EPRC Methods Workshop Series

Difference-in-Differences in Education Research

A Practical Workshop on Traditional and Staggered DID, with Hands-on R Applications

Wei Li **Yaosheng Lou** **Katherine Strickland**

University of Florida

Workshop website: lcds-lab.github.io/did-eprc-workshop/

Slides available at lcds-lab.github.io/training/

Acknowledgements

- We thank the **UF Education Policy Research Center** and **Dr. F. Chris Curran** for supporting this workshop.
- This work is supported by the **National Science Foundation** through a research grant titled *"CAREER: Statistical Power Analysis and Optimal Sample Size Planning for Longitudinal Studies in STEM Education"* (PI: Wei Li, Award Number DRL-2339353).
- This work is also supported by the **UF Lastinger Center Research Catalyst Award** for *"Enhancing Longitudinal Program Evaluations: Estimating Design Parameters and Illustrating Advanced Quasi-Experimental Methods Using UF Lastinger Center Data"* (PI: Wei Li).
- The opinions expressed herein are those of the authors and do not necessarily reflect the views of the funding agencies.

Meet the Workshop Team

Wei Li
Associate Professor, UF

Methods lecture

Director, Longitudinal Causal Data Science Lab. Research on causal inference, longitudinal design, and statistical power analysis.

Katherine J. Strickland
Postdoctoral Scholar, UF

Online questions and support

Causal inference, longitudinal methods, and staggered DID in education. Former elementary and special education teacher. Please call her Kat.

Yaosheng Lou
Postdoctoral Scholar, UF

Hands-on R lab

Longitudinal and multilevel methods, causal inference, and AI-assisted evaluation in education and the social sciences.

Workshop Overview

1. **Scope and purpose.** Participant backgrounds, conceptual focus, and software resources.
2. **Why DID in education?** Policies, programs, and nonrandom assignment.
3. **One treatment group.** The 2×2 comparison, parallel trends, TWFE, and event studies.
4. **Staggered adoption.** Different start dates and problems with already-treated comparison groups.
5. **Modern DID methods.** Focus on Callaway–Sant'Anna (CSA), aggregation, inference, and pretrend checks. Wooldridge ETWFE if time permits.
6. **R applications.** Yaosheng leads the TWFE and CSA lab from 1:00 p.m., followed by questions and discussion.

Further reading: papers and software documentation are listed in the references.

Participant Background and Workshop Scope

Registration snapshot

56 registrations

Prior traditional DID use	**37.5%**
No prior traditional DID use	**62.5%**

Software use

R	**73.2%**
Stata	**44.6%**
Neither	**1.8%**

R and Stata overlap: 11 report both.

Conceptual understanding first

We will focus on DID logic, assumptions, and interpretation. We will not work through detailed mathematical derivations.

Additional material for reference

I usually prepare more than we can cover. Some additional equations are included for reference. We may skip selected equations and optional slides depending on our progress.

Clarification questions are welcome

Please feel free to stop me for a brief clarification, or ask in the chat.

R now, Stata to follow: Hands-on examples use **R**. A **Stata tutorial** will be added to the workshop website later.

Registration update, October 22, 2026. Rounded percentages. One possible duplicate remains in the registration count pending confirmation.

Today's Schedule and Focus

12:00–1:00 p.m.	**Methods lecture with Wei** Traditional DID with TWFE, then Callaway–Sant'Anna (CSA) for staggered adoption. Extended TWFE if time permits.
From 1:00 p.m.	**Hands-on R lab with Yaosheng: 30–40 minutes** Guided TWFE and CSA examples. We will not cover R installation.
After the lab	**Questions and discussion until 2:00 p.m.**

We will stop the lecture at 1:00 p.m. to protect lab time.

The aim is conceptual understanding and practical use, not a complete survey of DID methods.

Questions and Future Sessions

An introductory DID methods workshop

- **Brief clarification questions are welcome.** Ask in person or use the chat, where Kat will assist.
- Given our limited time, we are unable to address questions specific to individual research contexts during the workshop. We kindly ask that questions stay focused on the methods and examples presented.
- We would be happy to discuss questions about your own DID study after the workshop.

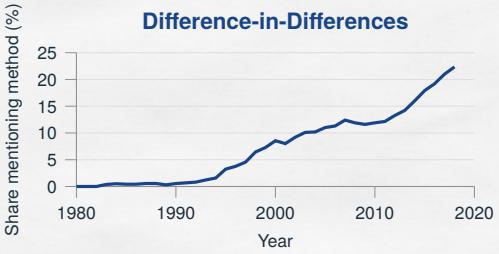
Help us plan what comes next

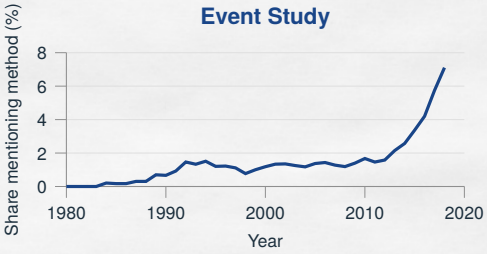
This is our first DID workshop. We plan to offer regular workshops and seminars and would welcome your interests and suggestions.

Interest survey: Future DID workshops and seminars

Importance of Difference-in-Differences

- DID is widely used in policy and program evaluation, supported by richer data and greater computing power.
- It is particularly useful when random assignment is not feasible and longitudinal data are available.





Source: Currie, Kleven, and Zwiers (2020), Figure IV. NBER applied-microeconomics papers, 1980–2018. Five-year moving averages of shares mentioning each method. The panels use different vertical scales.

DID in Education: District-Level G&T Examples

DID is also widely used in education for **policy and program evaluation**.

In this workshop, we use **gifted and talented (G&T) education** as the illustrative context (e.g., Strickland, Chan, & Gottfried, 2026).

Data structure: repeated annual measures nested within districts in **one state**. One row per district-year, with district mean achievement as the outcome.

Example 1: One treatment group and one control group

Treated districts adopt in Year 5. Comparison districts do not adopt.

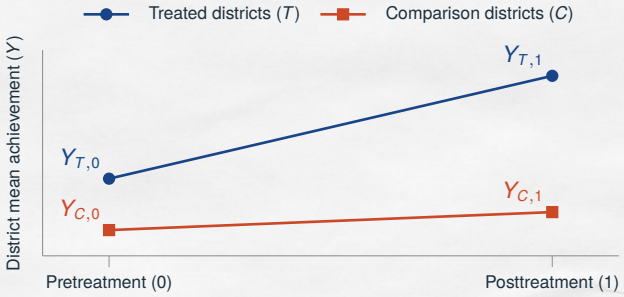
Example 2: Two treatment groups and one control group

Early districts adopt in Year 5, late districts in Year 6, and controls remain untreated.

The Classic 2 × 2 DID Design

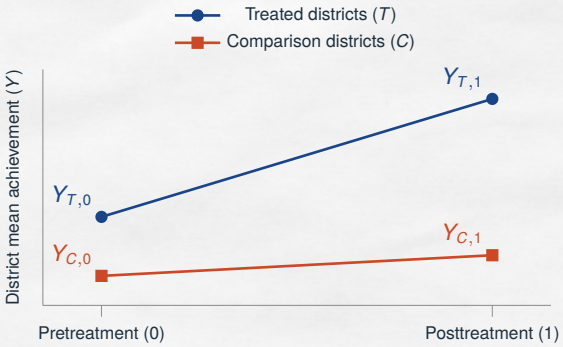
Follow **two groups of districts in one state**, before and after G&T adoption.

