

# Week 10: Modern Difference-in-Differences

Minimum wage revisited · PP5001 · Martinmas 2026

[Class slides](#) · [Slides PDF](#) · [Notes PDF](#)

## Learning Objectives

---

By the end of this session, you should be able to:

- Identify the comparison groups used by TWFE when treatment begins at different times.
- Explain why changes in treatment effects can contaminate comparisons with already-treated units.
- Interpret a group-time average treatment effect and construct it from a two-period comparison.
- Explain how averaging across cohorts, calendar years or event times answers different questions.
- Assess the identifying assumptions and empirical evidence in a modern DiD application.

## Before class

---

Read the empirical application in **Callaway and Sant'Anna (2021), Section 5, Tables 2–3 and Figure 1**, available in Moodle. Goodman-Bacon (2021) supplies the decomposition underlying our discussion. The technical proofs are advanced reading. Revisit the Week 9 distinction between calendar time and event time.

For textbook support, revisit *The Effect*, Chapter 18, and *Causal Inference: The Remix*, Chapter 10 (Complex Diff-in-Diff Designs). *Mastering 'Metrics*, Chapter 5, and *Mostly Harmless Econometrics*, Chapter 5, provide background on the original DiD logic. The papers develop the recent staggered-adoption issues.

These notes follow the sequence of the original lecture slides: timing groups, the comparisons inside TWFE, changing treatment effects, and explicit comparisons followed by aggregation.

## Recall: Two-Way Fixed Effects

---

$$Y_{it} = \alpha_i + \lambda_t + \delta D_{it} + u_{it}.$$

- $\alpha_i$ : time-invariant differences between units.
- $\lambda_t$ : common changes in calendar time.
- $D_{it}$ : whether unit  $i$  is treated in period  $t$ .

With different adoption dates, **which comparisons determine  $\hat{\delta}$** ?

With two groups and two periods, we could identify the four means behind the interaction coefficient. With multiple adoption dates, the regression combines several comparisons. To assess the coefficient, we need to know who supplies the counterfactual in each comparison. Throughout this session, treatment is absorbing: once a unit adopts, it remains treated.

## Multiple Treatment Times

---

Policies often begin in different places at different times:

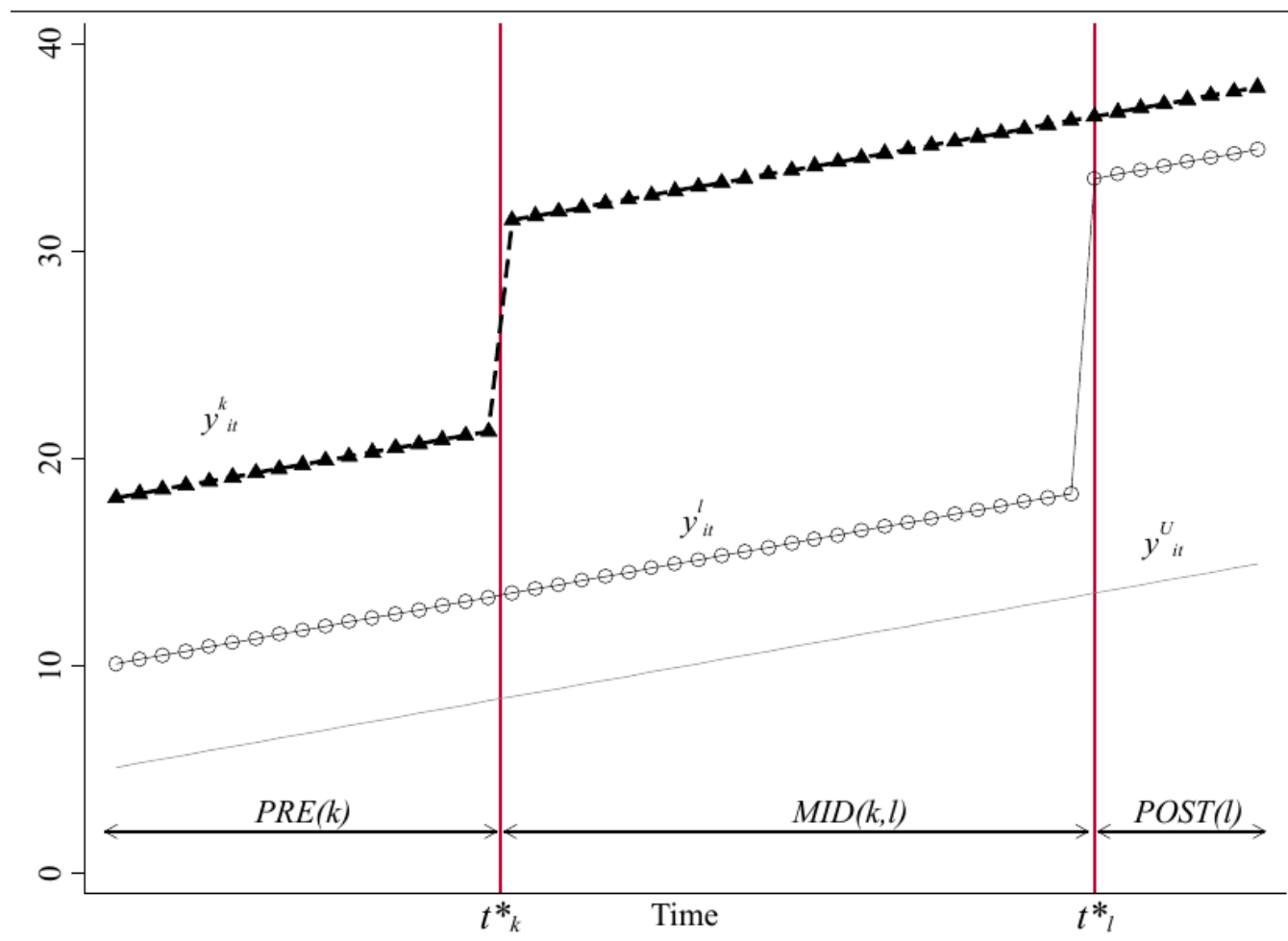
- Minimum-wage increases.

