# Causal Panel Analysis under Parallel Trends:

## Lessons from a Large Reanalysis Study

February 2025

Albert Chiu (Stanford) Xingchen Lan (NYU) Ziyi Liu (Berkeley) Yiqing Xu (Stanford)

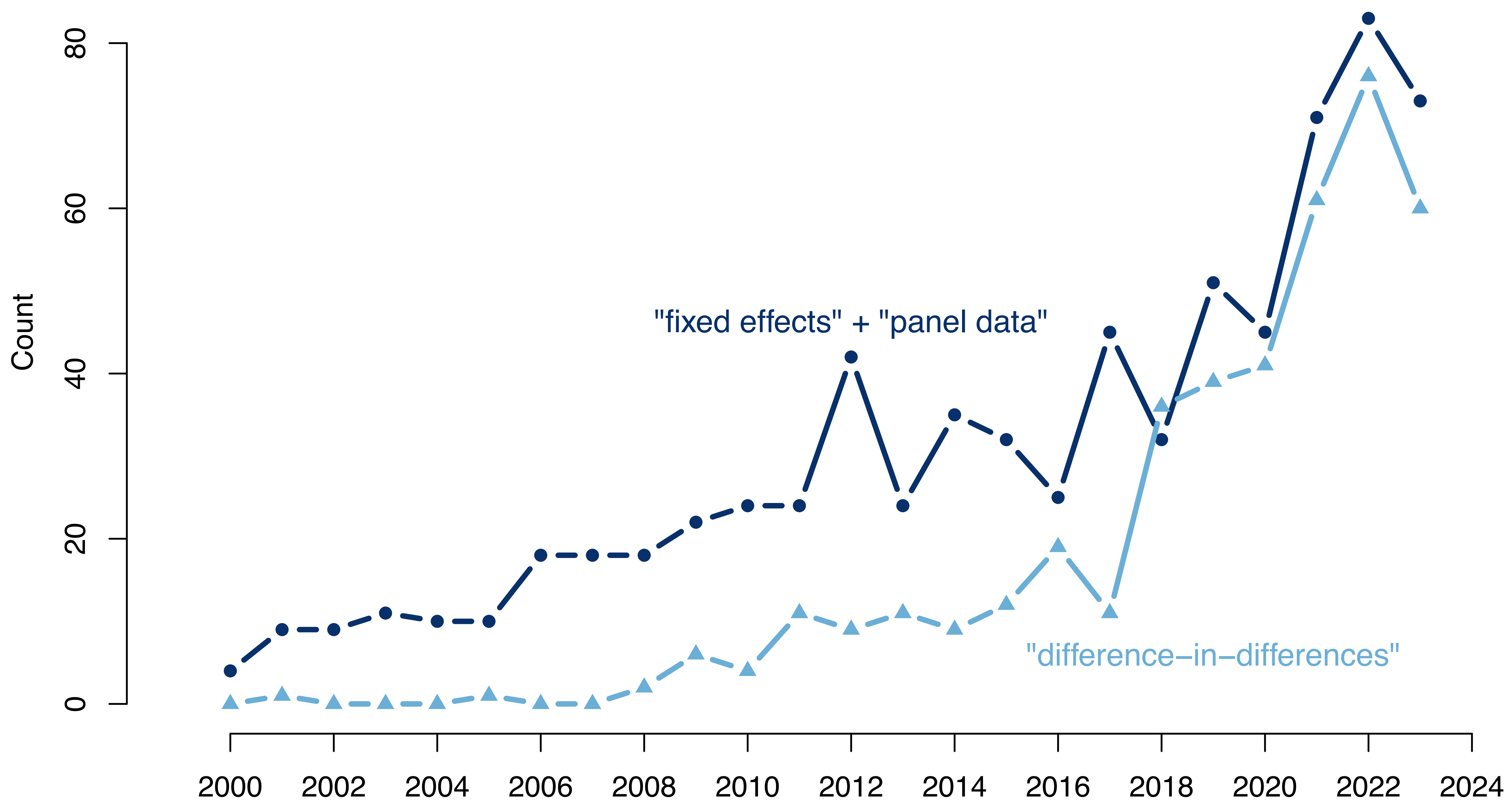
# Motivation

# Motivation

- Fact 1. Panel data are ubiquitous in today's social sciences. Two-way fixed effects (TWFE) models are the most commonly used to establish causality

# Motivation

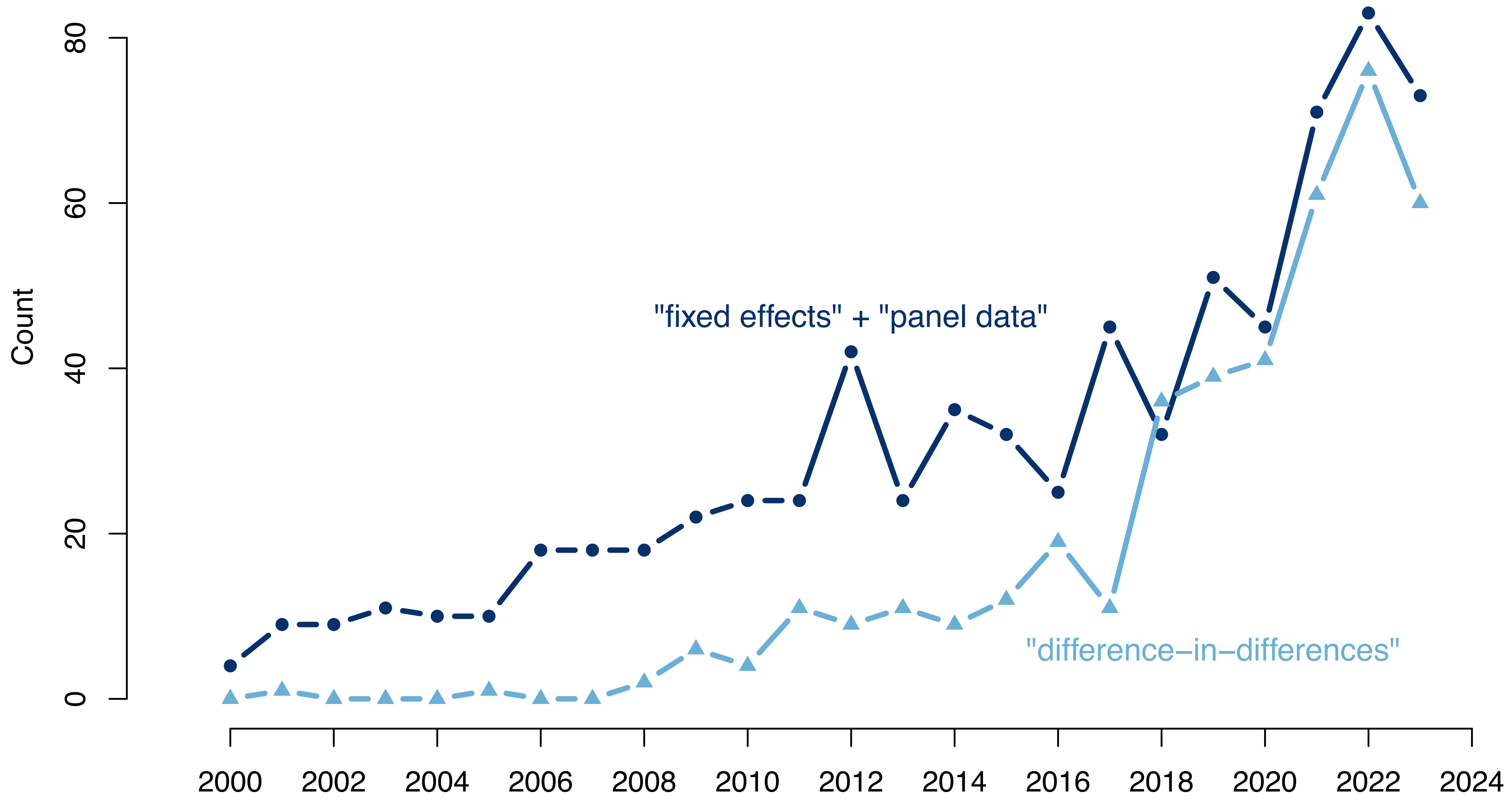
- Fact 1. Panel data are ubiquitous in today's social sciences. Two-way fixed effects (TWFE) models are the most commonly used to establish causality




#Paper Published in APSR/AJPS/JOP/CPS/IO

# Motivation

- Fact 1. Panel data are ubiquitous in today's social sciences. Two-way fixed effects (TWFE) models are the most commonly used to establish causality
  - 77% of the articles we review use FE models (mostly TWFE)




#Paper Published in APSR/AJPS/JOP/CPS/IO

# Motivation

- Fact 1. Panel data are ubiquitous in today's social sciences. Two-way fixed effects (TWFE) models are the most commonly used to establish causality
  - 77% of the articles we review use FE models (mostly TWFE)
  - TWFE models are synonymous to "difference-in-differences" (94%)

# Motivation

- Fact 1. Panel data are ubiquitous in today's social sciences. Two-way fixed effects (TWFE) models are the most commonly used to establish causality
  - 77% of the articles we review use FE models (mostly TWFE)
  - TWFE models are synonymous to "difference-in-differences" (94%)

- Fact 2. Existing and nascent literature casts doubts on FE/TWFE estimators

# Motivation

- Fact 1. Panel data are ubiquitous in today's social sciences. Two-way fixed effects (TWFE) models are the most commonly used to establish causality
  - 77% of the articles we review use FE models (mostly TWFE)
  - TWFE models are synonymous to "difference-in-differences" (94%)

- Fact 2. Existing and nascent literature casts doubts on FE/TWFE estimators
  - Inferential problems (e.g., Bertrand et al., 2004; Cameron et al, 2008)

# Motivation

- Fact 1. Panel data are ubiquitous in today's social sciences. Two-way fixed effects (TWFE) models are the most commonly used to establish causality
  - 77% of the articles we review use FE models (mostly TWFE)
  - TWFE models are synonymous to "difference-in-differences" (94%)

- Fact 2. Existing and nascent literature casts doubts on FE/TWFE estimators
  - Inferential problems (e.g., Bertrand et al., 2004; Cameron et al, 2008)
  - Unrealistic assumptions on assignment mechanism or lack of designs (e.g., Blackwell and Glynn 2018; Imai & Kim 2019)

# Motivation

- Fact 1. Panel data are ubiquitous in today's social sciences. Two-way fixed effects (TWFE) models are the most commonly used to establish causality
  - 77% of the articles we review use FE models (mostly TWFE)
  - TWFE models are synonymous to "difference-in-differences" (94%)

