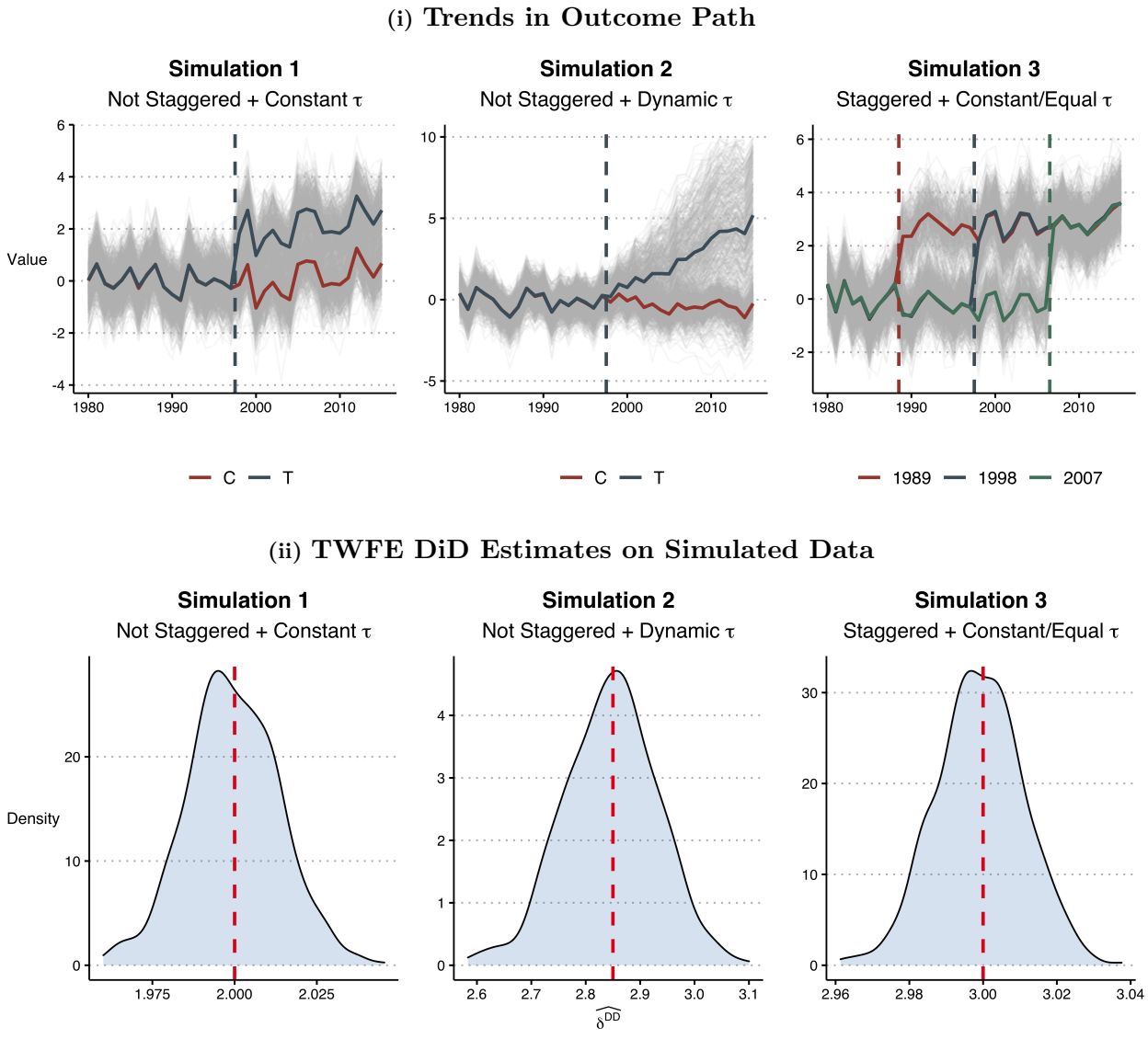


Fig. 3. Simulation: TWFE DiD Estimates Under Staggered Timing and Treatment Effect Heterogeneity

Panel (i) plots the outcome path by unit in the simulated data. The outcome is generated as the sum of firm and year fixed effects, drawn from $\mathcal{N}(0, 0.5^2)$, a treatment effect, and a random noise term $\epsilon_{it} \sim \mathcal{N}(0, 0.5^2)$. Each simulation in Figure 3 uses a staggered treatment design where firms are assigned randomly to one of the groups of states, receiving treatment in either 1989, 1998, or 2007. The treatment effects are either constant (Simulation 4) or dynamic and time-varying (Simulations 5 and 6). In addition, the treatment effects either vary across groups (Simulations 4 and 6) or are equal across groups (while being dynamic over time *within* a given firm) (Simulation 5). Panel (ii) plots the distribution of treatment effect estimates $\widehat{\delta^{DD}}$ from 500 Monte Carlo simulations of our three different data generating processes. The curve represents the distribution of the estimates, while the red vertical line is the true treatment effect imputed in the data. The standard TWDE DiD estimator is now different from the true average treatment effect for all three simulations. In Simulation 4, $\widehat{\delta^{DD}}$ does not equal the sample weighted average, because OLS weights by variance. The $\widehat{\delta^{DD}}$ estimates for Simulations 5 and 6 are much further away from the true value, because the use of prior treated groups as effective comparison units in the presence of dynamic treatment effects heavily biases the coefficients towards 0. In Simulation 6, because earlier treated groups have larger dynamic treatment effects than later ones, the estimates are even of the wrong sign of the true value (the red dotted line).

