Staggered Treatments in Law and Finance: Do State Antitakeover Provisions Matter?

Andrew C. Baker
Stanford Graduate School of Business

David F. Larcker
Stanford Graduate School of Business

Charles C.Y. Wang*
Harvard Business School

January 2020

Abstract

XXX

Keywords: Difference in differences; state antitakeover

JEL:

*First draft: January 2020. Baker (abaker2@stanford.edu) is a doctoral candidate at Stanford GSB. Larcker (dlarcker@stanford.edu) is the James Irvin Miller Professor of Accounting 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. 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 law and finance studies. A workhorse approach in this literature utilizes the passage of laws or market rules impacting one set of firms (treated) or market participants but not others (controls), typically by comparing the differences in the outcomes between treated and control units after the implementation of the law with the differences in the outcomes between treatment and control units before the law. XXXmight want to say something about the endogenous nature of policy adoption is a common flaw

A generalized version of this estimation approach relies on the staggered adoption of laws (e.g., across states or across countries) have become especially popular over the last two decades. However, recent advances in the econometrics literature show these "event studies" designs are not valid for the estimation of the usual estimands of interest—the average treatment effect or the average treatment effect on the treated—even under random assignment (cite athey and imbens, Abraham and Sun).

2 Background

2.1 Review of the DiD Method

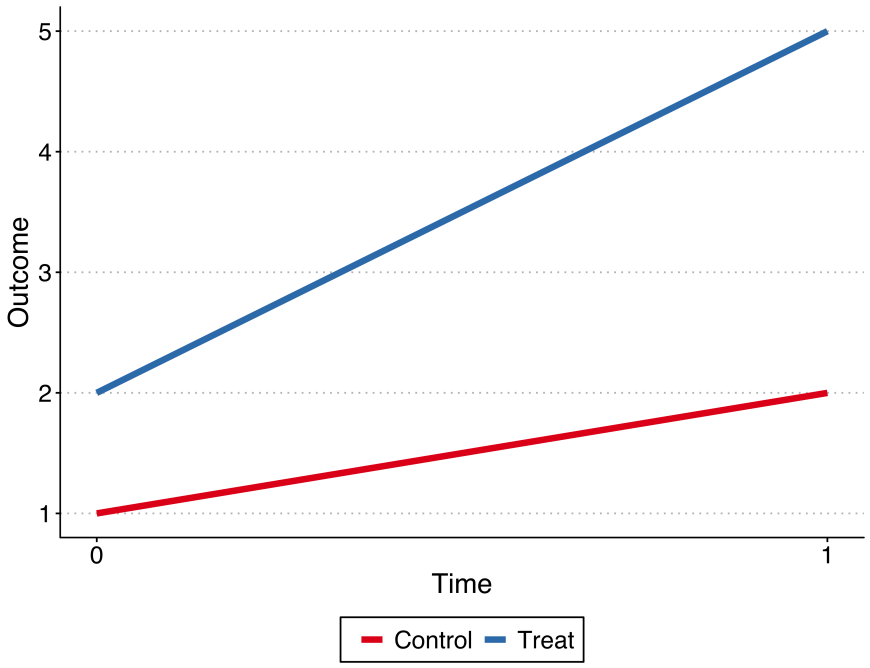
Difference-in-differences (DiD) is one of the most commonly used methods for identifying causal effects in applied microeconomics. [Maybe a cite here with just number of google scholar hits] As noted in [CITE ANGRIST PISCHKE P 227], the idea for DiD can arguably be credited to the physician John Snow, who famously studied cholera epidemics in London in the mid-nineteenth century. In order to establish that contaminated drinking water was the source of a cholera outbreak, Snow compared changes in death rate from cholera among citizens who received water treated by two different companies. While the companies had

previously obtained their water supply from the same source, one company later moved upstream to an area less contaminated with sewage. Snow credibly established the source of the contagion when the death rate for citizens serviced by the moving company subsequently declined sharply in comparison to that of citizens from the non-moving company.

DiD was later popularized in the modern empirical literature with an influential study of the minimum wage by David Card and Alan Kreuger [Cite Card and Kreuger 1994]. While classical theory in labor economics posited that an increase in the minimum wage would reduce employment given downward-sloping labor demand curves, there was little empirical support for the proposition. [Card and Krueger 1994] exploits a 1992 change in the state hourly minimum wage for New Jersey, from $4.25 to $5.05. The authors collected data on employment at fast food restaurants in February and March of 1992, for restaurants in New Jersey and eastern Pennsylvania (where the minimum wage did not change). They used their survey data to conduct a DiD estimate of the effect of the wage increase by comparing the change in employment (from February to November) in the fast food establishments in New Jersey to the change in employment for those located in Pennsylvania. The difference in these two changes (hence the "difference-in-differences") is the causal estimate of the effect of the minimum wage increase.

More generally, as [Angrist and Pischke p 228] note, we can think of simple two-by-two (2x2) DiD estimates as a version of a fixed effects estimator using aggregate data. Denote $Y_{i,t}^1$ as the value of the dependent variable in unit i in period t when the unit receives treatment, and $Y_{i,t}^0$ as the corresponding value of the dependent variable when unit i in period t does *not* receive treatment. Note that these are potential outcomes, as we only observe a unit in a state of either treatment or no-treatment. Under the DiD framework we assume that there is an additive structure to the potential outcome with a lack of treatment; the expected outcome for the untreated state is a linear function of unit and time fixed effects, or $E[Y_{i,t}^0] = \alpha_i + \alpha_t$. If we assume that the treatment effect is a constant δ, then the corresponding expectation

for a treated unit i in period t is equal to $E[Y^1_{i,t}] = \alpha_i + \alpha_t + \delta D_{st}$. Graphically, you can think of the observed relationship as the one presented in the figure below.



The goal of estimating a DiD is to get an unbiased measure of the treatment effect δ, which is esentially estimated by solving for an unknown variable in a system of equations. For the untreated control unit, we have two observations $Y^0_{C,0}$ and $Y^0_{C,1}$. If we look at the difference in the expectations for the control unit at times $t = 1$ and $t = 0$, we see that:

$$E[Y^0_{C,1}] = \alpha_1 + \alpha_C$$

$$E[Y^0_{C,0}] = \alpha_0 + \alpha_C$$

$$E[Y^0_{C,1}] - E[Y^0_{C,0}] = \alpha_1 - \alpha_0$$

Similarly, the difference in the expectations between $t = 1$ and $t = 0$ for the treated unit (where treatment occurs between periods 0 and 1) is:

$$E[Y^1_{T,1}] = \alpha_1 + \alpha_T + \delta$$

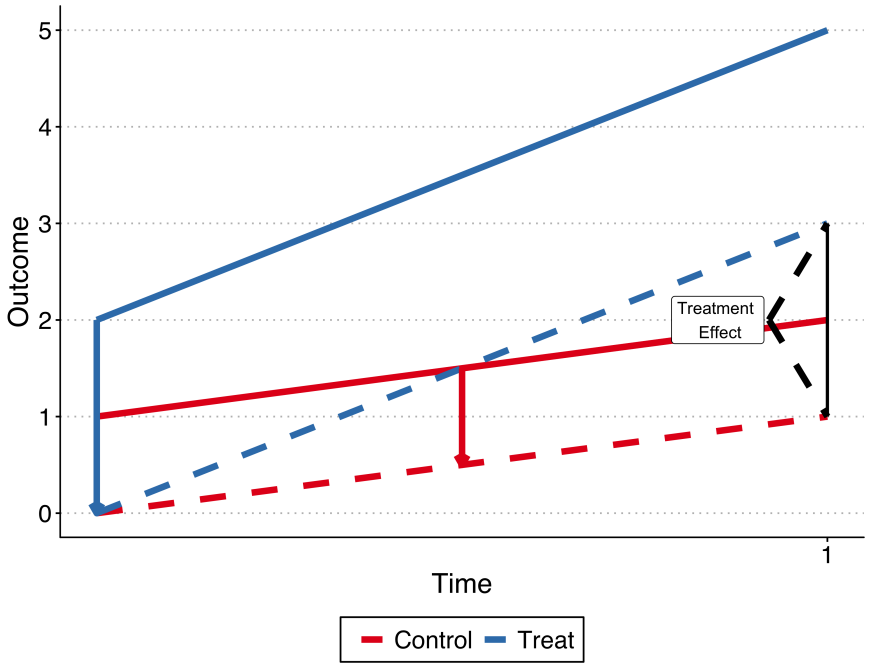
$$E[Y^1_{T,0}] = \alpha_0 + \alpha_T$$

$$E[Y^1_{T,1}] - E[Y^1_{T,0}] = \alpha_1 - \alpha_0 + \delta$$

If we then take the population difference in the differences between the treated and control units we will recover the constant treatment effect δ:

$$\big(E[Y^1_{T,1}] - E[Y^1_{T,0}]\big) - \big(E[Y^0_{C,1}] - E[Y^0_{C,0}]\big) = \delta$$

This double differencing can is represented graphically below:



The assumption required to identify δ is that the *trend* in the outcome variable from

period $t = 0$ to $t = 1$ would have been the same between the treatment and control units, even without treatment. This is satisfied if we assume that potential outcomes have the additive structure described above: the outcome is only a function of static differences between the treatment and control states and a common additive difference between the two time periods, both of which are removed by double differencing. However this may not be true in practice; if different units have both different baseline levels *and* different outcome trends, then the DiD estimate will be biased.

For practical purposes, the DiD can be estimated through a fixed effects regression model. Assume $TREAT_i$ is an indicator variable for the treated unit, and $POST_t$ is an indicator variable for observations in periods $t = 1$. We can write the estimate of the potential outcomes as:

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

The coefficients from the regression estimate in Equation 1 recover the same parameters as the double-differencing performed above:

$$\begin{aligned}
\alpha &= E[y_{it}|i = C, t = 0] = \alpha_0 + \alpha_C \\
\beta_1 &= E[y_{it}|i = T, t = 0] - E[y_{it}|i = C, t = 0] \\
&= (\alpha_0 + \alpha_T) - (\alpha_0 + \alpha_C) = \alpha_T - \alpha_C \\
\beta_2 &= E[y_{it}|i = C, t = 1] - E[y_{it}|i = C, t = 0] \\
&= (\alpha_1 + \alpha_C) - (\alpha_0 + \alpha_C) = \alpha_1 - \alpha_0 \\
\delta &= (E[y_{it}|i = T, t = 1] - E[y_{it}|i = T, t = 0]) - \\
&\quad (E[y_{it}|i = C, t = 1] - E[y_{it}|i = Ct = 0]) = \delta
\end{aligned}$$

An advantage of regression-based DiD is that it provides both estimates of δ and standard errors for the estimate. In addition, as [Angrist and Pishke] note "it's also easy to add additional states or periods to the regression setup ... [and] it's easy to add additional covariates." In settings where there are more than two units and two time periods, the regression DiD model is flexibly modified as:

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

where α_i and α_t are unit and time period fixed effects, and D_{it} is an indicator for a treated unit in treated time periods. Here the main effects for $TREAT_i$ and $POST_t$ are subsumed by the unit and time fixed effects. This two-way fixed effect (TWFE) regression model can be easily modified to include covariates, time trends, and dynamic treatment effect estimation, which has led to regression DiD becoming a workhorse model in empirical applied microeconomics over the past two decades.