- **Groups:** T = adopting districts. C = comparison districts.
- **Periods:** 0 = before adoption. 1 = after adoption.



$Y_{g,t}$ averages district mean achievement within group g at time t.

How to estimate the treatment effect



How each post outcome is formed

$$Y_{C,1} = Y_{C,0} + \Delta_C \quad (1)$$

$$Y_{T,1} = Y_{T,0} + \Delta_T^{\text{no G\&T}} + \text{TE} \quad (2)$$

Δ_C: comparison districts' observed change, $Y_{C,1} - Y_{C,0}$.

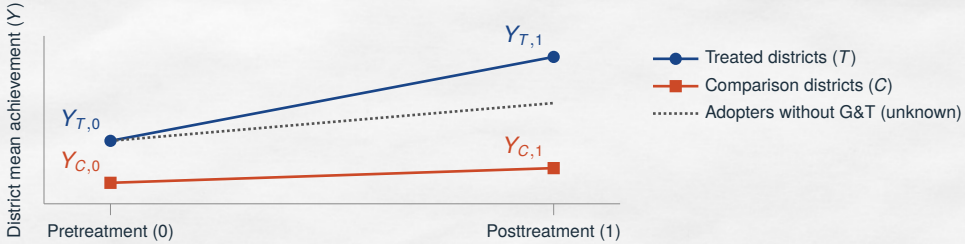
$\Delta_T^{\text{no G\&T}}$: adopters' change from pre to post **if they had not adopted G&T**.

Question: does the posttreatment gap $Y_{T,1} - Y_{C,1}$ isolate TE?

No anticipation: Future G&T adoption has not affected baseline achievement.

Baseline differences can still exist. Separately, the control group remains unaffected by G&T.

How to estimate the treatment effect



Subtract the two outcome equations from slide 11:

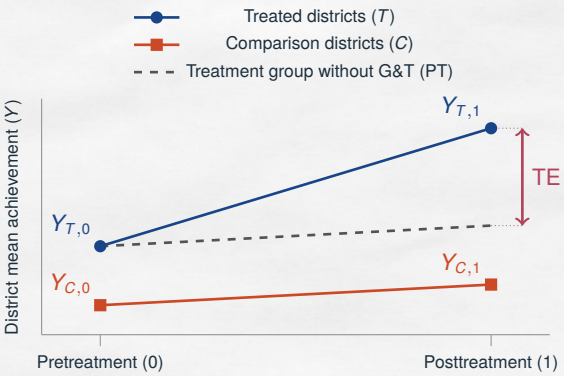
$$\underbrace{Y_{T,1} - Y_{C,1}}_{\text{Posttreatment difference}} = \underbrace{(Y_{T,0} - Y_{C,0})}_{\text{Baseline difference}} + \underbrace{\left(\Delta_T^{\text{no G\&T}} - \Delta_C\right)}_{\text{Difference in changes without treatment}} + \text{TE}. \quad (3)$$

Rearrange to isolate the treatment effect:

$$\text{TE} = (Y_{T,1} - Y_{C,1}) - (Y_{T,0} - Y_{C,0}) - \left(\Delta_T^{\text{no G\&T}} - \Delta_C\right). \quad (4)$$

The missing piece is $\Delta_T^{\text{no G\&T}}$: how adopters would have changed without G&T.

Parallel trends and the DID estimator



Parallel trends (PT)

Without G&T, the groups' **average changes would be equal**:

$$\Delta_T^{\text{no G\&T}} = \Delta_C. \tag{5}$$

$$\Delta_T^{\text{no G\&T}} - \Delta_C = 0. \tag{6}$$

The untreated gap is **constant**, not necessarily zero.

$$\begin{aligned} \text{TE} = &(Y_{T,1} - Y_{C,1}) \\ &- (Y_{T,0} - Y_{C,0}). \end{aligned} \tag{7}$$

Y denotes population means. Using sample means $\overline{Y}$ gives the DID estimator:

$$\widehat{\text{TE}}_{\text{DID}} = (\overline{Y}_{T,1} - \overline{Y}_{C,1}) - (\overline{Y}_{T,0} - \overline{Y}_{C,0}) = (\overline{Y}_{T,1} - \overline{Y}_{T,0}) - (\overline{Y}_{C,1} - \overline{Y}_{C,0}). \tag{8}$$

Why Our Treatment Effect Is ATT

$\mathrm{TE} = \mathrm{ATT}$: Average Treatment Effect on the Treated

The average effect of G&T adoption on achievement **among adopting districts**.

Districts that adopted G&T

We observe their achievement **with G&T**.

Under PT, no anticipation, and valid controls, the control districts' change helps reconstruct adopters' average achievement **without G&T**.

The difference gives **ATT**.

Districts that did not adopt

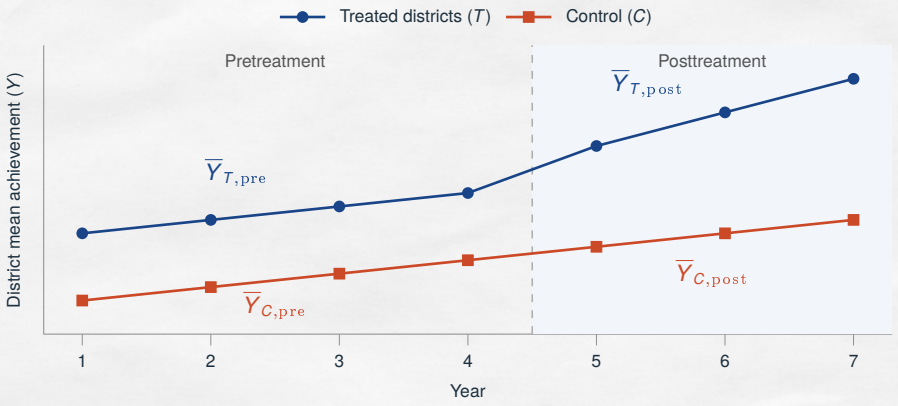
We observe their achievement **without G&T**.

We do not know what their achievement would have been **with G&T**.

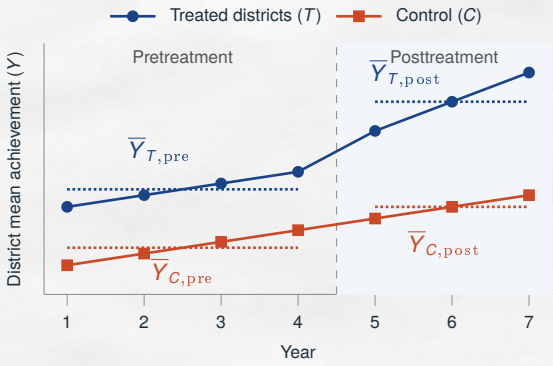
PT alone does not supply this missing outcome.

ATE for all districts also requires effects for nonadopters, using additional assumptions or evidence.

DID with Multiple Periods



DID with Multiple Periods



Dotted horizontal lines show pre/post averages.

Two groups, one start date
G&T starts in Year 5. The control group remains untreated.