- Health-insurance expansions.
- School reforms.
- Environmental regulations.

A **cohort** is a group of units first treated in the same calendar period.

The cohort label records the adoption date. It does not record how long a unit has been treated. A county in the 2004 cohort has event time zero in 2004 and event time three in 2007. At the same calendar date, a county in the 2007 cohort is just beginning treatment.

## Timing Groups



Timing groups, from the original lecture slides.

**For discussion:** Identify the early-treated, late-treated and never-treated groups. When can each group supply untreated observations?

## Four Comparisons

With early, late and never-treated groups, TWFE combines:

1. Early treated versus never treated.

2. Late treated versus never treated.
3. Early treated versus late treated, **before the late group adopts**.
4. Late treated versus early treated, **after the early group has adopted**.

In comparison 4, the comparison group's outcomes already include treatment effects.

Comparisons 1–3 can identify treatment effects under suitable parallel trends and no anticipation. Their availability alone does not establish those assumptions. Comparison 4 requires particular attention because changes in the early group's treatment effect enter its outcome change.

## The Goodman-Bacon Decomposition

For the basic TWFE regression in a balanced panel,

$$\hat{\delta}^{TWFE} = \sum_j s_j \hat{\delta}_j^{2 \times 2}, \quad s_j \geq 0, \quad \sum_j s_j = 1.$$

- Each component is a two-group, two-period comparison.
- Timing and treatment variation determine the weights.
- Both not-yet-treated and already-treated groups can serve as controls.

This is the Goodman-Bacon (2021) decomposition for the basic binary-treatment specification without additional time-varying covariates. The two periods may themselves be averages over several calendar periods. The weights on these component DiD estimates are nonnegative. This is distinct from rewriting the coefficient as a weighted average of underlying treatment effects, where negative weights can arise. Covariates and other specifications require additional care, so we use the basic setting to establish the logic.

## When a Common Effect Is Adequate

Suppose untreated outcomes follow parallel trends and there is no anticipation.

If treatment has the **same constant effect** for every treated cohort and period, an already-treated group's treatment effect cancels when its outcome is differenced.

The difficulty arises when that effect changes over the comparison period.

A constant effect is a useful benchmark. It is more restrictive than allowing a policy to have a small immediate effect and a larger effect after several years. It also rules out differences in the effect across cohorts. We should consider whether that benchmark fits the policy being evaluated.