2.2 Recent Literature on Staggered DiD

While the 2x2 DiD treatment effect can easily be calculated from Equation 1, most DiD applications exploit variation across groups and units that receive treatment at different points in time. The coefficient that comes from the two-way fixed effects (TWFE) estimator when there are more than two units and periods, and when there is variation in treatment timing, is not an easily interperable parameter in the same manner. Numerous papers have now documented that this coefficient is in fact a weighted average of many different treatment effects, and that these weights are often negative and non-intuitive [Cite Goodman-Bacon, Abraham-Sun, Borusyak and Jaravel 2016 Callaway Sant Anna, Imai and Kim, Streznev].

For the purposes of explaining the methodological issues with DiD with staggered treatment timing, I will focus on the derivation from [Goodman-Bacon (2019)] [hereinafter GB

(2019)] At this point, multiple econometric papers have derived a consistent result with varying terminology. According to GB (2019), the general estimator from the TWFE approach is actually a "weighted average of all possible two-group/two-period DiD estimators in the data." As explained in GB (2019), we know relatively little about what the TWFE DiD value δ^{DD} from Equation 3 measures when treatment timing varies, including how it compares mean outcomes across groups, when alternative specifications will work, or why alternative specifications change estimates.

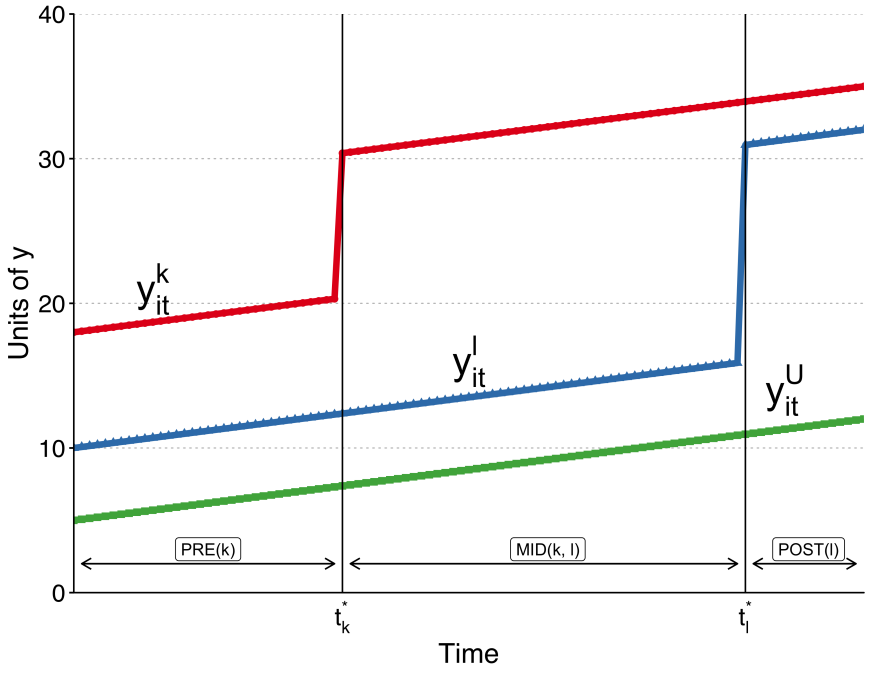
What the derivation in GB (2019) shows is that we can think of the TWFE estimator as a weighted average of individual 2x2 DiD estimators, with the weights proportional to group sizes and the variance of the treatment indicators in each pair, which is highest for units treated in the middle of the panel. In addition, GB (2019) shows how some of the 2x2 estimators use units treated at a particular time as the treatment group and never-treated units as a control group, while others use units treated at two different times, with later-treated groups being used as a control before treatment begins, and the earlier-treated group being used as control after its treatment begins.

Furthermore, GB (2019) shows that when the treatment effects do not change over time, δ^{DD} is the variance-weighted average of cross-group treatment effects, and all of the weights are positive. However, when the treatment effect does vary across time, some of these 2x2 estimates enter the average with *negative* weights. This is because already-treated units act as controls for later-treated units, and changes in a portion of their treatment effects over time are subtracted from the DiD estimate. To make the intuitions behind these results more transparent, we briefly provide their derivation below. Readers seeking more insight on the econometric theory behind these results should refer to Goodman-Bacon 2019 for a fuller explication.

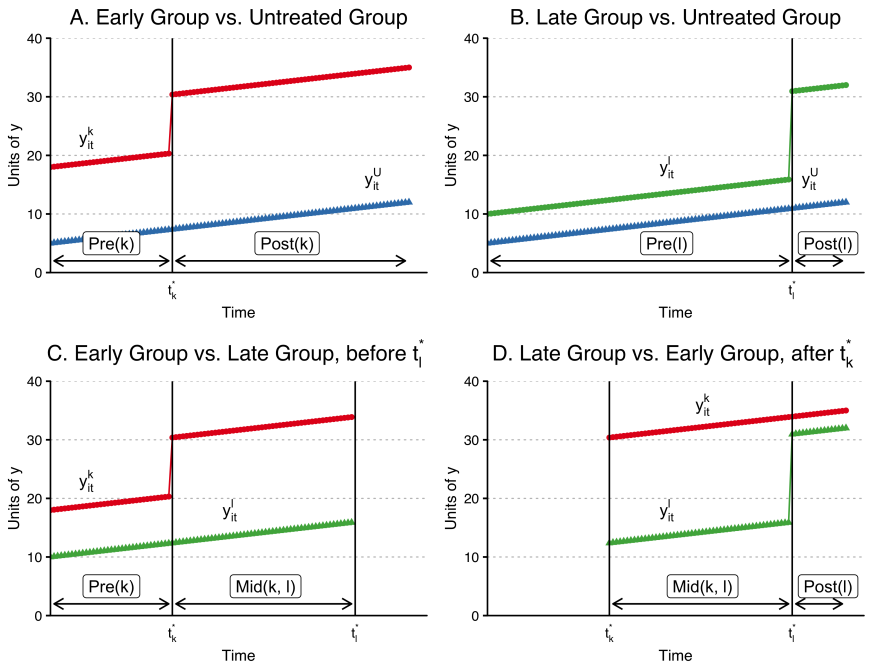
2.2.1 Goodman-Bacon (2019) Derivation

Assume a balanced panel with T periods t and N cross-sectional units i that belong to either an untreated group U, an early treatment group k which receives binary treatment at t_k^*, and late treatment group l that receives a binary treatment at $t_l^* > t_k^*$. For each group a, it's sample share is n_a and the share of time it spends treated is $\overline{D}_a$. Denote $\overline{y}_b^{POST(a)}$ to be the sample mean of y_{it} for units in group b during group a's post period $[t_a^*, T]$, with $\overline{y}_b^{PRE(a)}$ defined similarly.

GB (2019) shows how δ^{DD} is calculated in the stylized setting where there are just the three groups described above - U, k, and t. Here U is never treated, k - the early treatment group - is treated at $t_k^* = \frac{34}{100}T$, and the late treatment group l receives treatment at $t_l^* = \frac{85}{100}T$. Denote three sub-periods, the pre-period for group $k = PRE(k) = [0, t_k^* - 1]$, the middle period when group k is treated but group l is not $MID(k, l) = [t_k^*, t_l^* - 1]$, and the post-period for group $l = POST(l) = [t_l^*, T]$. Further assume that the treatment effect is equal to 10 for group k and 15 for group l. The path in the dependent variables for the groups described above are represented graphically below:



As GB (2019) note, the challenge is to show how the estimates from the TWFE estimation in Equation 3 map to the groups and times depicted in the figure above. GB (2019) shows that in the three group case here, we can form four possible 2x2 DiDs that can be estimated by Equation 1 on the subsamples of groups and times. The four possible 2x2 designs in the figure below.



In Panels A and B we see that, if we consider only one treatment group and the untreated group, the TWFE estimate reduces to the standard 2x2 DiD with the estimate equal to:

$$\hat{\delta}_{jU}^{2x2} = \left(\overline{y}_{j}^{POST(j)} - \overline{y}_{j}^{PRE(j)}\right) - \left(\overline{y}_{U}^{POST(j)} - \overline{y}_{U}^{PRE(j)}\right), j = k, l$$

However, if there are *no* untreated units in the sample, the TWFE estimator is only estimated by the difference in the *timing* of the treatments between treatment groups. This is represented in Panel C, where before t_l^*, the early unit k acts as the treatment group and the later treated unit l acts as a control. The 2x2 DiD estimator compares the differences in outcome between the window when treatment status varies ($MID(k,l)$) and the early group's pre-period ($PRE(k)$):

$$\hat{\delta}_{kl}^{2x2,k} = \left(\overline{y}_{k}^{MID(k,l)} - \overline{y}_{k}^{PRE(k)}\right) - \left(\overline{y}_{l}^{MID(k,l)} - \overline{y}_{l}^{PRE(k)}\right)$$

Panel D shows the opposite situation, where the later group changes treatment after t_l^* and

the earlier treated unit k acts as the control. Again, the 2x2 compares average outcomes between two periods, here $POST(l)$ and $MID(k,l)$:

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

What's important to note here is that the already-treated unit k acts as a control even though they are treated, because their treatment assignment indicator does not change over the relevant period. In addition, each one of the four 2x2 estimates in the figure above uses only a fraction of the available data. The DiDs with treated and untreated units use the entire time period, but only for the two respective groups, so their sample shares are (n_k+n_U) and (n_l+n_U). The timing-only DiDs (Panels C and D) also use only the observations from the two groups, but also use only a portion of the available time periods. $\hat{\delta}^{2x2,k}_{k,l}$ only uses group l's pre-period, so it's sample share is $(n_k+n_l)(1-\overline{D}_l)$, while $\hat{\delta}^{2x2,l}_{kl}$ uses group k's post-period, so it's share is $(n_k+n_l)\overline{D}_k$.