For each group, average outcomes across **pre** periods and across **post** periods.

$$\widehat{\text{TE}}_{\text{DID}} = (\overline{Y}_{T,\text{post}} - \overline{Y}_{T,\text{pre}}) - (\overline{Y}_{C,\text{post}} - \overline{Y}_{C,\text{pre}}). \quad (9)$$

Under PT, this estimates the **average treatment effect across the posttreatment periods**.

Same districts each year. Each year counts equally within its pre/post average.

The TWFE Model

Two-way fixed effects: unit effects and time effects. Here, Y_{it} is mean achievement for district i in year t, within one state.

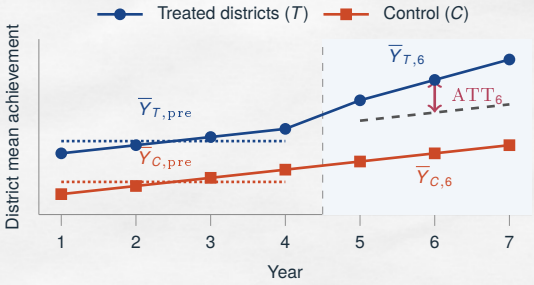
$$Y_{it} = \alpha_i + \lambda_t + \beta D_{it} + u_{it} \tag{10}$$

Term	Meaning in the G&T example
α_i	Stable differences across districts, such as persistent resource differences.
λ_t	Year-specific changes shared by both groups.
D_{it}	1 for adopting districts in Years 5–7. Otherwise 0, including all control observations.
β	One coefficient summarizing the posttreatment contrast.
u_{it}	Remaining variation not explained by the model.

In this example, the OLS estimate $\widehat{\beta}$ equals the DID estimate.

Two groups, one treatment start, balanced panel, equal observation weights, and no additional covariates. Causal interpretation requires PT, no anticipation, and a valid untreated comparison group.

Event Study: Pooled Preperiod Baseline



Keep the preperiod average.
Compare each post year $\ell = 5, 6, 7$ with Years 1–4, averaged within each group.

DID estimate for year ℓ

$$\widehat{\beta}_\ell = (\overline{Y}_{T,\ell} - \overline{Y}_{T,\mathrm{pre}}) - (\overline{Y}_{C,\ell} - \overline{Y}_{C,\mathrm{pre}}). \tag{11}$$

Under PT and no anticipation, this estimates ATT_ℓ. The dashed path shows outcomes without G&T under PT.

Balanced panel, equal observation weights, no additional covariates. The next slide writes out the regression.

Event Study: Posttreatment Effects by Year

$$
\begin{aligned}
Y_{it} = \alpha_i &+ \lambda_2 T_{2t} + \lambda_3 T_{3t} + \lambda_4 T_{4t} \\
&+ \lambda_5 T_{5t} + \lambda_6 T_{6t} + \lambda_7 T_{7t} \\
&+ \beta_5 (D_i^* \times T_{5t}) + \beta_6 (D_i^* \times T_{6t}) + \beta_7 (D_i^* \times T_{7t}) + u_{it}.
\end{aligned}
\tag{12}
$$

Reading the notation

Y_{it}: district i's mean achievement in year t.

α_i, λ_j: district and year fixed effects.

D_i^*: 1 for G&T adopters, 0 for controls.

T_{jt}: 1 in Year j, 0 in other years.

u_{it}: residual. Year 1's time dummy is omitted.

Interpreting the coefficients

G&T adoption begins in Year 5.

$\widehat{\beta}_5$ estimates ATT_5: the adoption year.

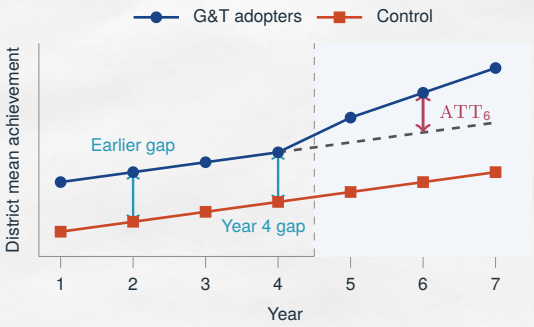
$\widehat{\beta}_6$ estimates ATT_6: one year after adoption.

$\widehat{\beta}_7$ estimates ATT_7: two years after adoption.

Same baseline: Each estimate compares that year's treated–control gap with the average gap in Years 1–4.

ATT interpretation requires PT, no anticipation, and an unaffected comparison group.

Event Study: Reference Year and Pretrends



Fix the reference: $r = t^* - 1 = 4$.
For every other year $\ell \neq r$:

$$\widehat{\beta}_\ell = (\overline{Y}_{T,\ell} - \overline{Y}_{T,r}) - (\overline{Y}_{C,\ell} - \overline{Y}_{C,r}). \quad (13)$$

Years 1–3: pretreatment placebo contrasts.

Years 5–7: year-specific ATT estimates under PT and no anticipation.

Interpretation: Compare each year's treated-control gap with the Year 4 gap.

Year 4 is the reference. Its contrast with itself is zero by construction.

The next slide writes out the regression and the joint pretrend test.

Event Study: Posttreatment Effects and Pretrend Testing

$$
\begin{aligned}
Y_{it} = \alpha_i &+ \lambda_2 T_{2t} + \lambda_3 T_{3t} + \lambda_4 T_{4t} \\
&+ \lambda_5 T_{5t} + \lambda_6 T_{6t} + \lambda_7 T_{7t} \\
&+ \beta_1 (D_i^* \times T_{1t}) + \beta_2 (D_i^* \times T_{2t}) + \beta_3 (D_i^* \times T_{3t}) \\
&+ \beta_5 (D_i^* \times T_{5t}) + \beta_6 (D_i^* \times T_{6t}) + \beta_7 (D_i^* \times T_{7t}) + u_{it}.
\end{aligned}
\tag{14}
$$

New reference: Year 4 ($k = -1$)

Each β_j compares Year j's treated–control gap with the gap in Year 4.

$\beta_1, \beta_2, \beta_3$: pretreatment placebo contrasts.

$\beta_5, \beta_6, \beta_7$: year-specific ATTs under the DID assumptions.

$\beta_4 = 0$: omitted interaction, not tested.

Joint pretrend test

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0.$

District-clustered joint Wald test.

Reject: evidence of differing pretrends.

Do not reject: insufficient evidence of differing pretrends.

Staggered DID in Education

Staggered adoption: different units begin treatment in different periods, rather than all starting together.

Our pilot literature review (2015–2025)

Among the eligible panel-DID studies:

Papers

63.6%

178 of 280 papers
included a staggered-adoption analysis.

Analyses

62.7%

188 of 300 analyses
used staggered adoption.

Staggered adoption was common in the education studies we reviewed.

Two Ongoing Evaluations with Staggered Timing

New Worlds Reading

Program background
Florida program providing free books and at-home reading resources to eligible children.

Our evaluation
Compare reading gains after admission with those of eligible untreated students.

Why staggered?
Students' **first admission dates differ**, creating earlier- and later-admission cohorts.

Literacy Matrix

Program background
Online professional learning on evidence-based reading instruction for teachers.

Our evaluation
Link teacher participation to their students' reading achievement.

