

# Canonical Research Designs I: Difference-in-Differences: Basics, Pre-trends, and Issues with Staggered Timing

## Revisiting Research Design

- Recall my attempt at a definition:
  - A (*causal*) *research design* is a statistical and/or economic statement of how an empirical research paper will estimate a relationship between two (or more) variables that is causal in nature: how  $X$  causes  $Y$ .
- Recall these research designs can be considered into two types:
  - *Model-based*: the estimand is identified using assumptions on the modeling of the potential outcomes conditional on treatment and additional variables
  - *Design-based*: the estimand is identified using assumptions on the treatment variable, conditional on the potential outcomes and additional variables

## Revisiting Research Design

- The design-based ideal is not necessarily achievable in all scenarios
- Recall the reason for wanting a randomized experiment: we want to estimate *counterfactual* outcomes
- Alternative way to do this: model the outcome (absent treatment) directly
  - The relationship between X and Y can be clearly articulated, but the “experimental design” analogy is more fraught/complicated
- This issue will become clear as we discuss our first topic

## Estimating causal effects in real settings

- In many applications, we want to estimate the effect of a policy across groups
- However, the policy assignment is *not* necessarily uncorrelated with group characteristics
- How can we identify the effect of the policy without being confounded by these level differences?

# Estimating causal effects in real settings

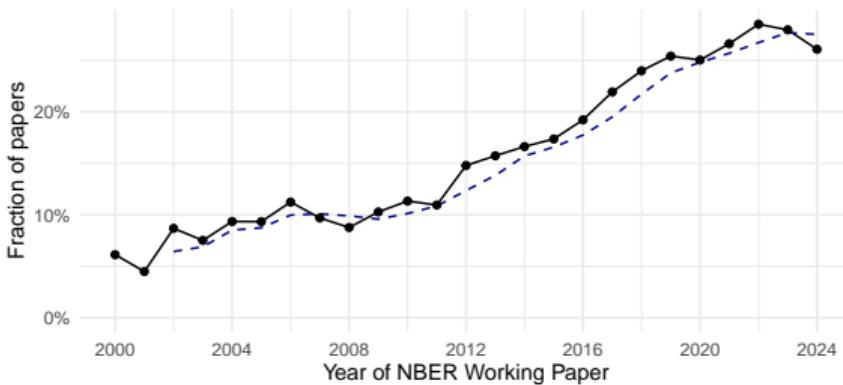
- In many applications, we want to estimate the effect of a policy across groups
- However, the policy assignment is *not* necessarily uncorrelated with group characteristics
- How can we identify the effect of the policy without being confounded by these level differences?

Difference-in-differences!  
(DiD)

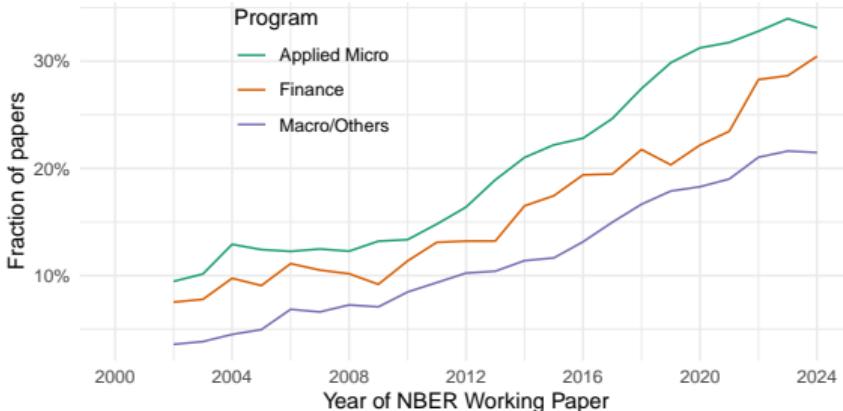
# Difference-in-differences

- Difference-in-differences has become the dominant methodology for estimating causal effects in applied work
- True across all fields  
(Goldsmith-Pinkham (2024))

Trends in difference-in-differences



Program



## First, a warning

- This literature has had a certain amount of upheaval over the past 5-6 years
- Tension: provide context for how people currently and historically have studied diff-in-diff
  - But also elaborate on concerns identified in recent papers
- The key issues boil down into two questions:
  1. *What is the counterfactual estimand?*
    - Does your estimator map to your estimand? (e.g. "Are you getting at what you meant to?")
  2. *What are your structural assumptions and their implications?*
    - Do you need to assume functional forms?
- Papers have both pointed out issues but also provided solutions to almost all of the problems that they've raised, so not something that should prevent you from using these tools

## Simple setup in basic 2x2 Diff-in-diff

- Assume we have  $n$  units ( $i$ ) and  $T = 2$  time periods ( $t$ )
- Consider a binary policy  $D_{it}$ , and we are interested in estimating its effect on outcomes  $Y_{it}$
- Consider the potential outcome notation for  $Y_{it}$ :
  - $Y_{it}(0, 0)$  is the outcome in period  $t \in \{1, 2\}$  if untreated in both periods
  - $Y_{it}(0, 1)$  is the outcome in period  $t \in \{1, 2\}$  if untreated in first period, treated in second
  - Can simplify to just  $Y_{it}(0)$  and  $Y_{it}(1)$ , but when we have many time periods, want to account for path of treatments
- The inherent problem is that  $D_{it}$  is *not* necessarily randomly assigned, but we still want to estimate the ATT in period 2:

$$\tau_2^{ATT} = E(Y_{i,2}(1) - Y_{i,2}(0) | D_i = 1)$$

## Basic 2x2 DiD setup

- Recall our typical estimand of interest is the ATE or the ATT:

$$\tau_{ATE} = E(Y_{it}(1) - Y_{it}(0)) = E(\tau_i)$$

$$\tau_{ATT} = E(Y_{it}(1) - Y_{it}(0) | D_{it} = 1) = E(\tau_i | D_{it} = 1)$$

- Since  $D$  is not randomly assigned, this model is inherently not identified without the additional assumptions (and two time periods).
  - Why?  $D_i$  could be correlated with  $\alpha_i$
  - Recall that our plug-in estimator approaches need estimates for  $E(Y_{it}(1))$  and  $E(Y_{it}(0))$
  - But the correlation prevents this without conditional exogeneity assumptions
- Using the assumptions and multiple time periods, we can make progress!

## $2 \times 2$ DinD estimation

	$t = 0$	$t = 1$
$D = 0$	$\gamma_0 + \alpha_i$	$\gamma_1 + \alpha_i$
$D = 1$	$\gamma_0 + \alpha_i + \tau_i$	$\gamma_1 + \alpha_i + \tau_i$

- Now consider the within unit difference:

$$Y_{i1} - Y_{i0} = (\gamma_1 - \gamma_0) + \tau_i(D_{i1} - D_{i0})$$

- Hence

$$E(Y_{i1} - Y_{i0}|D_{i1} - D_{i0} = 1) - E(Y_{i1} - Y_{i0}|D_{i1} - D_{i0} = 0) = E(\tau_i|D_{i1} - D_{i0} = 1)$$

- Wait, you say, that's a lot more notation than I was expecting.

- Simplifying assumption: treatment only goes one way in period 1
  - "absorbing adoption", e.g.  $D_{i0} = 0$

$$E(Y_{i1} - Y_{i0}|D_{i1} = 1) - E(Y_{i1} - Y_{i0}|D_{i1} = 0) = \underbrace{E(\tau_i|D_{i1} = 1)}_{ATT}$$

## Basic 2x2 Diff-in-diff

How can we identify the ATT?

