VANDERBILT
Peabody College

Advances in difference-in-differences: Implications for education research

Mark J. Chin
10/25/2023

Introduction

I’m giving you a short lecture today on (what I understand about) recent advances in differences-in-differences (DD) methodology and what they mean for education research. My goals are:

- Provide a conceptual overview of DD
- Introduce the (potential) problems that recent advances identify when using DD
- Discuss solutions to these problems—for executing DD correctly and for publishing

My Background

I am not:

- An econometrician
- An economist, though I do use econometric techniques and follow economics research closely

I am:

- An education researcher interested in using rigorous quantitative methods to identify causal relationships
- Always complaining about DD advances on Twitter

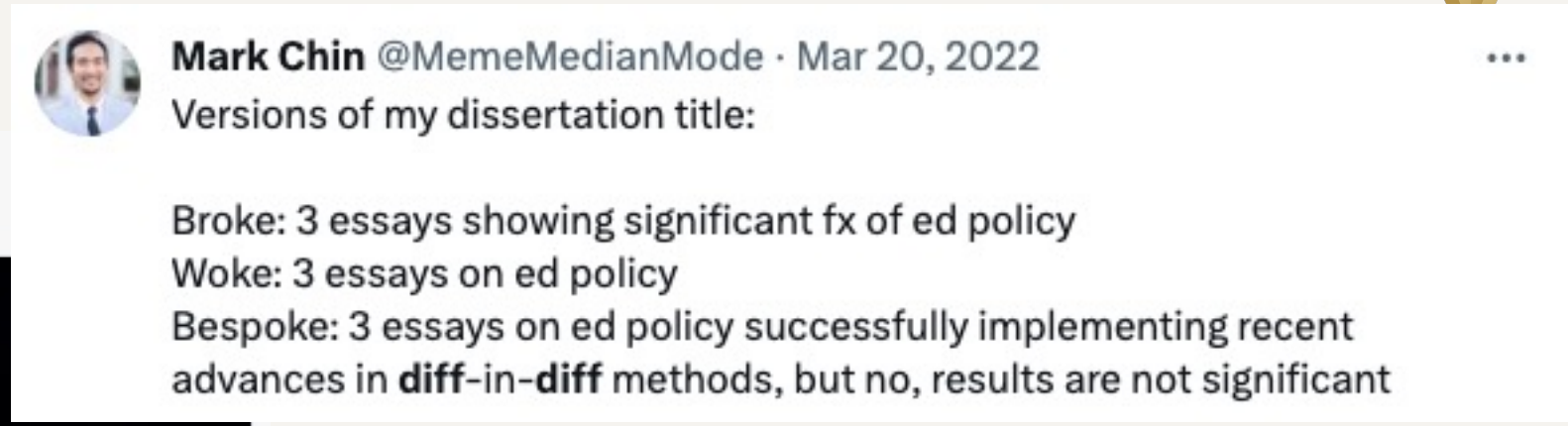
My Background

techniques and follow economics

rigorous quantitative methods to identify

- Always complaining about DD advances on Twitter

My Background



techniques and follow economics

rigorous quantitative methods to identify

- Always complaining about DD advances on Twitter

My Background

Mark Chin @MemeMedianMode · Aug 1, 2021
my reaction to the advances in the **diff**-in-**diff** literature

Mark Chin @MemeMedianMode · Mar 20, 2022
Versions of my dissertation title:

Broke: 3 essays showing significant fx of ed policy
Woke: 3 essays on ed policy
Bespoke: 3 essays on ed policy successfully implementing recent advances in **diff**-in-**diff** methods, but no, results are not significant

Mark Chin @MemeMedianMode · Dec 6, 2021
Applied economists after a new **diff** and **diff** paper drops

no context college football @nocontextcfb · Dec 6, 2021

- Always complaining about DD advan

Terminology

- This presentation is on DD models:
  - Some “comparative interrupted time series” models are basically DD models, but not all
  - I consider “event study” (ES) models to be extensions of DD models and will briefly discuss them today, but not all examples of ES models in the wild are actually DD models
  - This is all very confusing
- The presentation essentially applies to any quantitative model trying to get causal estimates by comparing trends in outcomes for two sets of observations: treated and not-treated (”comparison”) observations

Causal Identification

- Goal: We want to identify the causal impact of *X* on *Y*
- Ideal world: Compare *Y* for units in the universes where they either receive *X* or don't receive *X*
- Problem(s):
  - We never observe the counterfactual universe
  - The estimates of the causal impact of *X* based on comparisons of *Y* for units with different experiences of *X* will generally be plagued by omitted variable bias

DD Review: Intro

DD is a method to potentially address these problems:

- Widely used method in causal research
- Easy to understand
- Implementable in most settings, as you just need:
  - Many time periods
  - Treatment occurs for some units in some of the time periods

DD Intro: Canonical DD (2x2)

- 2 observations, 1 treated (T), 1 untreated (“comparison”, C) by some policy
- 2 time periods of data, time 0 (pre), and time 1 (post):
  - Pre: T and C both are untreated by the policy
  - Post: T is treated, C is still untreated
- Problem: how can you estimate the impact of the policy *X* on outcome *Y* when treatment is not randomly assigned?
- Solution: we estimate two differences in *Y*:
  - First difference: pre/post difference for T
  - Second difference: pre/post difference for C
  - Difference in differences: the first difference minus the second difference

DD Intro: Canonical DD (2x2)

- Hypothetical example: T = desegregating school districts, C = non-desegregating school districts, *Y* = suspension rate for Black students

	Pre-desegregation	Post-desegregation	Difference
Desegregating	8%	12%	4pp
Non-desegregating	7%	8%	1pp
Difference	1pp	4pp	3pp