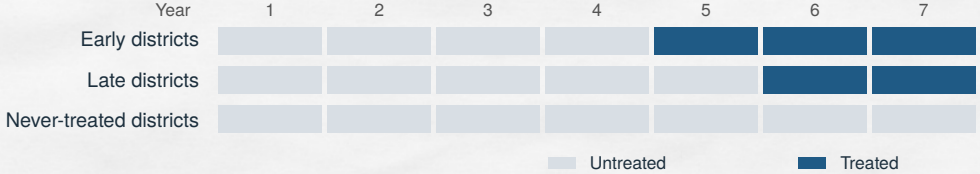
Why staggered?
Teachers **participate at different times**, creating earlier- and later-participation cohorts.

Different start dates, rather than one common start for everyone.

Program background: UF Lastinger Center, New Worlds Reading and Literacy Matrix.

Staggered G&T Adoption: Three District Groups

Follow districts in one state across Years 1–7. Group them by **first G&T adoption**.



Early adopters: first treated in Year 5.

Late adopters: first treated in Year 6.

Never-treated: no G&T adoption during the observed window.

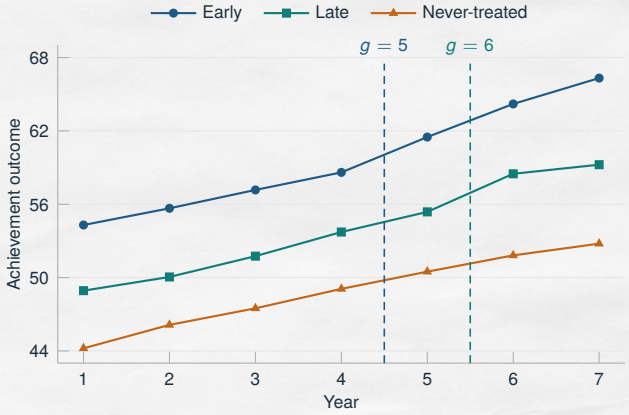
In Year 5, late adopters are **not yet treated**. In Year 6, they are treated.

Staggered DID: Can We Use TWFE?

Can we use TWFE to estimate:

- An overall treatment effect?
- Event-study effects over time?

Why, or why not?



TWFE Can Be Biased in Staggered DID

The TWFE estimator combines multiple 2×2 DIDs (Goodman-Bacon, 2021).

More precisely, it combines four types of comparisons:

1. Early adopters to never treated
2. Late adopters to never treated
3. Early adopters to late adopters (**not yet treated**)
4. Late adopters to early adopters (**already treated**)

Number 4 can mix the late group's effect with changing effects in the already-treated early group.

4. Late vs. already-treated Early

Compare Year 5 with Year 6, when Late first receives treatment.

Cohort	Year 5 Before Late's adoption	Year 6 After Late's adoption
Late, $g = 6$	$y_{6,5}$ Untreated	$y_{6,6}$ Treated
Early, $g = 5$	$y_{5,5}$ **Already treated**	$y_{5,6}$ **Already treated**

$y_{g,t}$: cohort mean of district outcomes. g = first adoption year, t = calendar year.
$\mathrm{TE}_{g,t}$: cohort g's effect in year t. Under PT and no anticipation:

$$\underbrace{(y_{6,6} - y_{6,5})}_{\text{Late's change}} - \underbrace{(y_{5,6} - y_{5,5})}_{\text{Early's change}} = \underbrace{\mathrm{TE}_{6,6}}_{\text{Late's effect}} - \underbrace{(\mathrm{TE}_{5,6} - \mathrm{TE}_{5,5})}_{\text{change in Early's effect}}. \tag{15}$$

Early is already treated in both years. Its change may include a changing treatment effect.
Subtracting that change can distort the estimate of Late's effect.

Why a TWFE Event Study Can Also Mislead

Adding leads and lags does not automatically solve the problem.

1. **One coefficient pools different cohorts.** Early and Late districts reach the same time since treatment in different calendar years, and their effects may differ.

2. **Effects from other event times can leak in.** A pooled TWFE coefficient need not equal the average effect at the event time shown on the horizontal axis.

3. **Even a pretreatment coefficient can be nonzero.** Sun and Abraham show that heterogeneous posttreatment effects can create apparent pretrends even when PT and no anticipation hold.

Their solution: estimate cohort-specific event-time effects using valid controls, then aggregate across cohorts.

This concerns conventional pooled TWFE event studies under staggered adoption, not every event-study method.

Sun and Abraham (2021), Journal of Econometrics.

Callaway–Sant'Anna DID: Key Ideas

Callaway and Sant'Anna (2021) organize staggered DID around cohort-year effects.

1. **Construct two-by-two comparisons: two key choices.**

   Reference period (When?): Year $g-1$, just before cohort g first receives treatment.

   Comparison group (Who?): Never-treated, or not-yet-treated by year t (including never-treated).

   Example: $\mathrm{ATT}(5,5)$**.** Compare Years 4 and 5.
   Controls: never-treated districts alone, or together with Year 6 adopters.

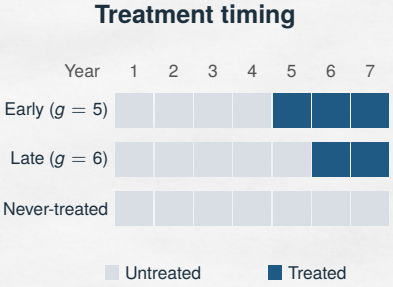
2. **Estimate** $\mathrm{ATT}(g,t)$**.** Compare changes from $g-1$ to t for cohort g and the selected controls.

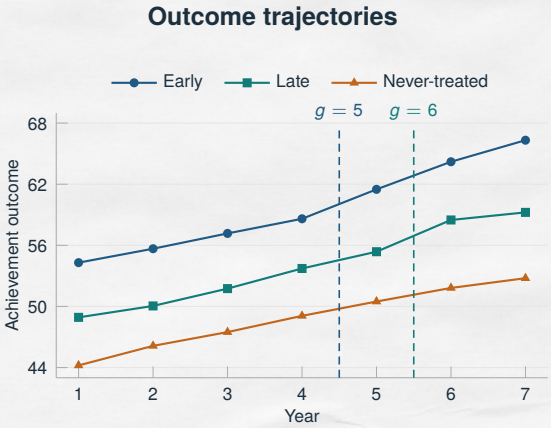
3. **Aggregate with explicit weights.** Summarize effects overall, by cohort, or by time since treatment.

4. **Conduct inference.** Use standard errors and confidence intervals with appropriate clustering. The multiplier bootstrap also supports simultaneous confidence bands for a set of effects.

Examples assume PT, no anticipation, and no covariates. Already-treated districts are excluded from the controls.

CSA: Groups and Treatment Timing





Same teaching trajectories as slide 25, not empirical G&T results. Early: Year 5. Late: Year 6.

CSA Example: Late's Effect in Year 6

Target: $\text{ATT}(6,6)$. Baseline: Year $g-1=5$. Outcome: Year 6.

Cohort	Year 5: baseline	Year 6: outcome
Late, $g=6$	$y_{6,5}$ Untreated	$y_{6,6}$ Treated
Never-treated, $g=\infty$	$y_{\infty,5}$ Untreated	$y_{\infty,6}$ Untreated
Early, $g=5$	Already treated	Already treated

Under PT and no anticipation:

$$\text{ATT}(6,6) = \underbrace{(y_{6,6} - y_{6,5})}_{\text{Late's change}} - \underbrace{(y_{\infty,6} - y_{\infty,5})}_{\text{Never-treated change}}. \tag{16}$$

Same cohort and years as slide 27, but Never-treated replaces Early.

y denotes population means. Substitute sample means to estimate the contrast.