## An Illustration

Suppose every group's untreated outcome would remain at 10.

| Group         | Period 0 | Period 1 | Period 2 |
|---------------|----------|----------|----------|
| Early treated | 10       | 12       | 16       |
| Late treated  | 10       | 10       | 12       |
| Never treated | 10       | 10       | 10       |

Early adopters have an effect of 2 initially and 6 later. Late adopters have an effect of 2 when they adopt.

These are illustrative group means. Parallel trends holds exactly because every untreated potential outcome remains at 10. There is no anticipation. Every treatment effect in the example is positive. This isolates the consequence of using outcomes affected by treatment as the comparison.

## The Same Illustration Graphically



**For discussion:** From period 1 to period 2, whose observed outcome changes most? What would happen if we used that group as the control for late adopters?

## Using an Untreated Comparison

From period 0 to period 1:

$$\hat{\delta}_{Early,Late} = (12 - 10) - (10 - 10) = 2.$$

From period 1 to period 2, using never-treated units:

$$\hat{\delta}_{Late,Never} = (12 - 10) - (10 - 10) = 2.$$

Each calculation recovers the newly treated group's effect.

## Using an Already-Treated Comparison

Compare late with early adopters from period 1 to period 2:

$$\hat{\delta}_{Late,Early} = (12 - 10) - (16 - 12) = -2.$$

The late group's treatment effect is **+2**, but this comparison subtracts the **+4 change in the early group's treatment effect**.

**For discussion:** Why does a negative comparison arise even though every treatment effect is positive?

## What Is Being Subtracted?

Under parallel trends in untreated outcomes,

$$DiD_{Late,Early} = \text{late group's new effect} - \text{change in early group's effect}.$$

Growing effects in the early cohort reduce this component of TWFE.

This is the central issue in the example. The late cohort's untreated trend can be parallel to the early cohort's untreated trend while their observed changes differ because the early cohort's treatment effect is evolving. A small TWFE coefficient can therefore reflect offsetting components. The regression's precision does not resolve this interpretation problem.

## Weights on Treatment Effects

---

The weights on the component  $2 \times 2$  comparisons are positive.

But an already-treated comparison **subtracts changes in treatment effects**. When TWFE is expressed in terms of underlying cohort-period effects, some of those effects can receive negative weight.

A TWFE coefficient can consequently lie outside the range of the underlying treatment effects.

The distinction between these two sorts of weights matters. We need no further weighting algebra here. The numerical example shows how an effect enters with a minus sign inside a positively weighted comparison. The full coefficient averages all the component comparisons, so the negative value of one component does not imply that the overall coefficient must be negative.

## Event-Study Coefficients

---

In a conventional TWFE event study with staggered adoption:

- Units at different event times are observed in the same calendar period.
- Treatment effects can differ across cohorts and exposure lengths.
- An estimated lead or lag can combine effects from other event times.

Even an apparent pre-treatment coefficient can reflect post-treatment effects elsewhere.

Our Week 9 event-study indicators still describe the data correctly. The concern is the causal interpretation of their regression coefficients under heterogeneous effects. Sun and Abraham (2021) establish this contamination result. We cite the result here to motivate explicit cohort-specific comparisons. Flat pre-treatment estimates alone cannot establish that a conventional TWFE event study has isolated each dynamic effect.

## Make the Comparisons Explicit

---

The approach we will use has three steps:

1. Define an untreated comparison group.
2. Estimate effects separately for each adoption cohort and calendar period.
3. Average those effects to answer a stated policy question.

Callaway and Sant'Anna (2021) organise their method around these steps.

## A Group-Time Treatment Effect

---

Let  $G_i = g$  mean that unit  $i$  first receives treatment in period  $g$ .

$$ATT(g, t) = E[Y_{it}^g - Y_{it}^\infty \mid G_i = g], \quad t \geq g.$$

- $Y_{it}^g$ : outcome at  $t$  if treatment first begins at  $g$ .
- $Y_{it}^\infty$ : outcome at  $t$  if the unit remains untreated.
- $ATT(g, t)$ : effect at  $t$  for the cohort first treated at  $g$ .