- Difference 1 = 12% - 8% = 4pp (pre-post difference for desegregating districts, accounts for different starting level of *Y* for T)
- Difference 2 = 8% - 7% = 1pp (pre-post difference for non-desegregating districts, accounts for different starting level of *Y* for C)
- Difference 1 – Difference 2 = 4pp – 1pp = 3pp (accounts for trends in *Y* over time)
  - Desegregation led a 3pp increase in suspension rates for Black students

DD Intro: Key Assumption

- Parallel trends: absence of treatment, pre-post difference in *Y* for T and C would have been the same

	Pre-desegregation	Post-desegregation	Difference
Desegregating	8%	12%	4pp
Non-desegregating	7%	8%	1pp
Difference	1pp	4pp	3pp

- E.g., desegregating districts would have had a 1pp increase in Black student suspension rates absent desegregation

DD Intro: Key Assumption

Parallel trends:

- Does NOT assume baseline equivalence of T and C
- Inherently unobservable

A whole set of econometrics literature focuses on the parallel trends assumption—not going to spend time on that today

DD Intro: Estimation

Classic approach, regression equation ("two way fixed effects" or "TWFE"):

$$Y_{it} = \beta_0 + \beta_1 Treated_{it} + \delta_i + \gamma_t + \varepsilon_{it}$$

Where

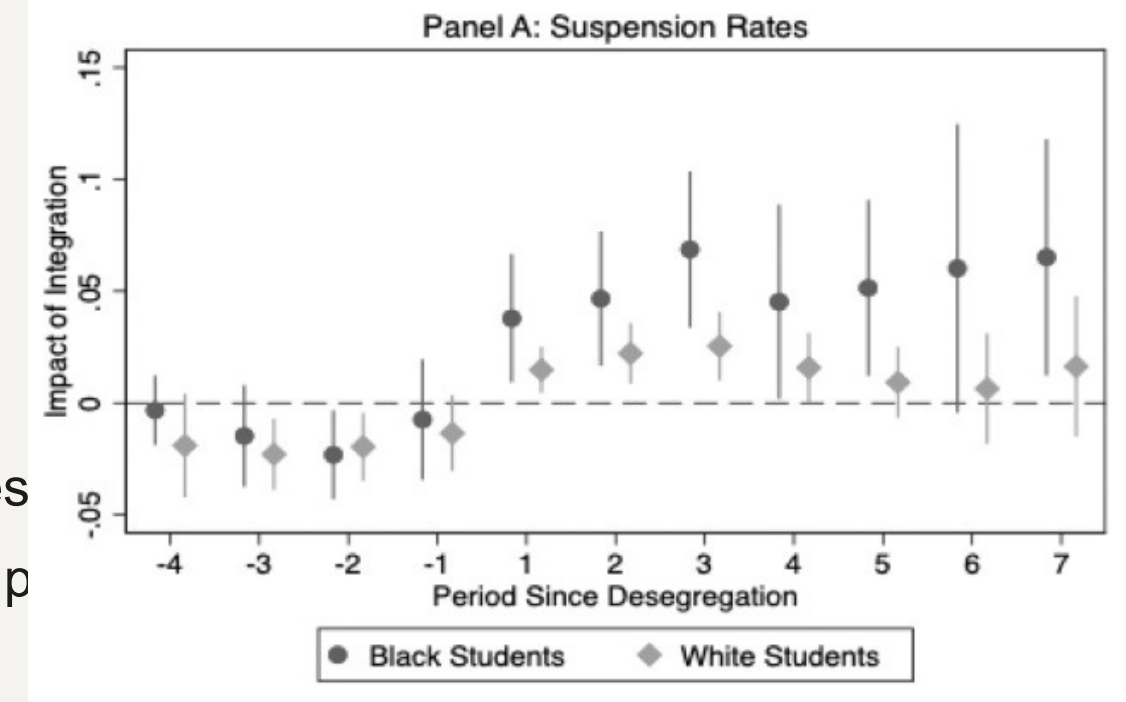
- Y_{it} is the outcome for unit *i* in time *t*
- $Treated_{it}$ indicates whether T units are in a post-treatment period (0/1)
- δ_i are unit fixed effects (accounts for between-unit differences)
- γ_t are time fixed effects (accounts for time trends)
- β_1 is the coefficient of interest

DD Intro: ES

Event study (ES) models are essentially extensions of DD:

- Instead of two time periods pre/post treatment, you have multiple time periods pre/post treatment
- You calculate the T/C DD estimate for each period (pre or post) relative to a baseline period, typically the period immediately preceding treatment
- Can be used to ”provide evidence” that the parallel trends assumption is met, i.e., if T/C DD estimates for periods pre-treatment are not statistically significant
- Can estimate “dynamic” post DD estimates (i.e., changing over time)

DD Intro: ES



Event study (ES) models are es

- Instead of two time periods p ds pre/post treatment
- You calculate the T/C DD estimate for each period (pre or post) relative to a baseline period, typically the period immediately preceding treatment
- Can be used to ”provide evidence” that the parallel trends assumption is met, i.e., if T/C DD estimates for periods pre-treatment are not statistically significant
- Can estimate “dynamic” post DD estimates (i.e., changing over time)

DD Advances

Generally, two types of DD advances

- Identify problems with some applications of DD (”How bad is it”)
- Provide solutions to these problems (“How can we solve it”)

DD Advances: “How bad is it” Papers

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

DD Advances: “How can we solve it” Papers

Gardner, J. (2022). Two-stage differences in differences. *arXiv preprint arXiv:2207.05943*.