CSA Example: Early's Effect in Year 6

Target: $\text{ATT}(5,6)$. Baseline: Year $g-1=4$. Outcome: Year 6.

Cohort	Year 4: baseline	Year 6: outcome
Early, $g=5$	$y_{5,4}$ Untreated	$y_{5,6}$ Treated
Never-treated, $g=\infty$	$y_{\infty,4}$ Untreated	$y_{\infty,6}$ Untreated
Late, $g=6$	Untreated	Treated: excluded

Under PT and no anticipation:

$$\text{ATT}(5,6) = \underbrace{(y_{5,6}-y_{5,4})}_{\text{Early's change}} - \underbrace{(y_{\infty,6}-y_{\infty,4})}_{\text{Never-treated change}}. \tag{17}$$

Use Year 4, before Early first receives treatment, rather than Year 5.

Late is treated by Year 6. Both control options therefore use only Never-treated here.

CSA Example: Two Control Options for Early

Target: ATT$(5,5)$. Baseline: Year 4. Outcome: Year 5.

Cohort	Year 4: baseline		Year 5: outcome	
Early, $g = 5$	$y_{5,4}$	Untreated	$y_{5,5}$	Treated
Late, $g = 6$	$y_{6,4}$	Untreated	$y_{6,5}$	Not yet treated
Never-treated, $g = \infty$	$y_{\infty,4}$	Untreated	$y_{\infty,5}$	Untreated

Under PT for the selected controls and no anticipation:

Option 1: Never-treated only

$$\text{ATT}(5,5) = (y_{5,5} - y_{5,4}) - (y_{\infty,5} - y_{\infty,4}). \tag{18}$$

Option 2: Not-yet-treated pool, P = Late + Never-treated

$$\text{ATT}(5,5) = (y_{5,5} - y_{5,4}) - (y_{P,5} - y_{P,4}). \tag{19}$$

P pools eligible units with the same weights at both dates, rather than weighting cohorts equally.

Each option requires its relevant PT assumption. Sample estimates can differ.

Aggregation 1: Group ATT

Question: What is the average post-treatment effect for one adoption group?

Average its group-time effects across the included post-treatment years:

$$\text{GATT}(g) = \frac{1}{T_g^+} \sum_{t \in \mathcal{T}_g^+} \text{ATT}(g,t). \tag{20}$$

- g is the group's first treatment year.
- $\mathcal{T}_g^+$ contains its included post-treatment years.
- T_g^+ is the number of those years. With all years through T, $T_g^+ = T - g + 1$.

In our seven-year design: Group 5 averages Years 5–7. Group 6 averages Years 6–7.

Equal year weights within each group in this balanced-panel example.

Aggregation 2: Average Across Groups

AGATT: First average within each group, then average the Group ATTs.

$$\text{AGATT} = \sum_{g \in \mathcal{G}} p_g \, \text{GATT}(g), \qquad \sum_{g \in \mathcal{G}} p_g = 1. \tag{21}$$

Here $\mathcal{G}$ contains treated groups, and p_g is group g's share of all treated units. In the sample, use $\widehat{p}_g = N_g / \sum_h N_h$.

Two equally sized groups:

$$\text{AGATT} = \frac{1}{2}\,\text{GATT}(5) + \frac{1}{2}\,\text{GATT}(6). \tag{22}$$

Each group receives half the total weight, even though Group 5 has three post-treatment years and Group 6 has two.

Unequal group sizes generally require unequal weights. Never-treated units receive no treatment-effect weight.

Aggregation 3: Simple ATT

Question: What is the average effect across all treated unit-periods?

Simple ATT averages the post-treatment group-time effects directly:

$$\text{ATT}^S = \frac{\sum_{g \in \mathcal{G}} p_g \sum_{t \in \mathcal{T}_g^+} \text{ATT}(g,t)}{\sum_{h \in \mathcal{G}} p_h T_h^+}. \tag{23}$$

For a balanced sample without sampling weights, each included cell receives

$$\widehat{w}_{g,t}^S = \frac{N_g}{\sum_{h \in \mathcal{G}} N_h T_h^+}. \tag{24}$$

Key difference: A group with more post-treatment years contributes more cells and therefore more total Simple ATT weight.

N_g counts distinct units in group g. The denominator counts treated unit-periods, not just treated units.

Group-Time Estimates: Workshop Example

Simulated district panel: outcome in SD units. Never-treated controls.

There are **208 districts in each treated group**, followed through Year 7.

Year	Early adopters ($g = 5$)	Late adopters ($g = 6$)
5	$\widehat{\text{ATT}}(5,5) = 0.1501$	
6	$\widehat{\text{ATT}}(5,6) = 0.2082$	$\widehat{\text{ATT}}(6,6) = 0.1926$
7	$\widehat{\text{ATT}}(5,7) = 0.3932$	$\widehat{\text{ATT}}(6,7) = 0.2903$

For example, Group 5's Year 7 effect uses Year 4 as its baseline:

$$\widehat{\text{ATT}}(5,7) = (\overline{Y}_{5,7} - \overline{Y}_{5,4}) - (\overline{Y}_{\infty,7} - \overline{Y}_{\infty,4}) \approx 0.3932. \tag{25}$$

$g = \infty$ denotes never-treated districts. N_g denotes a group size. Aggregation slides 38–40 use these same five estimates.
Source: EPRC DID workshop tutorial, Sections 8.1–8.4.

Compute the Two Group ATTs

Group 5: Average its three post-treatment effects, Years 5–7.

$$\widehat{\mathrm{GATT}}(5) = \frac{\widehat{\mathrm{ATT}}(5,5) + \widehat{\mathrm{ATT}}(5,6) + \widehat{\mathrm{ATT}}(5,7)}{3} \approx \frac{0.1500799 + 0.2082406 + 0.3932432}{3} \approx \boxed{0.2505}. \tag{26}$$

Group 6: Average its two post-treatment effects, Years 6–7.

$$\widehat{\mathrm{GATT}}(6) = \frac{\widehat{\mathrm{ATT}}(6,6) + \widehat{\mathrm{ATT}}(6,7)}{2} \approx \frac{0.1925952 + 0.2902764}{2} \approx \boxed{0.2414}. \tag{27}$$

All values are in SD units. These group averages cover different lengths of exposure, not a common follow-up window.

Source: Workshop, Section 8.2.

Compute the Average Across Groups

AGATT averages the Group ATTs, not all five cells equally.

With $N_5 = N_6 = 208$, each cohort receives $208/(208 + 208) = 1/2$.

$$\begin{aligned} \widehat{\text{AGATT}} &= \frac{1}{2}\widehat{\text{GATT}}(5) + \frac{1}{2}\widehat{\text{GATT}}(6) \\ &\approx \frac{1}{2}(0.2505212) + \frac{1}{2}(0.2414358) \approx \boxed{0.2460}. \end{aligned} \tag{28}$$

Group	Total weight	Weight per cell
Early: three post years	$1/2$	$(1/2)/3 = 1/6$
Late: two post years	$1/2$	$(1/2)/2 = 1/4$

In R: `aggte(model_csa, type = "group")`. Workshop Section 8.2.