1. Parallel Trends: “in the absence of the treatment, the average outcomes would have evolved in parallel”

$$E(Y_{i,2}(0) - Y_{i,1}(0)|D_i = 1) = E(Y_{i,2}(0) - Y_{i,1}(0)|D_i = 0)$$

- Absent the policy, units may have different *levels*, but their changes would be the same
- A sufficient parametric formulation:  $Y_{it}(0) = \gamma_t + \alpha_i + \epsilon_{it}$

2. No-anticipation: policy has no effect prior to treatment

$$E(Y_{i,1}(0)) = E(Y_{i,1}(1))$$

## Another way to see this, assuming absorbing treatment

- Rewrite the parallel trends assumption:

$$E(Y_{i,2}(0)|D_i = 1) = E(Y_{i,1}(0)|D_i = 1) + E(Y_{i,2}(0) - Y_{i,1}(0)|D_i = 0)$$

- In other words, the counterfactual “untreated” state is the untreated outcome in the pre-period for the treated group, plus the change from the other untreated group
- Then, thanks to no-anticipation, we can replace  $E(Y_{i,1}(0)|D_i = 1)$  with  $E(Y_{i,1}(1)|D_i = 1)$ , which has an empirical analog:

$$E(Y_{i,2}(0)|D_i = 1) = E(Y_{i,1}|D_i = 1) + E(Y_{i,2} - Y_{i,1}|D_i = 0)$$

- So parallel trends + no-anticipation generates our counterfactual outcome for us!
- Pop quiz: where does no-anticipation show up in the parametric formulation?

## An aside on our simplifying assumption on absorbing treatment

- The choice of focusing on take-up of a policy, such that  $D_{i1} \geq D_{i0}$ , is well-grounded in many policy settings
- However, there are cases where policies turn on, and then turn off, and this can vary across units
- This can be challenging and potentially problematic with heterogeneous effects
- De Chaisemartin and d'Haultfoeuille (2020) allow for this, but not innocuous
  - Additional assumption required (will discuss in subsequent classes)
- Is  $D_i$  turning on identical (but opposite sign) to  $D_i$  turning off?
  - Hull (2018) working paper on mover designs discusses this
- For today, will ignore this issue

## Estimation using linear regression

- A simple linear regression will identify  $E(\tau_i | D_{i1} = 1)$  with two time periods:

$$Y_{it} = \alpha_i + \gamma_t + D_{it}\beta + \epsilon_{it} \quad (1)$$

- This setup is sometimes referred to as the Two-way Fixed Effects estimator (TWFE)
- Note: we could have also estimated  $\tau$  directly:

$$\hat{\tau} = n^{-1} \sum_i D_i \underbrace{(Y_{i1} - Y_{i0})}_{\Delta \bar{Y}_1} - \underbrace{(1 - D_{i1})(Y_{i1} - Y_{i0})}_{\Delta \bar{Y}_0}$$

- Intuitively, we generate a counterfactual for the treatment using the changes in the untreated units:  $E(Y_{i1} - Y_{i0} | D_i = 0)$
- Necessary condition: two time periods! What if we have more?

## Multiple time periods in basic setup

- Let's consider a policy that occurs all at  $t_0$  (e.g. single timing rolled out to treated units)
- More time periods helps in several ways:
  1. If we have multiple periods *before* the policy implementation, we can partially test the underlying assumptions
    - Sometimes referred to as "pre-trends"
  2. If we have multiple periods *after* the policy implementation, we can examine effect timing
    - Is it an immediate effect? Does it die off? Is it persistent?
    - If you pool all time periods together into one "post" variable, this estimates the average effect.  
If sample is not balanced, can have unintended effects!
- How do we implement this?

$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=1, t \neq t_0}^T \delta_t D_{it} + \epsilon_{it},$$

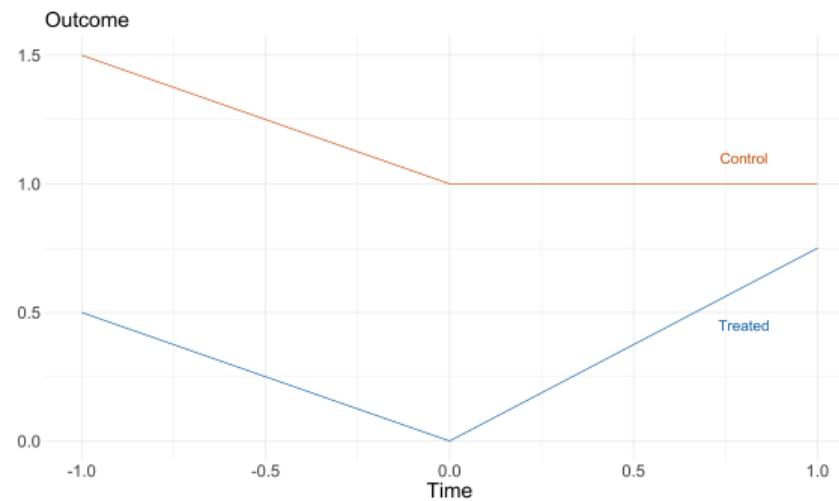
- One of the coefficients is fundamentally unidentified because of  $\alpha_i$
- All coefficients measure the effect *relative* to period  $t_0$ .

## Pre-testing and structural assumptions

- Note that for the above model, we made a stronger assumption about trends
  - We assumed that  $Y_{it}(d) - Y_{i,t-k}(d) = \gamma_t - \gamma_{t-k}$  for all  $k$  and  $d$
  - This is testable pre-treatment (hence the pre-test)
- This is very powerful and has helped spark the growth in DiD regressions
  - Visual demonstration of “pre-trends” helps support the validity of the design
  - Worth doing!
- Two key issues:
  1. Pre-testing can cause statistical problems
  2. What does parallel trends even mean?

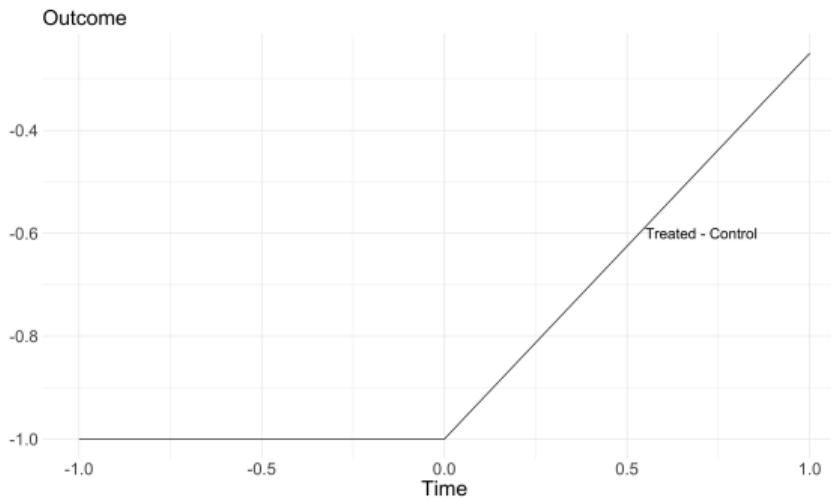
## Pre-testing issues (Roth 2020)

- Consider  $T = 3$  and think about what a pre-trend test is trying to do
- Testing whether the difference relative to  $t = 0$  for  $t = -1$  is significant



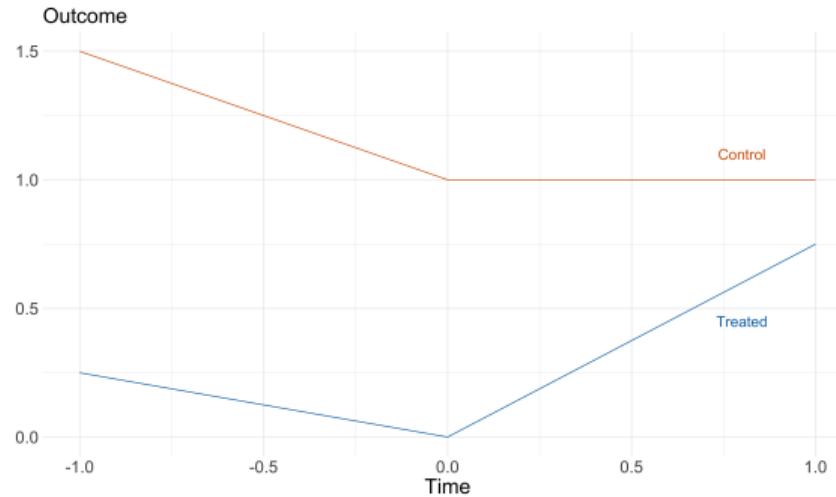
## Pre-testing issues (Roth 2020)

- Consider  $T = 3$  and think about what a pre-trend test is trying to do
- Testing whether the difference relative to  $t = 0$  for  $t = -1$  is significant