- Fact 2. Existing and nascent literature casts doubts on FE/TWFE estimators
  - Inferential problems (e.g., Bertrand et al., 2004; Cameron et al, 2008)
  - Unrealistic assumptions on assignment mechanism or lack of designs (e.g., Blackwell and Glynn 2018; Imai & Kim 2019)
  - Consequence of heterogeneous treatment effects (HTE) (e.g., Imai & Kim 2019; Athey and Imbens, 2018; Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020; Strezhnev, 2018; Callaway and Sant'Anna, 2021; Sun and Abraham 2021; Borusyak, Jaravel and Spiess, 2023)

# Motivation

- Fact 1. Panel data are ubiquitous in today's social sciences. Two-way fixed effects (TWFE) models are the most commonly used to establish causality
  - 77% of the articles we review use FE models (mostly TWFE)
  - TWFE models are synonymous to "difference-in-differences" (94%)

- Fact 2. Existing and nascent literature casts doubts on FE/TWFE estimators
  - Inferential problems (e.g., Bertrand et al., 2004; Cameron et al, 2008)
  - Unrealistic assumptions on assignment mechanism or lack of designs (e.g., Blackwell and Glynn 2018; Imai & Kim 2019)
  - Consequence of heterogeneous treatment effects (HTE) (e.g., Imai & Kim 2019; Athey and Imbens, 2018; Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020; Strezhnev, 2018; Callaway and Sant'Anna, 2021; Sun and Abraham 2021; Borusyak, Jaravel and Spiess, 2023)

    ➡ Many new estimators have been proposed...

# What TWFE Assumptions Entail

| | |
|---|---|
| Functional Form | $Y_{it} = \delta^{TWFE} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \epsilon_{it}$ |
| Strict Exogeneity | $D_{it} \perp\!\!\!\perp \epsilon_{js} \,\vert\, \mathbf{X}^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s$ |

**Related work:** Blackwell & Glynn (2018); Imai & Kim (2019); Athey & Imbens (2022); Liu, Wang & Xu (2022)

# What TWFE Assumptions Entail

| | |
|---|---|
| Functional Form | $Y_{it} = \delta^{TWFE} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \epsilon_{it}$ |
| Strict Exogeneity | $D_{it} \perp\!\!\!\perp \epsilon_{js} \mid \mathbf{X}^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s$ |

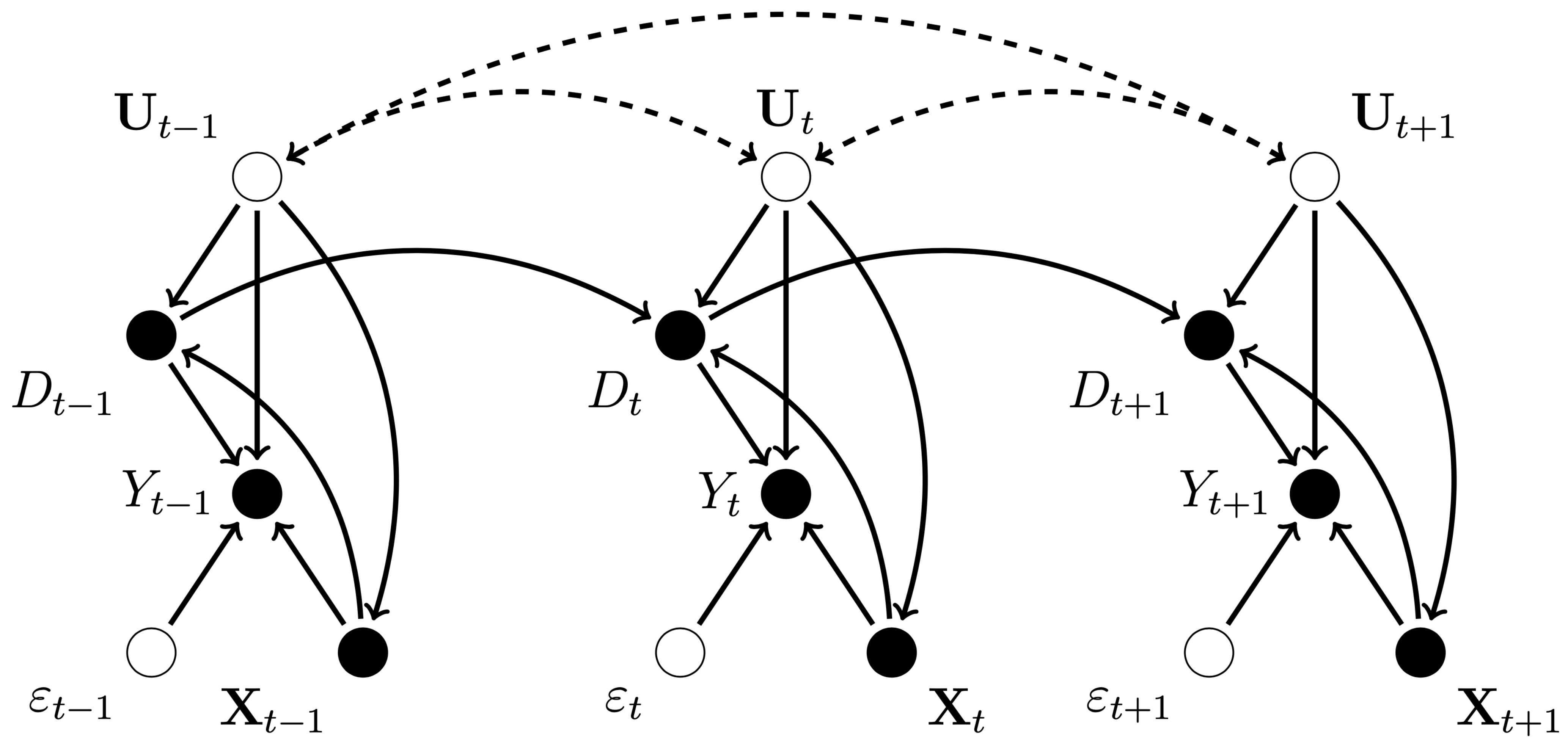
- On treatment assignment
  - Additive unobserved confounding
  - No "feedback"
- On interference (SUTVA)
  - No spatial spillover
  - No anticipation effects
  - No carryover effects
- On HTE
  - Constant treatment effect

**Related work:** Blackwell & Glynn (2018); Imai & Kim (2019); Athey & Imbens (2022); Liu, Wang & Xu (2022)

# What TWFE Assumptions Entail

| | |
|---|---|
| Functional Form | $Y_{it} = \delta^{TWFE} D_{it} + X'_{it}\beta + \alpha_i + \xi_t + \epsilon_{it}$ |
| Strict Exogeneity | $D_{it} \perp\!\!\!\perp \epsilon_{js} \mid \mathbf{X}^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s$ |

- On treatment assignment
  - Additive unobserved confounding
  - No "feedback"
- On interference (SUTVA)
  - No spatial spillover
  - No anticipation effects
  - No carryover effects
- On HTE
  - Constant treatment effect

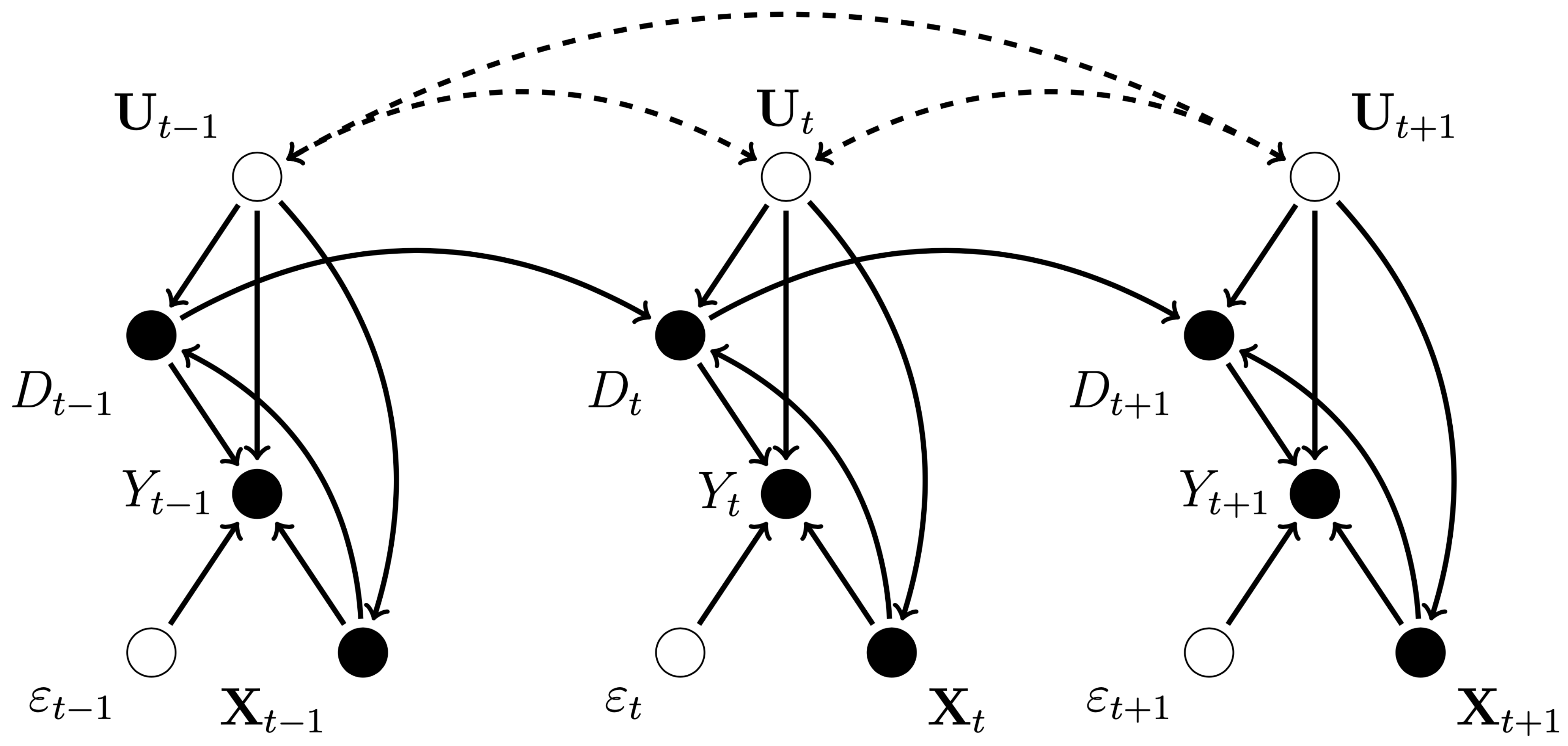



**Related work:** Blackwell & Glynn (2018); Imai & Kim (2019); Athey & Imbens (2022); Liu, Wang & Xu (2022)

# What TWFE Assumptions Entail

| | |
|---|---|
| Functional Form | $Y_{it} = \delta^{TWFE} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \epsilon_{it}$ |
| Strict Exogeneity | $D_{it} \perp\!\!\!\perp \epsilon_{js} \,|\, \mathbf{X}^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s$ |