Earlier weeks used  $Y_i^1$  and  $Y_i^0$  to distinguish treatment states. Here we must also distinguish treatment histories: receiving treatment for three years can differ from receiving it for one year. The superscript  $g$  records the date treatment starts. The infinity symbol denotes never receiving treatment. This extends the potential-outcome notation to make that history explicit.

## Reading the Two Dates

$$ATT(2004, 2006)$$

means the effect **in calendar year 2006**, for units **first treated in 2004**.

Their event time is

$$k = t - g = 2006 - 2004 = 2.$$

**For discussion:** Which effect would describe the 2006 cohort in its first treatment year? Which would describe that same cohort one year later?

## Identify the Counterfactual

Use  $g - 1$ , the last pre-treatment period, as the baseline.

With never-treated units as controls, assume

$$E[Y_{it}^\infty - Y_{i,g-1}^\infty | G_i = g] = E[Y_{it}^\infty - Y_{i,g-1}^\infty | G_i = \infty].$$

Also assume no anticipation at the baseline.

This is parallel trends over the particular comparison window. It allows the two groups to have different outcome levels. No anticipation means that the treated cohort's outcome at  $g - 1$  has not yet been affected by its future treatment. We also maintain well-defined treatment and no relevant spillovers into comparison units.

## Back to a Two-by-Two Difference

Under those assumptions,

$$ATT(g, t) = E[Y_{it} - Y_{i,g-1} | G_i = g] - E[Y_{it} - Y_{i,g-1} | G_i = \infty].$$

The sample estimator replaces each expectation with its corresponding sample mean.

We already know how to calculate this comparison.

Add and subtract the treated cohort's untreated outcome at  $t$ . Its observed change is the treatment effect plus its untreated change. Parallel trends supplies that untreated change from the comparison group. The calculation is the same logic as Week 8, applied to a particular cohort and follow-up date. Using the same units in both periods avoids changing the sample composition between the two means.

## Choosing the Comparison Group

---

**Never treated:** remain untreated throughout the observation period.

**Not yet treated:** remain untreated through the outcome period  $t$ .

For the latter, use units with  $G_i > t$ , including never-treated units where available, and maintain the corresponding parallel-trends and no-anticipation assumptions.

Units that adopt between the baseline and the follow-up date cannot supply untreated outcomes for this comparison. If behaviour changes before formal adoption, the comparison window must also account for anticipation. Once every unit is treated, this strategy has no untreated group left for subsequent periods. More observations after that point do not restore the missing comparison.

## Conditional Parallel Trends

---

Sometimes parallel trends is plausible only among units with similar baseline characteristics  $X_i$ .

Estimate comparisons conditional on those characteristics, then average over the treated cohort's characteristics.

Overlap is needed: the comparison group must contain suitable counterparts for the treated cohort.

This connects to Week 2. The identifying assumption now concerns untreated outcome changes conditional on  $X_i$ . Callaway and Sant'Anna provide regression, weighting and doubly robust implementations. For this module, understand what the covariate adjustment is intended to achieve. The application uses a doubly robust implementation, but we do not derive that estimator. Covariates should support the substantive parallel-trends argument and be measured before they could be affected by treatment.

## Which Average Do We Want?

---

Once we have  $ATT(g, t)$ , we can average:

- **Within cohorts:** how did each adoption cohort fare?
- **Within calendar years:** what was the effect among treated units in a particular year?
- **At a common event time:** how does the effect evolve with exposure?

These averages can differ because they combine different effects with different weights.

## An Event-Time Average

---

At event time  $k \geq 0$ , combine available cohorts at  $t = g + k$ :

$$\theta_k = \sum_{g \in \mathcal{G}_k} w_{g,k} ATT(g, g + k), \quad w_{g,k} \geq 0, \quad \sum_{g \in \mathcal{G}_k} w_{g,k} = 1.$$

$\mathcal{G}_k$  contains cohorts observed at event time  $k$  with a valid comparison group. Cohort-size weights give larger cohorts more influence.

This is a transparent average of effects at the same exposure length. A reader can ask which cohorts contribute and how much each contributes. The set of available cohorts often shrinks as  $k$  grows because later adopters have fewer post-treatment observations.