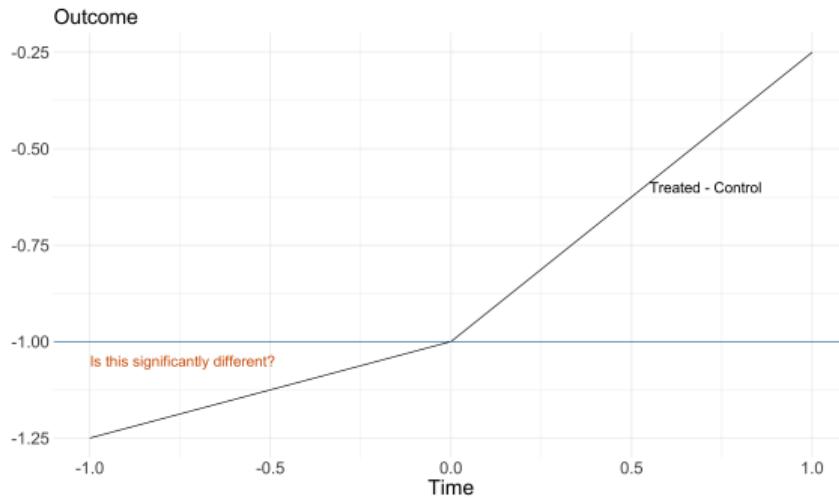
## Pre-testing issues (Roth 2020)

- Consider  $T = 3$  and think about what a pre-trend test is trying to do
- Testing whether the difference relative to  $t = 0$  for  $t = -1$  is significant
- Unconditionally, this is reasonable. However, Roth (2020) highlights that this is a form of *pre-testing*, and that low power in detecting pre-trends can be problematic



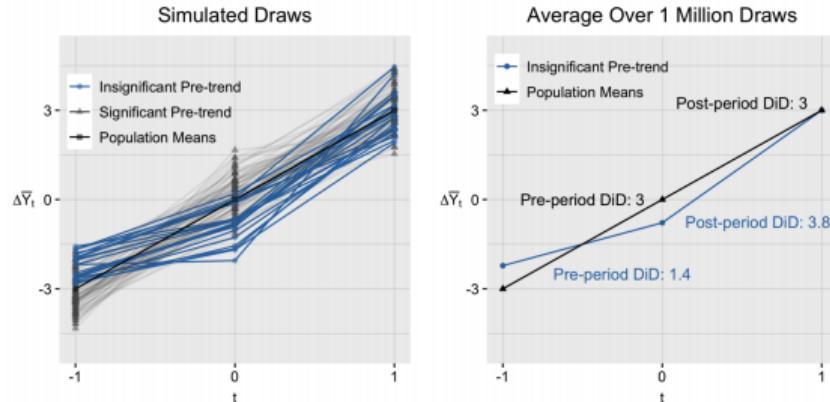
## Pre-testing issues (Roth 2020)

- Consider  $T = 3$  and think about what a pre-trend test is trying to do
- Testing whether the difference relative to  $t = 0$  for  $t = -1$  is significant
- Unconditionally, this is reasonable. However, Roth (2020) highlights that this is a form of *pre-testing*, and that low power in detecting pre-trends can be problematic



## Pre-testing issues (Roth 2020)

- Consider  $T = 3$  and think about what a pre-trend test is trying to do
- Testing whether the difference relative to  $t = 0$  for  $t = -1$  is significant
- Unconditionally, this is reasonable. However, Roth (2020) highlights that this is a form of *pre-testing*, and that low power in detecting pre-trends can be problematic
- By selecting on pre-trends that “pass”, will tend to choose baseline realizations that satisfy pre-trends, but induce *bias* in the effect



## How to interpret this caution?

- First, don't panic. Examining pre-trend is still important diagnostic
- Important to realize that selecting your design based on pre-trend is *constructing* your counterfactual
  - Pre-tests will cause you to potentially contaminate your design
- Suggested solution from Roth (2020): incorporate robustness to pre-trends into your analysis. Rambachan and Roth (2020) present results on testing sensitivity of DinD results to pre-trends
  - Brief intuition follows

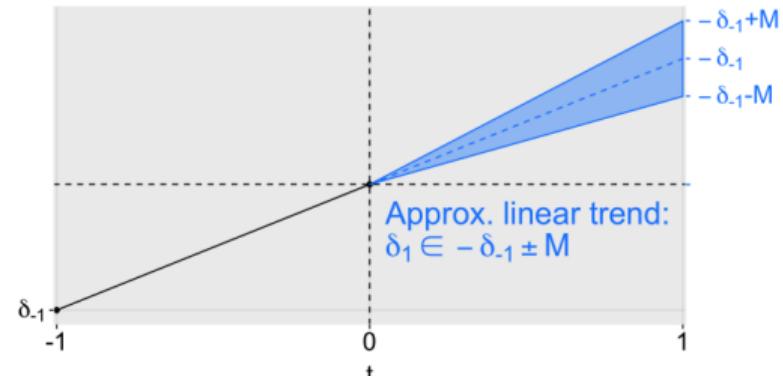
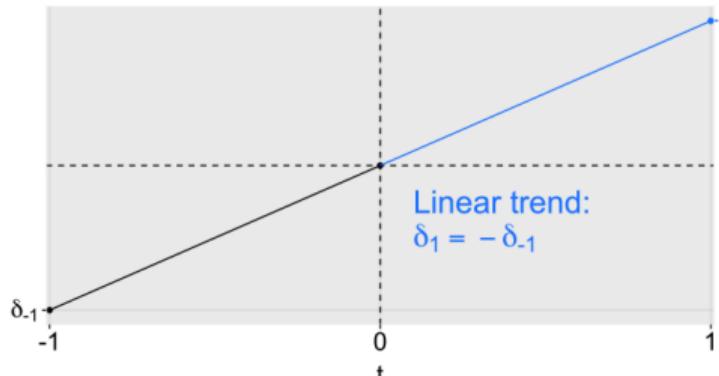
## Rambachan and Roth (2020) suggestion

- Intuitive proposed solution for robustness. Note the post and pre effects:

$$\mathbb{E} \left[ \hat{\beta}_1 \right] = \tau_{ATT} + \underbrace{\mathbb{E} [Y_{i,1}(0) - Y_{i,0}(0) | D_i = 1] - \mathbb{E} [Y_{i,1}(0) - Y_{i,0}(0) | D_i = 0]}_{\text{Post-period differential trend} =: \delta_1},$$

$$\mathbb{E} \left[ \hat{\beta}_{-1} \right] = \underbrace{\mathbb{E} [Y_{i,-1}(0) - Y_{i,0}(0) | D_i = 1] - \mathbb{E} [Y_{i,-1}(0) - Y_{i,0}(0) | D_i = 0]}_{\text{Pre-period differential trend} =: \delta_{-1}}.$$

- parallel trends assumes these  $\delta$  are zero. But pre-trends may not be zero.
  - R&R say: we can use the info from our pre-trends to bound post-trend
  - Use a *smoothness assumption*,  $M$ , on the second derivative. E.g. simple case:



## This approach adds more work but also more validity

- Need to select  $M$ , and will likely have less strong results
- However, very powerful way to address concerns about pre-trends
- Code for applying this technique is available in R:  
<https://github.com/asheshrambachan/HonestDiD>

## Parallel trends in what?

- A known issue that was historically not formalized is the question of what the outcome is specified as: logs, or levels?
- Hopefully it's clear that if something satisfies pre-trends in logs, it seems unlikely to satisfy in levels
- Recall that this is the issue of *invariance* we discussed with quantile treatment effects
  - In our parametric setting, if there are time trends in the outcomes, the parallel trends are likely not to hold for all transformations of the variables.
  - That could be problematic if you wanted to be agnostic about the model!
- Roth and Sant'Anna (2021) directly discuss this issue. Their suggestion:  
*Our results suggest that researchers who wish to point-identify the ATT should justify one of the following: (i) why treatment is as-if randomly assigned, (ii) why the chosen functional form is correct at the exclusion of others, or (iii) a method for inferring the entire counterfactual distribution of untreated potential outcomes.*