In addition, each of the 2x2 DiDs are identified by the treatment indicator variation in the sub-sample over which it is estimated, and this varies by sub-sample:

$$\begin{aligned}
\widehat{V}^D_{jU} &= n_{jU}(1-n_{JU})\overline{D}_j(1-\overline{D}_j), \quad j=k,l \\
\widehat{V}^{D,k}_{kl} &= n_{kl}(1-n_{kl})\frac{\overline{D}_k-\overline{D}_l}{1-\overline{D}_l}\frac{1-\overline{D}_k}{1-\overline{D}_l} \\
\widehat{V}^{D,l}_{kl} &= n_{kl}(1-n_{kl})\frac{\overline{D}_l}{\overline{D}_k}\frac{\overline{D}_k-\overline{D}_l}{\overline{D}_k}
\end{aligned}$$

where $n_{ab} \equiv \frac{n_a}{n_a+n_b}$ is the relative size of groups within each comparison group. The first portion of each variance measure is the concentration, or total size of the groups, while the second portion comes from when the treatment occurs in each sample, and is the variance of the treatment indicator variable scaled by the size of the relevant window. The central

result from GB (2019) is that any TWFE estimator is just an average of the 2x2 estimators in Panels A-D, with weights that are based on subsample shares n and variances $\hat{V}$.

With K timing groups, you can form $K^2 - K$ timing only estimates comparing earlier and later treated groups. With an untreated group U you can form K treated/untreated 2x2 DiDs for a total of K^2 DiD estimates. The weights on each of these 2x2 estimates used to construct δ^{DD} combine the absolute size of the subsample, the relative size of the treatment and control groups in the subsample, and the timing of the treatment in the subsample.

To put this in context, we can derive the weights for δ^{DD} from the example in the figures above. t_k^* and t_l^* were set so that $\overline{D}_k = 0.66$ and $\overline{D}_l = 0.16$. For the 2x2 DiDs in Panels A and B the weights given to the 2x2 DiD estimate for the earlier treated group, s_{kU} is greater than the weight given to the DiD for the later-treated group, s_{lU} because k is treated closer to the middle of the panel and has a higher treatment-indicator variance. This is also true for the timing-only 2x2 DiD's (Panels C and D), where the weights are higher for the earlier treated groups both because it uses more data and because it has a higher treatment variance. If we calculate the weights for our four estimates here we get:

Table 1. DiD Weights

weights	value
s_{kU}	0.37
s_{lU}	0.22
s_{kl}^k	0.28
s_{kl}^l	0.13

What is clear from the derivation is that panel length alone can change the DiD estimates substantially, even when each 2x2 DiD estimate $\delta^D D$ is constant. This seems normatively undesirable. In addition, the weights assigned to each 2x2 estimate when aggregating through TWFE OLS are a result of the size of the subsample and the magnitude of the treatment

variance. Groups treated closer to the middle of the panel get more weight, which has not been previously acknowledged in literature using TWFE to estimate δ^{DD}.

2.2.2 Identifying Assumptions

The derivation of the mechanics of TWFE DiD explains the coefficient is calculated, but does not explain map the calculation to the potential outcomes framework for causal inference. GB (2019) follows Callaway and Sant'Anna (2018) and defines the average treatment effect on the treated (ATT) for timing groups k (all firms that receive treatment during a certain period) at time τ (the "group-time average treatment effect") as

$$ATT_k(\tau) \equiv E[Y_{i\tau}^1 - Y_{i\tau}^0 | k]$$

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

$$ATT_k(W) \equiv \frac{1}{T_W} \sum_{t \in W} E[Y_{it}^1 - Y_{it}^0 | k]$$

GB (2019) shows that you can use the DiD decomposition to express the probability limit of the TWFE DiD estimator δ^{DD} (assuming T is fixed and N grows) as

$$\underset{N \to \infty}{plim} \hat{\delta}^{DD} = VWATT + VWCT - \Delta ATT$$

In this equation VWATT is the "variance-weighted average treatment effect on the treated", and 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 δ^{DD}, as derived in the earlier 3-unit example.

VWCT is the "variance-weighted common trend" term, which generalizes the parallel

trend assumption of DiD to a setting with timing variation. VWCT is the average of the difference in counterfactural 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 $\hat{\delta}^{DD}$ estimate. This captures the reality that different groups might not have the same underlying trend in outcome dynamics, which will inherently bias any DiD estimate. Note that the parallel trends assumption has been long-known and is inherently untestable.

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 controls 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. This can represent a substantial bias in the estimate.

When the treatment effect is constant in every period, then $ATT_k(W) = ATT, \Delta ATT = 0$, and $VWATT = ATT$. Any justification for the use of DiD implicitly assumes that $VWCT = 0$, or that the parallel trends assumption is satisfied. Bias will arise in δ^{DD} generally under two forms of treatment effect heterogeneity. First, if treatment effects vary across units, but not over time, then $\Delta ATT = 0$ and $ATT_k(W) = VWATT = \sum_{k \neq U} ATT_k \times w_k^T$. Here w_k^T is a function of the decomposition weights, and is a combination of sample shares and treatment variance. In general $w_k^T \neq n_k^*$. In other words, the weights are not equal to the sample shares, so $\hat{\delta}^{DD}$ will not equal the sample ATT. As explained in GB (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. Again, 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.

In addition, the coefficient will be biased when the treatment effect is time-varying within a treated unit. That is, instead of a constant additive effect (where y_{it} is shifted by a constant τ), there are dynamics to the treatment effect so that τ is a function of years since treatment. This is likely a frequent occurrence, as most "event study" DiD estimates document post-treatment trends in the estimated effects. In this case, time-varying treatment effects generate heterogeneity across the 2x2 DiDs by averaging over different post-treatment windows, up-weighting short run effects most likely to appear in the small windows between timing groups, and biasing the estimates away from the VWATT because $\Delta ATT \neq 0$. If treatment causes a trend-shift or other form of time-varying treatment effect, then δ^{DD} will use already-treated units as controls and will yield estimates that are too small or even wrong-signed.

2.2.3 Remedies

While the econometric literature has settled on an agreement of the methodological challenge posed by TWFE estimation of DiD with staggered treatment timing, a number of alternative DiD estimation techniques have been proposed to remedy the problem. Although each proposal deals with the weighting and bias issues inherent to TWFE DiD by modifying the set of units that can act as effective controls, they differ in terms of sample exclusions and dealing with covariates. Below we describe a few of the more promising proposals for unbiased DiD estimation that we will apply in later sections to simulated and real data.

Goodman-Bacon (2019)

In addition to the decomposition results above, GB (2019) proposes a series of diagnostic tests to examine the robustness of the TWFE DiD estimate. Most importantly, the paper shows how to test the stability of the DiD coefficients by plotting each 2x2 DiD against its weight w_k^*. You can calculate a variety of different conditional expectations over the

subgroups, including the average effect and total weights for treated/untreated comparisons, and late/early treatment and early/late treatment comparisons. Adding the weights on the timing-based coefficients from the decomposition show how much of δ^{DD} comes from timing variation. Finally, you can identify influential observations by comparing the weights, and calculate what percentage of the total 2x2 coefficients drive 50% of the estimate, or whether influential observations have substantially different levels of τ. However, these decomposition plots are only available for balanced panel data, which is often not available in typical corporate finance or accounting applications.

Callaway and Sant'Anna (2018)

Callaway and Sant'Anna (2018) [hereinafter CS (2018)] also consider 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. They propose a two-step estimation strategy with a bootstrap procedure to conduct asymptotically valid inference which can adjust for autocorrelation and clustering.

CS (2018) define the causal parameters of interest in a staggered DiD framework as functionals of the ATE for group g at time t, where a group is defined by when units are first treated (e.g. all firms treated in 2006, 2009, etc.). These causal parameters are called "group-time average treatment effects". This setting allows you to aggregate the treatment effects by either relative time (i.e. the event study approach) or by calendar time.

To keep the same notation as in CS (2018), assume there are $\mathcal{T}$ periods where $t = 1, \ldots, \mathcal{T}$, with D_{it} a binary variable equal to 1 if a unit is treated and 0 otherwise. In addition, define G_g to be a binary variable that is equal to 1 when an individual 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_g 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.

CS (2018) frames its identification strategy within the potential outcomes framework. Let $Y_t(1)$ and $Y_t(0)$ be the potential outcomes at time t with and without treatment, respectively. The observed outcome in each period can thus be expressed as $Y_t = D_t Y_t(1) + (1 - D_t)Y_t(0)$. The authors focus on the average treatment effect for individuals first treated in period g at time period t, called the group-time average treatment effect, denoted by:

$$ATT(g,t) = \mathbb{E}[Y_t(1) - Y_t(0)|G_g = 1]$$

The assumptions in the CS (2018) framework are of random sampling, parallel trends conditional on covariates, irreversability of treatment (after a unit is treated it is treated for the remainder of the panel), and overlap (the propensity scores are greater than 0 and less than 1).

Given the parameter of interest $ATT(g,t)$, CS (2018) develop a non-parametric identification strategy for the group-time average treatment effect, which allows for treatment effect heterogeneity and does not make functional form assumptions about the evolution of potential outcomes. Under the assumptions above, the group-time average treatment effect is nonparametrically identified as:

$$ATT(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - T_{g-1})\right] \tag{3}$$

This is just a simple weighted average of the "long difference" of the outcome variable, with the weights depending on the propensity score, which are normalized 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. This reweighting ensures that the covariates of the treatment and control group are balanced. The authors also explain in an appendix how to use their reweighting scheme where future-treated units can be used as controls.

With $ATT(g,t)$ we can aggregate the group-time treatment effects into fewer intrepetable causal effect parameters, which makes interpretation easier, and also increases statistical power and reduces estimation uncertainty. CS (2018) propose doing the following:

1. Aggregating $ATT(g,t)$ by timing-group, $\tilde{\theta}_s(g) = \frac{1}{\mathcal{T}-g+1}\sum_{t=2}^{\mathcal{T}} 1\{g \leq t\} ATT(g,t)$, and combining the group average treatment effect by the size of each group $\theta_s = \sum_{g=2}^{\mathcal{T}} \tilde{\theta}_s(g) P(G = g)$.

2. Aggregating $ATT(g,t)$ by the length of exposure to treatment (denoted e) to test whether there are dynamic treatment effects (i.e. the treatment effect is explicitly a function of the time since treatment). They consider $\tilde{\theta}_D(e) = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} 1\{t-g+1 = e\} ATT(g,t) P(G = g | t-g+1 = e)$. Here $\tilde{\theta}_D(1)$ would be equal to the average (based on group size) treatment effect in one ($e = 1$) period. This aggregation gives you a properly weighted event-study similar to pre-existing practice, which might be desirable for continuity. In addition, you could then average over all possible values of e to get $\theta_D = \frac{1}{\mathcal{T}-1}\sum_{e=1}^{\mathcal{T}-1} \tilde{\theta}_D(e)$.