## Changing Cohort Composition

Suppose data end in 2007.

| First treated | Last observed event time |
|---------------|--------------------------|
| 2004          | 3                        |
| 2006          | 1                        |
| 2007          | 0                        |

An event-time graph can change because effects evolve **and** because the contributing cohorts change.

A balanced-cohort graph keeps the same cohorts at every displayed exposure length.

## Minimum Wages Revisited

Callaway and Sant’Anna study county-level teenage employment during 2001–2007.

- The federal minimum wage stays at \$5.15 over the study window.
- Treated states raise their minimum wage above that level.
- Cohorts are defined by the first increase.
- The main comparison group is counties in states that remain at the federal minimum.

**For discussion:** Explain the policy, data and source of identifying variation. What differs from Card and Krueger’s comparison?

The paper’s main sample contains 2,284 counties. It combines employment from the Quarterly Workforce Indicators with pre-treatment county characteristics. The policy changes at the state level, while outcomes are measured at the county level. Their application provides a comparison of estimators under maintained identifying assumptions. It also offers evidence with which to scrutinise those assumptions.

## Table 2: The Comparison Groups

**Table 2**

Summary statistics for main dataset.

Source: Quarterly Workforce Indicators and 2000 County Data Book.

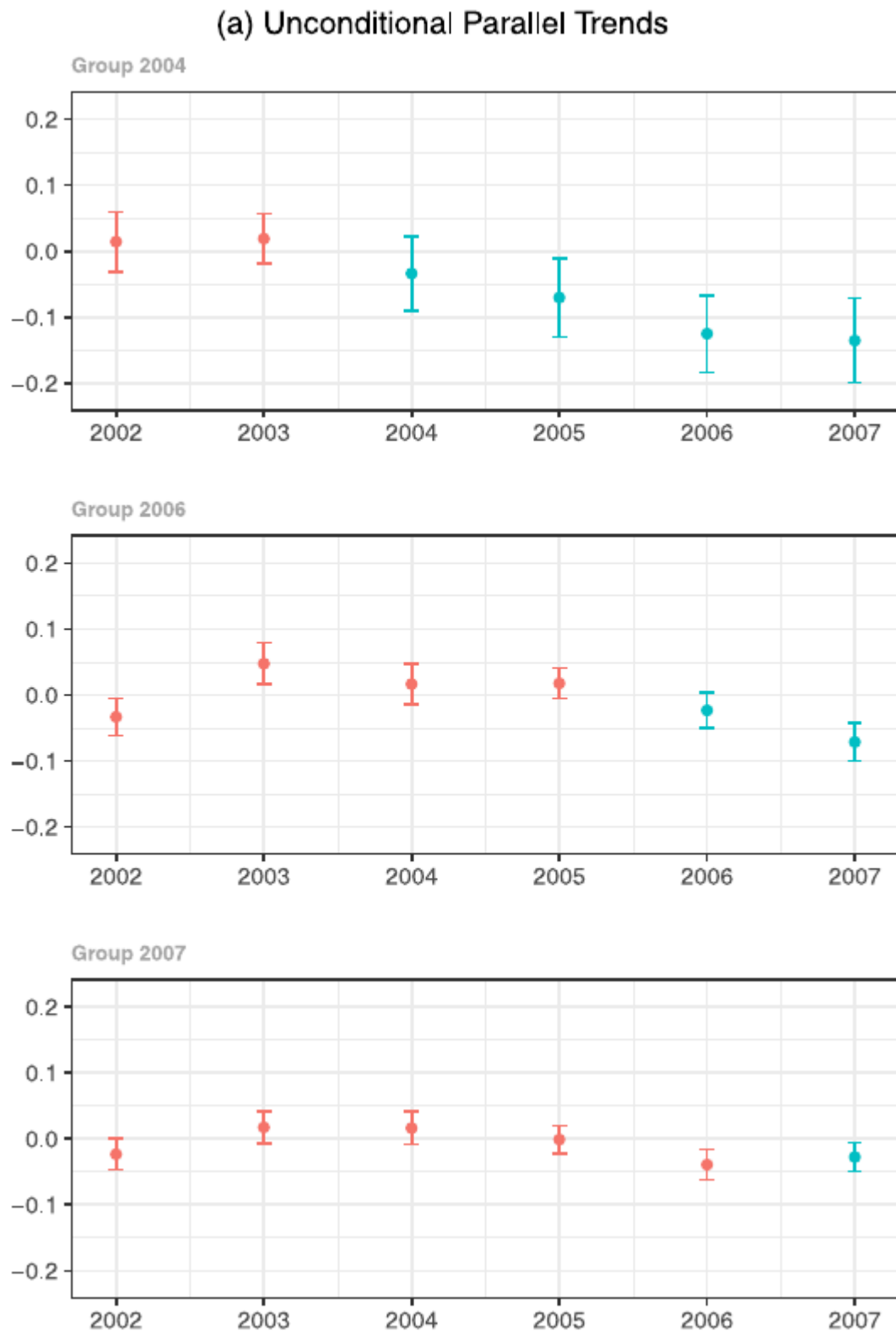
|                     | Treated counties | Untreated counties | Diff. | P-val on Diff. |
|---------------------|------------------|--------------------|-------|----------------|
| Midwest             | 0.59             | 0.34               | 0.25  | 0.00           |
| South               | 0.27             | 0.59               | −0.32 | 0.00           |
| West                | 0.14             | 0.07               | 0.07  | 0.00           |
| Population (1000s)  | 94.32            | 53.43              | 40.89 | 0.00           |
| White               | 0.89             | 0.83               | 0.06  | 0.00           |
| HS Graduates        | 0.59             | 0.55               | 0.04  | 0.00           |
| Poverty Rate        | 0.13             | 0.16               | −0.03 | 0.00           |
| Median Inc. (1000s) | 33.91            | 31.89              | 2.02  | 0.00           |