Compute Simple ATT: Weight the Treated Cells

The denominator counts treated district-years: $208 \times 3 + 208 \times 2 = 1{,}040$.

Each of the five cells receives weight $208/1{,}040 = 0.20$.

$$\begin{aligned} \widehat{\text{ATT}}^{S} &\approx 0.20(0.1500799 + 0.2082406 + 0.3932432) \\ &\quad + 0.20(0.1925952 + 0.2902764) \\ &\approx \boxed{0.2469}. \end{aligned} \tag{29}$$

Group	AGATT weight	Simple ATT weight
Early: three post years	0.50	$3(0.20) = 0.60$
Late: two post years	0.50	$2(0.20) = 0.40$

The answers are close here, but they summarize different targets. Use joint inference, not an average of standard errors. `aggte(model_csa, type = "simple")`. Workshop Section 8.4.

Why Do We Cluster Standard Errors?

Fixed effects remove stable differences, not every source of dependence.

District fixed effects

Account for a district's stable achievement level. A district can consistently score higher than another before and after G&T adoption.

What can remain?

A local teacher shortage can affect several years. These changing shocks can remain correlated even after district and year fixed effects.

Clustered standard errors allow this remaining within-district dependence when calculating uncertainty. Ignoring it can overstate precision.

AR(1) is one example, not a requirement. Clustering does not impose a particular pattern of correlation across years.

Fixed effects address the outcome model. Clustered SEs address uncertainty, not bias or invalid parallel trends.

Cameron and Miller (2015).

Choosing the Clustering Level in Education Studies

Start with **how treatment is assigned** and **how units are sampled**.

Education study	Cluster by
District panel, program adopted by district (our example)	District
Student panels, program assigned by school	School
Student/school panels, policy assigned by district	District
District panels, policy assigned by state	State

Adjust for additional sampling or dependence structure as needed.

Many students do not compensate for very few independent clusters. Count treated clusters and clusters in each adoption cohort, too.

Our one-state example clusters by district, assuming adequate independence across districts. One state is not enough for state-clustered inference.

Abadie, Athey, Imbens, and Wooldridge (2023).

CSA Inference: One Treatment Effect

Question: Is Early's effect in Year 5 different from zero?

$$H_0 : \ \mathrm{ATT}(5,5) = 0 \qquad \text{versus} \qquad H_1 : \ \mathrm{ATT}(5,5) \neq 0. \tag{30}$$

Test: divide the estimate by its standard error, allowing dependence within districts.

$$Z = \frac{\widehat{\mathrm{ATT}}(5,5)}{\widehat{\mathrm{SE}}[\widehat{\mathrm{ATT}}(5,5)]}, \qquad \text{reject } H_0 \text{ at 5\% if } |Z| > 1.96. \tag{31}$$

This is a two-sided, large-sample normal test for **one prespecified effect**.

Result in our simulated workshop data:

$$\widehat{\mathrm{ATT}}(5,5) = 0.1501, \quad \widehat{\mathrm{SE}} = 0.0379, \quad Z = 3.96, \quad p < 0.001. \tag{32}$$

The pointwise 95% interval is $0.1501 \pm 1.96(0.0379) \approx [0.0758,\ 0.2244]$ SD.

Interpretation: reject the zero-effect null. The estimate is positive, with uncertainty assessed for this one effect, not all effects together.

Same public workshop data. District-level multiplier bootstrap, 9,999 draws. CSA att_gt documentation.

CSA Inference: One Overall ATT

Question: Is the average effect across treated district-years zero?

$$H_0 : \ \mathrm{ATT}^S = 0 \qquad \text{versus} \qquad H_1 : \ \mathrm{ATT}^S \neq 0. \tag{33}$$

Test: use the aggregate's standard error, **not an average of cell SEs**. For fixed weights w and joint covariance $\widehat{V}$ of the cell estimates:

$$\widehat{\mathrm{SE}}(\widehat{\theta}) = \sqrt{w' \widehat{V} w}, \qquad Z = \frac{\widehat{\theta}}{\widehat{\mathrm{SE}}(\widehat{\theta})}, \qquad \theta = \mathrm{ATT}^S. \tag{34}$$

Result: `aggte(model_csa, type="simple")` gives

$$\widehat{\mathrm{ATT}}^S = 0.2469, \quad \widehat{\mathrm{SE}} = 0.0320, \quad Z = 7.72, \quad p < 0.001. \tag{35}$$

The pointwise 95% interval is $[0.1842,\ 0.3096]$ SD. Reject H_0 at 5%.

One overall ATT is still one hypothesis, even though it combines five cells. It does not establish that every cell effect is positive.

Same simulated data. CSA aggte uses joint influence contributions and accounts for weight estimation when needed.

Pointwise Intervals vs. Simultaneous Bands

Pointwise: approximately 95% coverage for *one* prespecified effect.

Simultaneous: approximately 95% coverage for *all effects in a chosen set together*.

Five separate 95% intervals do not automatically provide 95% joint coverage. If they were independent, joint coverage would be $0.95^5 \approx 77\%$. Our DID estimates are correlated, so this is an illustration, not their actual coverage.

How the bootstrap chooses the critical value c^*:

1. Generate estimation errors jointly, preserving dependence within districts and correlation across effects.
2. In each draw, find the **largest absolute standardized error** across the selected effects.
3. Use the 95th percentile of those maxima as c^*.

$$\text{Band for effect } j: \quad \widehat{\theta}_j \pm c^* \widehat{\mathrm{SE}}(\widehat{\theta}_j). \tag{36}$$

The estimates and SEs stay the same. The **joint critical value is typically larger than 1.96**.

CSA tutorial: multiple-testing adjustment. Specify the family of effects before examining the results.

Simultaneous Inference: Test and Results

Let $\theta_1, \ldots, \theta_5$ be the five post-treatment group-time ATTs below.

$$H_0 : \theta_1 = \cdots = \theta_5 = 0 \quad \text{versus } H_1 : \text{ at least one } \theta_j \neq 0,$$

$$M = \max_j \left| \frac{\widehat{\theta}_j}{\widehat{\mathrm{SE}}(\widehat{\theta}_j)} \right| \qquad \text{reject at 5\% if } M > c^*. \tag{37}$$

Result: $M = 6.84 > c^* = 2.492$. Reject the joint zero-effect null.

Effect	Estimate	SE	Pointwise 95% CI	Simultaneous 95% band
ATT(5, 5)	0.1501	0.0379	[0.0758, 0.2244]	[0.0556, 0.2446]
ATT(5, 6)	0.2082	0.0499	[0.1104, 0.3061]	[0.0838, 0.3326]
ATT(5, 7)	0.3932	0.0575	[0.2806, 0.5059]	[0.2500, 0.5365]
ATT(6, 6)	0.1926	0.0413	[0.1117, 0.2735]	[0.0897, 0.2954]
ATT(6, 7)	0.2903	0.0495	[0.1932, 0.3873]	[0.1669, 0.4137]

Joint rejection means **at least one effect is nonzero**. Here, additionally, every simultaneous band lies above zero, supporting a positive effect in each cell.