DD Advances: “How bad is it and how can we solve it” Papers

A few canonical ones:

- Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates?. *Journal of Financial Economics*, *144*(2), 370-395.
- Callaway, B., & Sant’Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200-230.
- Goodman-Bacon, A. (2019). *So you’ve been told to do my difference-in-differences thing: A guide.*
- Roth, J., Sant’Anna, P. H., Bilinski, A., & Poe, J. (2023). What’s trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*.
- Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175-199.

DD Advances: The Problem

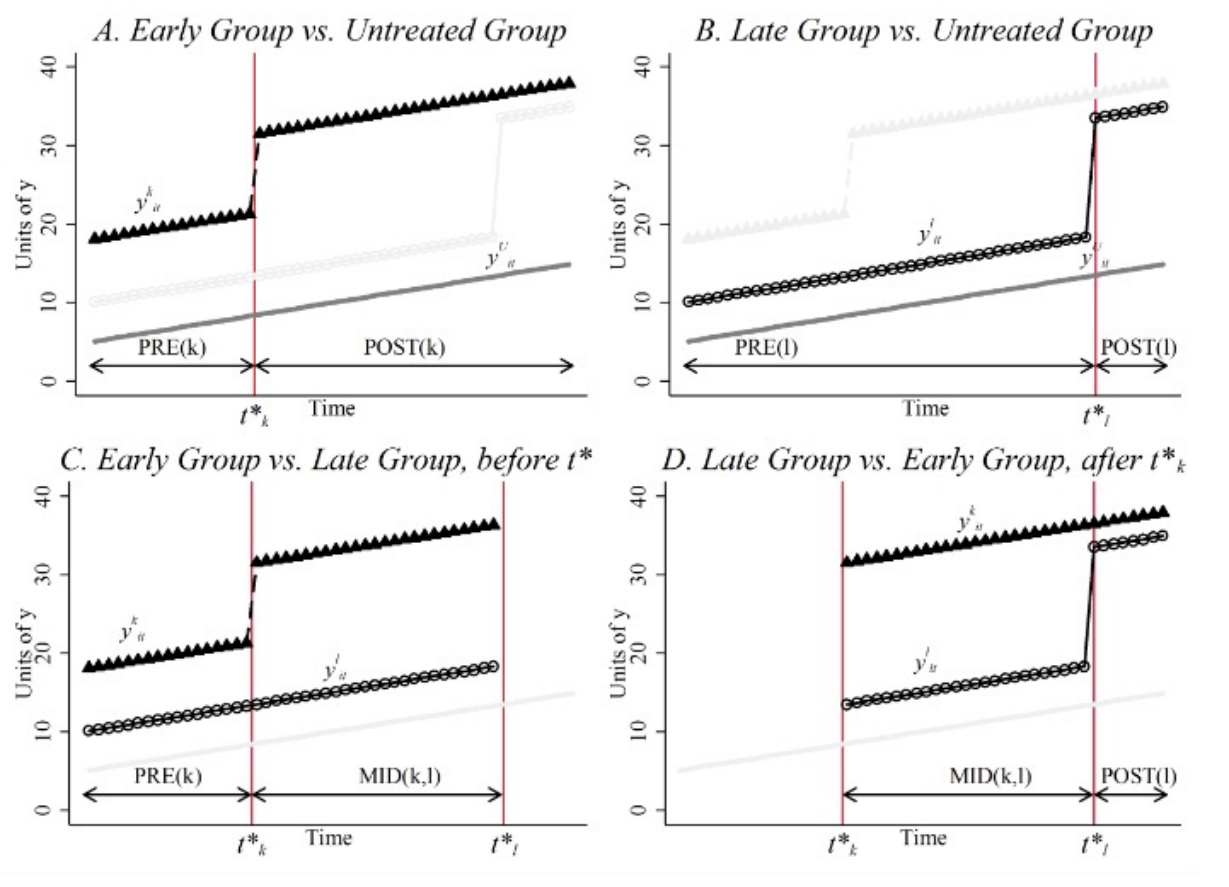
Most applications of DD no longer follow the canonical 2x2 structure:

- Multiple treated and comparison units
- Treatment does not happen all in the same time period (e.g., rollout of policy and programs across geographies), “differential timing” or “staggered DD”

These changes can lead to biased estimates

DD Advances: The Problem

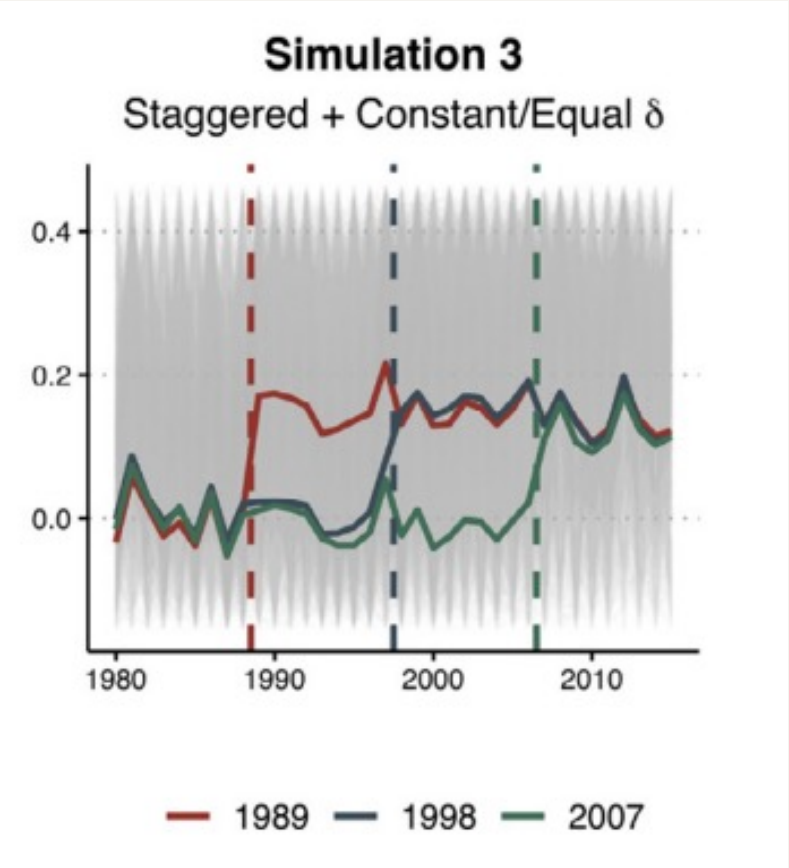
- Goodman-Bacon (2021) shows that TWFE DD estimates are weighted averages of all possible 2x2 estimators in the data
- Consider a simplified example where there are 3 units: early treated unit, later treated unit, untreated unit