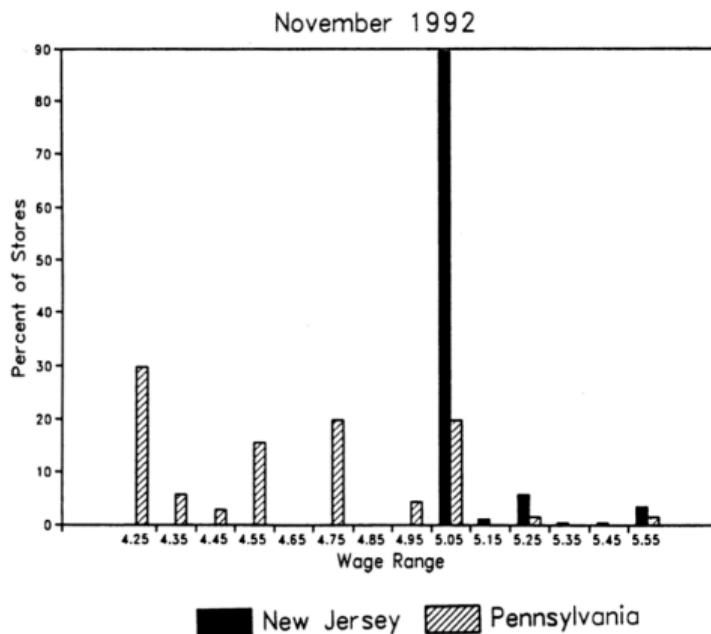
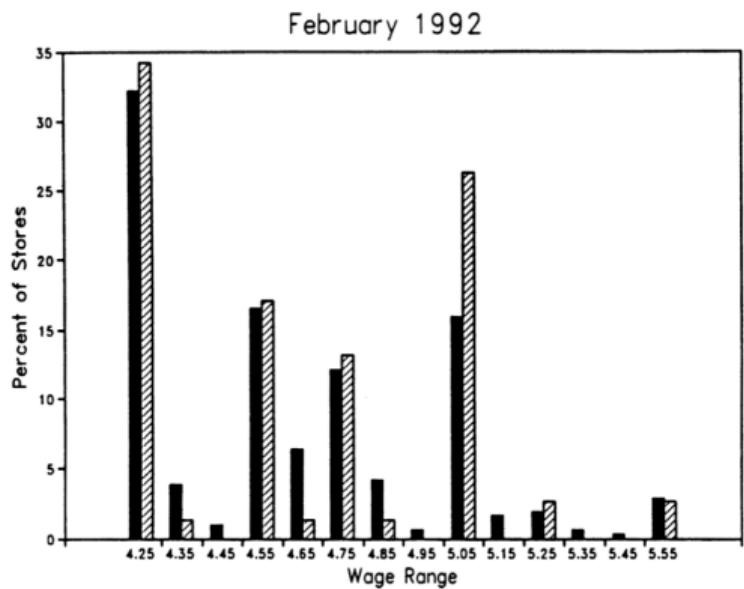
## Cases of DinD

- 1 treatment timing, Binary treatment, 2 periods
  - Card and Krueger (AER, 1994)
- 1 treatment timing, Binary treatment, T periods
  - Yagan (AER, 2015)
- Staggered treatment timing, Binary treatment
  - Bailey and Goodman-Bacon (AER, 2015)

## Card and Krueger (1994)

- Card and Krueger (1994) study the impact of New Jersey increasing the minimum wage 4.25 to 5.05 dollars an hour on April 1, 1992
- Key question is what impact does this have on employment?
  - Need a counterfactual for NJ, and use Pennsylvania as a control
- Collected data in 410 fast food restaurants
  - Called places and asked for employment and starting wage data
  - Sample data from Feb 1992 and Nov 1992
- Hence,  $D_i$  is NJ vs PA, and  $t = 0$  is Feb 1992 and  $t = 1$  is Nov 1992

# Stark Effect on Wages in Card and Krueger (1994)



# Effect on Employment in Card and Krueger (1994)

- Despite a large increase in wages, seemingly no negative impact on employment
  - In fact, marginally significant positive impact
- Looking at raw data, this positive impact is driven by a decline in PA
  - This decline is reasonable if you think that PA is a good counterfactual, since 1992 is in the middle of a recession
- A second comparison can be run with stores whose starting wage in pre-period was above treatment cutoff
  - These stores perform similarly to PA

Variable	Stores by state		
	PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Stores in New Jersey <sup>a</sup>		
Wage = \$4.25 (iv)	Wage = \$4.26–\$4.99 (v)	Wage ≥ \$5.00 (vi)
19.56 (0.77)	20.08 (0.84)	22.25 (1.14)
20.88 (1.01)	20.96 (0.76)	20.21 (1.03)
1.32 (0.95)	0.87 (0.84)	-2.04 (1.14)

## Key considerations for thinking about Card and Krueger (1994)

- The treatment can't really be thought of as randomly assigned
  - Treatment is completely correlated within states
  - As a result, any within-state correlation of errors will be correlated with treatment status
- Given the limited number of states, time periods, and treatments, more valuable to view this as a case study
  - Under strong parametric assumptions, can infer causality!
  - Card acknowledges (Card and Krueger interview with Ben Zipperer):

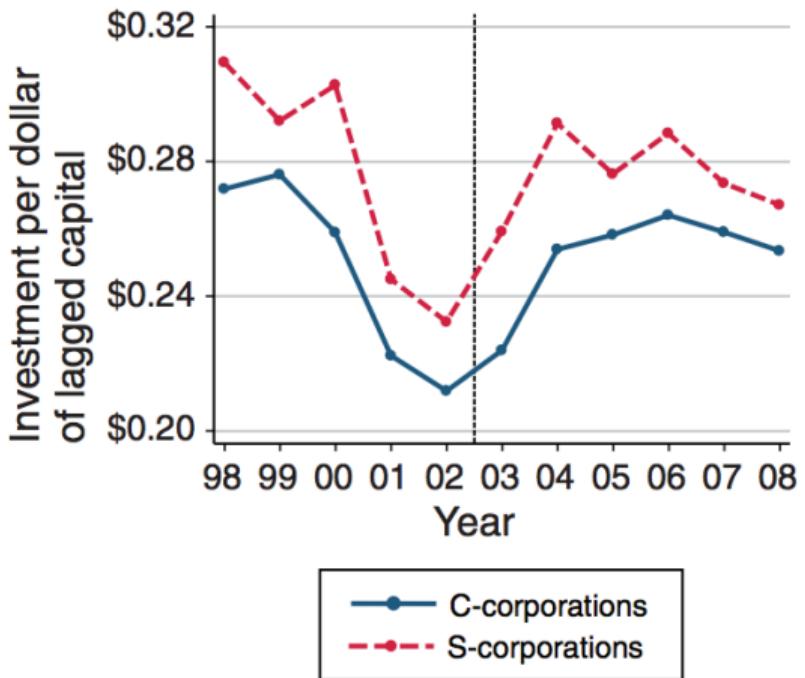
So the great advantage of a quasi-experiment or natural experimental like minimum wage is that it's a real intervention. It's real firms that are all affected. You get part of the general equilibrium effect. That's pretty important for understanding the overall story. The disadvantage is that someone can always say, well, it isn't truly random. And the number of units might be small. So you might only have two states. At some abstract level, there's only two degrees of freedom there. And so that's a problem.