*Notes:* Summary statistics for counties located in states that raised their minimum wage between Q2 of 2003 and Q1 of 2007 (treated) and states whose minimum wage was effectively set at the federal minimum wage for the entire period (untreated). The sample consists of 2284 counties.

Callaway and Sant'Anna (2021), Table 2, p. 217.

**For discussion:** Describe the differences between treated and untreated counties. What do these differences imply for the parallel-trends argument and the choice of controls?

**Figure 1(a): Unconditional Parallel Trends**

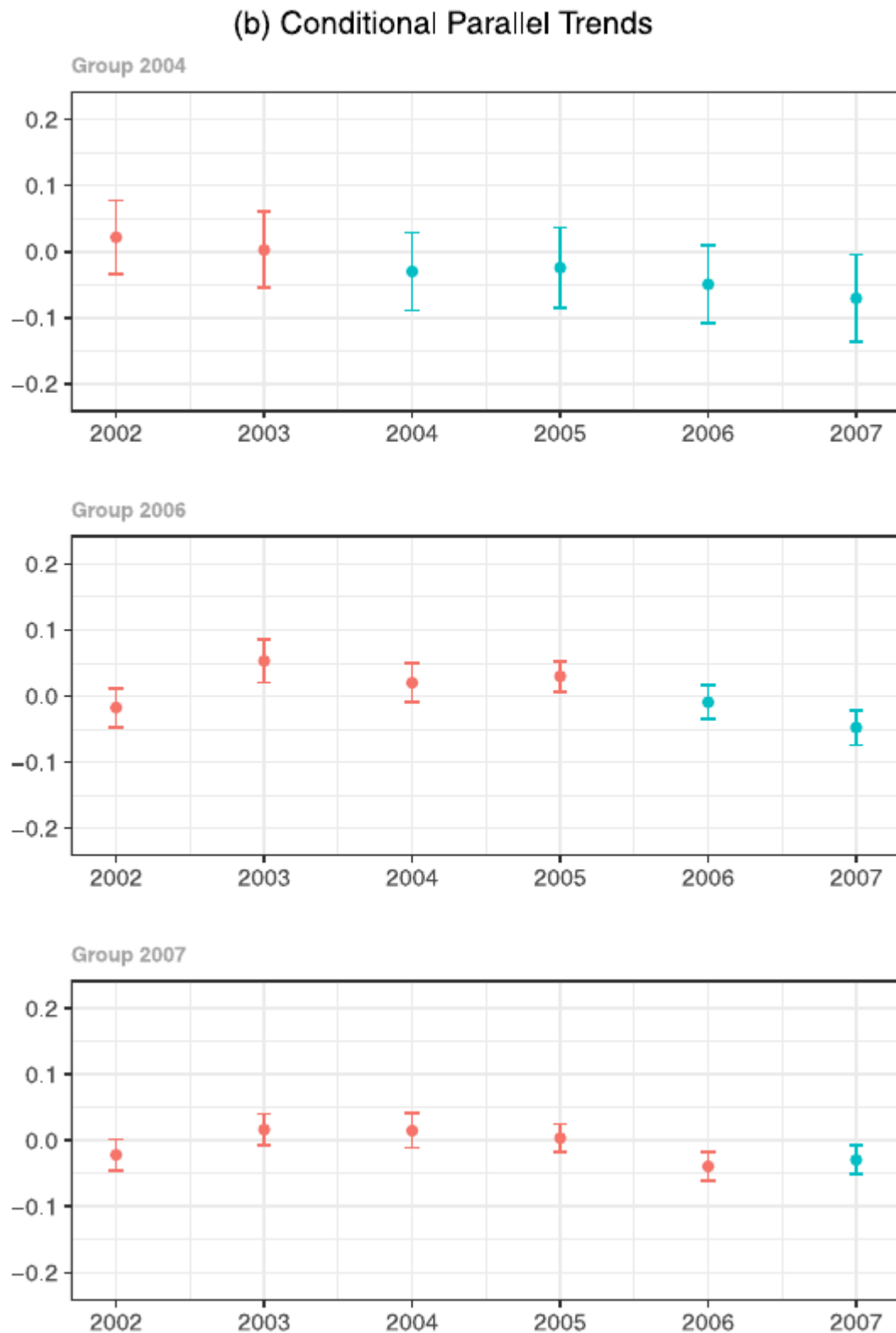


Callaway and Sant'Anna (2021), Figure 1(a), p. 218.

**For discussion:** Identify the cohorts and axes. Interpret one post-treatment estimate. What concerns arise from the pre-treatment evidence?

The horizontal axis is calendar year. Each row represents a different adoption cohort. Red points are pre-treatment comparisons and blue points are post-treatment effects. The published uncertainty intervals are simultaneous 95% confidence bands. Pre-treatment comparisons assess changes before adoption, rather than representing causal effects of a treatment that has not yet occurred.

**Figure 1(b): Conditional Parallel Trends**



Callaway and Sant'Anna (2021), Figure 1(b), p. 218.

**For discussion:** What changes after adjustment for baseline characteristics? How convincing is the identifying assumption now?

The controls include region, population, racial composition, educational attainment, poverty and income. Adjustment changes the comparisons used to construct the counterfactual. Examine both the effect estimates and the pre-treatment evidence. The method's ability to accommodate heterogeneous effects does not by itself establish conditional parallel trends.

## Table 3(a): Different Averages

**Table 3**

Minimum wage aggregated treatment effect estimates.