- On treatment assignment
  - Additive unobserved confounding
  - No "feedback"
- On interference (SUTVA)
  - No spatial spillover
  - No anticipation effects
  - No carryover effects
- On HTE
  - Constant treatment effect

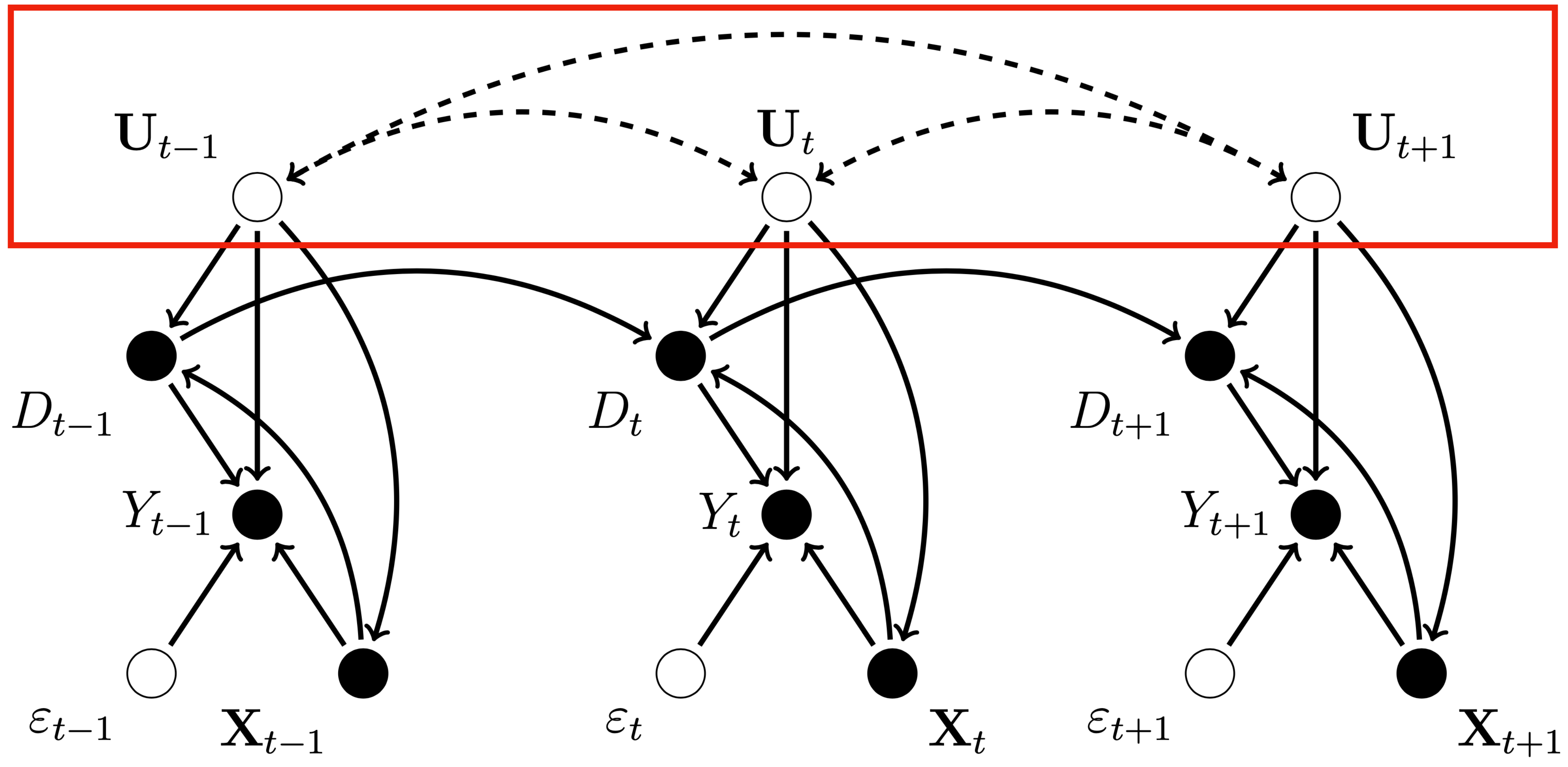



**Related work:** Blackwell & Glynn (2018); Imai & Kim (2019); Athey & Imbens (2022); Liu, Wang & Xu (2022)

# What TWFE Assumptions Entail

| | |
|---|---|
| Functional Form | $Y_{it} = \delta^{TWFE} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \epsilon_{it}$ |
| Strict Exogeneity | $D_{it} \perp\!\!\!\perp \epsilon_{js} \mid \mathbf{X}^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s$ |

- On treatment assignment
  - Additive unobserved confounding
  - No "feedback"
- On interference (SUTVA)
  - No spatial spillover
  - No anticipation effects
  - No carryover effects
- On HTE
  - Constant treatment effect




**Related work:** Blackwell & Glynn (2018); Imai & Kim (2019); Athey & Imbens (2022); Liu, Wang & Xu (2022)

# What TWFE Assumptions Entail

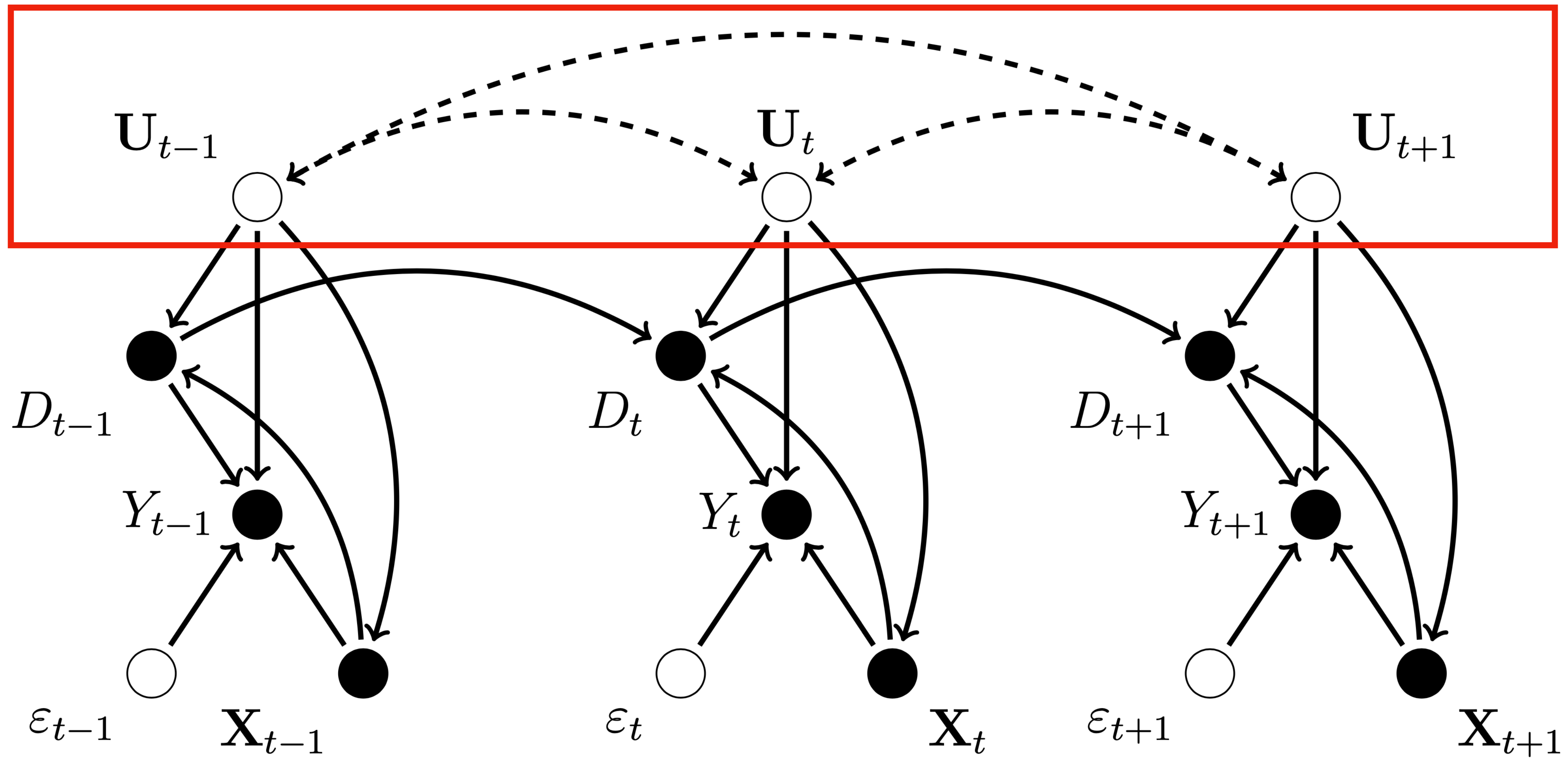
| | |
|---|---|
| Functional Form | $Y_{it} = \delta^{TWFE} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \epsilon_{it}$ |
| Strict Exogeneity | $D_{it} \perp\!\!\!\perp \epsilon_{js} \mid \mathbf{X}^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s$ |

$\Rightarrow$

**Parallel Trends (PT)**

$$\mathbb{E}[Y_{it}(0) - Y_{is}(0) \mid \Delta X_{i,ts} = x_0] = \mathbb{E}[Y_{jt}(0) - Y_{js}(0) \mid \Delta X_{j,ts} = x_0] \quad \forall i, j, t, s$$

- On treatment assignment
  - Additive unobserved confounding
  - No "feedback"
- On interference (SUTVA)
  - No spatial spillover
  - No anticipation effects
  - No carryover effects
- On HTE
  - Constant treatment effect