Same simulated data, SD units. 9,999 multiplier draws, joint coverage of five post-treatment ATTs.

PT Assessment: The Null Hypothesis

Question: Before treatment, did each group's gap relative to Never-treated remain constant?

Use each cohort's last untreated year $g-1$ as its reference:

$$\pi(g,t) = (y_{g,t} - y_{g,g-1}) - (y_{\infty,t} - y_{\infty,g-1}), \qquad t < g-1. \tag{38}$$

These are **pre-treatment placebo contrasts**, not treatment effects.

Cohort	Reference year	Contrasts tested
Early: $g=5$	Year 4	$\pi(5,1), \pi(5,2), \pi(5,3)$
Late: $g=6$	Year 5	$\pi(6,1), \ldots, \pi(6,4)$

$$H_0: \ \pi(g,t) = 0 \text{ for all seven contrasts} \quad \text{versus} \quad H_1: \ \text{at least one is nonzero.} \tag{39}$$

The reference-year contrasts equal zero **by normalization**. They are not tested.

CSA pre-testing guide. Never-treated controls, no anticipation, and base_period="universal".

PT Assessment: Test and Interpretation

Test: combine the seven estimated preperiod contrasts in $\widehat{\pi}$, using their joint covariance $\widehat{V}_{\text{pre}}$.

$$W = \widehat{\pi}' \widehat{V}_{\text{pre}}^{-1} \widehat{\pi} \overset{H_0}{\approx} \chi^2_7, \qquad \text{reject at 5\% if } W > 14.07. \tag{40}$$

This Wald test accounts for correlation among the contrasts and repeated observations within districts. It is **not seven separate tests at 5%**.

Result in our simulated workshop data:

$$W = 9.282, \qquad \text{df} = 7, \qquad p = 0.2331 > 0.05. \tag{41}$$

Decision: do not reject the joint zero-pretrend null.

Interpretation: the test does not detect a departure from constant pre-treatment gaps. **It does not prove PT after treatment.**

- A test may miss meaningful departures when power is low.
- Inspect the pretrend plot and substantive design evidence as well.

In R: summary(model_csa). CSA pre-testing guide. No anticipation, Never-treated controls, district-level dependence.

Choosing Covariates in Education DID Studies

Three questions before adding a control:

1. Was it measured before treatment or anticipation could affect it?
2. Could it help explain selection and differences in untreated trends?
3. Are comparable untreated observations available at those values?

Possible variable	What to consider
Pretreatment poverty or school resources	Possible controls with a substantive rationale.
Post-adoption attendance, engagement, or enrollment	The program may change them. They are not automatic controls for its total effect.
Prior achievement	Potentially useful, but measurement error and regression to the mean need attention.

More controls do not necessarily produce a more credible DID estimate.

Wooldridge (2025), Sections 4 and 10.1.

How CSA Uses Covariates

Compare changes among units with similar baseline characteristics. Then average those adjusted comparisons for the treated group.

Example: districts with similar baseline poverty may have more comparable untreated trends.

Covariate data	**Balanced-panel example in `did`**
A time-constant characteristic	Use the same value in each comparison.
A characteristic recorded each year	Use the relevant baseline value for the post-treatment comparison.

Our timing example: Early uses Year 4. Late uses Year 5. These are not the current post-treatment year's values.

Assumes no anticipation and a balanced panel. Covariates do not remove the need for conditional PT and overlap. In R: `xformla = ~X`.

Official did guide.

Wooldridge ETWFE: The Key Idea

Keep district and year fixed effects. Allow treatment effects to vary.

Traditional TWFE estimates one treatment coefficient:

$$Y_{it} = \alpha_i + \lambda_t + \beta D_{it} + u_{it}. \tag{42}$$

Extended TWFE replaces βD_{it} with a separate interaction for each treated group and post-treatment year.

Our workshop data	Effects to estimate
Early: $g = 5$	$\text{ATT}(5,5),\ \text{ATT}(5,6),\ \text{ATT}(5,7)$
Late: $g = 6$	$\text{ATT}(6,6),\ \text{ATT}(6,7)$
Never: $g = \infty$	Untreated throughout the seven observed years.

Same simulated panel as CSA: 832 districts, Years 1–7, outcome in SD units.

Wooldridge (2025), Sections 5.2 and 6. No covariates in this teaching example.

ETWFE: Five Post-treatment Interactions

G_{5i} and G_{6i} indicate district i's treatment group. T_{jt} indicates calendar Year j.

$$
\begin{aligned}
Y_{it} = \alpha_i + \lambda_t \\
&+ \beta_{5,5}(G_{5i} \times T_{5t}) + \beta_{5,6}(G_{5i} \times T_{6t}) \\
&+ \beta_{5,7}(G_{5i} \times T_{7t}) \\
&+ \beta_{6,6}(G_{6i} \times T_{6t}) + \beta_{6,7}(G_{6i} \times T_{7t}) + u_{it}.
\end{aligned}
\tag{43}
$$

- α_i: district fixed effects. λ_t: calendar-year fixed effects.
- Each post-treatment cell has its own coefficient. Under the DID assumptions, $\beta_{g,t} = \text{ATT}(g,t)$.
- No pre-treatment interactions: untreated observations help fit a common time path with fixed group differences.

ETWFE: Compute an Effect from the Fitted Model

Post-treatment-only estimates from the same workshop dataset

Year	Early ($g = 5$)	Late ($g = 6$)
5	$\widehat{\beta}_{5,5} = 0.1167$	
6	$\widehat{\beta}_{5,6} = 0.1861$	$\widehat{\beta}_{6,6} = 0.2263$
7	$\widehat{\beta}_{5,7} = 0.3711$	$\widehat{\beta}_{6,7} = 0.3240$

Early adopters in Year 7:

$$\widehat{\beta}_{5,7} = \underbrace{-0.5338}_{\text{Observed mean}} - \underbrace{(-0.9049)}_{\text{Fitted untreated mean}} \tag{44}$$

$$\approx 0.3711 \quad \text{SD.}$$

The fitted untreated mean comes from district and year effects estimated using untreated observations. It is not an observed outcome for Early in Year 7.

Verified locally with fixest on the public workshop CSV. Point estimates, not significance claims.

ETWFE with Pre-treatment Interactions

Alternatively, allow separate pre-treatment deviations as well as post effects. Use Year 4 as Early's reference and Year 5 as Late's reference.

$$
\begin{aligned}
Y_{it} = \alpha_i + \lambda_t \\
&+ \gamma_{5,1}(G_{5i} \times T_{1t}) + \gamma_{5,2}(G_{5i} \times T_{2t}) + \gamma_{5,3}(G_{5i} \times T_{3t}) \\
&+ \gamma_{5,5}(G_{5i} \times T_{5t}) + \gamma_{5,6}(G_{5i} \times T_{6t}) + \gamma_{5,7}(G_{5i} \times T_{7t}) \\
&+ \gamma_{6,1}(G_{6i} \times T_{1t}) + \gamma_{6,2}(G_{6i} \times T_{2t}) + \gamma_{6,3}(G_{6i} \times T_{3t}) \\
&+ \gamma_{6,4}(G_{6i} \times T_{4t}) + \gamma_{6,6}(G_{6i} \times T_{6t}) + \gamma_{6,7}(G_{6i} \times T_{7t}) + u_{it}.
\end{aligned}
\tag{45}
$$

12 interactions: 7 pre-treatment coefficients and 5 post-treatment coefficients.