3. Aggregating $ATT(g,t)$ by calendar time involves computing an average treatment effect for all individuals that are treated in period t and then averaging across all periods. Here, $\tilde{\theta}_C(t) = \sum_{g=2}^{\mathcal{T}} 1\{g \leq t\} ATT(g,t) P(G = g | g \leq t)$. $\tilde{\theta}_C(t)$ can be interepreted as the average treatment effect in period t for all groups treated by period t, and you can further aggregate to $\theta_C = \frac{1}{\mathcal{T}-1}\sum_{t=2}^{\mathcal{T}} \tilde{\theta}_C(t)$, which would be the average treatment effect when calendar time matters. This final summary statistic naturally puts the

most weight on groups that are treated in the earliest periods, because they enter more of the $ATT(g,t)$ estimates.

4. Aggregating $ATT(g,t)$ in the case where the timing of treatment matters, and where there are dynamic treatment effects (probably the most likely reality in real world scenarios). Here CS (2018) considers dynamic treatment effects only for $e \leq e^{'}$ and for groups with at least $e^{'}$ periods of post-treatment data, which removes the impact of selective treatment timing by keeping the same set of groups across all values of e. There is a tradeoff here between the amount of groups you have in your final estimation sample (larger with smaller values of $e^{'}$) and the length of the exposure that you can estimate (smaller with smaller values of $e^{'}$). Let $\delta_{gt}(e,e^{'}) = 1\{t-g+1=e\}1\{T-g+1 \geq e^{'}\}1\{e \leq e^{'}\}$. Thus $\delta_{gt}(e,e^{'})$ is equal to one in the period where group g has been treated for exactly e periods, if group g has at least $e^{'}$ post-treatment periods available, and if the length of exposure e is less than the post-treatment period requirement $e^{'}$. The average treatment effect for groups in $\delta_{gt}(e,e^{'})$ is given by $\tilde{\theta}_{SD}(e,e^{'}) = \sum_{g=2}^{\mathcal{T}}\sum_{t=2}^{\mathcal{T}}\delta_{gt}(e,e^{'})ATT(g,t)P(G=g|\delta_{gt}(e,e^{'})=1)$. With $\tilde{\theta}_{SD}(e,e^{'})$ we can calculate the average treatment effect for groups with at least $e^{'}$ periods of post-treatment data as $\theta_{SD}(e^{'}) = \frac{1}{\mathcal{T}-e^{'}}\sum_{e=1}^{\mathcal{T}-e^{'}}\tilde{\theta}_{SD}(e,e^{'})$.

Abraham and Sun (2019)

This paper is tightly linked with the Callaway and Sant'Anna paper, but focuses exclusively on the event-study context, where you include leads and lags of the treatment variable instead of a single binary indicator variable. The authors confirm that in the event study context, where the timing of treatment varies across units, lead/lag regressions can also produce causally uninterpretable results because they assign non-convex weights to cohort-specific treatment effects. Their proposed method estimates the dynamic effect for each

cohort (equivalent to group G_g from CS (2018)), and then calculates the weighted average of cohort-specific estimates.

In the event study context, the TWFE regression with leads and lags of treatments takes the form of:

$$y_{it} = \alpha_i + \alpha_t + \sum_{l=-K}^{-2} \delta_l D_{it}^l + \sum_{l=0}^{L} \delta_l D_i t^l + \epsilon_{it}$$

.

where D_{it}^l is an indicator for being l time periods relative to i's initial treatment (treatment is $l = 0$), and α_i and α_t are unit and time fixed effects, as before. AS (2019) focus on the "cohort-specific average treatment effects on the treated" l periods from initial treatment. This is denoted $CATT_{el} = E\left[y_{i,e+l}^e - y_{i,e+l}^\infty | E_i = e\right]$ where E_i is the time period of initial treatment, and a cohort e is a set of units for which $E_i = e$; y_{it}^∞ is the counterfactual outcome of unit i if it never received treatment.

The key theoretical result in this paper is that, even when doing 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 are hard to interpret and need not be positive without assuming treatment effect homogeneity. Specifically, the FE estimands for l periods relative to treatment can be written as non-convex averages of not only $CATT_{e,l}$ from that period, but also $CATT_{e,l'}$ from other periods. This is similar to the result in GB(2019) that $\Delta ATT \neq 0$ with dynamic treatment effects, although the event study framework does solve some of the OLS variance-weighted issues brought up by GB (2019) in the binary indicator context.

The proposed alternative estimation technique in AS (2019) is to use an interacted specification that is saturated in relative time indicators D_{it}^l and cohort indicators $1\{E_i = e\}$ to estimate each $CATT_{el}$, which they call an "interaction-weighted" (IW) estimator. The DiD

under the IW estimator is equivalent to the difference between the average change in outcomes for cohort e, which is exactly l periods relative to treatment, and the average change for cohorts that have not been treated by $t = e + l$ and is estimated simply by

$$y_{it} = \alpha_i + \alpha_t + \sum_e \sum_{l \neq -1} \delta_{el}(1\{E_i = e\} \cdot D_{it}^l) + \epsilon_{it}$$

Finally, you can re-create the standard event-study plots by taking the weighted average over cohorts e for time period l, with the weights equal to the share of each cohort in the relevant periods.

Cengiz, Dube, Lindner, and Zipperer (2019)

CDLZ (2019) 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, CDLZ (2019) notes that there are issues in aggregating discrete DiD estimates through OLS, and as a robustness check seperates and plots the distribution of the *individual* treatment effect for each of the events.

To do this, the authors create 138 event h-specific datasets including the outcome variable and controls for the treated state h 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 $t = 4$). For each event, they they then run a one-treated panel DiD based on their baseline strategy (with their syntax):

$$y_{sjth} = \sum_{\tau=-3}^{4} \sum_{k=-4}^{4} \alpha_{\tau kh} \mathcal{I}_{sjth}^{\tau k} + \mu_{sjh} + \rho_{sjth} + u_{sjth}$$

In essence this is just a saturated panel DiD with a lead/lag treatment indicator $\mathcal{I}$ and state μ and time ρ fixed effects. CDLZ (2019) then plots the distribution of the α treatment

effects, along with their confidence intervals using the Ferman and Pinto (forthcoming) method for heteroskedastic robust cluster residual bootstrapping more appropriate for single-treated units.

CDLZ (2019) also stacks the event-specific data sets to calculate an average effect across all 138 events using a single set of treatment effects. 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 negative weighting of some events that may occur with a staggered design. Moreover, by dropping all control states with any state-level minimum wage increases within the 8 year event window, this method guards against bias due to heterogeneous treatment effects that show up in the ΔATT term from GB (2019).

3 Simulation Analysis

In this section we will simulate a dataset that has a similar data structure to those frequently used in corporate finance and accounting settings, with a panel of firms and years corresponding to annual financial reporting data. We will then simulate a treatment that occurs across units at different points in time, and compare how the proposed alternative estimation techniques perform under a situation of known heterogeneous treatment effects. This simulation approach is obviously a simplified representation of the true data-generating processes likely to be found in real-world situations, but it is the *implicit* data generating process assumed through a TWFE DiD estimation.

Assume we are modeling an outcome variable y_{it} on a balanced panel dataset with $T = 36$ years from $t = 1980$ to 2015, and 1000 units (firms) i. There are unit and year effects on the

outcome variable, which in the data are both drawn from $\mathcal{N}(0,1)$. In the panel every firm is eventually treated, but the timing of the treatment varies within the sample. Firms are incorporated in one of 50 randomly drawn states, and the states are randomly assigned into five treatment groups G_g based on treatment being initiated in 1985, 1991, 1997, 2003, and 2009. All treatment groups are equal-sized (e.g. 10 states adopt a hypothetical law in each year).

The simulated treatment effects τ are all positive in expectation. To make the results less deterministic, each individual firm's treatment effect is drawn from a normal distribution with different means depending on treatment group. For all treatment firms incorporated in a state in treatment group G_g, firms receive a treatment effect $\tau_i \sim \mathcal{N}(\tau_g, .2^2)$. This is a "dynamic treatment effect" as referred to in GB (2019); instead of adding a constant τ_i to each post-treatment observation, the treatment effect is cumulative (if $\tau_i = 0.2$ then in the first period the treatment effect is 0.2, in the second period it is 0.4, etc.).[1] The average yearly additive treatment effects for each treatment group are given in Table 2 below:

Table 2. Treatment Effect Averages

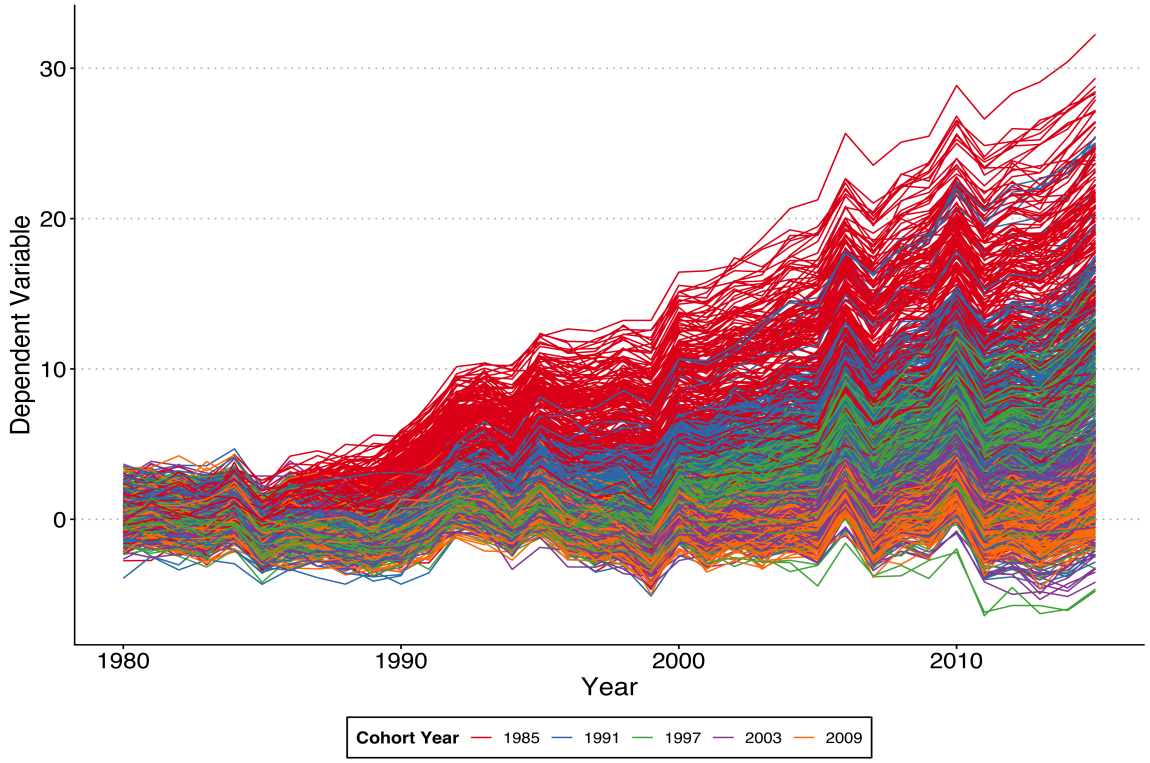
G_g	τ_g
1985	0.5
1991	0.4
1997	0.3
2003	0.2
2009	0.1

First, we can plot the paths in the outcome variable by treatment cohort. In Figure 1 we see the outcome paths of the $N = 1000$ firms in our simulated data, with colors corresponding to different treatment cohorts. The outcome variables is a function of unit

[1]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 GB (2019), but we could limit the treatment effect accumulation to a set number of years after treatment

and time fixed effects, as well as the cumulative sum of the average treatment effect by cohort group and a simulated random noise term $\epsilon_{it} \sim \mathcal{N}(0, 0.5^2)$. We see that there is an upward trend in the outcome path in the units of all treatment cohorts, which is driven by the consistently positive treatment effect average. However, the early-treated units, which both have a longer post-treatment period and a higher average treatment effect multiple τ_G ultimately have higher levels of the outcome variable by the end of the panel.