DD Advances: The Problem

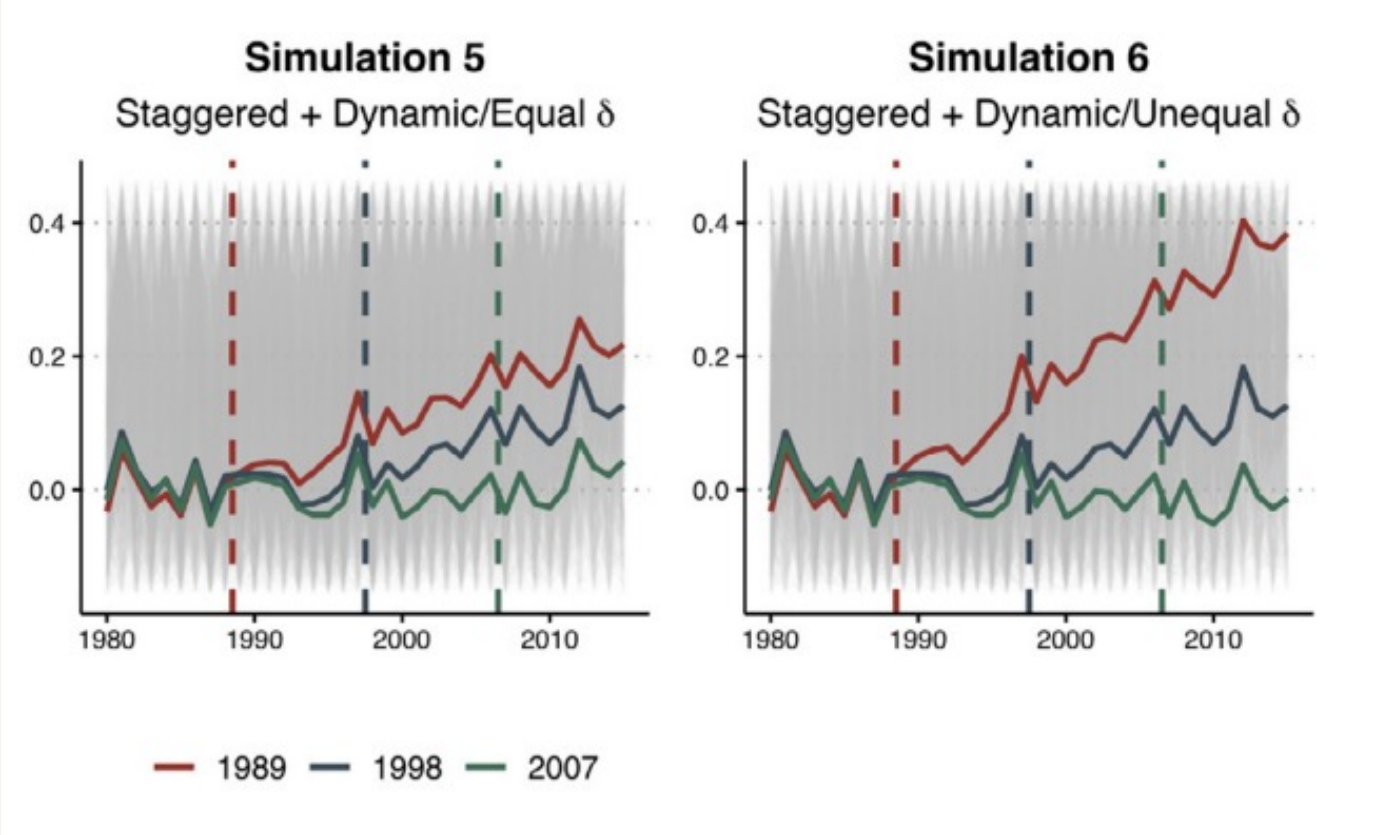
- How TWFE constructs its DD estimates matters when treatment effects are dynamic (i.e., not just a level change post treatment) and treatment effects are heterogeneous

- Consider a simulated staggered DD with constant and equal and homogenous treatment effects (Baker et al., 2022)