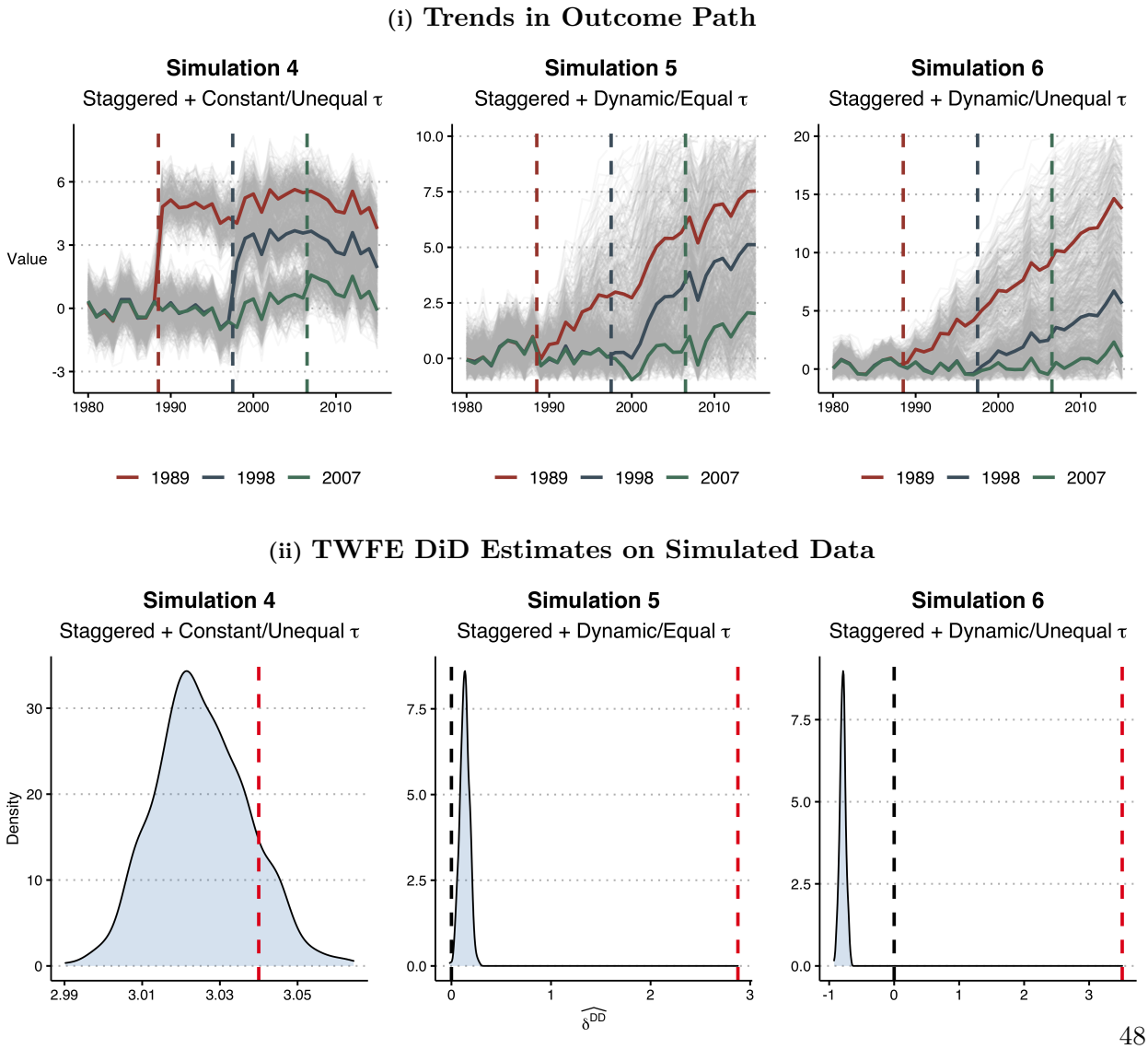


Fig. 4. Staggered DiD: A Three-Group Example

Figure 4 presents the stylized example from Goodman-Bacon (2019) which decomposes $\widehat{\delta^{DD}}$ in a setting with just three treatment groups: a never-treated group (denoted U), an early-treatment group (k) that is treated at time t_k^*, and a late-treatment group (l) that is treated at t_l^*. There are three sub-periods in this set up: the pre-period for group k (denoted $T1 = [0, t_k^* - 1]$), the middle period when group k is treated but group l is not (denoted $T2 = [t_k^*, t_l^* - 1]$), and the post-period for group l (denoted $T3 = [t_l^*, T]$). Here the true treatment effect is equal to 10 for group k and 15 for group l. Figure 4i depicts each group's dependent variable path over time, and Figure 4ii shows the constituent 2x2 DiD estimates, from which $\widehat{\delta^{DD}}$ is just a weighted average.

(i) Staggered treatment setting with three treatment groups.

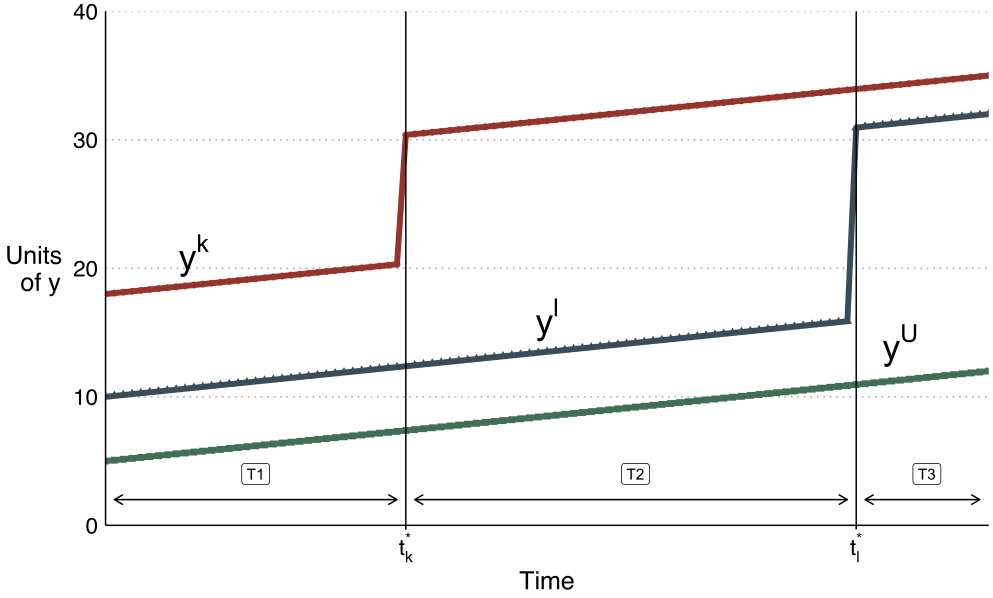


Fig. 4. [Continued]