## Yagan (2015)

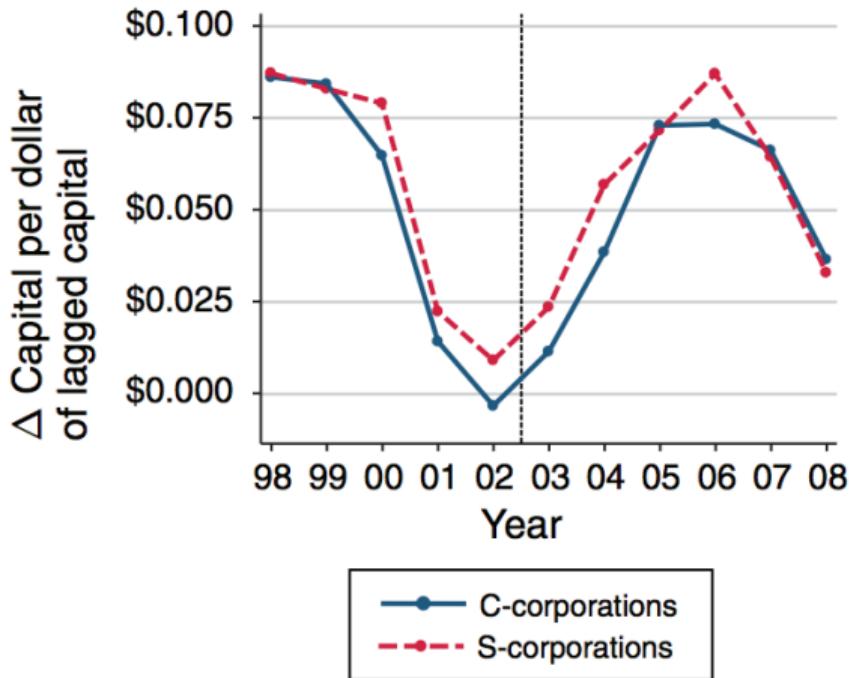
- Yagan (2015) tests whether the 2003 dividend tax cut stimulated corporate investment and increased labor earnings
- Big empirical question for corporate finance and public finance
- No direct evidence on the real effects of dividend tax cut
  - real corporate outcomes are too cyclical to distinguish tax effects from business cycle effects, and economy boomed
- Paper uses distinction between “C” corp and “S” corp designation to estimate effect
  - Key feature of law: S-corps didn’t have dividend taxation
- Identifying assumption (from paper):  
*The identifying assumption underlying this research design is not random assignment of C- versus S-status; it is that C- and S-corporation outcomes would have trended similarly in the absence of the tax cut.*

## Investment Effects (none)

Panel A. Investment

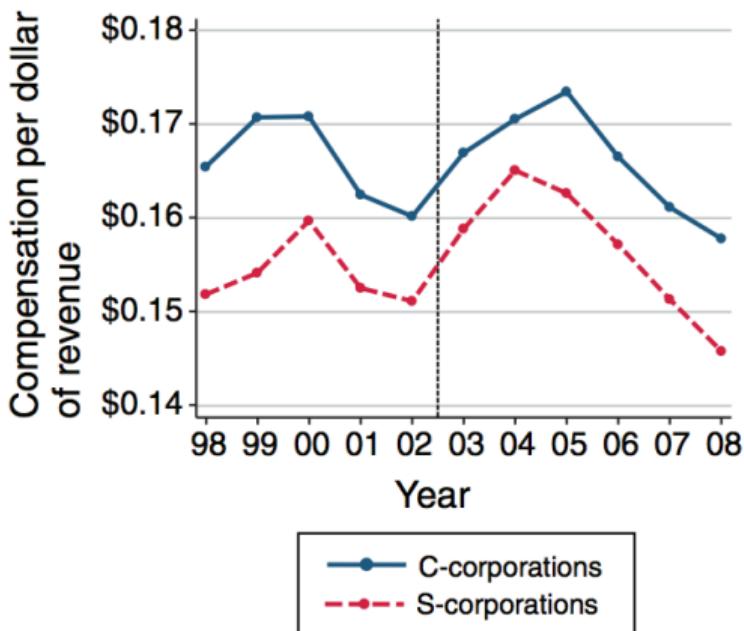


Panel B. Net investment

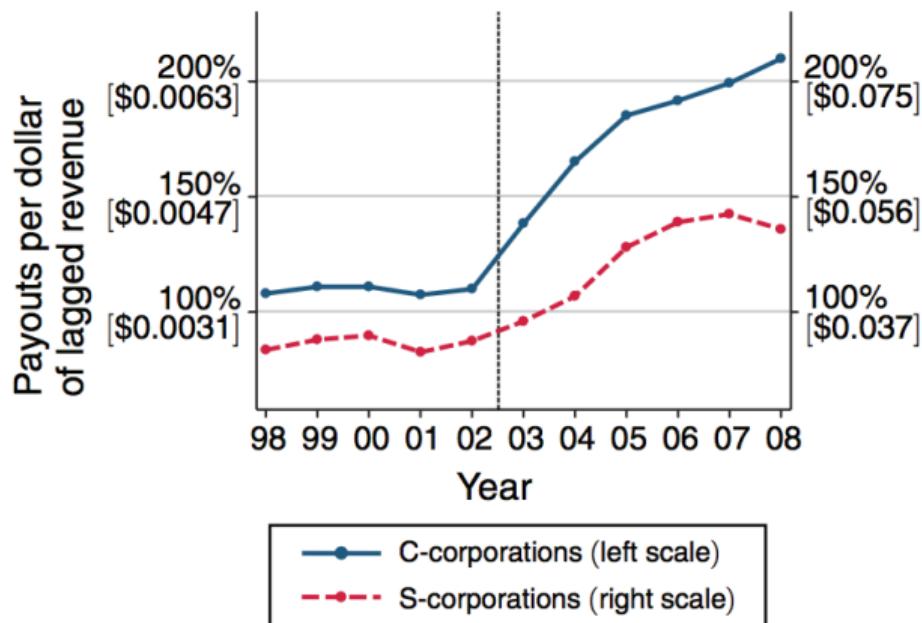


## Employee + Shareholder effects (big)

Panel C. Employee compensation



Panel D. Total payouts to shareholders



## Key Takeaway + threats

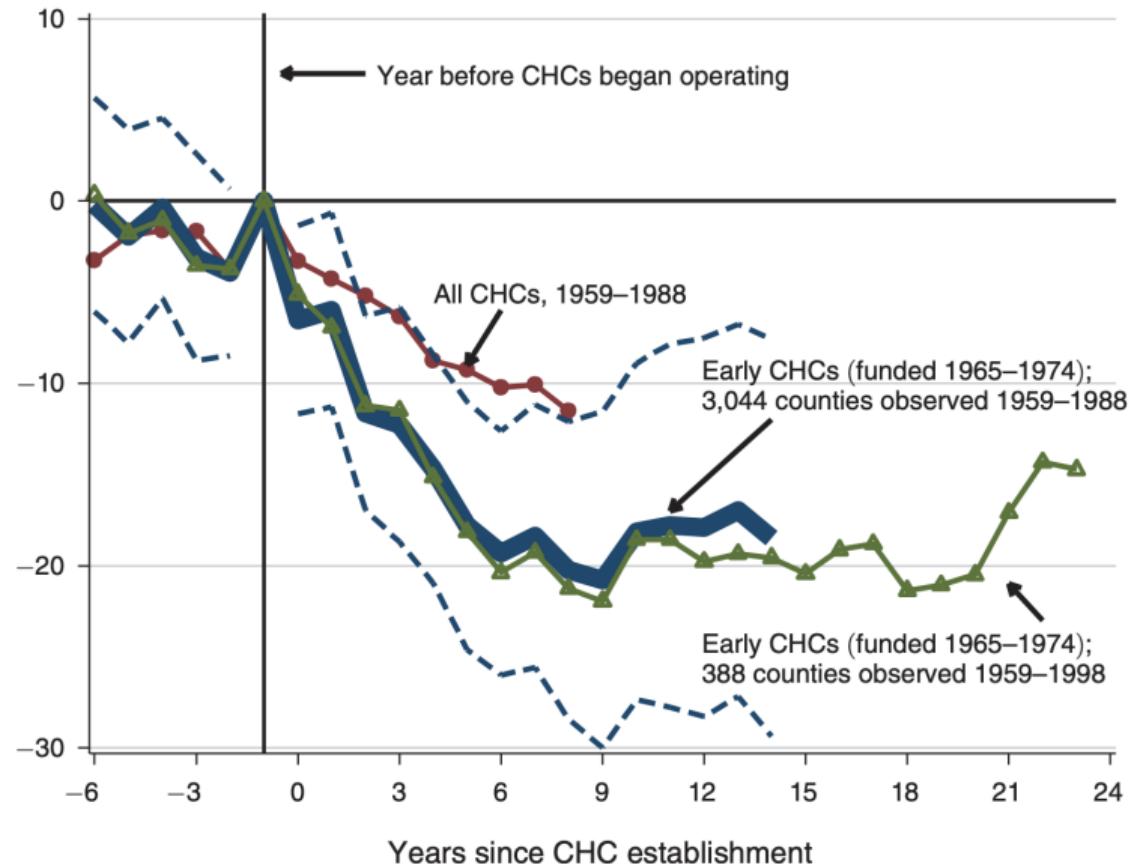
- Tax reform had zero impact on differential investment and employee compensation
- Challenges orthodoxy on estimates of cost-of-capital elasticity of investment
- What are underlying challenges to identification?
  1. Have to assume (and try to prove) that the only differential effect to S- vs C-corporations was through dividend tax changes
  2. During 2003, could other shocks differentially impact?
    - Yes, accelerated depreciation – but Yagan shows it impacts them similarly.
- Key point: you have to make *more* assumptions to assume that zero **differential** effect on investment implies zero **aggregate** effect.