**Related work:** Blackwell & Glynn (2018); Imai & Kim (2019); Athey & Imbens (2022); Liu, Wang & Xu (2022)

# What TWFE Assumptions Entail

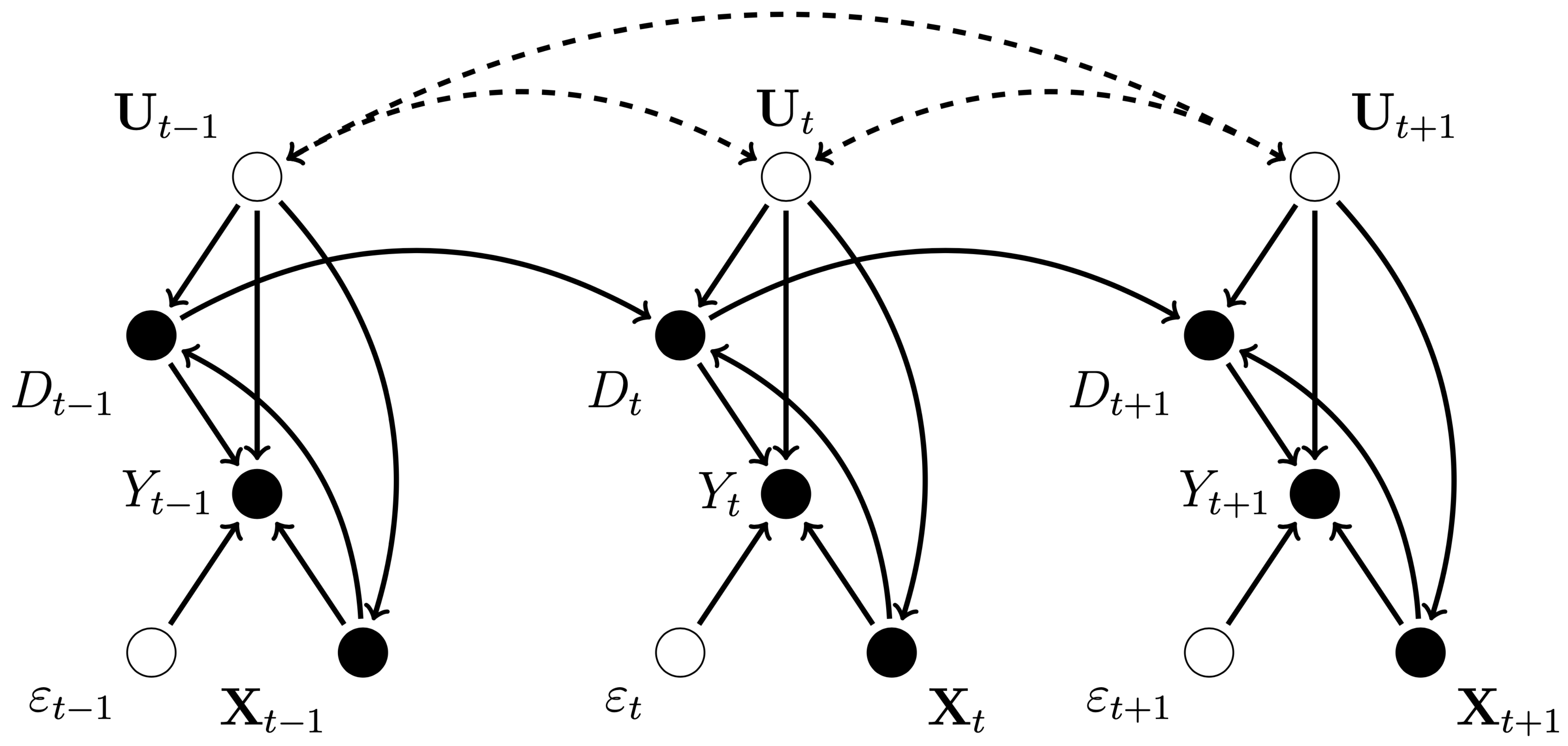
| | |
|---|---|
| Functional Form | $Y_{it} = \delta^{TWFE} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \epsilon_{it}$ |
| Strict Exogeneity | $D_{it} \perp\!\!\!\perp \epsilon_{js} \mid \mathbf{X}^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s,$ |

$\Rightarrow$

Parallel Trends (PT)

$$\mathbb{E}[Y_{it}(0) - Y_{is}(0) \mid \Delta X_{i,ts} = x_0] = \mathbb{E}[Y_{jt}(0) - Y_{js}(0) \mid \Delta X_{j,ts} = x_0] \quad \forall i, j, t, s$$

- On treatment assignment
  - Additive unobserved confounding
  - No "feedback"
- On interference (SUTVA)
  - No spatial spillover
  - No anticipation effects
  - No carryover effects
- On HTE
  - Constant treatment effect




**Related work:** Blackwell & Glynn (2018); Imai & Kim (2019); Athey & Imbens (2022); Liu, Wang & Xu (2022)

# What TWFE Assumptions Entail

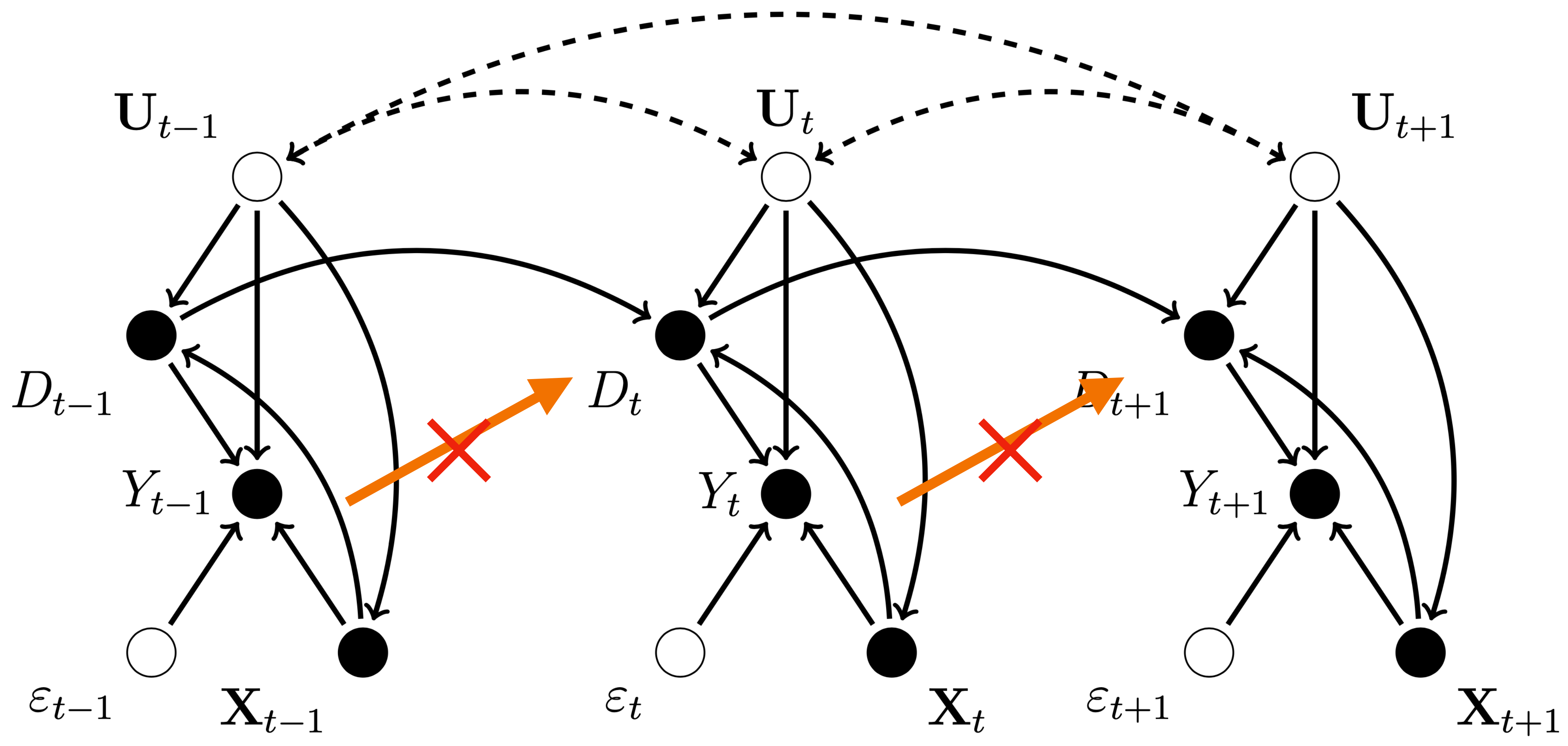
| | |
|---|---|
| Functional Form | $Y_{it} = \delta^{TWFE} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \epsilon_{it}$ |
| Strict Exogeneity | $D_{it} \perp\!\!\!\perp \epsilon_{js} \mid \mathbf{X}^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s,$ |

$\Rightarrow$

Parallel Trends (PT)

$$\mathbb{E}[Y_{it}(0) - Y_{is}(0) \mid \Delta X_{i,ts} = x_0] = \mathbb{E}[Y_{jt}(0) - Y_{js}(0) \mid \Delta X_{j,ts} = x_0] \quad \forall i, j, t, s$$

- On treatment assignment
  - Additive unobserved confounding
  - No "feedback"
- On interference (SUTVA)
  - No spatial spillover
  - No anticipation effects
  - No carryover effects
- On HTE
  - Constant treatment effect