(ii) **Four constituent 2x2 designs.**

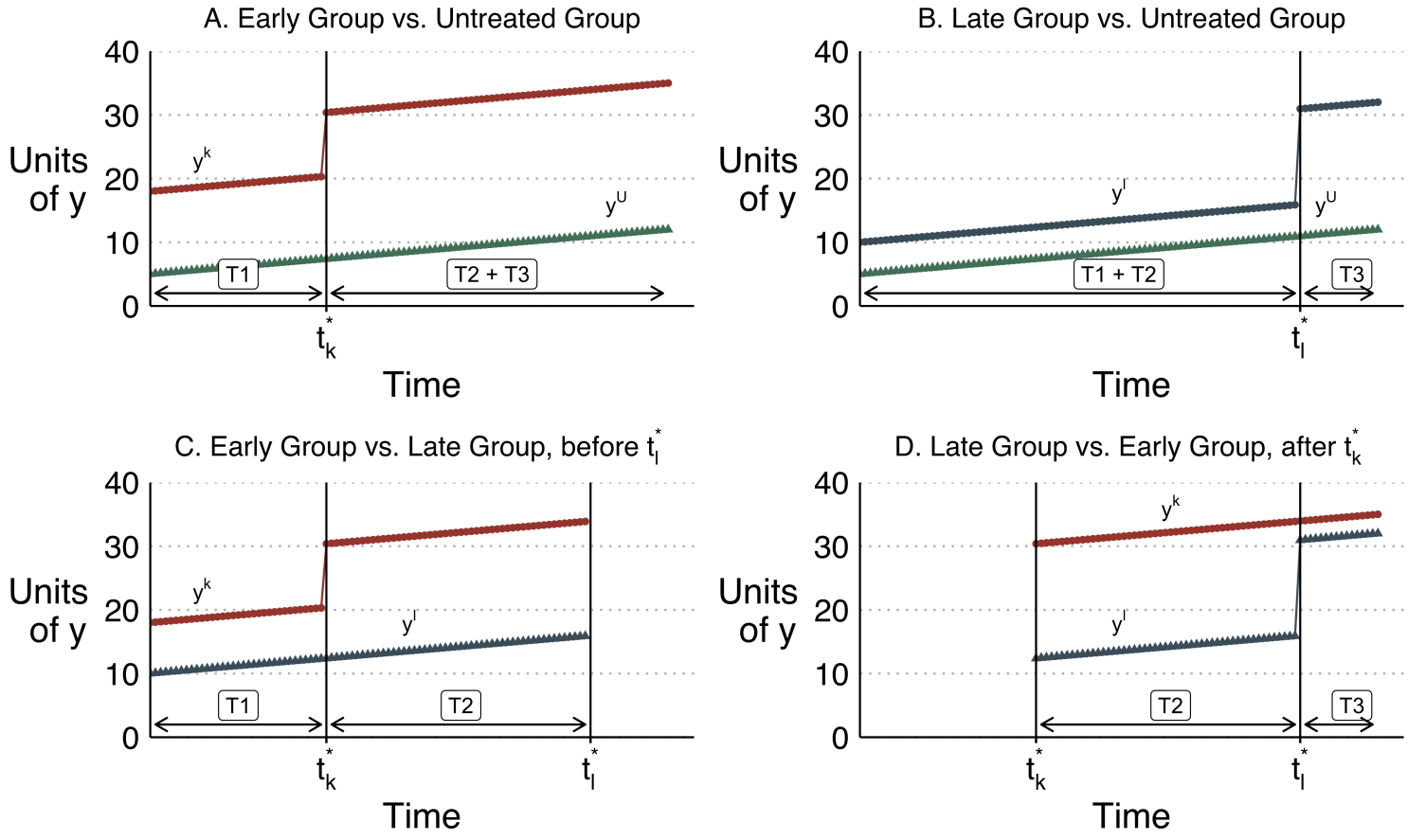


Fig. 5. Simulation: Diagnostics

Figure 5 plots the implicit weight given to each 2x2 treatment cohort comparison in the simulated data, and the associated estimate of the treatment effect. The red dots represent the unbiased estimates using later treated units as effective comparison units, while the blue triangles are the biased estimates using prior treated units as effective comparison units for future treated units. The shapes without a fill show the *true* value in the data. In Simulation 4 we see that all of the constituent 2x2 estimates are unbiased, as predicted, but that the variance weighted estimates move the aggregate measure away from the sample-share based estimate. In Simulations 5 and 6, the Earlier v. Later treated estimates still equal the treatment effect in the data, but the Later v. Earlier treated comparisons are highly negatively biased. The negative estimates from comparisons of prior treated units as effective comparison units drive the overall negative estimate $\widehat{\delta^{DD}}$ using Simulation 6. In the bottom panel we show graphically one constituent 2x2 comparison—a comparison of firms treated in 2007 to firms treated in 1989 in Simulation 6.

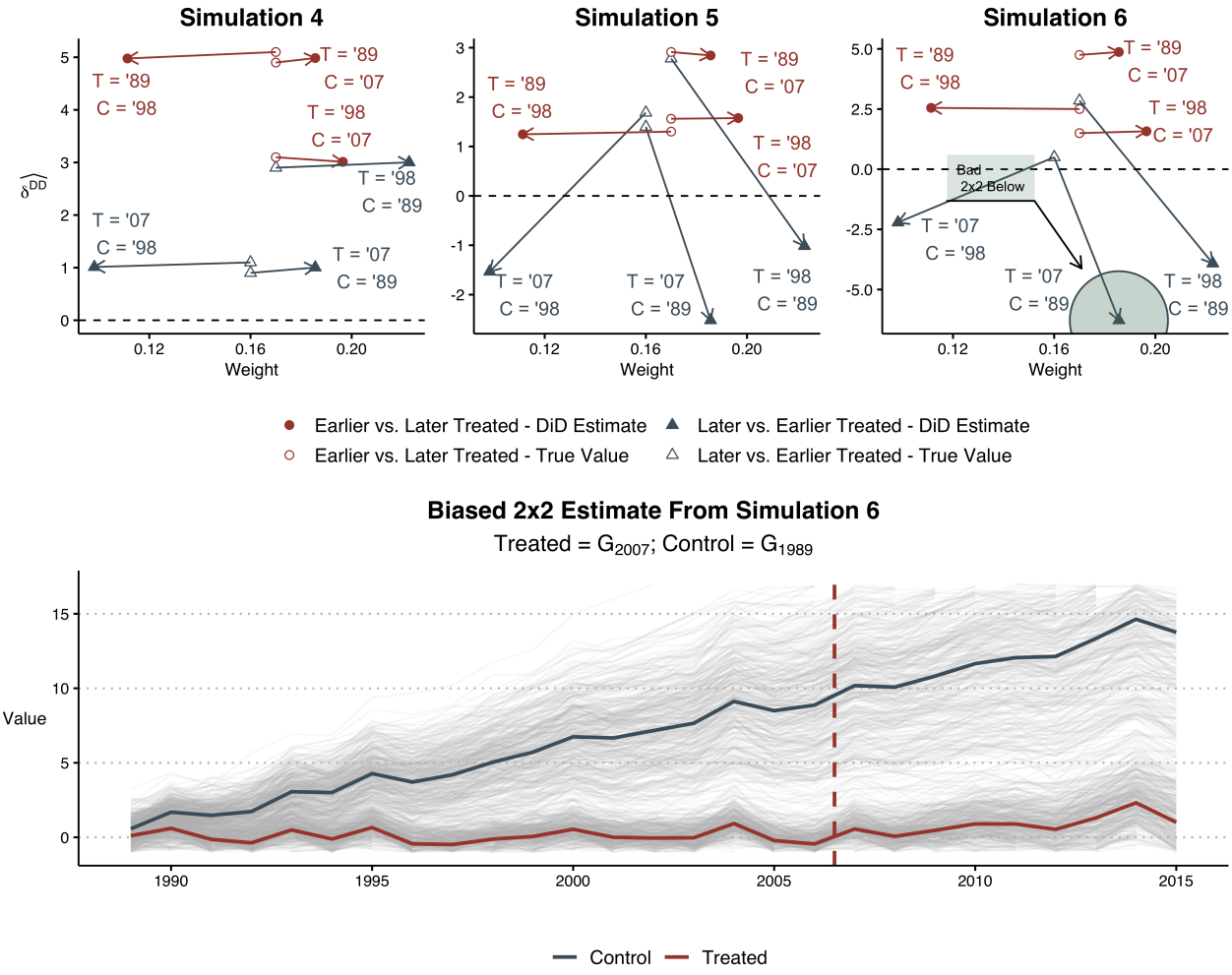


Fig. 6. Robust DiD Methods with Staggered Treatment Assignment and Dynamic Treatment Effects

Figure 6 depicts the true treatment path and estimated effects for Simulation 6 using robust DiD estimators. The TWFE DiD estimate $\widehat{\delta^{DD}}$ was shown to be highly biased in the presence of staggered treatment timing and time-varying treatment effects that are larger for earlier treated cohorts. In Simulation 6, even where the true treatment effect was positive in expectation for every treated firm, $\widehat{\delta^{DD}}$ was negative and statistically significant. However, the estimators from Callaway and Sant'Anna (2020) and Sun and Abraham (2020), as well as stacked regression, provide unbiased estimates of the true treatment effects here.