| (a) Unconditional parallel trends |                                    |                                    |                                    |                                    |                   |
|-----------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|-------------------|
|                                   | Partially aggregated               |                                    |                                    |                                    | Single parameters |
| TWFE                              |                                    |                                    |                                    |                                    | −0.037<br>(0.006) |
| Simple weighted average           |                                    |                                    |                                    |                                    | −0.052<br>(0.006) |
| Group-specific effects            | <u>g=2004</u><br>−0.091<br>(0.019) | <u>g=2006</u><br>−0.047<br>(0.008) | <u>g=2007</u><br>−0.028<br>(0.007) |                                    | −0.039<br>(0.007) |
| Event study                       | <u>e=0</u><br>−0.027<br>(0.006)    | <u>e=1</u><br>−0.071<br>(0.009)    | <u>e=2</u><br>−0.125<br>(0.021)    | <u>e=3</u><br>−0.136<br>(0.023)    | −0.090<br>(0.013) |
| Calendar time effects             | <u>t=2004</u><br>−0.034<br>(0.019) | <u>t=2005</u><br>−0.071<br>(0.02)  | <u>t=2006</u><br>−0.055<br>(0.009) | <u>t=2007</u><br>−0.050<br>(0.006) | −0.052<br>(0.013) |
| Event study<br>w/ Balanced groups | <u>e=0</u><br>−0.027<br>(0.009)    | <u>e=1</u><br>−0.071<br>(0.009)    |                                    |                                    | −0.049<br>(0.008) |

Callaway and Sant'Anna (2021), Table 3(a), p. 219.