Fig. 1. Trends in Outcome Path



The average treatment effect, under the parallel trends assumption, as calculated through TWFE DiD is equivalent to the value of $\hat{\delta}$ that one attains using maximum-likelihood estimation of the fixed effects regression from Equation 3:

$$y_{it} = \alpha_i + \alpha_t + \delta^{DD} D_{it} + \epsilon_{it}$$

This is a simple regression model with unit (firm) fixed effects α_i, which captures time-invariant differences at the unit level, and time (year) fixed effects α_t which allows for secular changes in the outcome variable over time. Because the treatment assigned at the state level, the standard errors should be clustered at the state of incorporation level. If we estimate the treatment effect through TWFE fixed effects regression, we get the estimate in Table 3 below. Here we see the bias identified in the literature with staggered adoption and treatment effect heterogeneity; although *every* treatment effect is positive in expectation, the sign of the DiD treatment effect from TWFE estimation is negative and statistically significant when prior-treated units are used as effective controls.

Table 3. TWFE Estimates

	Estimate	Cluster s.e.	t value	$Pr(>\|t\|)$
$\hat{\delta}$	-1.04	0.28	-3.76	0

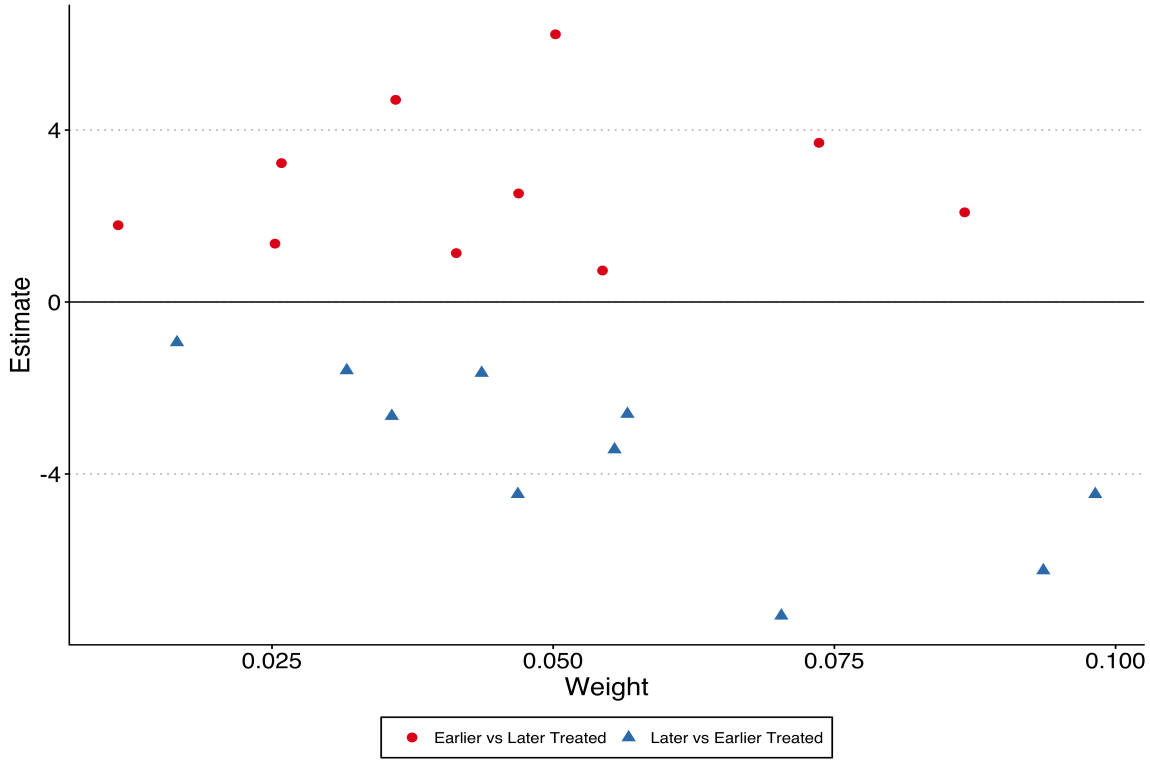
3.1 Remedies

Table 3 shows that TWFE estimation can be highly biased when used to estimate average treatment effects through DiD with staggered adoption and dynamic treatment effects. While the remedies discussed above in Goodman-Bacon 2019, AS, CS, and CLDZ have been theoretically justified, in order to be applicable to empirical corporate governance research, it should be verified that the methods recover true treatment effect paths in data similar in structure to that used in the field.

First, we can use the Goodman-Bacon (2019) decomposition theorem to illustrate the

variation in treatment effect estimation in the sample. Because every unit is located in a state that ultimately adopts the pseudo-legislation during the panel, the identification of δ^{DD} here is timing-only; that is there are no never-treated units, and the variation is dependent on the timing rather than the existence of treatment. As Figure 6 in Goodman-Bacon (2019) demonstrates, a useful diagnostic test is to plot the 2x2 group DiD estimates by the implicit assigned weight (which is a function of first-treatment timing and group size) and by timing group (either earlier vs. later treated states, or later vs. earlier treated states), which is reported in Figure 2.

Fig. 2. Goodman-Bacon (2019) Decomposition



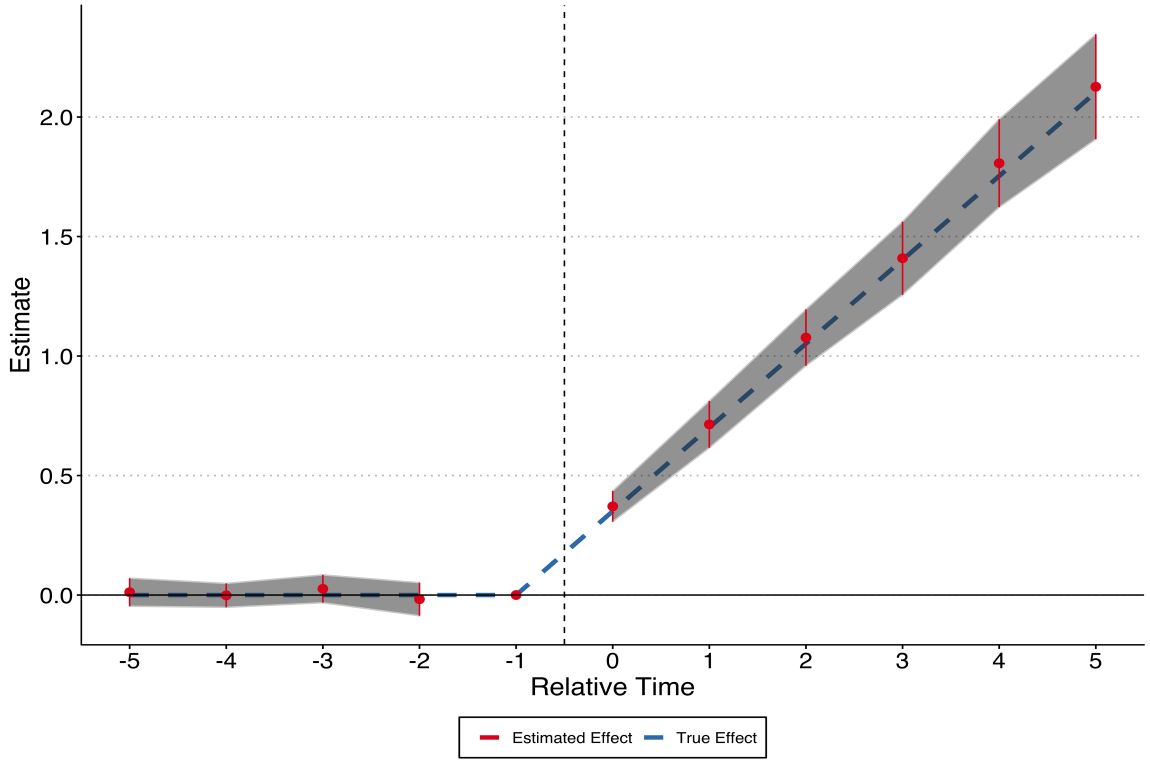
The potential for bias in the δ^{DD} estimate derived above is most extreme in the later vs. earlier treated units, where dynamic treatment effects are subtracted in the second dif-

ferencing. This concern is evidenced in Figure 2; although every imputed treatment effect τ_G is positive in expectation, the later vs. earlier treated groups show negative estimated treatment effects, because the large positive effect of earlier treatments are *subtracted* from the estimate for future treated groups. These large negative weights bias downwards the estimate of the TWFE DiD, and is so extreme is this stylized setting that the overall estimated treatment effect is both of the wrong sign and robustly statistically significant.

Next, we can estimate the DiD model using the two-step estimation strategy from Callaway Sant'Anna (2019). The CS (2019) method can be estimated with or without covariates. With the inclusion of covariates, the CS method is the inverse-propensity weighted long-difference in cohort-specific average treatment effects between treated and untreated units for a given a treatment cohort, as described in Equation 3 above. Without covariates, as in the simulated data here, the propensity scores are normalized to one, and the method calculates the simple long difference between all treated units i in relative year k with all potential control units that have not yet been treated by year k. The dynamic average treatment effect using CS 2019 is positive and statistically significant (.61, s.e. = 0.03), and the dynamic event-study for the treatment effects consistent with the data-generating process, as presented in Figure 3.