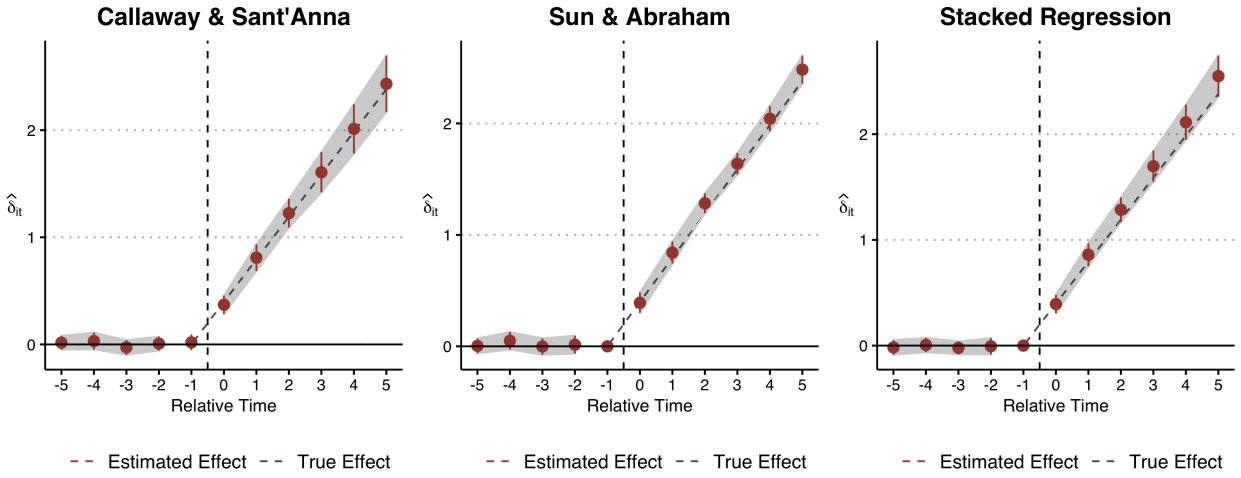
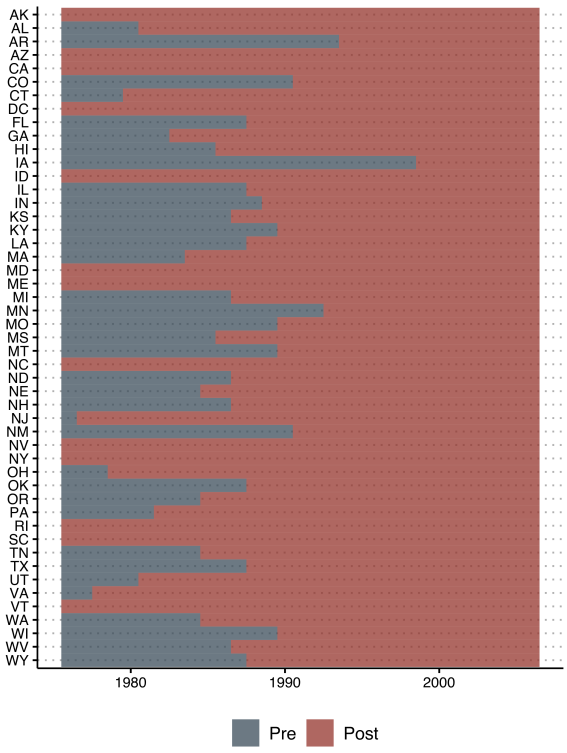


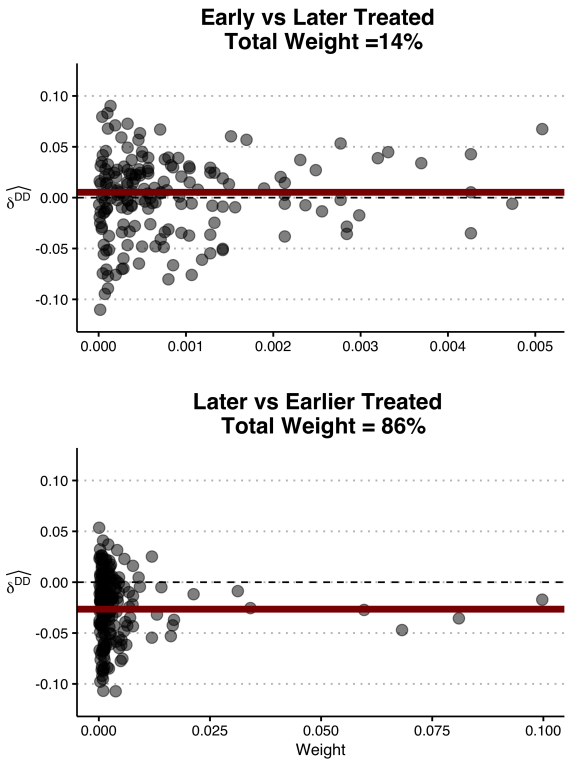
Fig. 7. BLL: Goodman-Bacon Decomposition Diagnostic

Figure 7 decomposes the overall TWFE DiD estimate into its subcomponents. The table decomposes the overall ATT into the average and total weights contributed by earlier vs. later treated comparisons and later vs. earlier treated comparisons. While the former comparisons do not suffer from the bias issues documented in the literature, the latter do. Figure 7i visually portrays the variation in treatment timing used to identify $\widehat{\delta^{DD}}$, while Figure 7ii plots the weights and 2x2 DiD estimates for each treatment timing cohort, broken down by early vs. later and later vs. earlier treated states. Each dot is a unique comparison between treatment timing cohorts (e.g., states treated in 1990 compared to states treated in 1985), and the bold red line represents the weighted average within each comparison type. The overall ATT is the weighted sum of each weighted average.

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



(i) Staggered Treatment Timing



(ii) Weights and Estimates By Timing Type

Fig. 8. Event Study DiD Plots BLL

Figure 8 plots the coefficient values and 95% confidence interval for the lead/lag indicator variables for time periods from $t = -10$ to $t = t + 15$ around deregulation. In Panel (A) we report the estimates as published in BLL, where the average value of the pre-adoption coefficients are subtracted from all of the estimates, so that the pre-trend coefficients are centered at zero. Panel B presents identical estimates to the published results, but does not subtract the average of the pre-treatment coefficients. Panel C uses indicators for all but two relative time indicators, rather than binning the indicators at $t = -10$ and $t = +15$. Finally, Panel D drops all observations where deregulation occurred before the beginning of the panel or once all states have deregulated bank branching.