**For discussion:** Focus on TWFE and the group-specific effects. Compare the TWFE estimate with the overall group-specific effect in the rightmost column. How similar are they?

## Table 3(b): With Covariate Adjustment

| (b) Conditional parallel trends   |                   | Partially aggregated |                   |                   |  | Single parameters |
|-----------------------------------|-------------------|----------------------|-------------------|-------------------|--|-------------------|
| TWFE                              |                   |                      |                   |                   |  | −0.008<br>(0.006) |
| Simple weighted average           |                   |                      |                   |                   |  | −0.033<br>(0.007) |
| Group-specific effects            | <u>g=2004</u>     | <u>g=2006</u>        | <u>g=2007</u>     |                   |  | −0.031<br>(0.007) |
|                                   | −0.044<br>(0.020) | −0.029<br>(0.008)    | −0.029<br>(0.008) |                   |  |                   |
| Event study                       | <u>e=0</u>        | <u>e=1</u>           | <u>e=2</u>        | <u>e=3</u>        |  | −0.046<br>(0.013) |
|                                   | −0.024<br>(0.006) | −0.041<br>(0.009)    | −0.050<br>(0.022) | −0.071<br>(0.026) |  |                   |
| Calendar time effects             | <u>t=2004</u>     | <u>t=2005</u>        | <u>t=2006</u>     | <u>t=2007</u>     |  | −0.033<br>(0.012) |
|                                   | −0.030<br>(0.022) | −0.025<br>(0.021)    | −0.030<br>(0.009) | −0.049<br>(0.007) |  |                   |
| Event study<br>w/ Balanced groups |                   | <u>e=0</u>           | <u>e=1</u>        |                   |  | −0.028<br>(0.008) |
|                                   |                   | −0.016<br>(0.010)    | −0.041<br>(0.009) |                   |  |                   |

Callaway and Sant'Anna (2021), Table 3(b), p. 219.

**For discussion:** Compare TWFE with the overall group-specific effect in the rightmost column. What changes relative to panel (a) after covariate adjustment? Would your policy conclusion depend on the estimator?

The outcome is log teenage employment. Interpret a coefficient  $b$  approximately as a  $100b$  percent change, or use  $100(\exp(b) - 1)$  for the exact conversion. The group-specific aggregation first averages within each cohort over its post-treatment periods and then combines cohorts. A simple average over available cohort-time effects gives more influence to cohorts observed for more post-treatment periods. The table illustrates why the target average must be specified.

## Assessing the Policy Evidence

- Are untreated employment trends plausibly comparable?
- What do the pre-treatment estimates suggest?
- Minimum-wage increases differ in size. What does the binary treatment represent?

The authors explicitly acknowledge pre-treatment evidence against parallel trends and variation in the size of the policy change. The published application uses county-clustered bootstrap standard errors. Observations from the same county in different years may share common shocks. Clustering allows for this dependence when measuring uncertainty about the estimates. The exercise shows how to request county-clustered standard errors in Python.

## Choosing and Reporting an Estimate

1. State the policy effect and population of interest.
2. Identify the adoption cohorts and untreated comparisons.
3. Explain parallel trends, anticipation and possible spillovers.
4. Report how cohort-time effects are averaged.
5. Show uncertainty and the evidence used to assess the design.

**For discussion:** Which of these choices would matter most if you were advising a government about another minimum-wage increase?

## What We Take Forward

---

The basic DiD logic remains: use an untreated group's change to construct a counterfactual.

With staggered adoption, make the comparison separately for each cohort and period, then average deliberately.

Next week: constructing a comparison for a single treated country or region using synthetic control.

## After class

---

The [ungraded exercise](#) uses the authors' public teaching subset. You will estimate TWFE, calculate two group-time comparisons directly, and examine how aggregation changes the question. The subset has fewer counties and years than the published application, so report its results as a separate exercise.

## References

---

Callaway, B., and P. H. C. Sant'Anna (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., and S. Abraham (2021). [Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects](#). *Journal of Econometrics*, 225(2), 175–199.