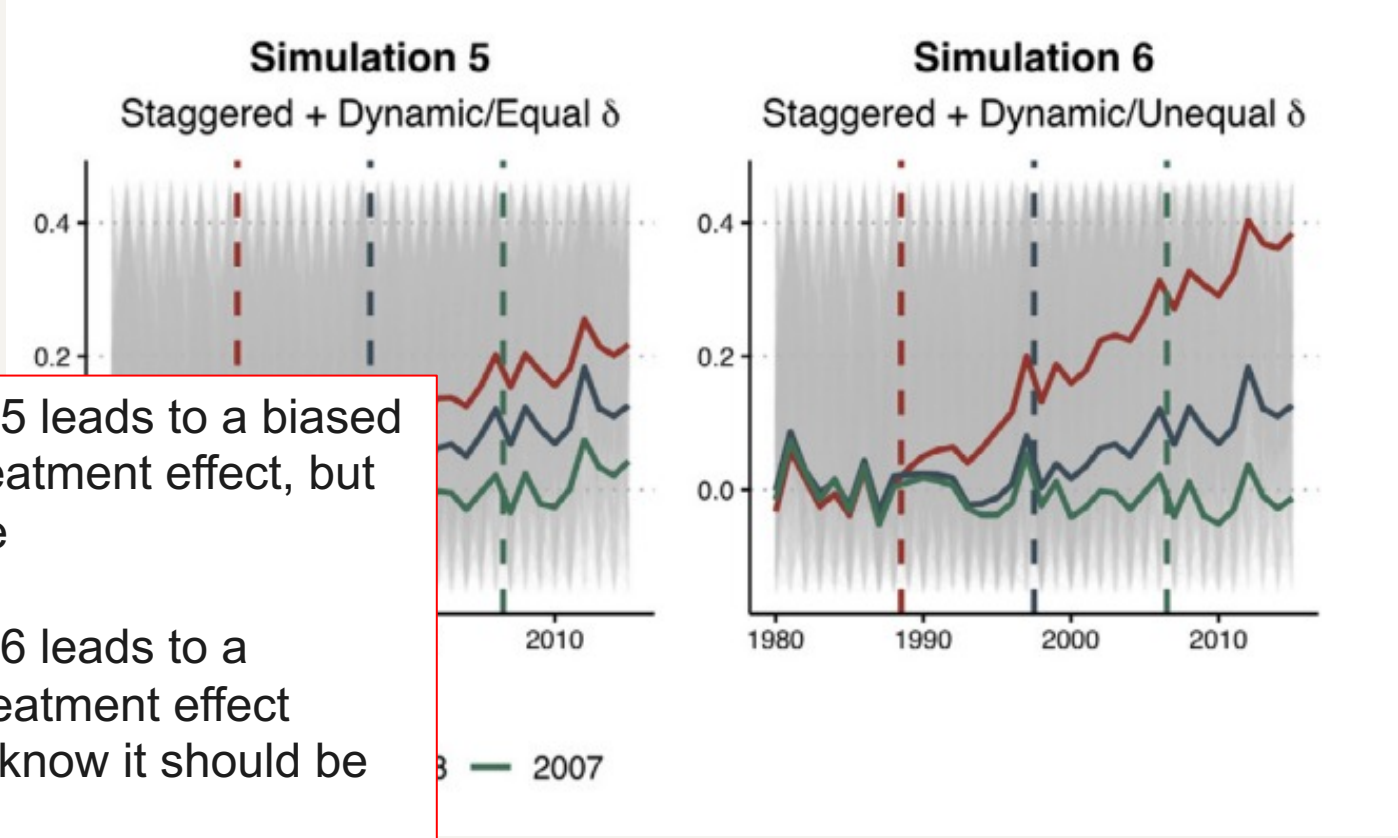
DD Advances: The Problem

Now consider the simulated staggered DD with dynamic and/or heterogeneous treatment effects



DD Advances: The Problem

Now consider the simulated staggered DD with dynamic and/or heterogeneous treatment effects



Simulation 5 leads to a biased average treatment effect, but still positive

Simulation 6 leads to a negative treatment effect though we know it should be positive!

DD Advances: Solutions

Is everyone treated at the same time? If yes, don’t worry about this

DD Advances: Solutions

Is everyone treated at the same time? If no, use TWFE, and restrict treatment effect heterogeneity