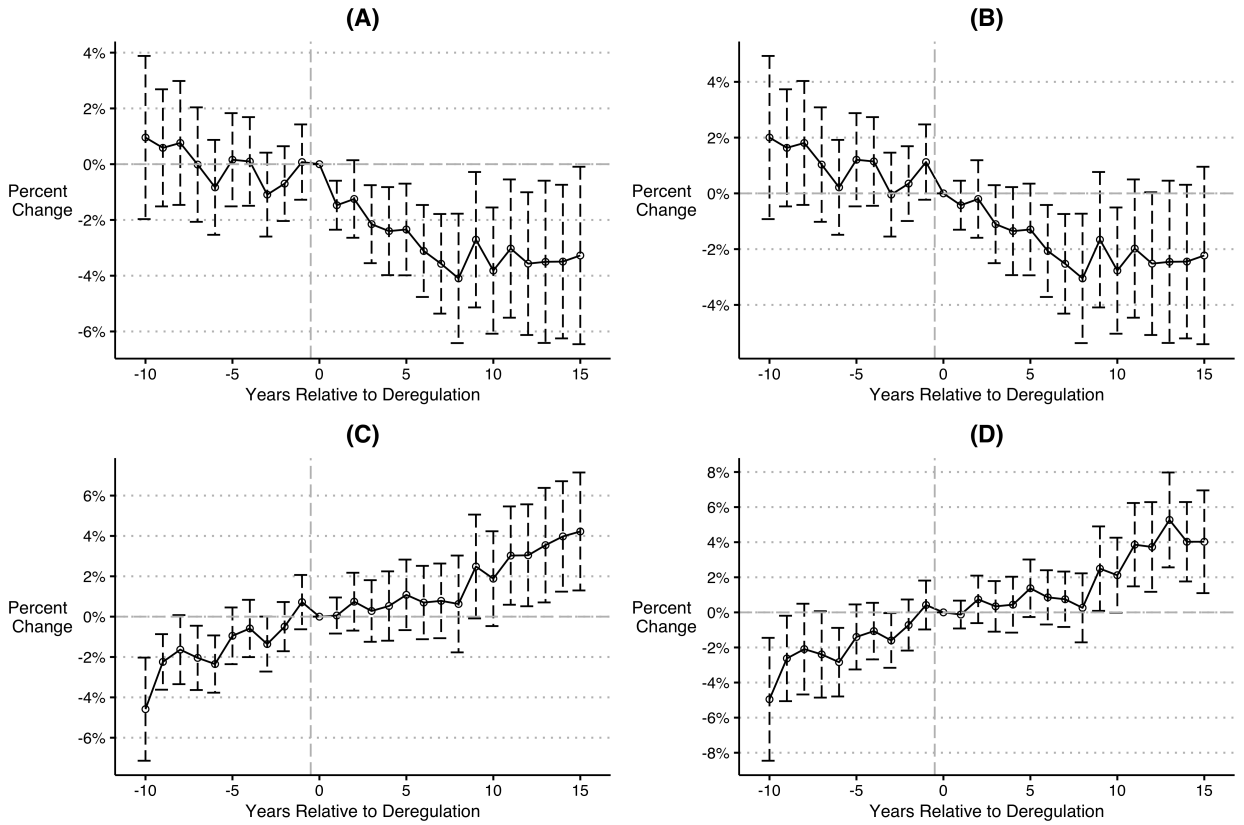
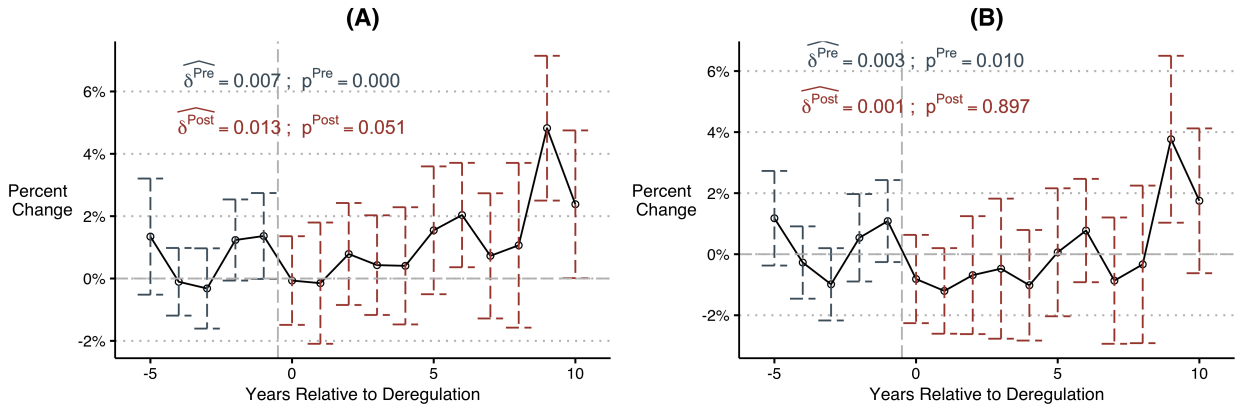


Fig. 9. BLL: CS Estimator and Stacked Regression Event Study Plots

Figure 9 plots the event study coefficients from the estimator in Callaway and Sant'Anna (2020) and the stacked regression approach. Figure 9i plots the event study coefficients from the Callaway and Sant'Anna (2020) estimator. Panel (A) includes only never-treated firms as effective comparison units, while Panel (B) uses not-yet-treated firms. Figure 9ii Panel (A) stacks cohort-specific datasets that include observations from states that deregulate in a certain year, and all states that do not deregulate within 10 years. Panel (B) stacks cohort-specific datasets that include all states that deregulate in that year and all other state-year observations that are pre-treatment for later treated units. The only difference between Panels (A) and (B) in Figure 9ii is that Panel B allows for more observations to act as control firm observations. We also report the aggregated averages of the coefficients for the pre- and post-period event indicators using either the aggregation approach in Callaway and Sant'Anna (2020), or the coefficients and p-values from standard binary indicator regressions on data divided into pre and post-observations for treated units.

(i) Callaway & Sant'Anna Estimator



(ii) Stacked Regression Approach

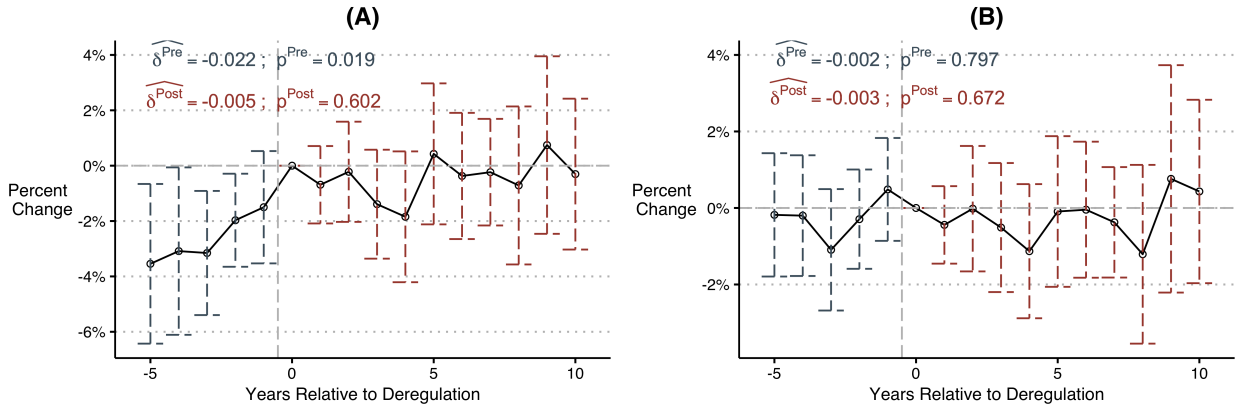


Fig. 10. FHLT: Regulatory Timing and Event Study Plots

Figure 10 plots the timing of board reforms (both the major reforms and the first reforms) across countries in the data (top panel). Blue squares are pre-reform, red squares are post-reform, and empty squares are missing data. The number of firm-year observations for each country are indicated by the shade of the square. The bottom panel presents the event study results using the full set of indicators for all but two relative time indicators. Panel A reports the event study results for the major reforms, and Panel B reports the event study results for the first reforms.