**Related work:** Blackwell & Glynn (2018); Imai & Kim (2019); Athey & Imbens (2022); Liu, Wang & Xu (2022)

# What TWFE Assumptions Entail

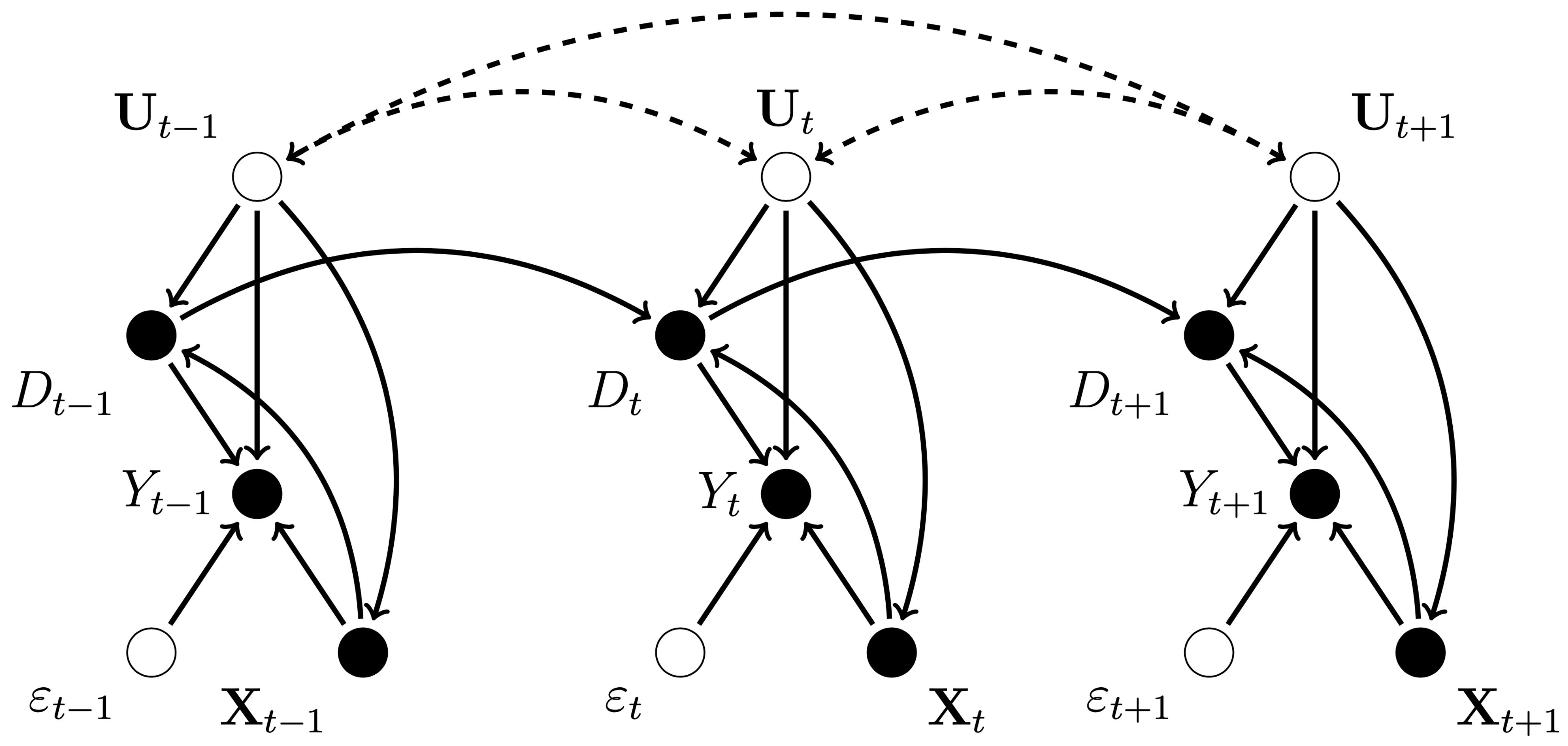
| | |
|---|---|
| Functional Form | $Y_{it} = \delta^{TWFE} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \epsilon_{it}$ |
| Strict Exogeneity | $D_{it} \perp\!\!\!\perp \epsilon_{js} \mid \mathbf{X}^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s$ |

$\Rightarrow$

Parallel Trends (PT)

$$\mathbb{E}[Y_{it}(0) - Y_{is}(0) \mid \Delta X_{i,ts} = x_0] = \mathbb{E}[Y_{jt}(0) - Y_{js}(0) \mid \Delta X_{j,ts} = x_0] \quad \forall i, j, t, s$$

- On treatment assignment
  - Additive unobserved confounding
  - No "feedback"
- On interference (SUTVA)
  - No spatial spillover
  - No anticipation effects
  - No carryover effects
- On HTE
  - Constant treatment effect




**Related work:** Blackwell & Glynn (2018); Imai & Kim (2019); Athey & Imbens (2022); Liu, Wang & Xu (2022)

# What TWFE Assumptions Entail

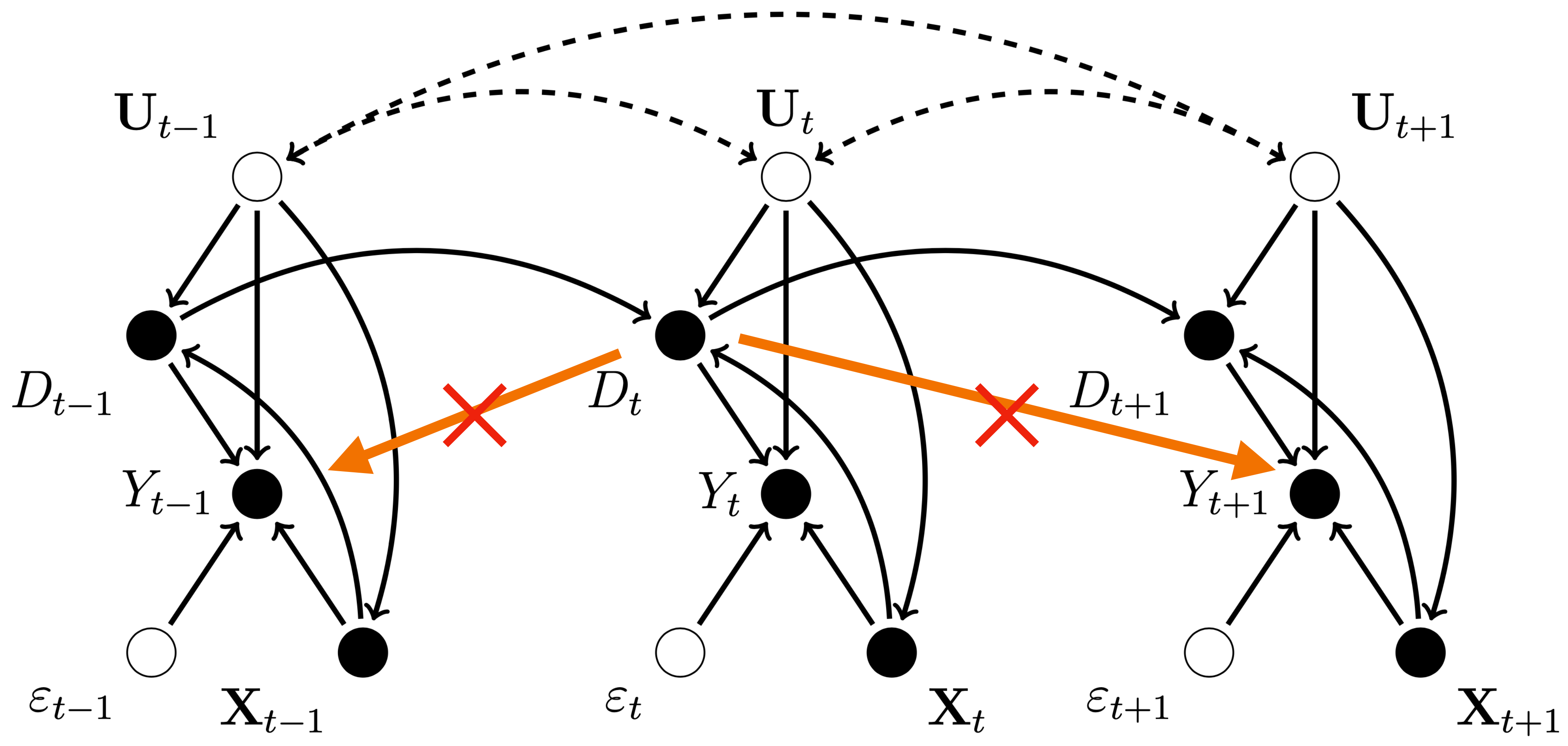
| | |
|---|---|
| Functional Form | $Y_{it} = \delta^{TWFE} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \epsilon_{it}$ |
| Strict Exogeneity | $D_{it} \perp\!\!\!\perp \epsilon_{js} \mid \mathbf{X}^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s$ |

$\Rightarrow$

Parallel Trends (PT)

$$\mathbb{E}[Y_{it}(0) - Y_{is}(0) \mid \Delta X_{i,ts} = x_0] = \mathbb{E}[Y_{jt}(0) - Y_{js}(0) \mid \Delta X_{j,ts} = x_0] \quad \forall i, j, t, s$$

- On treatment assignment
  - Additive unobserved confounding
  - No "feedback"
- On interference (SUTVA)
  - No spatial spillover
  - No anticipation effects
  - No carryover effects
- On HTE
  - Constant treatment effect