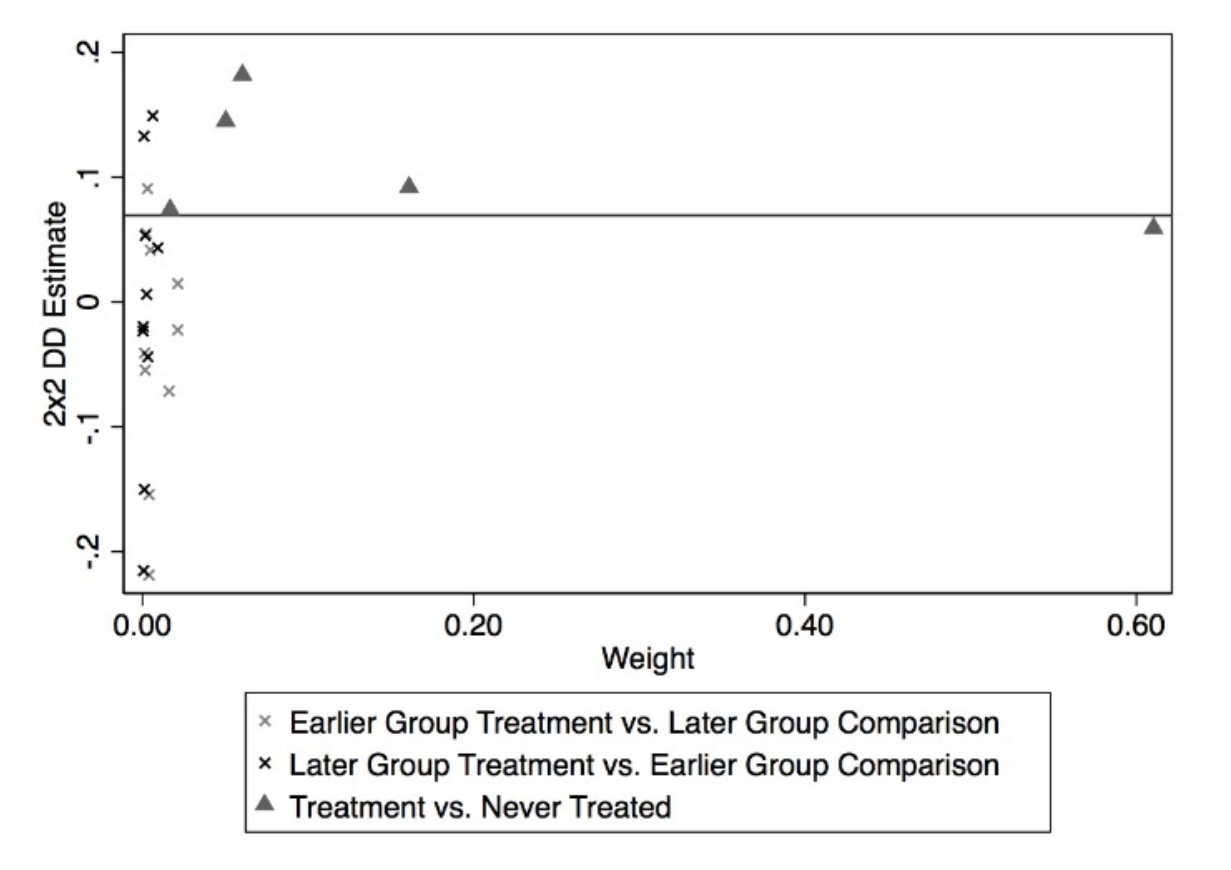
DD Advances: Solutions

Is everyone treated at the same time? ~~If no, use TWFE, and restrict treatment effect heterogeneity~~

Actually, you should not do this (very strong assumption)

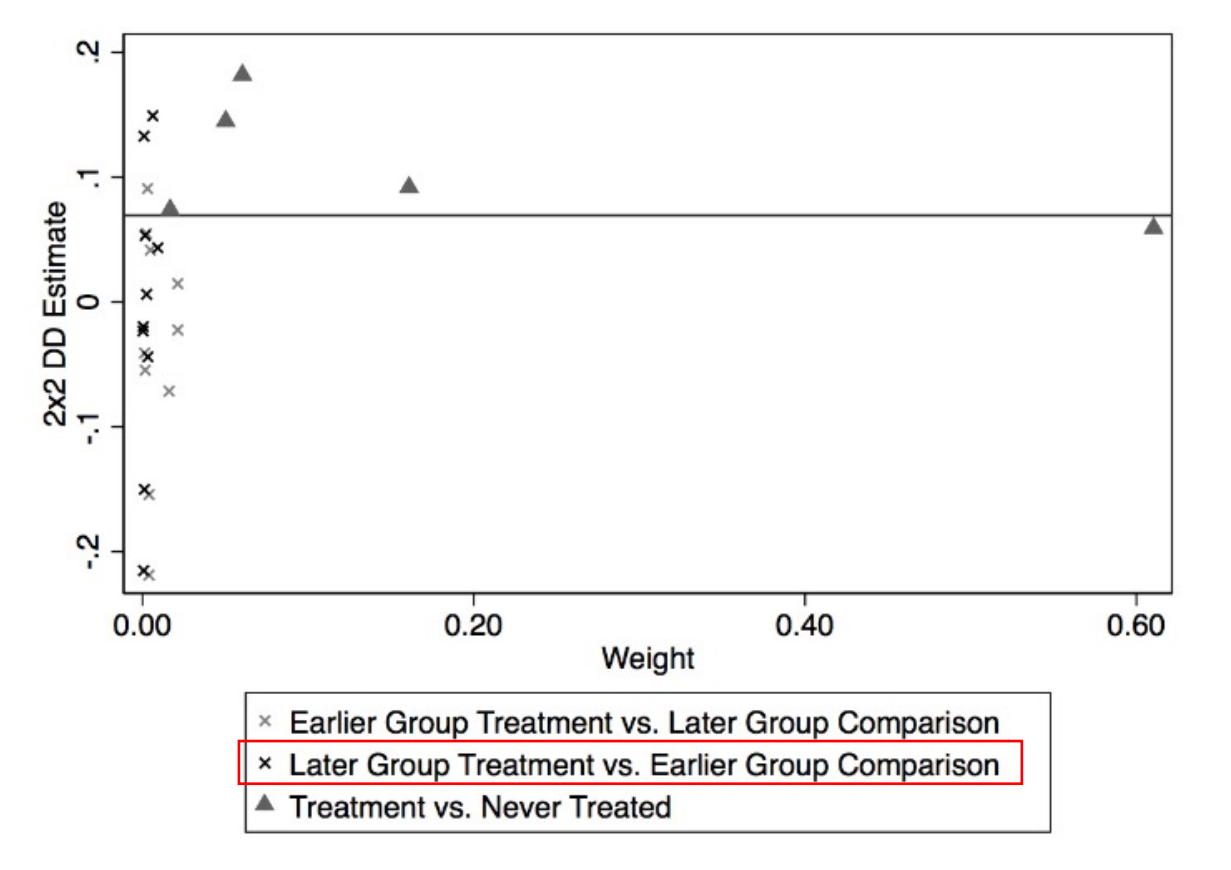
DD Advances: Solutions

- Is everyone treated at the same time? If no, you might consider implementing the Bacon decomposition of DD estimator with variation in treatment timing (Goodman-Bacon, 2021)
- In Stata, `bacondecomp` produces a figure that decomposes the TWFE DD estimate into each individual 2x2 (estimate and weight) in the data



DD Advances: Solutions

- Ideally find that your TWFE DD estimate is NOT driven by “bad” 2x2s, i.e., those that identify the effect of treatment on later groups using early groups as a comparison



DD Advances: Solutions

Is everyone treated at the same time? If no, you might even just skip straight to using "heterogeneity-robust" estimators

Table 2
Statistical packages for recent DiD methods.