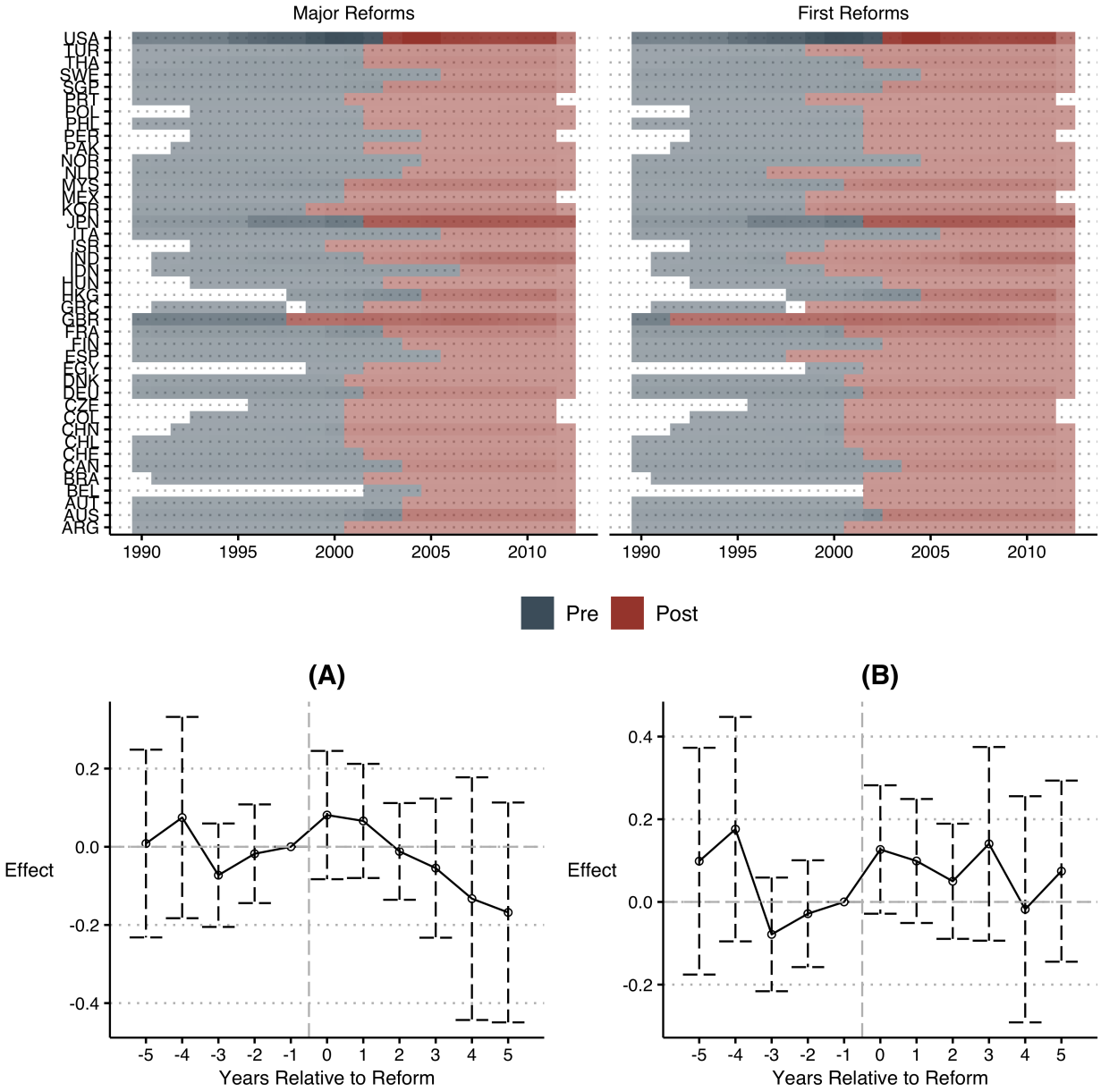


Fig. 11. FHLT: CS Estimator and Stacked Regression Event Study Plots

Figure 11 plots the event study DiD results for four estimators for the data in Fauver et al. (2017). Panel A reports results from the Callaway and Sant'Anna (2020) estimator using using not-yet-treated units as effective comparison units for Major Reforms and Panel B plots the estimator for First Reforms. In Panel C we report stacked regression results where the effective comparison units are all firms in the window from $t-5$ to $t+5$ who have not yet been treated using Major Reforms; Panel D is the same stacked regression using the First Reforms. We report the aggregated averages of the coefficients for the pre- and post-period event indicators using either the aggregation approach in Callaway and Sant'Anna (2020), or the coefficients and p-values from standard binary indicator regressions on data divided into pre and post-observations for treated units.