## Bailey and Goodman-Bacon (2015)

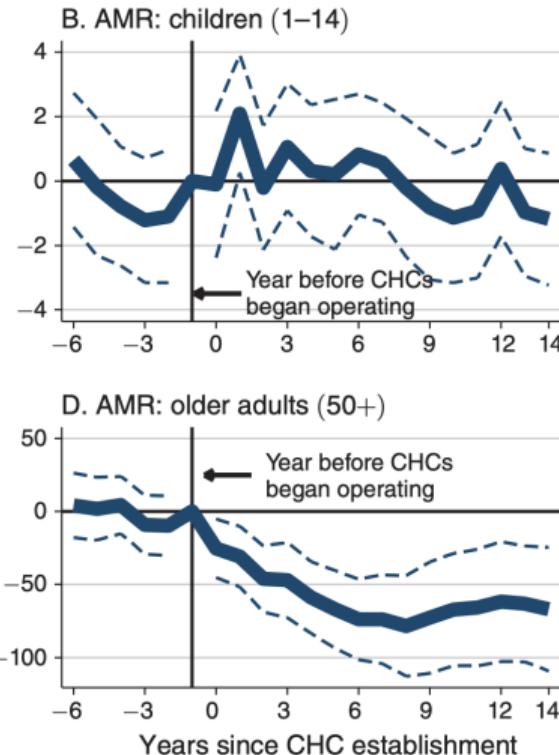
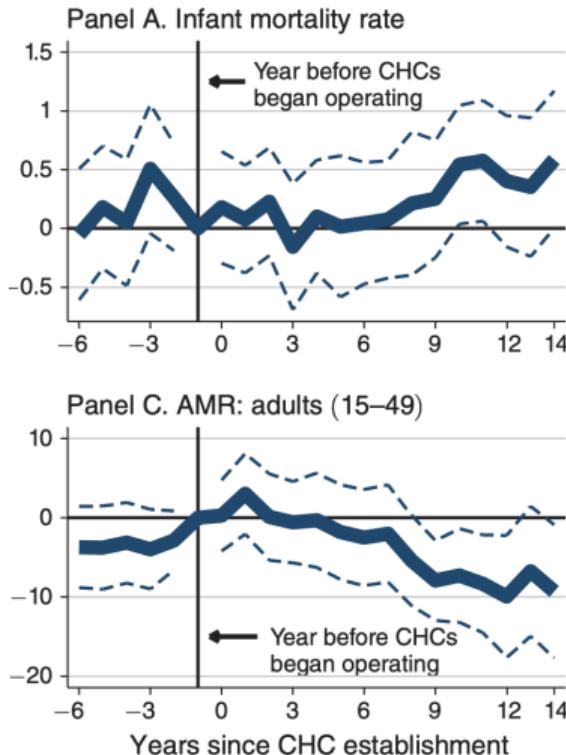
- Paper studies impact of rollout of Community Health Centers on mortality
  - Idea is that CHCs can help lower mortality (esp. among elderly) by providing accessible preventative care
- Exploit timing of implementation of CHCs

*Our empirical strategy uses variation in when and where CHC programs were established to quantify their effects on mortality rates. The findings from two empirical tests support a key assumption of this approach—that the timing of CHC establishment is uncorrelated with other determinants of changes in mortality.*
- Issue is that CHCs tend to be done in places
- Since CHCs are started in different places in different time periods, we estimate effects in event-time, e.g. relative to initial rollout.

## Negative effect on mortality



# Negative effect on mortality, particularly among elderly



## Key takeaways

- Since the policy changes are staggered, we are less worried about effect driven by one confounding macro shock.
- Easier to defend story that has effects across different timings
  - Also allows us to test for heterogeneity in the time series
- Still makes the exact same identifying assumptions – parallel trends in absence of changes

## But a big issue emerges when we exploit differential timing

- We have been extrapolating from the simple pre-post, treatment-control setting to broader cases
  - multiple time periods of treatment
- In fact, in some applications, the policy eventually hits everyone – we are just exploiting differential timing.
- If we run the “two-way fixed effects” model for these times of DiD

$$y_{it} = \alpha_i + \alpha_t + \beta^{DD} D_{it} + \epsilon_{it} \quad (2)$$

what comparisons are we doing once we have lots of timings?

- Key point: is our *estimator* mapping to our *estimand*?
- Well, what's our estimand?

## What is our estimand with staggered timings?

- There are a huge host of papers touching on this question
- Callaway and Sant'anna (2020) propose the following building block estimand:

$$\tau_{ATT}(g, t) = E(Y_{it}(1) - Y_{it}(0) | D_{it} = 1 \forall t \geq g), \quad (3)$$

the ATT in period  $t$  for those units whose treatment turns on in period  $g$ .

- In the 2x2 case, this was exactly our effect!
- This paper assumes absorbing treatment, but can be weakened in other papers (de  
Chaisemartin and d'Haultfoeuille (2020) discuss this)
- It seems very reasonable that for our overall estimand, we want some weighted  
combination of these ATTs
- Callaway and Sant'anna (2020) highlight two ways to identify the above estimand:
  1. Parallel trends of treatment group with a group that is “never-treated”,  $G_i = \infty$
  2. Parallel trends of treatment group with the group of the “not yet treated”,  $G_i \geq t + 1$
- Using these estimands, C&S provide a very natural set of potential ways to aggregate  
these estimands up

## Wait what happened to TWFE?

- It turns out that the logic of the TWFE does not naturally extend to differential timings
  - Recall that from our discussion of linear regression, regression is great because it does a variance weighted approximation:

$$\tau = \frac{E(\sigma_D^2(W_i)\tau(W_i))}{E(\sigma_D^2(W_i))}, \quad \sigma_D^2(W_i) = E((D_i - E(D_i|W_i))^2|W_i)$$

- It turns out that in the panel setting with staggered timings, these weights are not necessarily positive
- Why? The predicted value of  $D_{i,t}$ ,  $\hat{D}_{it}$ , may be greater than 1 because  $E(D_i|W_i)$  is the model is incorrectly specified by the two-way fixed effects:

$$\hat{\beta}_{post} = \frac{\sum_{i,t}(D_{it} - \hat{D}_{i,t})Y_{i,t}}{\sum_{i,t}(D_{it} - \hat{D}_{i,t})^2}$$

## Misspecification in conditional expectation

- An example from GPHK (2022): Three time periods, and three groups of units: treated early in period 2 ( $n_E$ ), treated late in period 3 ( $n_L$ ), and never ( $n_N$ )
- The weights in the estimands on the treatment effects will be