**Related work:** Blackwell & Glynn (2018); Imai & Kim (2019); Athey & Imbens (2022); Liu, Wang & Xu (2022)

# What TWFE Assumptions Entail

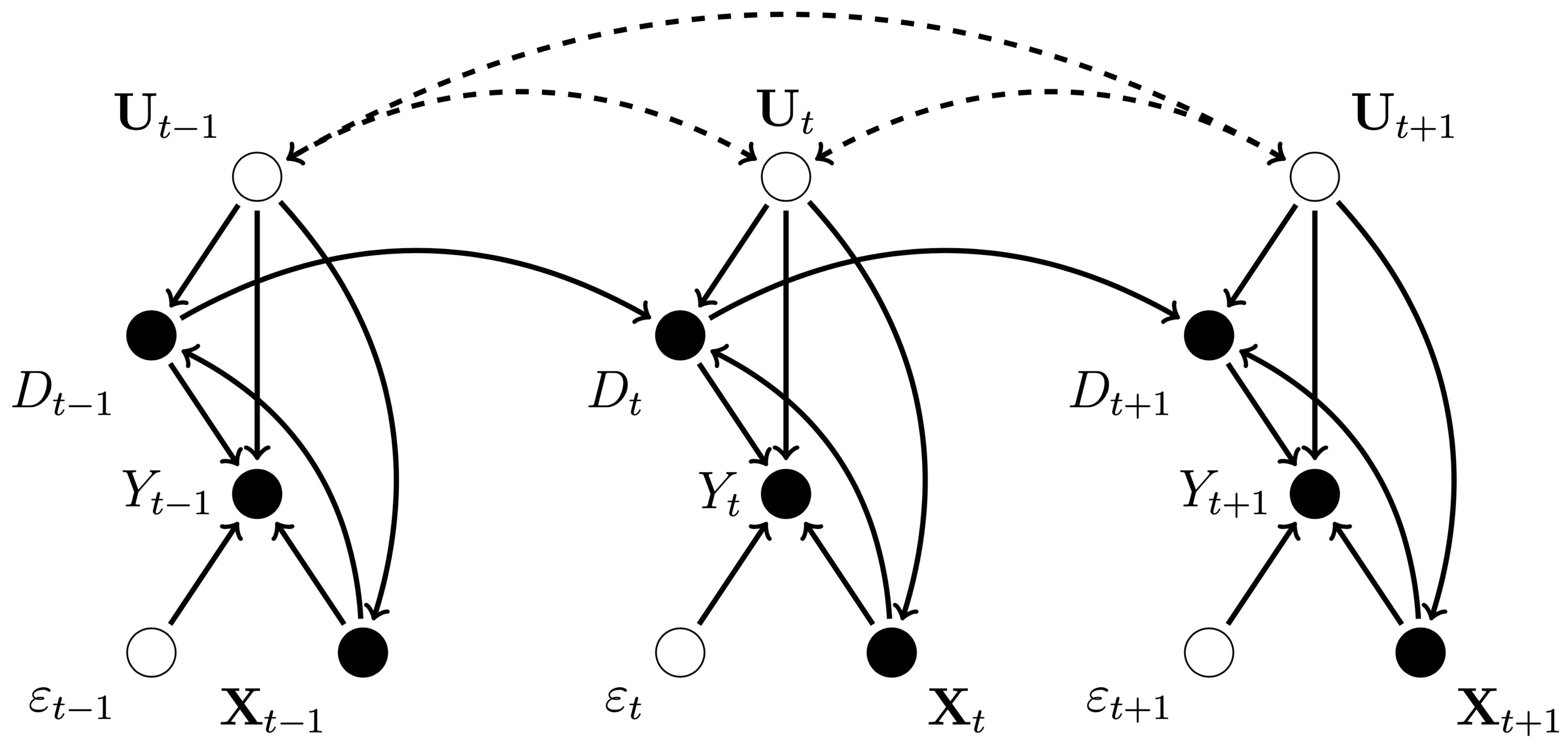
| | |
|---|---|
| Functional Form | $Y_{it} = \delta^{TWFE} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \epsilon_{it}$ |
| Strict Exogeneity | $D_{it} \perp\!\!\!\perp \epsilon_{js} \mid \mathbf{X}^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s$ |

$\Rightarrow$

Parallel Trends (PT)

$$\mathbb{E}[Y_{it}(0) - Y_{is}(0) \mid \Delta X_{i,ts} = x_0] = \mathbb{E}[Y_{jt}(0) - Y_{js}(0) \mid \Delta X_{j,ts} = x_0] \quad \forall i, j, t, s$$

- On treatment assignment
  - Additive unobserved confounding
  - No "feedback"
- On interference (SUTVA)
  - No spatial spillover
  - No anticipation effects
  - No carryover effects (can be relaxed)
- On HTE
  - Constant treatment effect (more to follow)




**Related work:** Blackwell & Glynn (2018); Imai & Kim (2019); Athey & Imbens (2022); Liu, Wang & Xu (2022)

# Review: The Consequence of HTE (Goodman-Bacon 2021 and others)

# Review: The Consequence of HTE (Goodman-Bacon 2021 and others)

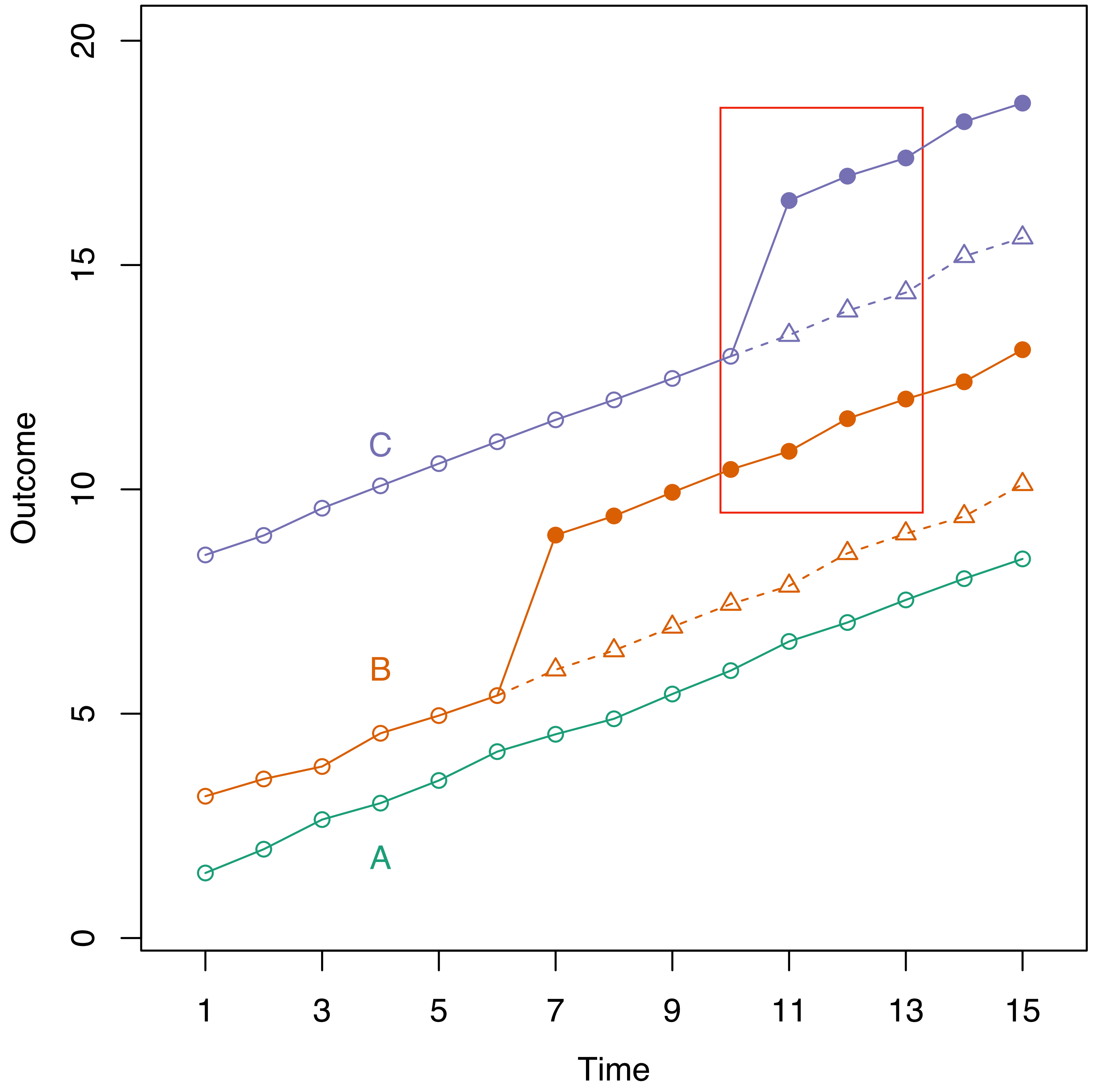
- **Result:** With HTE, TWFE cannot always arrive at some convex combination of individualistic treatment effect when the PT is valid

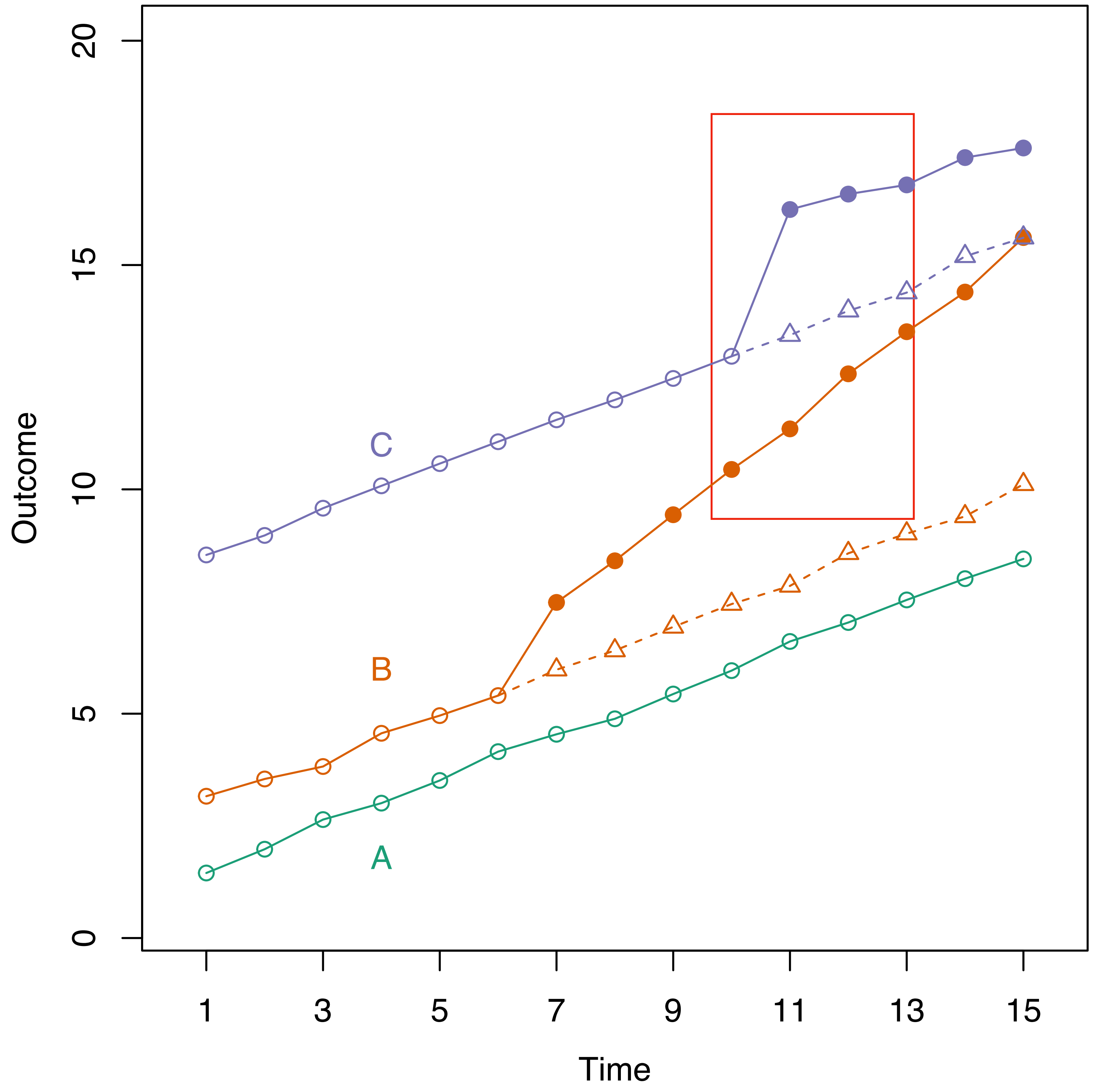
# Review: The Consequence of HTE (Goodman-Bacon 2021 and others)