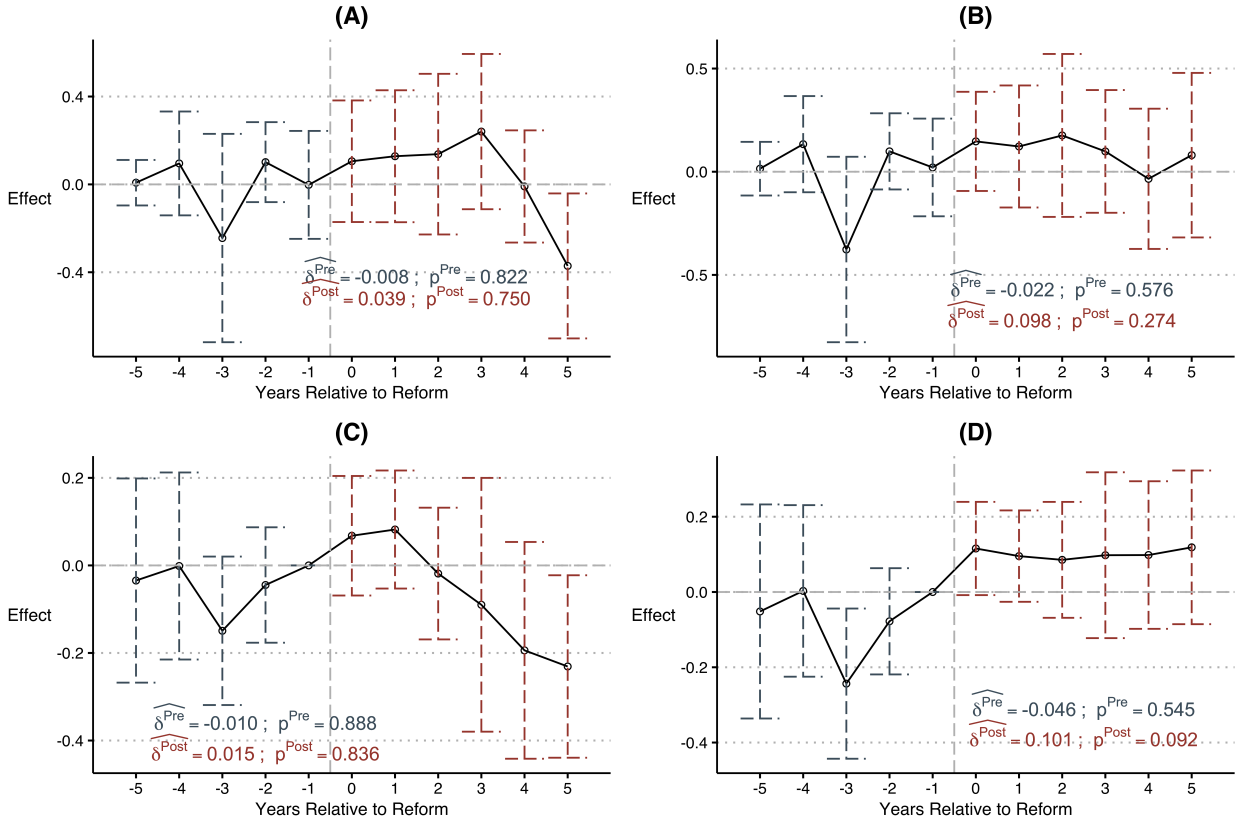
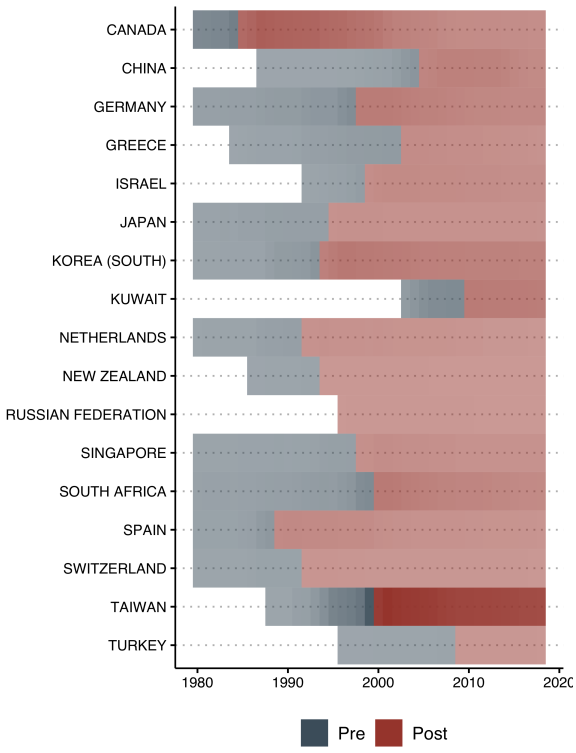
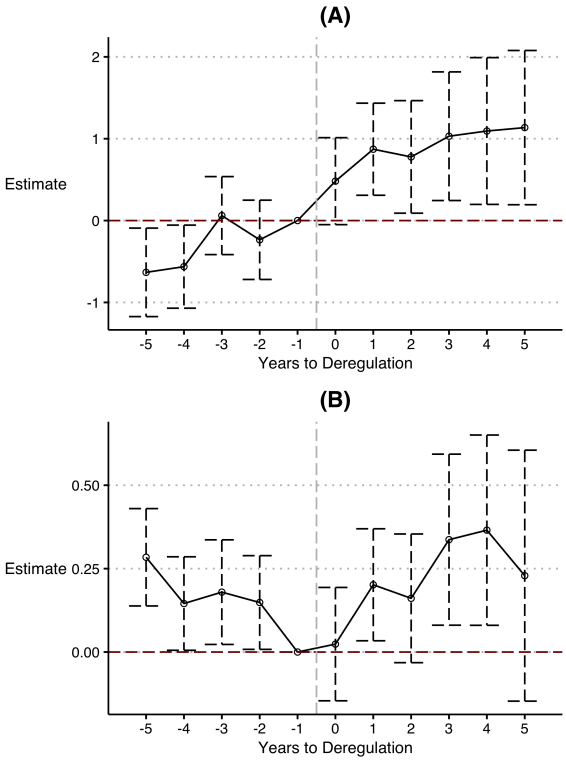


Fig. 12. WYY: Share Repurchase Legalization Timing and Event Study Plots

Figure 12i plots the timing of the legalization of open market share repurchases across countries in the data. Blue squares represent the pre-legalization data while the red squares represent the post-legalization data, and empty squares denote no available data. The number of firm-year observations for each country are indicated by the shade of the square. Figure 12ii presents the event study results using the full set of indicators for all but two relative time indicators (the most negative and the year before treatment). Panel A presents the results for CAPEX, and Panel B presents the results for R&D.



(i) Staggered Treatment Timing



(ii) Event Study Plots

Fig. 13. WYY: CS Estimator and Stacked Regression Event Study Plots

Figure 13 presents the results from CS and stacked regression event study DiD estimates for the data in Wang et al. (2021). Panel A and B report results from the Callaway and Sant'Anna (2020) estimator for CAPEX and R&D, respectively, using using not-yet-treated units as effective comparison units. Panels C and D report stacked regression results for CAPEX and R&D, respectively, where the effective comparison units are all firms in the window from $t-5$ to $t+5$ who have not yet been treated. We report the aggregated averages of the coefficients for the pre- and post-period event indicators using either the aggregation approach in Callaway and Sant'Anna (2020), or the coefficients and p-values from standard binary indicator regressions on data divided into pre- and post-observations for treated units.