$$\lambda(j, t = 3) = \frac{n_E + 2n_N}{\kappa}, j \in L \quad (4)$$

$$\lambda(j, t = 2) = \frac{n_N + 2n_L}{\kappa}, j \in E \quad (5)$$

$$\lambda(j, t = 3) = \frac{n_N - n_L}{\kappa}, j \in E, \quad (6)$$

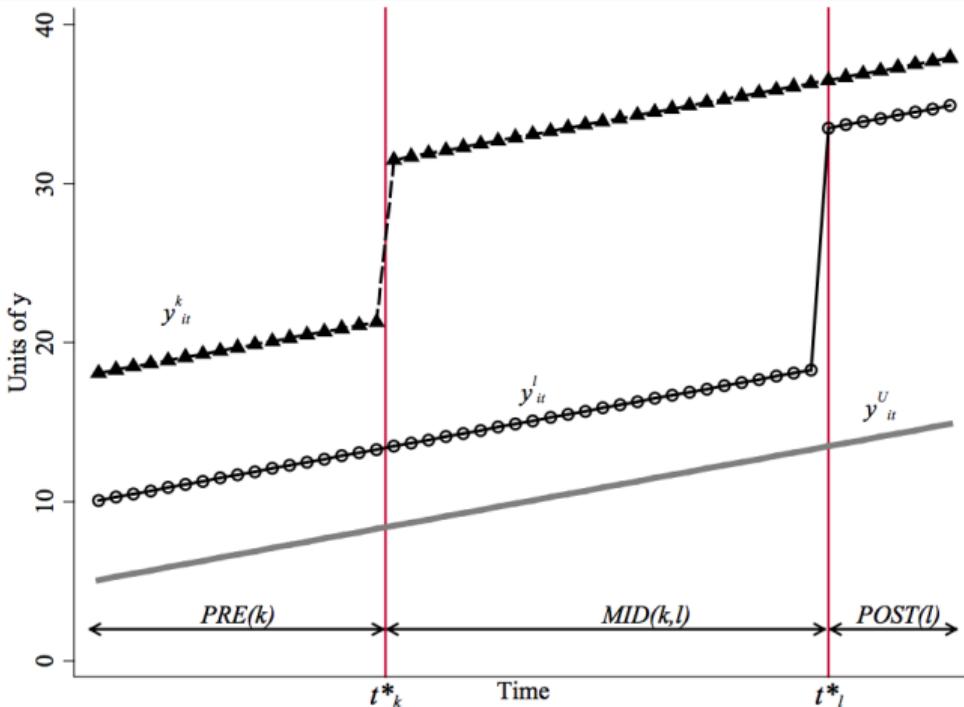
$\kappa$  just scaling. Hence, first two always get positive weight, but the ATT from period 3 of the early adopters is negative if the never-adopters is small relative to late adopters

- Key insight from several papers: with staggered timings + heterogeneous effects, the TWFE approach to DiD can put negative weight on certain groups' TE
  - Serious issue for interpretability
- This is *solveable*. Merely a construct of being overly casual with estimator definition

# Goodman-Bacon 2x2 comparisons

- Consider two staggered treatments and a never-treated group
- What does the TWFE estimator estimate?

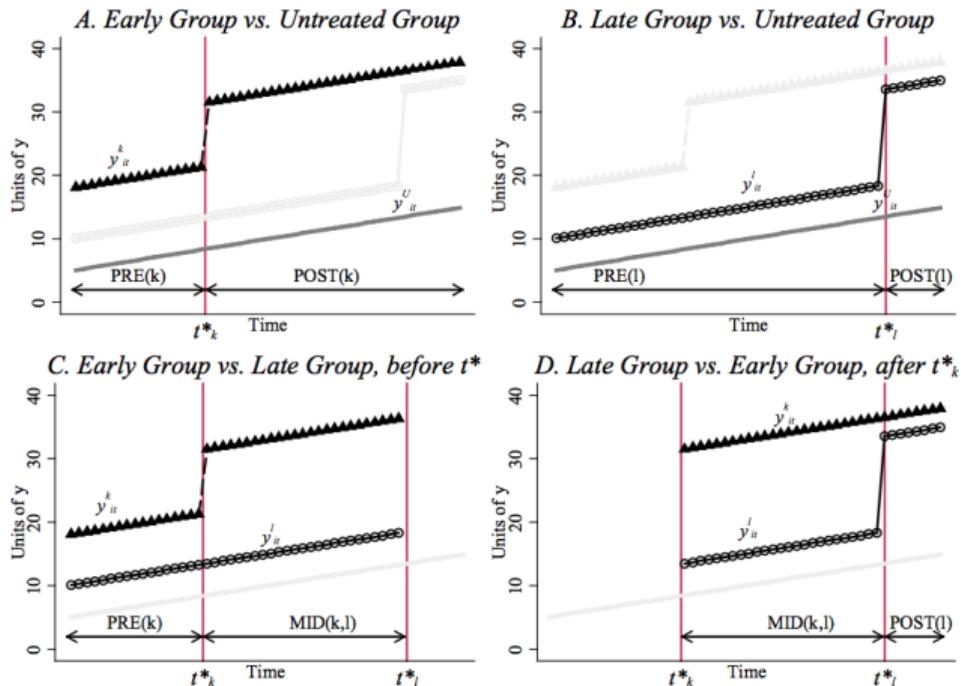
Figure 1. Difference-in-Differences with Variation in Treatment Timing: Three Groups



# Goodman-Bacon 2x2 comparisons

- Four potential comparisons that can be made
- turns out that TWFE DD estimator (pooled) is the weighted average of all 2x2 comparisons
- These weights end up putting a high degree of weight on units treated in the middle of the sample (since they have the highest variance in the treatment indicator!)

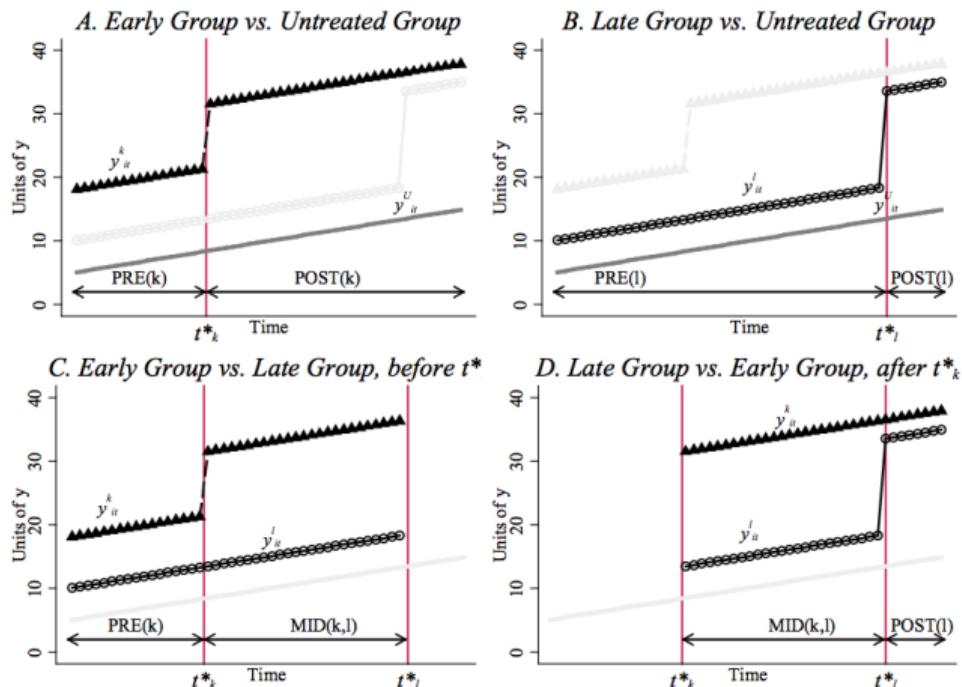
Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



# Goodman-Bacon 2x2 comparisons

- The weighting becomes problematic if the effects vary over time – if the effects are instantaneous and time-invariant, the weights are all positive

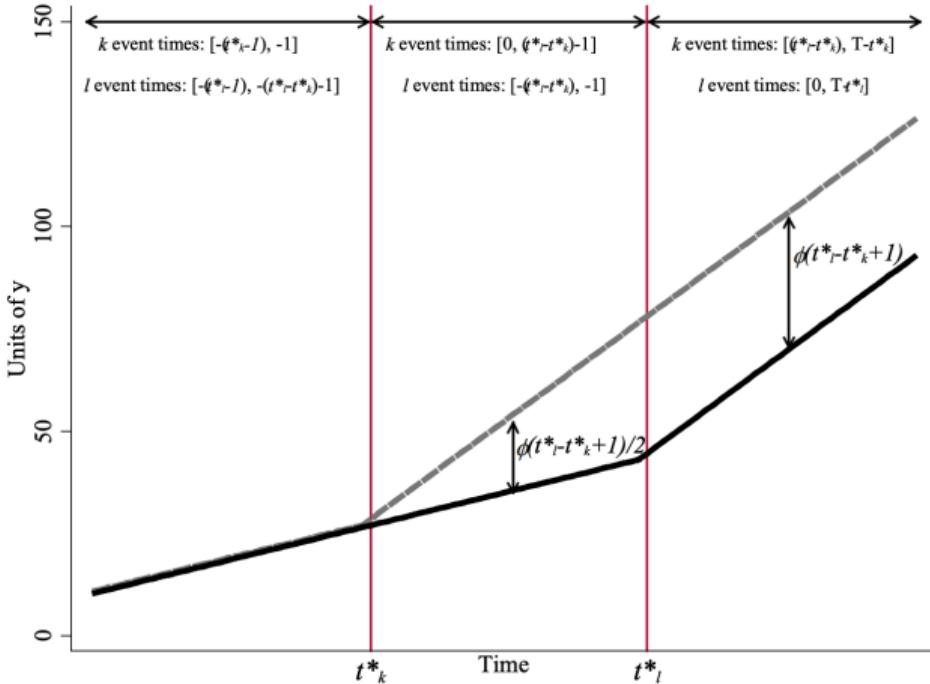
Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