- **Result:** With HTE, TWFE cannot always arrive at some convex combination of individualistic treatment effect when the PT is valid
- **Intuition:** Treated observations of early adopters serve as controls for treated observations of late adopters, or "forbidden comparison"

# Review: The Consequence of HTE (Goodman-Bacon 2021 and others)

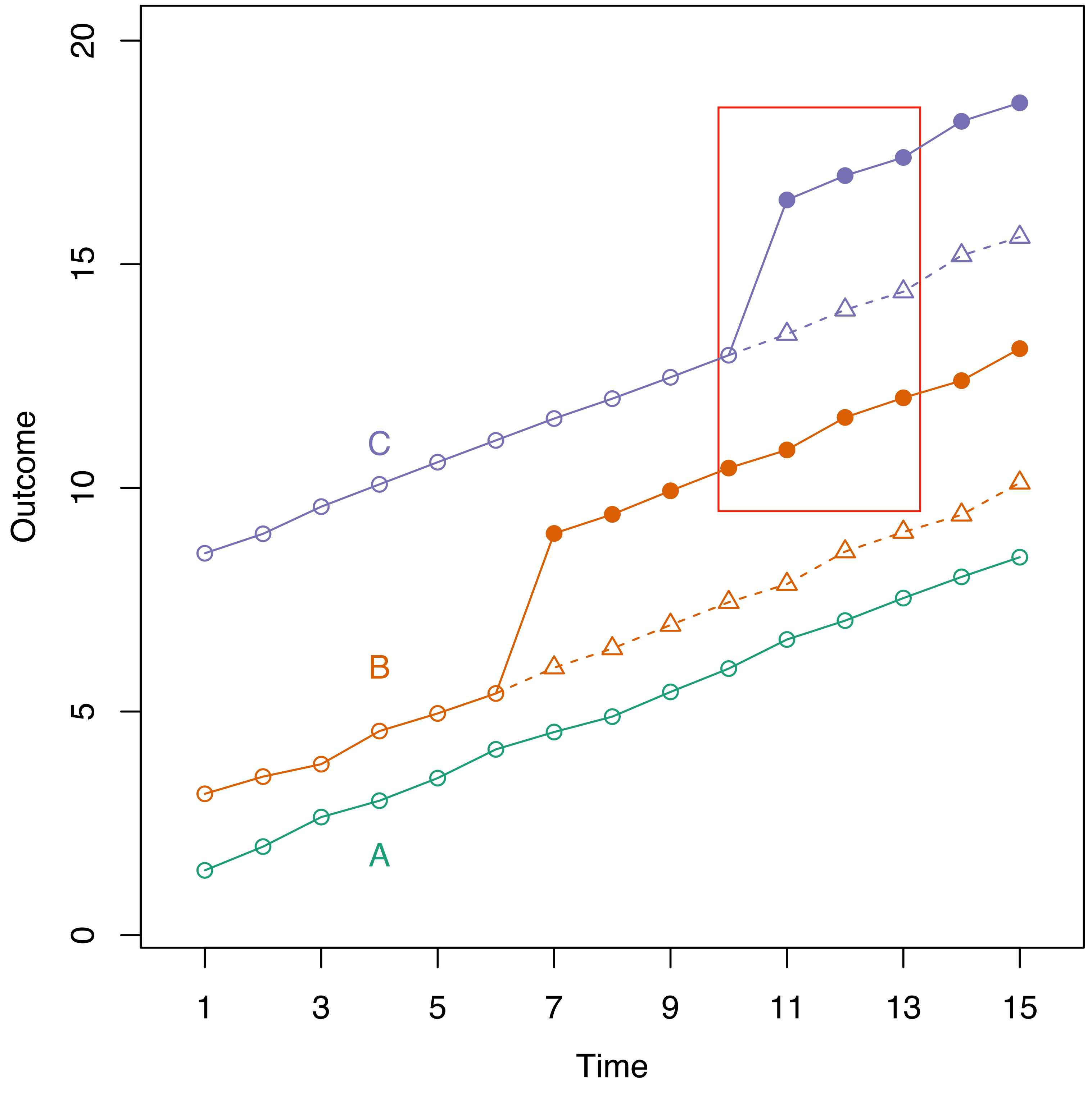
- **Result:** With HTE, TWFE <u>cannot always</u> arrive at some <u>convex combination</u> of individualistic treatment effect <u>when the PT is valid</u>
- **Intuition:** Treated observations of early adopters serve as controls for treated observations of late adopters, or "forbidden comparison"

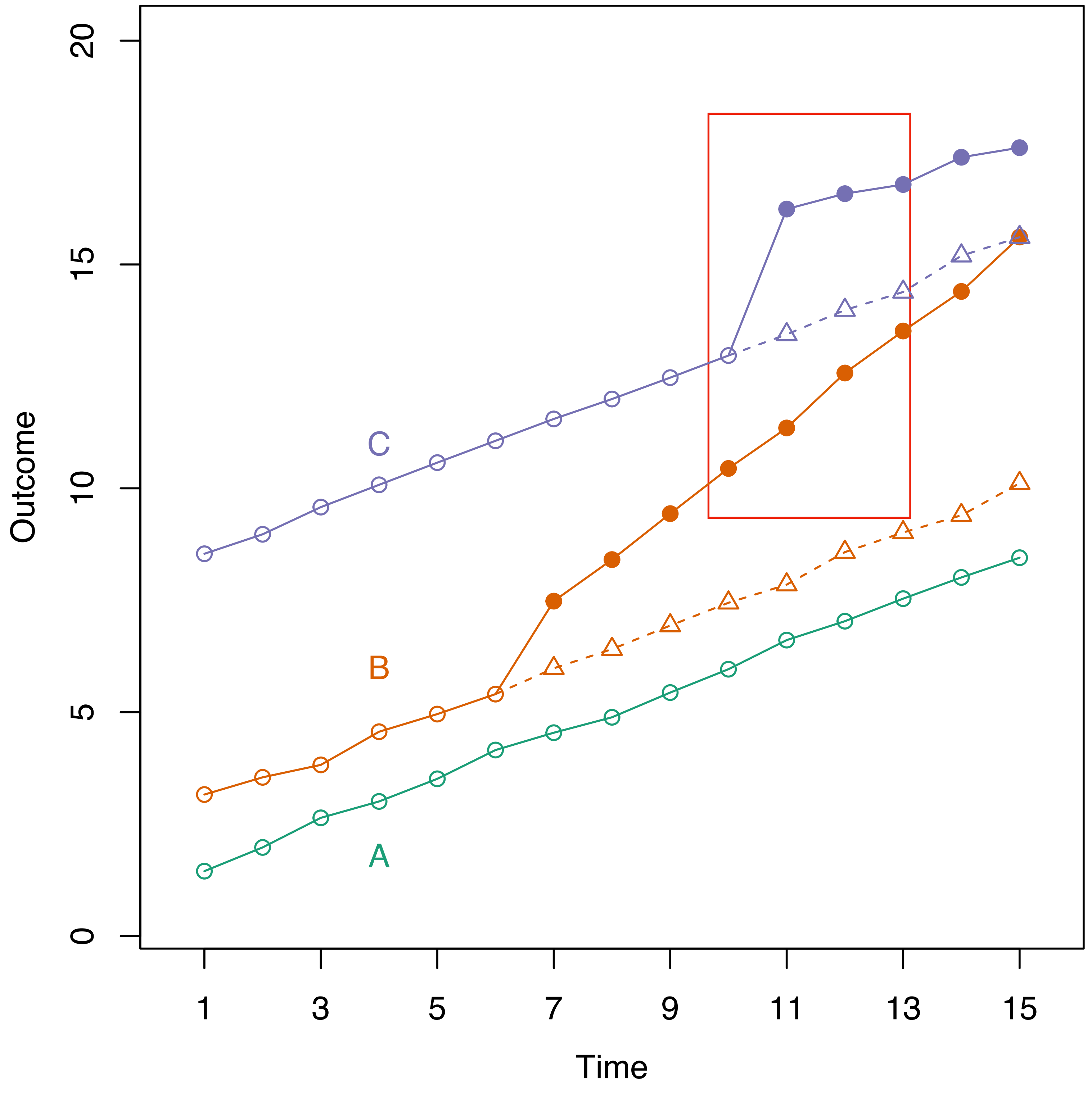


# Review: The Consequence of HTE (Goodman-Bacon 2021 and others)

- **Result:** With HTE, TWFE <u>cannot always</u> arrive at some <u>convex combination</u> of individualistic treatment effect <u>when the PT is valid</u>
- **Intuition:** Treated observations of early adopters serve as controls for treated observations of late adopters, or "forbidden comparison"
- **Complexity:** How important this issue is depends on many factors