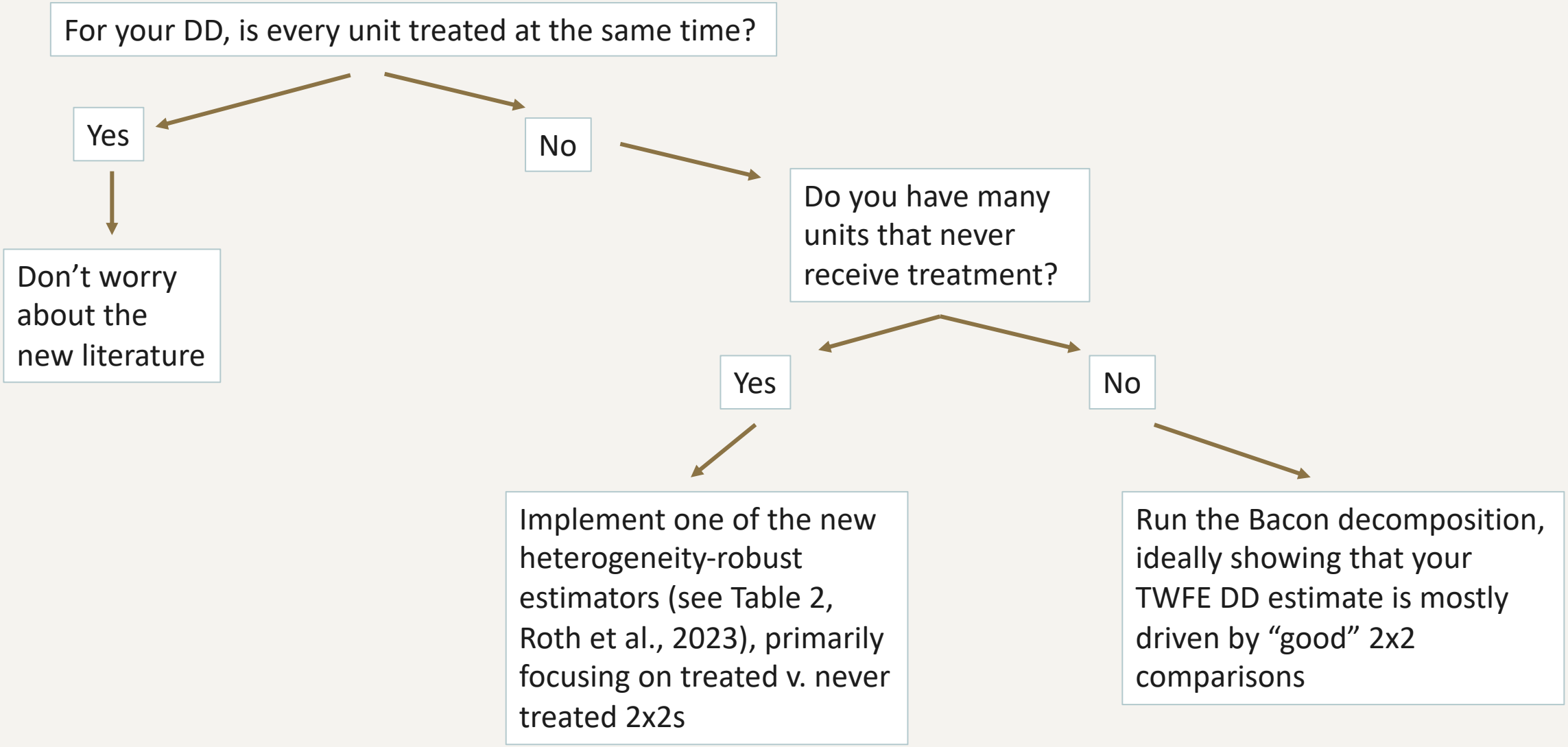
Heterogeneity Robust Estimators for Staggered Treatment Timing		
Package	Software	Description
did, csdid	R, Stata	Implements Callaway and Sant'Anna (2021)
did2s	R, Stata	Implements Gardner (2021), Borusyak et al. (2021), Sun and Abraham (2021), Callaway and Sant'Anna (2021), Roth and Sant'Anna (2021)
didimputation, did_imputation	R, Stata	Implements Borusyak et al. (2021)
DIDmultiplegt, did_multiplegt	R, Stata	Implements de Chaisemartin and D'Haultfoeuille (2020)
eventstudyinteract	Stata	Implements Sun and Abraham (2021)
flexpaneldid	Stata	Implements Dettmann (2020), based on Heckman et al. (1998)
fixest	R	Implements Sun and Abraham (2021)
stackedev	Stata	Implements stacking approach in Cengiz et al. (2019)
staggered	R	Implements Roth and Sant'Anna (2021), Callaway and Sant'Anna (2021), and Sun and Abraham (2021)
xtevent	Stata	Implements Freyaldenhoven et al. (2019)

Roth et al., 2023

DD Advances: Solutions

- These estimators make it more transparent how weights are being assigned to the individual 2x2 estimates, and exclude forbidden 2x2s—those that compare later treated units to already treated units
- For publishing, if you have staggered DD, my opinion is that you should default to using a heterogeneity-robust estimator (and I typically show robustness of results when using at least one other estimator)

DD Advances: Simple Recommended Flowchart



DD Advances: Solutions

- In my (limited) experience, differences across the three estimators I have employed—the Callaway and Sant'Anna (2021) estimator, the two-stage DD estimator (Gardner, 2022), and stacked approaches (e.g., Cengiz et al., 2019)—have not resulted in vastly different results
  - Two-stage DD and stacked DD approaches can easily implement parametric and non-parametric ES models, in general appear more flexible to code up specific adaptations, and can clearly be reproduced through simple regressions
- Today I'm going to walk through the stacked DD approach, which I think is the most intuitive/transparent