The pre coefficients are placebo contrasts, not causal effects before treatment. The post coefficients identify $\text{ATT}(g,t)$ under the relevant DID assumptions.

Wooldridge (2025), Section 6. Different letters distinguish this fit from the post-only model.

Wooldridge: A Joint Pretrend Test

Null hypothesis: all seven estimated pre-treatment deviations are zero.

$$H_0: \begin{cases} \gamma_{5,1} = \gamma_{5,2} = \gamma_{5,3} = 0, \\ \gamma_{6,1} = \gamma_{6,2} = \gamma_{6,3} = \gamma_{6,4} = 0. \end{cases} \tag{46}$$

The alternative is that at least one differs from zero. Early's Year 4 and Late's Year 5 are omitted references, not tested coefficients.

Test: collect these coefficients in $\widehat{\gamma}_{\text{pre}}$. Use their joint **district-clustered** covariance $\widehat{V}_{\text{pre}}$.

$$W = \widehat{\gamma}'_{\text{pre}} \widehat{V}^{-1}_{\text{pre}} \widehat{\gamma}_{\text{pre}} \overset{H_0}{\approx} \chi^2_7. \tag{47}$$

Workshop result: $W = 9.242$, $p = 0.2358 > 0.05$.

Decision: do not reject. No detected preperiod departure, but this does **not establish PT after treatment**.

The post-only fit has no freely estimated leads. In `etwfe`, `cgroup="notyet"` fixes pre effects at zero. That is not a passed test.

Wooldridge (2025), Section 6. Verified with the displayed full fixest model. etwfe control-group documentation.

Why Can the Two ETWFE Fits Differ?

Effect	Post only: $\widehat{\beta}$	Pre + post: $\widehat{\gamma}$
ATT(5, 5)	0.1167	0.1501
ATT(5, 6)	0.1861	0.2082
ATT(5, 7)	0.3711	0.3932
ATT(6, 6)	0.2263	0.1926
ATT(6, 7)	0.3240	0.2903

Post only: pool untreated periods to fit the counterfactual path.

Pre + post: compare with each group's last untreated year:

$$\widehat{\gamma}_{g,t} = (\overline{Y}_{g,t} - \overline{Y}_{g,g-1}) - (\overline{Y}_{\infty,t} - \overline{Y}_{\infty,g-1}), \quad t \geq g. \tag{48}$$

Here, the right column reproduces CSA's point estimates. Removing pre interactions changes the fitted model, not just what is reported.

ETWFE Aggregation: Use the Same Target Weights

With two equally sized cohorts, retain the weights defined earlier.

Post-only model:

$$\widehat{\mathrm{GATT}}(5) = 0.2246, \qquad \widehat{\mathrm{GATT}}(6) = 0.2751,$$
$$\widehat{\mathrm{AGATT}} = 0.50\widehat{\mathrm{GATT}}(5) + 0.50\widehat{\mathrm{GATT}}(6) \approx 0.2499. \tag{49}$$

$$\widehat{\mathrm{ATT}}^{S} = 0.20(0.1167 + 0.1861 + 0.3711 + 0.2263 + 0.3240) \approx 0.2448. \tag{50}$$

Model	AGATT	Simple ATT
Post-only ETWFE	0.2499	0.2448
Pre + post ETWFE / CSA here	0.2460	0.2469

Inference: combine the estimates with their full joint covariance, not an average of their standard errors. Cluster by district here.

How Wooldridge Uses Covariates

Baseline characteristics can help explain different outcome trends. For example, the association between baseline poverty and achievement may change across years.

Control	Practical guidance
Time-constant	Allow its relationship with outcomes to vary over time using the required interactions. Adding one common slope is not enough.
Time-varying	It must not respond to treatment. For the unit-FE approach, outcome shocks must not feed back into future covariate values.

Education caution: post-program engagement, staffing, or enrollment may be part of the program's effect rather than an appropriate control.

Both approaches still require credible untreated comparisons and an appropriate conditional outcome model. The exact regression matters.

Wooldridge (2025), Sections 4–5 and 10.1.

Future DID Workshops and Discussion Groups

Which topics interest you? Select all that apply, or suggest another topic.

- DID foundations, assumptions, and traditional TWFE limitations
- Staggered DID: CSA, Sun–Abraham, imputation/two-stage approaches (BJS, Gardner), ETWFE, de Chaisemartin–D'Haultfœuille, and stacked DID
- PT diagnostics, pretesting, and sensitivity analysis
- Covariate adjustment, conditional PT, and time-varying covariates
- Treatment-effect heterogeneity and alternative DID estimators
- Inference, serial correlation, and few treated groups
- Repeated cross-sections and changes in sample composition
- Hands-on implementation using R or Stata
- DID applications in education research
- Power analysis for DID study design

Share your interests: Future DID workshops and seminars

Thank You

Contact

wei.li@coe.ufl.edu

Lab Website

lcds-lab.github.io

References: Core DID Methods

- Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics, 225*(2), 200–230.
- Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics, 225*(2), 254–277.
- Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics, 225*(2), 175–199.
- Wooldridge, J. M. (2025). Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators. *Empirical Economics, 69*, 2545–2587.

Blue titles link to the papers.

References: Context and Education Applications

- Currie, J., Kleven, H., & Zwiers, E. (2020). Technology and big data are changing economics: Mining text to track methods. *AEA Papers and Proceedings, 110*, 42–48. The motivation charts use Figure IV in the NBER working-paper version (No. 26715).

- Baker, A. C., Larcker, D. F., & Wang, C. C. Y. (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics, 144*(2), 370–395.

- Strickland, K., Chan, W., & Gottfried, M. A. (2026). Longitudinal effects of NYC's gifted and talented program: Evidence from a staggered difference-in-differences approach. *Educational Evaluation and Policy Analysis*. Advance online publication.

Our education pilot-review counts are preliminary project results, not a published prevalence estimate.

References: Clustering and Statistical Inference

- Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics, 138*(1), 1–35.
- Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317–372.
- Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics, 36*(4), 672–683. See also the 2023 corrigendum.
- MacKinnon, J. G., & Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *The Econometrics Journal, 21*(2), 114–135.

References: Additional Staggered DID Approaches

- Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *The Review of Economic Studies, 91*(6), 3253–3285.
- de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review, 110*(9), 2964–2996.
- Gardner, J. (2022). Two-stage differences in differences. Preprint, arXiv:2207.05943.
- Wing, C., Freedman, S. M., & Hollingsworth, A. (2024). Stacked difference-in-differences. NBER Working Paper No. 32054.

Background and possible future-session readings. These methods are not all covered in today's lecture.

References and Resources: Software and Lab

- Callaway, B., & Sant'Anna, P. H. C. `did`: Treatment effects with multiple periods and groups. Official R package documentation, including `att_gt`, `aggte`, and pretesting guidance.
- Bergé, L. `fixest`: Fast fixed-effects estimations. Official R package documentation for estimation, interactions, and covariance options.
- McDermott, G. `etwfe`: Extended two-way fixed effects. Official R package documentation and vignette for `etwfe()` and `emfx()`.
- EPRC DID workshop: R tutorial and example data. The worked examples use this public workshop dataset.