# Goodman-Bacon 2x2 comparisons

- The weighting becomes problematic if the effects vary over time – if the effects are instantaneous and time-invariant, the weights are all positive
- However, time-varying effects create bad counterfactual groups, and create negative weights
- Goodman-Bacon provides a way to assess the weights in a given TWFE design

Figure 3. Difference-in-Differences Estimates with Variation in Timing Are Biased When Treatment Effects Vary Over Time



## What to do with staggered timing in DiD?

- There's really no reason to use the baseline TWFE in staggered timings
  - A perfect example wherein the estimator does not generate an estimate that maps to a meaningful estimand
- There are different approaches proposed in the literature that are just as good!
  - E.g. de Chaisemartin and d'Haultfoeuille (2020), Callaway and Sant'anna (2020)
- These all are robust to this issue. I find Callaway and Sant'anna quite intuitive, but your circumstances may vary slightly. Key piece to keep in mind that differs a bit:
  - Is my treatment absorbing?
- Irrespective of the exact paper, the key point is that we are generating a counterfactual and need to be careful that our estimator does so correctly

# Taking a step back on design

- The previous discussion all sits on the “model-based” approach for difference-in-differences
  - E.g. Correctly modeling  $E(Y(0))$  (the counterfactual)
  - Challenging when thinking about logs vs. levels, for example
- Alternative approach: random timing?



Journal of Econometrics

Volume 226, Issue 1, January 2022, Pages 62-79



## Design-based analysis in Difference-In-Differences settings with staggered adoption ★

Susan Athey <sup>a, b</sup>, Guido W. Imbens <sup>a, b, c, d</sup>

Show more ▾

+ Add to Mendeley Share Cite

<https://doi.org/10.1016/j.jeconom.2020.10.012>

Get rights and content

### Abstract

In this paper we study estimation of and inference for average treatment effects in a setting with panel data. We focus on the staggered adoption setting where units, e.g., individuals, firms, or states, adopt the policy or treatment of interest at a particular point in time, and then remain exposed to this treatment at all times afterwards. We take a design perspective where we investigate the properties of estimators and procedures given assumptions on the assignment process. We show that under random assignment of the adoption date the standard Difference-In-Differences (DID) estimator is an unbiased estimator of a particular weighted average causal effect. We characterize the exact finite sample properties of this estimand, and show that the standard variance estimator is conservative.

## Taking a step back on design

- The previous discussion all sits on the “model-based” approach for difference-in-differences
    - E.g. Correctly modeling  $E(Y(0))$  (the counterfactual)
    - Challenging when thinking about logs vs. levels, for example
  - Alternative approach: random timing?
  - Key point: we can define a propensity score where the treatment is considered “random”

Define  $W : \mathbb{A} \times \mathbb{T} \mapsto \{0, 1\}$ , with  $W(a, t) = \mathbf{1}_{a \leq t}$  to be the binary indicator for the adoption date  $a$  preceding  $t$ , and define  $W_{it}$  to be the indicator for the policy having been adopted by unit  $i$  prior to, or at, time  $t$ , so that:

$$W_{it} \equiv W(A_i, t) = \mathbf{1}_{A_i \leq t}.$$

The  $N \times T$  matrix  $\mathbf{W}$  with typical element  $W_{it}$  has the form:

## Taking a step back on design

- In the panel setting, it's worth considering whether you can just consider the policy "as-if" random, conditional on some controls
- Can you make units sufficiently comparable?
  - E.g. if all firms greater than 100 employees are treated, could you compare firms just below 100 and just above 100 (to come in RD!)
  - This would *not* require a functional form assumption
- Are there controls such that the groups appear "observationally identical" in the pre-period?
  - Synthetic control + propensity score methods (to come)

## Finally, a discussion on inference

- First, let's start with the old school fact that you must know if you are working with panel data and Dind
- **You must cluster on the unit of policy implementation if possible.** See Bertrand, Duflo and Mullainathan (2004)
  - Why? Outcomes and the treatment tend to be severely autocorrelated within unit
  - I say "if possible" since clearly in Card and Krueger that is infeasible
- If the policy variation is implemented at the industry level, you should not cluster at the firm level
- If the policy variation is implemented at the firm level, you cannot use robust standard errors

## Small clusters

- The Card and Krueger case was too extreme, but there are approaches for dealing with a small number of clusters
- This approach typically involves bootstrapping, and can handle small number of treated groups relative to the overall population
- See Andreas Hagemann's work for a place to start

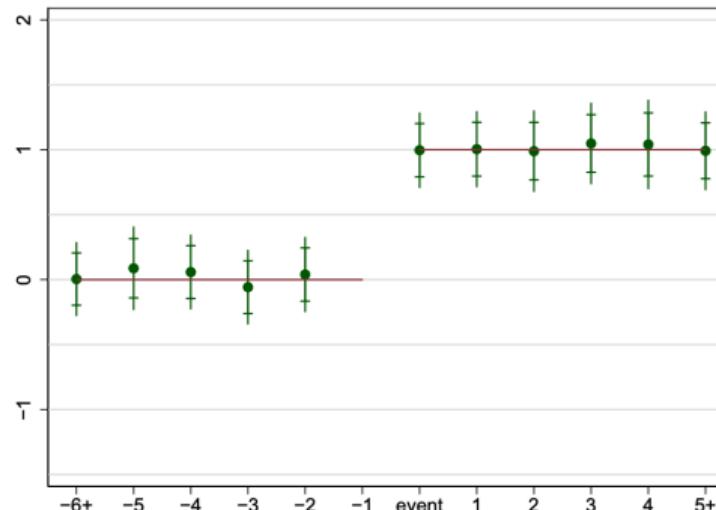
## Uniform confidence intervals

- Finally, when considering event study graphs, pre-trend graphs should use uniform confidence intervals, rather than pointwise confidence intervals
  - Advocated for by Freyaldenhoven et al (2018):
- Code available here thanks to Ryan Kessler:  
[https://github.com/paulgp/simultaneous\\_confidence\\_bands](https://github.com/paulgp/simultaneous_confidence_bands)

Figure 2 shows both pointwise 95% confidence intervals and uniform 95% sup-t confidence bands (Olea and Plagborg-Møller 2019). Applied papers commonly include pointwise confidence intervals in event plots.<sup>12</sup> These permit testing only of preselected pointwise hypotheses. Uniform bands such as those we show here are designed to contain the true path of the coefficients 95% of the time, and are therefore arguably more useful for giving readers a sense of what kinds of pre-trends are consistent with the data.

## Uniform confidence intervals

- Finally, when considering event study graphs, pre-trend graphs should use uniform confidence intervals, rather than pointwise confidence intervals
  - Advocated for by Freyaldenhoven et al (2018):
- Code available here thanks to Ryan Kessler:  
[https://github.com/paulgp/simultaneous\\_confidence\\_bands](https://github.com/paulgp/simultaneous_confidence_bands)



## Conclusion

- Difference in difference is hugely powerful in applied settings
- Does not require random assignment, but rather implementation of policies that differentially impacts different groups and is not confounded by other shocks at the same time.
- Can be a great application of big data, with convincing graphs that highlight your application
- Also allows for partial tests of identifying assumptions
- Worth carefully thinking about what your identifying assumptions are in each setting, and transparently highlighting them.
- Important to note that this always identifies a *relative* effect, and to aggregate, you will typically need a model and additional strong assumptions (see Auclert, Dobbie and Goldsmith-Pinkham (2019) for an example in a macro setting).

## My takeaways from new literature

- Beware weak tests of pre-trends. Consider using R&R's partial identification tests to assess robustness of results.
- Do not worry about the new literature on staggered timings if you only have one rollout period!
- Think carefully about your estimand if you're using a staggered timing DinD – what's your counterfactual in each case?
  - Software exists for many of these papers. This is doable!
- When plotting confidence intervals in event studies, you should plot uniform confidence intervals.