DD Advances: Stacked DD

- Canonical examples:
  - Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, *134*(3), 1405-1454.
  - Deshpande, M., & Li, Y. (2019). Who is screened out? Application costs and the targeting of disability programs. *American Economic Journal: Economic Policy*, *11*(4), 213-248.
  - Chin, M. J. (2021). Desegregated but still separated? The impact of school integration on student suspensions and special education classification. *Journal of Urban Economics*. ☺

DD Advances: Stacked DD

- Example is based off of the analysis in “School district consolidation in North Carolina: Impacts on school composition and finance, crime outcomes, and educational attainment” (Chin, 2023)
- Main RQ: What is the impact of school district mergers (or consolidations) on outcomes?
- Approach: Staggered DD, analyzed using Stacked DD
  - Compare outcomes before and after consolidation for those districts that consolidate (N=18)
  - Compare these differences to analogous ones from districts in other counties that *never* consolidated or consolidated in the far past over the same period (N=59)

DD Advances: Stacked DD

- Before “stacking” the data, each row in the full dataset is at the county-year level
- To “stack” the data, loop over the following steps for each county with consolidating districts (*n* = 1:18):
  1. Preserve the full data set; keep all the county-year data only for consolidating district *n*; assign them `consolidated=1`; save this as a temporary file; restore the full data set
  2. Preserve the full data set; keep all the county-year data for all 59 non-consolidating districts; assign them the treatment year of consolidating district *n*; assign them `consolidated=0`; append the dataset from (1); save this treat *n*+control dataset as a temporary file; restore the full dataset
- Append all 18 treat *n*+control datasets into a single “stacked” dataset

DD Advances: Stacked DD

Estimate a regression model represented by the following equation:

$$Y_{ijt} = \beta_0 + \beta_1 POST_{ijt} + \beta_2 (POST_{ijt} \times CONSOLIDATED_i) + \delta_i + \gamma_j + \varepsilon_{ijt}$$

Where

- Y_{ijt} is an outcome for county *i,* in stacked dataset *j* (1-18), at time *t*
- $POST_{ijt}$ is a dummy variable that indicates if time *t* is after the consolidation year of the consolidating district in dataset *j*
- $CONSOLIDATED_i$ is a dummy variable that indicates if the county *i* has consolidating districts (i.e., β_2 is the coefficient of interest)
- δ_i, γ_j are county and time fixed effects, respectively

DD Advances: Stacked DD

Note that, within stacked dataset, the only untreated units are those never consolidating (i.e., no comparison made with earlier treated units), and $POST_{ijt}$ is specifically determined off of the treatment year of the treated unit in the stacked dataset!

Estimate a regression model represented by the following equation:

$$Y_{ijt} = \beta_0 + \beta_1 POST_{ijt} + \beta_2 (POST_{ijt} \times CONSOLIDATED_i) + \delta_i + \gamma_j + \varepsilon_{ijt}$$

Where

- Y_{ijt} is an outcome for county *i,* in stacked dataset *j* (1-18), at time *t*
- $POST_{ijt}$ is a dummy variable that indicates if time *t* is **after the consolidation year of the consolidating district in dataset *j***
- $CONSOLIDATED_i$ is a dummy variable that indicates if the county *i* has consolidating districts (i.e., β_2 is the coefficient of interest)
- δ_i, γ_j are county and time fixed effects, respectively

DD Advances: Stacked DD

Impact (SE) of consolidation on current expenditures PP on district services:

- TWFE: -225*** (57)
- Stacked DD: -238*** (59)
- Callaway & Sant'Anna: -305*** (38)
- Two-stage DD: -202*** (62)

In general, if non-treated units make up a large proportion of your sample (here 59 v. 18), the TWFE DD estimate will not be too far off from the heterogeneity-robust ones (Baker et al. 2021 also notes this)

Final Considerations

- These DD Advances solutions are only solving one problem—potential bias due to heterogeneous and/or dynamic treatment effects in staggered DD designs; you still need to convince your reader that the parallel trends assumption is met!
- This presentation does not cover:
  - Continuous treatments (i.e., dosages) in a DD framework
  - Repeated treatments (or treatments that turn on and off)
  - Some newer work is starting to think more deeply about this—let me know if you are interested!

(More) Final Considerations

- Reminder: DD does not require baseline equivalence to recover causal effects
  - It can be nice though ☺ since reviewers seem to still like to see this; maybe combine DD and matching
- In general, avoid controlling for within-unit time-variant covariates because they might be endogenous to treatment (collider bias)
- You can compare early treated units to later treated units (comparison) as a “valid” 2x2, but sometimes it is easier just to make just treated vs. untreated 2x2 comparisons
  - This is essentially what synthetic control (SC) is doing!
  - If you have very few treated observations, you might be tempted to use SC, but you honestly can still use DD, you just need to adjust standard errors appropriately (wild bootstrap; Ferman & Pinto, 2019)

(Even More) Final Considerations

- No reason why these new DD approaches cannot be applied to data/program evaluations of all types (e.g., “big data”, educational policies or not)
- I am starting to think that the concerns around staggered DD might apply to other TWFE models that we don’t typically conceive of as DD models?
  - Example: student-year level dataset, key predictor is student-teacher race match (which varies from year-to-year), fixed effects for student and year
  - Relies on within-student variation in student-teacher race match to identify effects… but what is this actually estimating?
  - Lavar Edmonds has a paper on exposure to HBCU-trained teachers that uses this approach and considers the DD advances

Thank you!

- Website: markjchin.com
- Email: mark.chin@Vanderbilt.edu
- SM: @MemeMedianMode (Twitter, Bluesky)
- Sample DD code/data: https://github.com/MemeMedianMode/Staggered-Difference-in-Differences-Example