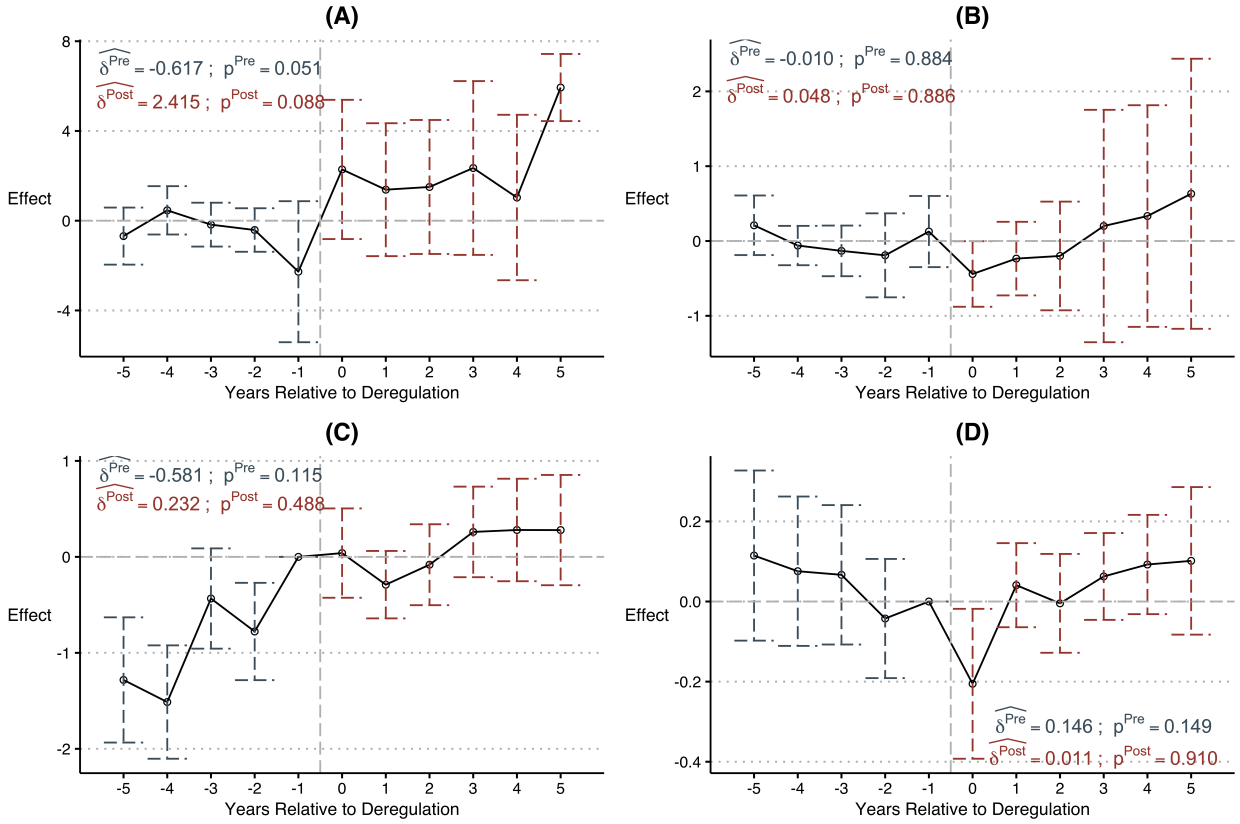


Table 1. Use of DiD and Staggered DiD in Finance and Accounting: 2000-2019

	(1) DiD	(2) Staggered DiD	(3) Staggered DiD / DiD (%)
Journal of Finance	54	29	53.70%
Journal of Financial Economics	162	79	48.77%
Review of Financial Studies	139	66	47.48%
Review of Finance	28	12	42.86%
Journal of Financial and Quantitative Analysis	56	32	57.14%
Finance	**439**	**218**	**49.66%**
Journal of Accounting Research	52	21	40.38%
Journal of Accounting and Economics	63	34	53.97%
The Accounting Review	108	52	48.15%
Review of Accounting Studies	46	24	52.17%
Contemporary Accounting Research	43	17	39.53%
Accounting	**312**	**148**	**47.44%**
Finance and Accounting	**751**	**366**	**48.74%**

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 in the 10 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 of the downloaded papers 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 journal and for finance, accounting, ad all ten journals during the 2000-2019 period.

Table 2. The Impact of Deregulation on Income Inequality

	No Controls	With Controls
	Log Gini	Log Gini
Bank deregulation	-0.022***	-0.018***
	(0.008)	(0.006)
Observations	1519	1519
Adj. R2	0.51	0.54

Note: Table 2 replicates the estimate for Log Gini from Table II of Beck et al. (2010). The table reports results for the impact of bank deregulation on inequality, using the natural logarithm of the Gini index as a proxy. There are minor differences in the standard errors between our replication and the published results, which is a function of different variance calculations between regression functions in Stata and R. The regression includes fixed effects for state and year, and robust standard errors are clustered at the state level. We report the results both with and without controls found in Table II of their paper. *,**, and *** denote two-tailed significance tests at the 10%, 5%, and 1% levels, respectively.

Table 3. The Impact of Board Reforms on Firm Value

	Full Sample			
	With Covariates		Without Covariates	
Variable	Major Reform	First Reform	Major Reform	First Reform
Post	0.096***	0.149***	0.110**	0.136**
	(2.82)	(3.21)	(2.22)	(2.02)
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

Note: Table 3 replicates the estimate from Table 4B of Fauver et al. (2017) for the DiD for board reforms with Tobin's Q as the dependent variable. The first two columns of results replicate the published values with firm and country covariates. In the third and fourth columns we present the results without the inclusion of covariates. All estimates use firm and year fixed effects, and robust standard errors are clustered at the country level. *,**, and *** denote two-tailed significance tests at the 10%, 5%, and 1% levels, respectively.

Table 4. Sources to Finance Share Repurchases

	CAPEX			R&D		
	Short	Long	No Covs	Short	Long	No Covs
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Legalization	-1.088***	-0.847***	-1.196***	-0.133**	-0.236***	-0.145**
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.03)
Total Assets	-0.039	-0.340		-0.147**	-0.590***	
	(0.79)	(0.24)		(0.03)	(0.00)	
Net Sales	0.147***	0.537*		0.087***	0.528***	
	(0.01)	(0.06)		(0.00)	(0.00)	
Net Income	0.079	0.036		-0.011	-0.018	
	(0.25)	(0.71)		(0.45)	(0.36)	
Leverage	0.016***	0.012*		-0.002	-0.001	
	(0.01)	(0.09)		(0.62)	(0.80)	
ROA	0.057***	0.091***		-0.005	-0.011	
	(0.00)	(0.00)		(0.48)	(0.18)	
Sales Growth		0.005***			-0.001***	
		(0.00)			(0.01)	
EBIT / Sales		-0.008			0.002	
		(0.17)			(0.25)	
PPE / Sales		0.005***			0.001***	
		(0.00)			(0.00)	
Quick Ratio		-0.191***			-0.024	
		(0.00)			(0.16)	
Market Share		-0.003			-0.001	
		(0.64)			(0.70)	
Observations	14,593	11,311	18,198	14,914	11,278	18,679
Adj. R2	0.507	0.487	0.461	0.787	0.831	0.752

Note: Table 4 replicates the estimates from Table 5 of Wang et al. (2021) for the DiD of repurchase legalization on capital expenditures (CAPEX) and research and development (R&D) as the dependent variables of interest. Columns 1-3 report the the results for CAPEX, with the inclusion of a short and long set of covariate controls (columns 1 and 2) and without the inclusion of covariate controls (column 3). Columns 4-6 present the analogous estimates for R&D. All estimates use firm and year fixed effects, and robust standard errors are clustered at the firm level. *,** , and *** denote two-tailed significance tests at the 10%, 5%, and 1% levels, respectively.

about ECGI

The European Corporate Governance Institute has been established to improve *corporate governance through fostering independent scientific research and related activities.*

The ECGI will produce and disseminate high quality research while remaining close to the concerns and interests of corporate, financial and public policy makers. It will draw on the expertise of scholars from numerous countries and bring together a critical mass of expertise and interest to bear on this important subject.

The views expressed in this working paper are those of the authors, not those of the ECGI or its members.

ECGI Working Paper Series in Finance

www.ecgi.global/content/working-papers

Electronic Access to the Working Paper Series

The full set of ECGI working papers can be accessed through the Institute's Web-site (www.ecgi.global/content/working-papers) or SSRN:

Finance Paper Series	http://www.ssrn.com/link/ECGI-Fin.html
Law Paper Series	http://www.ssrn.com/link/ECGI-Law.html