We can also estimate the DiD through the method of Abraham and Sun (2019). The AS 2019 method calculates event-study estimates for each treatment cohort through a saturated fixed effects model, and then calculates the weighted average of the estimates for each relative time indicator using the sample share of each treatment cohort as the relevant weights. With this method we have to drop lead/lag indicators for the last treated cohort (units treated in 2009), because by the end of the panel there are no available units to use as controls. Thus, units in the last treatment cohort are effectively only used as controls. We drop all observations more than five years following treatment to ensure that prior treated units

Fig. 3. Callaway & Sant'Anna (2019) DiD Design

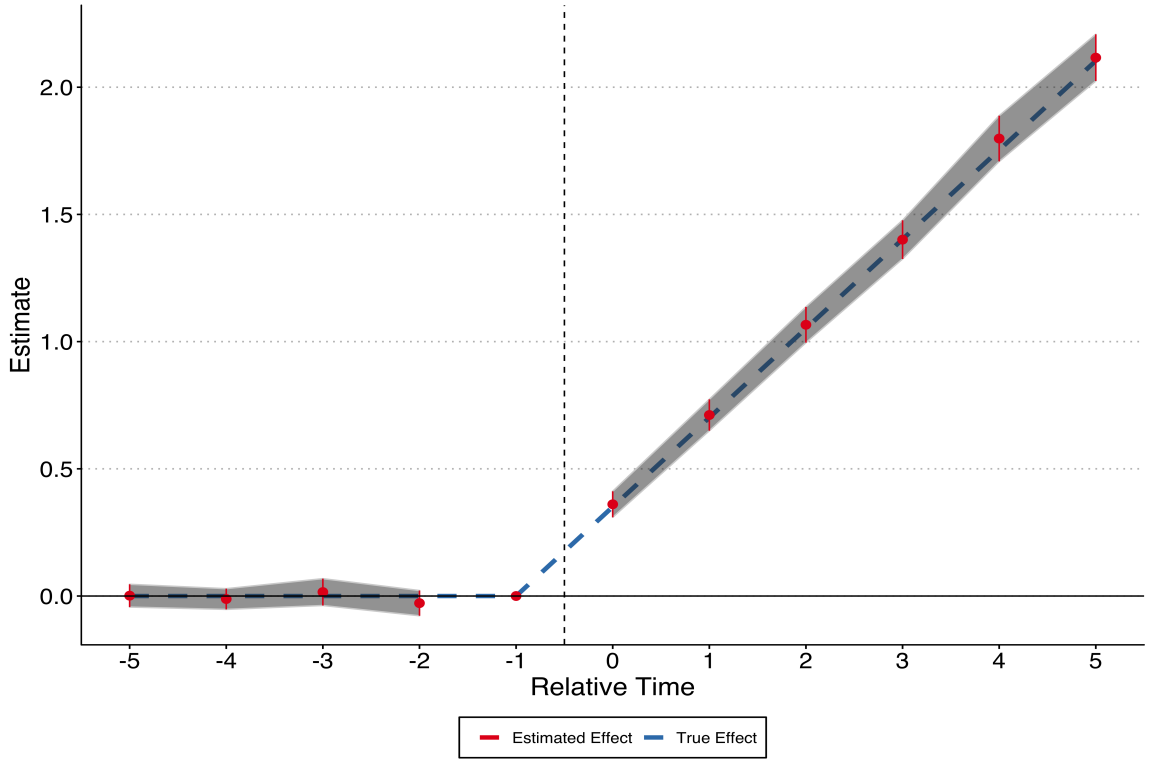


cannot bias any of the estimates. The results are presented in Figure 4 and show a similar positive treatment effect beginning

Finally, the method from Appendix D of CLDZ 2019 can be used to test the treatment effects on the simulated data. First, CLDZ proposes to run individual level DiD estimates of the treatment effect for each treated unit. For each unit treated before the 2009 treatment cohort[2] we create a clean dataset with the observations for the treated unit within the estimation window (t = -5 to 5), as well as all other units not treated by t = + 5, which are used as controls. We then estimate the unit-specific DiD estimate δ_i^{DD} and plot the sorted

[2]For units in the last treatment cohort (2009), there are no controls units to estimate with the DiD. These units are only used as effective controls

Fig. 4. Abraham & Sun (2019) DiD Design

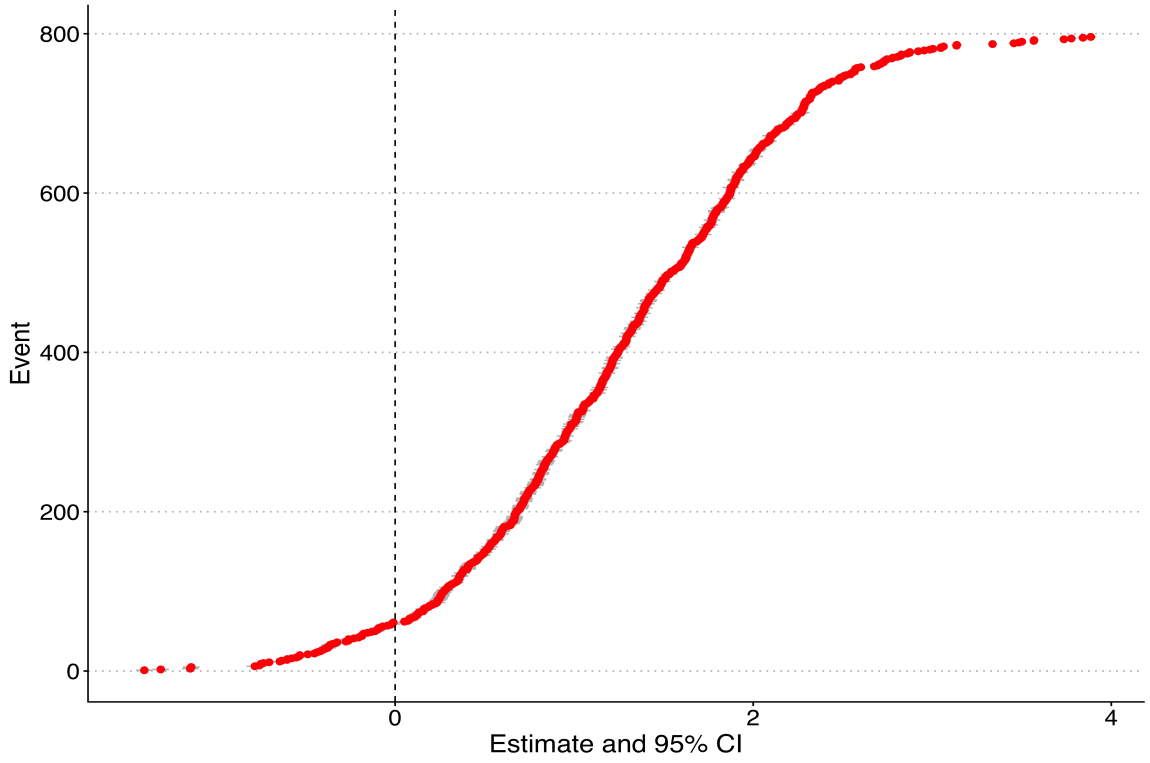


estimates and their confidence interval in Figure 5[3]. As is evident from the distribution of the individual DiD estimates, the clean-sample approach from CLDZ does effectively remove the bias from the standard TWFE estimates, as the large majority of individual estimates are correctly signed (showing positive treatment effects).

CDLZ (2019) also proposes to stack these event-specific data sets in relative time and calculate an average effect across all events using a single set of treatment effects. We stack our data at the cohort-specific rather than unit-specific level in order to reduce redundant observations. For each cohort $G_g \in \{1985, 1991, 1997, 2003\}$ we create a cohort specific dataset with all observations for units that are either i) in the treatment group G_g, or ii) are

[3]Note, these confidence intervals are likely too narrow, as CLDZ propose to use the Ferman and Pinto (2019) bootstrapped standard errors for inference which are not implemented here.

Fig. 5. CLDZ (2019) Individual DiD Estimates



not treated by year $G_g + 5$. The four cohort-specific datasets are stacked in relative time, and the following modified TWFE event-study regression is estimated:

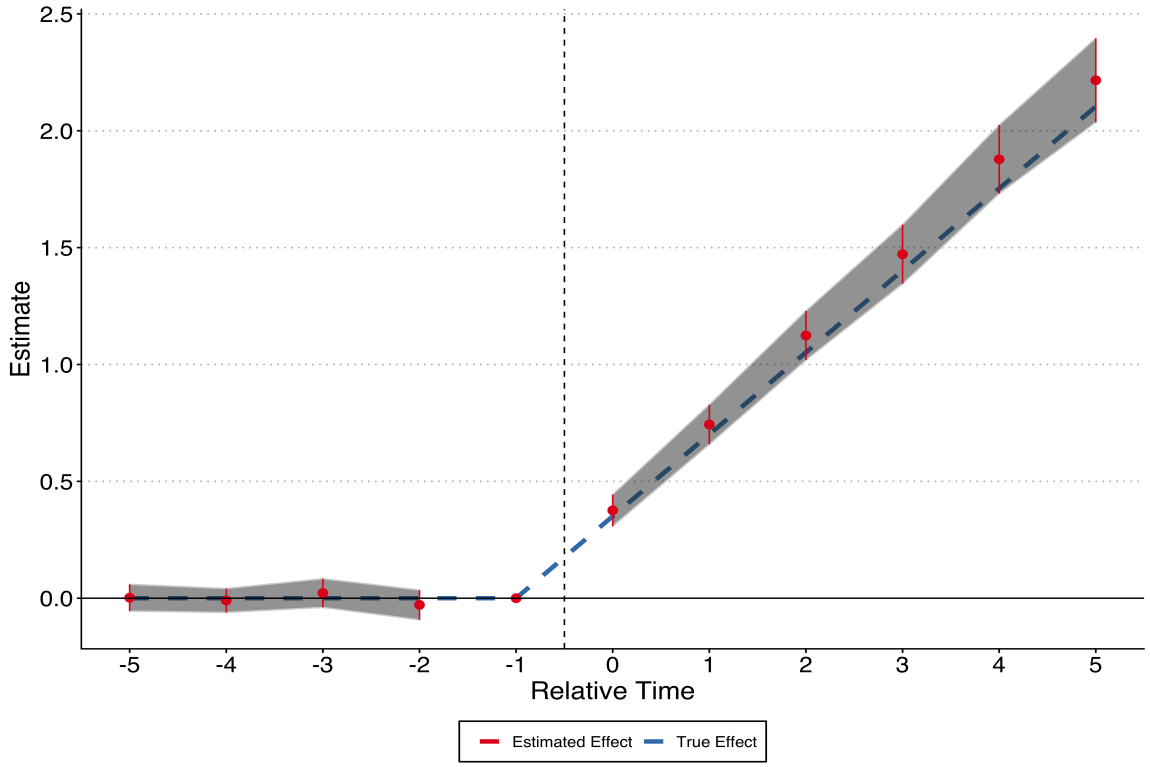
$$y_{it} = \alpha_{ig} + \alpha_{tg} + \sum_{k=-5}^{-2} \gamma_k D_{it} + \sum_{k=0}^{5} \gamma_k D_{it} + \epsilon_{it} \tag{4}$$

This is similar in structure to the standard TWFE event-study DiD, except that the fixed effects are now at the unit × cohort-dataset and year × cohort-dataset level, and the standard errors are clustered at the state of incorporation × cohort-dataset level, to adjust for potential repeated observations from the stacking process.[4] The results are reported in Figure 6 and show a similar trend in the treatment path with a lack of significant pre-trends

[4]Note, given the amount of year between treatment in our simulated data sample this would not be a problem. However, in most corporate governance contexts where there is temporal bunching in treatment adoption this correction would be necessary

and an increasing treatment effect in years post-adoption.

Fig. 6. CLDZ (2019) Stacked DiD Design



In conclusion, 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 no reason to believe that treatment effects are homogeneous across time or units. In addition, the modal 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 controls, that will either shrink or even flip the sign

of the treatment coefficient. However, all of the proposed models by CS, AS, CLDZ are sufficient to recover the true treatment path in the data.

4 Application to State antitakeover Laws

4.1 Prior Literature

As KW note, "empirical corporate finance research is plagued by endogeneity." Financial market outcomes are set in equilibrium, and the outcomes frequently modeled in the literature are a function of both the economic environment and a firm's effort to maximize profits. The result of the equilibrium nature of the data generating process is that firm characteristics and outcomes are endogenous to the economic system, making causal interpretations of corporate policies challenging. As a result, researchers frequently assess the impact of notionally exogenous changes to the competitive environment through legislative, regulatory, or legal-interpretive changes that "shock" the equilibrium process. The identification strategy assumes that national or sector-level shocks are plausibly exogeneous, given that the changes are initiated *outside of* firms' control.

Legal and regulatory changes have been used to asses to the impact of tax changes on firm investment choice, securities laws and regulations on firm value, and banking deregulation on economic growth and innovation [Cite papers from KW. This might be too close in spirit to their writing, maybe we can bring up other examples] One line of research that has been particularly contentious among corporate governance scholars is the effect (or lack thereof) of state antitakeover statues on firm value or profitability. Beginning with Garvey and Hanka (1999), Bertrand and Mullainathan (1999), and most notably Bertrand and Mullainathan (2003), researchers have exploited ostensibly exogenous changes in state antitakeover provisions to examine the relationship between manager-entrenchment and firm

outcomes. Bertrand and Mullainathan (2003) found that business combination statutes (BC) in particular had a marked impact on firm-level governance, and used shocks to BC law coverage to examine the relationship between governance and value. In a survey of the field, KW found 78 additional papers that use antitakeover law adoption to identify the effects of firm governance changes on operating performance, innovation, and stakeholder relations.

Bertrand and Mullainathan (2003), relying on legal arguments set forth in Romano (1987), argues that business combination statutes are exogeneously determined for firms covered by their authority. They use the passage of a BC law as an unconfounded increase in takeover protection or manager entrenchment, which represents a decrease in firm governance under standard governance theory. Given that these antitakeover protections are adopted at different times, and set at the level of the state of incorporation rather than the headquarters state of a corporation, BM (2003) argues that the results of a difference-in-differences analysis around the adoption of the laws is unlikely to be driven by contemporaneous changes that affect all firms in a given state. Using this identification strategy, BM (2003) finds that the passage of BC laws is associated with increases in employee wages, a decrease in the rate of plant closures, and reduced levels of productivity and profitability.

While BM (2003) focus on BC laws as the exogenous source of variation in firm antitakeover protection, researchers have identified other forms of legislation that act in a similar manner. These include poison pill laws that explicitly allow a firm to adopt a poison pill to thwart a takeover attempt, directors duties laws which mandate that directors take into consideration the interest of all firm stakeholders in addition to shareholders' interests, and control share acquisition laws that revoke the voting rights of a bidder unless a majority of the other shareholders vote to restore them. It has been argued that all of these antitakeover statutes theoretically provide protection of unsolicited takeovers, thereby increasing manager entrenchment and reducing firm governance.[5] As a result, a robust literature has developed

[5]Assuming you believe an open market for corporate control is unequivocally good for firm governance,

using the range of antitakeover provisions that been devised at the state level.

While scholars in economics, finance, and accounting have used the adoption of anti-takeover provisions to examine the effect of firm governance, others have raised objections to the theoretical basis for determining that the legal adoptions provide exogenous sources of variation. Karpoff and Malatesta (1989) and Gartman (2000) identify firms that lobbied for specific state antitakeover laws, violating the theoretical exogeneity implied in the DiD design. Coates (2000) and Klausner (2013) argue that essentially all firms can adopt a poison pill at any time, both before and after a takeover attempt. Thus, even firms that do not have an effective poison pill in place have a "shadow" poison pill, which in conjunction with a classified board acts as an effective deterrent in most takeover attempts. Cain, McKeon, and Solomon (2017) argues that the legal context matters, and advocates the construction of a "Takeover Index" that includes controls for court decisions, macroeconomic conditions, and firm characteristics. Finally, Catahan and Kahan (2016) provide a broad-ranging critique of the literature on state antitakeover protections, building upon Coates (2000) and Klausner (2013) to argue that all other takeover defenses above and beyond the shadow pill are legally redundant, and as a result inconsequential in practice.

KW (2018) provides a theoretical justification for the potential relevance of antitakeover statutes, and an empirical re-evaluation of the literature in light of the critiques described above. KW (2018) argues that the Catahan-Kahan critique is an extreme view, which would require that "poison pills are costless for managers to implement and, if combined with classified boards, offer takeover protection that is so unambiguously effective as to render other types of defenses irrelevant." It finds this argument uncompelling given that the adoption of a poison pill is professionally costly for managers, and that the resolution of uncertainty regarding the legal validity of poison pills took years to develop. In addition, while agreeing with CK that some prior findings are the result of model misspecification,

a contentious yet commonly adopted position in academic finance and accounting.

KW (2018) argues that "one would have to maintain unrealistically strong priors to therefore infer that all previous results are spurious."

4.2 Replication of Karpoff-Wittry (2018) Results

Separate from the theoretical disputes described above, every study in the empirical antitakeover literature assumes that DiD through linear fixed effects (often but not always using two-way fixed effects) is a methodologically appropriate mechanism to test for treatment effects in this setting. Given our simulation results, as well as the staggered nature of antitakeover legislation adoption documented in KW (2018), this assumption may be unwarranted. In this section we replicate the results from KW (2018).[6] Table IV of KW (2018) presents the results for the impact of business combination laws on seven different outcome variables (return on assets (ROA), capital expenditure (CAPEX), growth in Plants, Property, and Equipment (PPE Growth), asset growth, cash holdings, selling, general and administrative expenses (SGA Expense), and leverage). For each of the seven dependent variables, KW (2018) tests the effect of business combination statutes using two DiD specifications - what the authors call the "short regression model" and the "full regression model."

The short regression model is estimated as:

$$y_{it} = \alpha_{i,lt,jt} + \delta BC_{st} + \theta_1^{'} X_{it}^1 + \epsilon_{it} \tag{5}$$

where y_{it} is one of the seven dependent variables, $\alpha_{i,lt,jt}$ are firm, state-year, and industry-year fixed effects, BC_{st} is an indicator for whether the state of incorporation s for firm i has passed a business combination statute, and X_{it}^1 is a matrix of firm-level covariates (including controls for firm size, the square of firm size, age, and the square of firm age).

In light of the issues raised concerning the coexistence of multiple overlapping antitakeover

[6] We thank Jonathan Karpoff and Michael Wittry for generously providing their replication files and data.

provisions and violations of plausible exogeneity, KW (2018) argues that the short regression model in Equation 5 suffers from omitted variable bias. Due to the presence of other state laws, court cases, and firm-level defenses, the correctly specified model according to KW (2019) is:

$$y_{it} = \alpha_{i,lt,jt} + \delta BC_{st} + \theta_1^{'} X_{it}^1 + \theta_2^{'} X_{it}^2 + \epsilon_{it} \tag{6}$$

This "full regression model" is identical to the short model, with the addition of a matrix of additional control variables X_{it}^2. These additional controls include indicators for the presence of "first generation" antitakeover laws, poison pill (PP) laws, control share acquisition laws (CS), directors' duties laws (DD), and fair price laws (FP), as well as indicators for firms covered by control share acquisition laws after the U.S. Supreme Court upheld Indiana's CS law, firms covered by BC laws after the 7th Circuit upheld Wisconsin's BC law, and indicators for firms covered by BC laws and which helped motive the business combination law through lobbying.[7]

The results of the models are replicated in Table 4 below,[8] showing that, across choice in dependent variable, the standard DiD estimate for staggered BC law adoption is statistically significant when using the short regression model. However, once you control for preexisting statutes and characteristics the results are attenuated and only statistically significant when used to test changes in asset growth, SG&A expense, and leverage. In addition, the coefficient for the presence of a poison pill law is statistically significant in the full regression model for three variables, which is consistent with the observation in KW that "poison pill laws are associated with a larger average share price reaction than business combination laws." As a result, KW argue that "researchers seeking measures of exogenous variation in takeover

[7]We draw no conclusions on whether this is an effective method to control for the potential exogeneity issues at hand.

[8]The standard errors in our replication table are slightly different than the published results due to differences in how R and Stata calculate cluster-robust standard errors.

vulnerability should consider poison pill laws."

Table 4. Table IV from Karpoff-Wittry (2018)

	Dependent Variable													
	(1) ROA		(2) Capex		(3) PPE Growth		(4) Asset Growth		(5) Cash		(6) SGA Expense		(7) Leverage	
	Short Regression	Full Model	Short Regression	Full Model	Short Regression	Full Model	Short Regression	Full Model	Short Regression	Full Model	Short Regression	Full Model	Short Regression	Full Model
Business combination law (BC)	-0.017*	-0.008	0.003**	0.002	-0.016**	-0.008	-0.044*	-0.041**	-0.008***	-0.005	0.018**	0.011**	0.023**	0.020***
	(0.009)	(0.007)	(0.002)	(0.002)	(0.007)	(0.012)	(0.023)	(0.019)	(0.003)	(0.003)	(0.008)	(0.005)	(0.009)	(0.008)
First-generation law		-0.030**		0.004		-0.001		-0.026		0.004		0.017		-0.019
		(0.015)		(0.003)		(0.012)		(0.021)		(0.006)		(0.011)		(0.021)
Poison Pill law (PP)		-0.011**		0.002		-0.005		-0.027		-0.002		0.012**		0.027***
		(0.005)		(0.001)		(0.012)		(0.018)		(0.003)		(0.005)		(0.009)
Control share acquisition law (CS)		-0.021**		0.002		-0.003		0.018		0.008*		0.011		0.020
		(0.009)		(0.003)		(0.017)		(0.028)		(0.004)		(0.008)		(0.013)
Directors' duties law (DD)		0.002		-0.001		0.010		0.022		0.005		-0.004		-0.007
		(0.008)		(0.001)		(0.011)		(0.022)		(0.004)		(0.008)		(0.011)
Fair price law (FP)		-0.009		0.003		-0.017		0.014		-0.003		-0.001		0.000
		(0.010)		(0.002)		(0.012)		(0.024)		(0.003)		(0.008)		(0.013)
CS x CTS		0.000		0.000		-0.012		-0.029		0.000		0.021**		0.004
		(0.011)		(0.002)		(0.016)		(0.028)		(0.006)		(0.010)		(0.020)
BC x Amanda		-0.026*		0.002		-0.016		-0.024		-0.003		0.026**		0.015
		(0.015)		(0.002)		(0.030)		(0.026)		(0.006)		(0.011)		(0.015)
BC x MF (motivating firms)		0.201***		0.012		0.060*		0.142***		-0.041***		-0.114***		-0.057
		(0.046)		(0.009)		(0.035)		(0.054)		(0.013)		(0.038)		(0.038)

The critical assumption in any DiD application is that of "parallel trends", which holds that, in the absence of treatment, both treatment and control units would have the same trend in outcome path. The two-way differencing will only recover an unbiased measure of δ^{DD} if the assumption is satisfied. Unfortunately, the parallel trends assumption is inherently untestable; we only see one realization of the world. Faced with this uncertainty, empiricists have chosen to "use deduction as a second best for checking the assumption" by replacing the binary indicator TWFE DiD with an "event-study" specification (Cunningham Mixtape). The event study specification replaces a binary treatment variable with a set of lead and lag indicators for individual time periods. The goal of the event-study design is to test the dynamics of the treatment effect over periods around treatment adoption. In particular, if there is evidence of different relative changes in the outcome variable in periods *preceding* the adoption of the treatment, then it is more challenging to argue that, *cetirus paribus*, the treated and control outcomes would have changed in a corresponding manner without treatment.

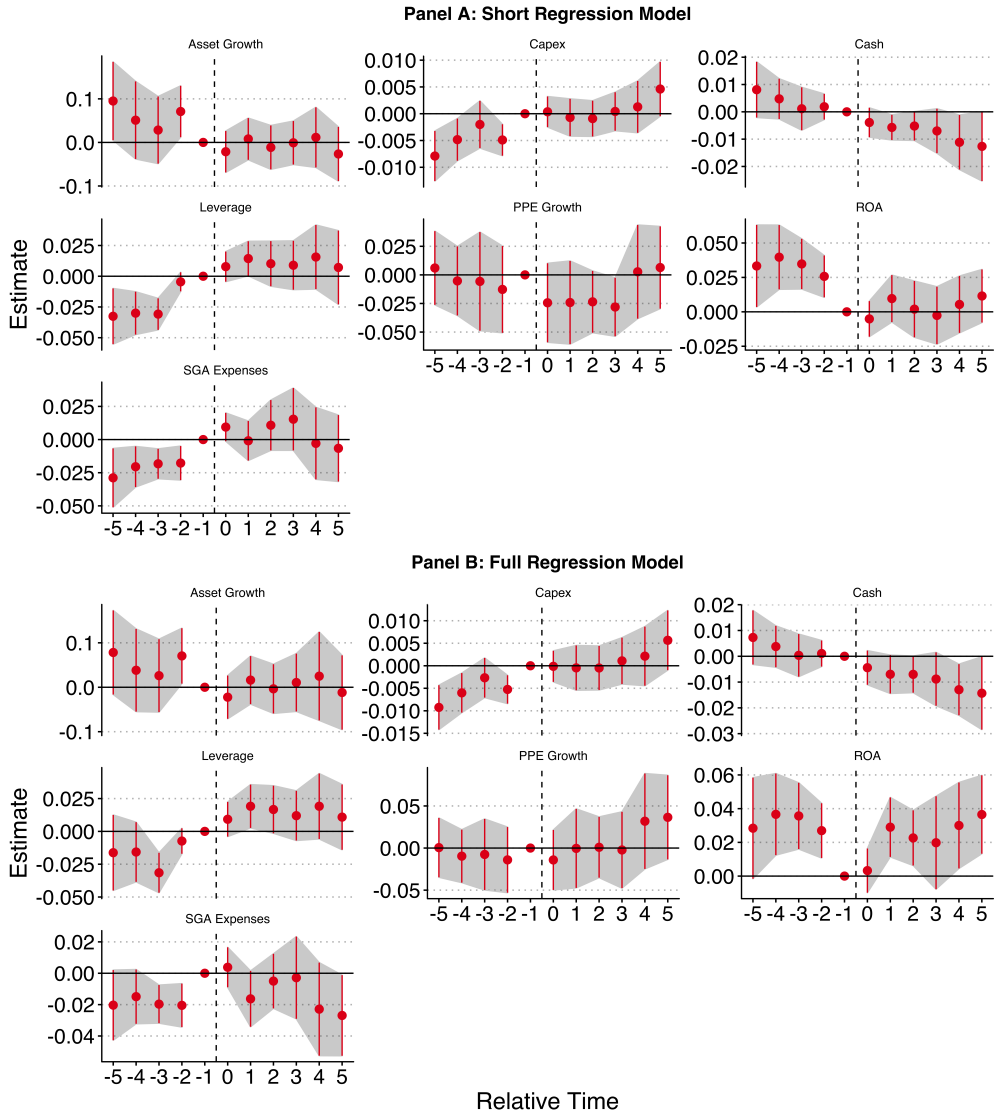
The event-study DiD approach is a simple extension to the standard TWFE fixed effects model. Assume that that the DiD estimation controls for a set of fixed effects and covariates $(\alpha_k; X_{it}) \equiv \Omega$. The event study method estimates the DiD as:

$$y_{it} = \mu_{Pre} + \sum_{k=l_*}^{-2} \gamma_k D_{st} + \sum_{k=0}^{l^*} D_{st} + \mu_{Post} + \Omega + \epsilon_{it} \tag{7}$$

Here we substitute a set of lead and lag indicators for time periods between l_* and l^* (the treatment estimation window) from the first year of treatment, as well as indicators for all years less than l_* (Pre) and more than l^* periods (Post) from treatment, for the binary treatment indicator. One period needs to be excluded to avoid collinearity; it is common practice to exclude the indicator for $k = -1$, the year before treatment. All coefficients on γ_k are therefore measured in relation to the fixed effect for the year before treatment.

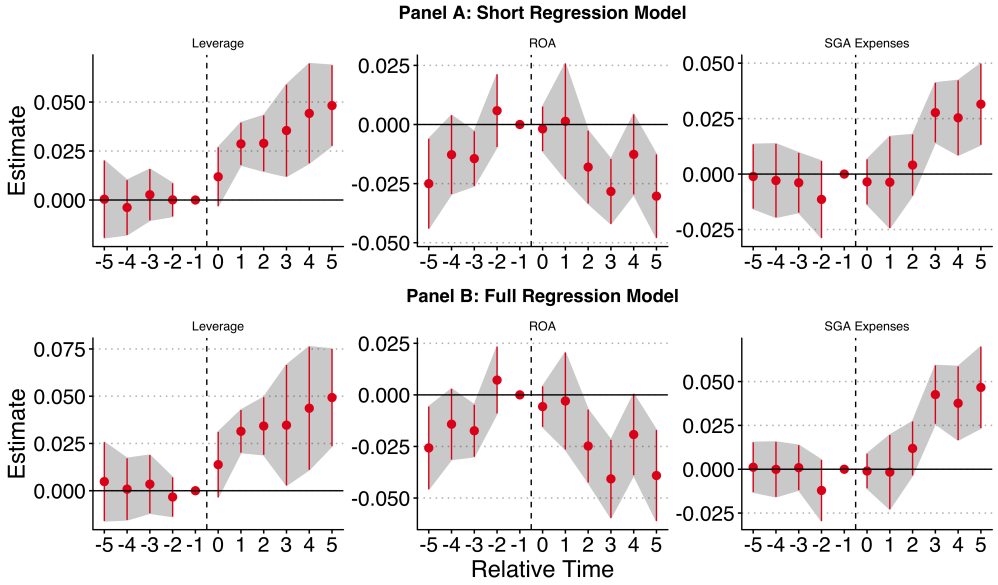
The event-study estimates for BC law adoption are reported in Figure 7 below. Panel A presents the dynamics for each outcome variable around BC law adoption for the "short regression model" (with a limited set of covariates), while Panel B presents the results for the preferred model in KW (2018). In both specifications, there is either i) no statistically significant evidence of a change in outcome variable, or ii) evidence of significant reaction in advance of the legislation. This suggests that the differences between the post and pre-period measured by a the TWFE DiD estimate is likely confounded by secular differences between control and treatment states, and/or evidence of endogenous treatment adoption.

Fig. 7. Event Study Estimates for BC Law



KW (2018) also stress the robust effects of poison pill statute adoption using the preferred full specification, and advocate for scholars to analyze PP law adoption in addition to BC laws. The same event study estimates are presented in Figure 8, where we substitute in the lead and lag indicators for years around adoption of PP laws for firms incorporated in covered states, focusing on three significant dependent variables from Table IV in KW (2018) (ROA, SGA Expense, and Leverage).

Fig. 8. Event Study Estimates for PP Law



Here the event study results more closely approximate the experimental ideal. The pre-trend identifiers are mostly clustered around zero[9] and not statistically significant from zero (i.e. not different from the year prior to treatment). While a nascent literature has developed discussing how to treat DiD estimates with evidence of a violations in the parallel trends assumptions (cite Roth here), because the purpose of this paper is to address the specific issues of staggered treatment DiD, we will focus on the PP law results for Leverage, ROA, and SGA Expenses, where these concerns are not present.

4.3 Remedies

5 Conclusion

[9]This is particularly true when estimating the relationship for leverage and SGA expenses, while ROA here exhibits some evidence of mean reversion.

References

Fig. 9. Placebo Tests by Time

Figure 9 displays...

Appendix A Description of Variables

This table defines variables used in our analysis.

Variable	Description	Computation
Accruals to Assets	Ratio of total accruals to total assets	(Net Income [WC07250] – Funds from Operations [WC04201]) / Total Assets [WC02999]

Appendix A Continued

Variable	Description	Computation
Log Market Cap	Natural logarithm of market capitalization	ln(Market Value [MV])

Table 1. BLANK

Columns 1 and 2 report... Standard errors, clustered at the firm level, are reported in parentheses. Significance levels are indicated by *, **, and *** for 10%, 5%, and 1% respectively.