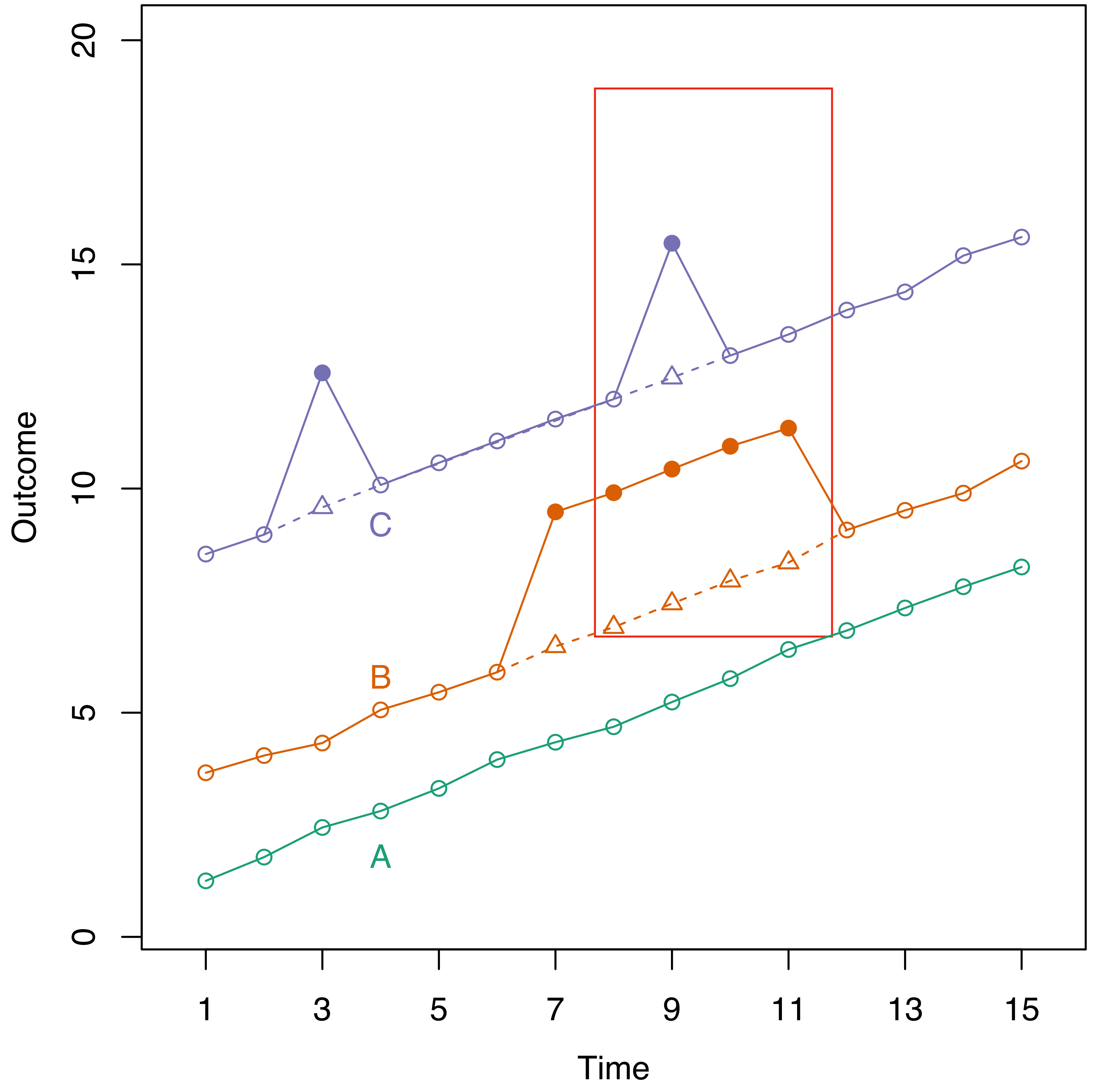
# HTE Compounded with Other Issues...

# HTE Compounded with Other Issues...

- Treatment reversal (majority of PoliSci studies)

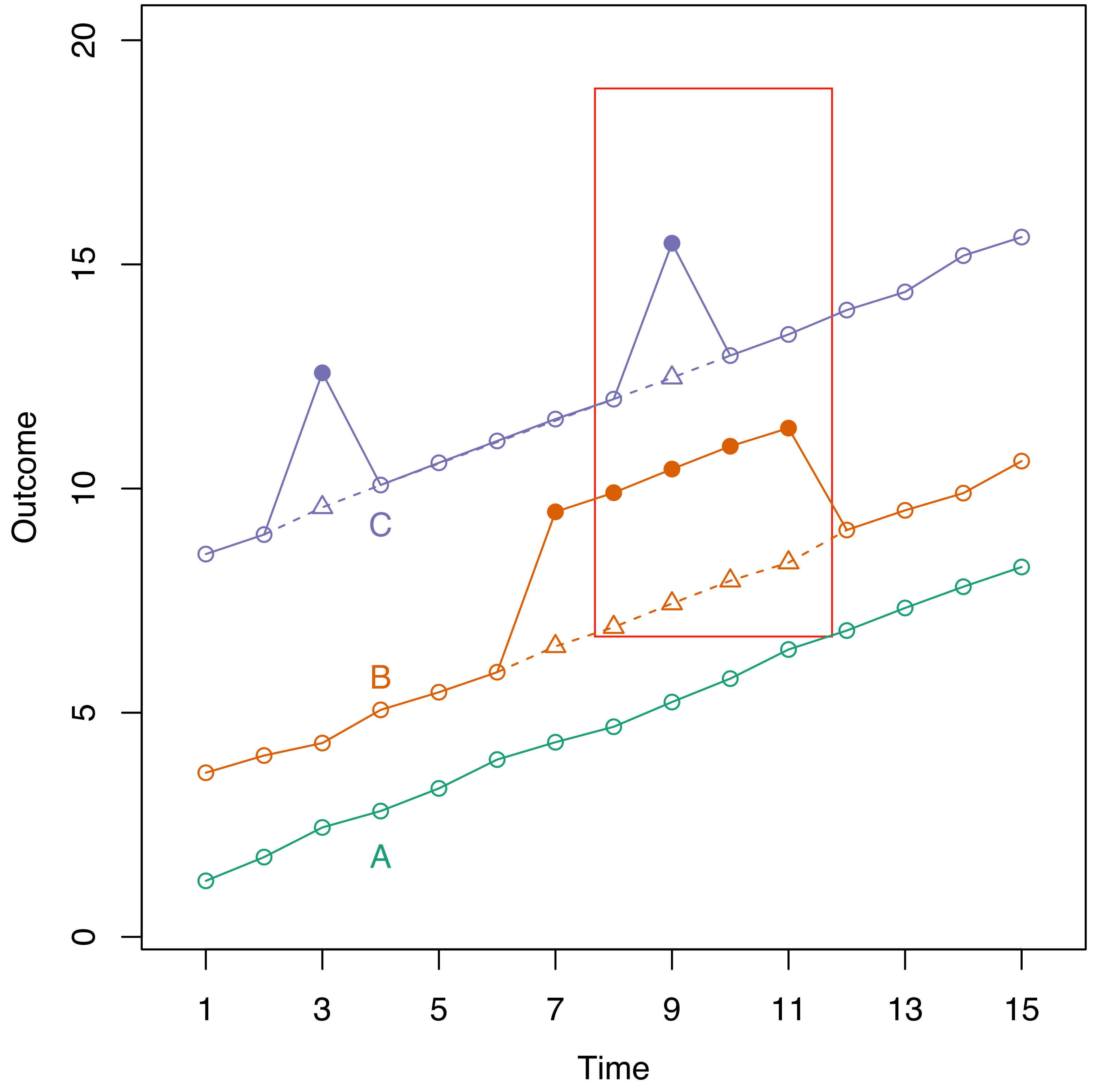
# HTE Compounded with Other Issues...

- Treatment reversal (majority of PoliSci studies)



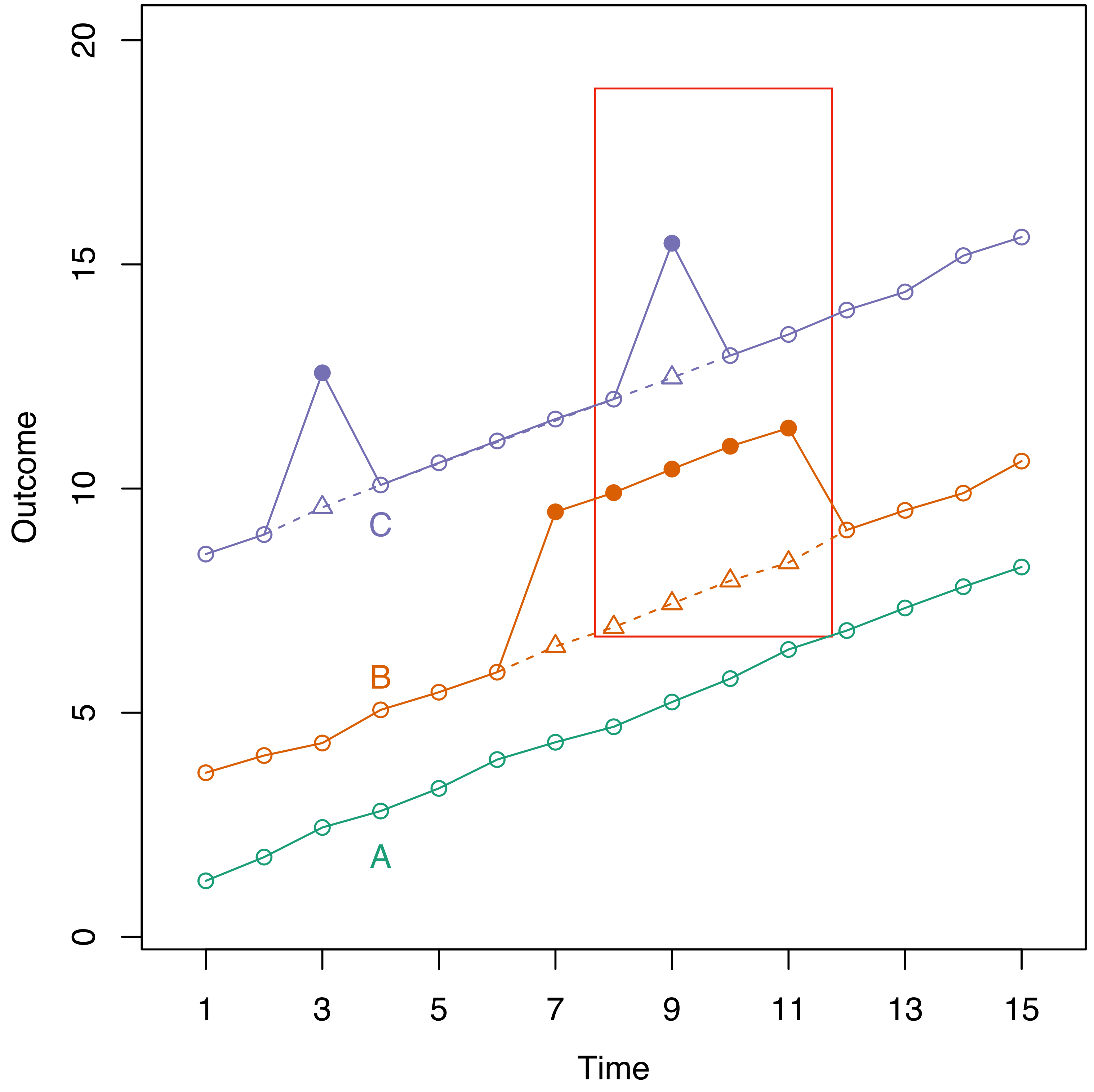
# HTE Compounded with Other Issues...

- Treatment reversal (majority of PoliSci studies)
- PT violations



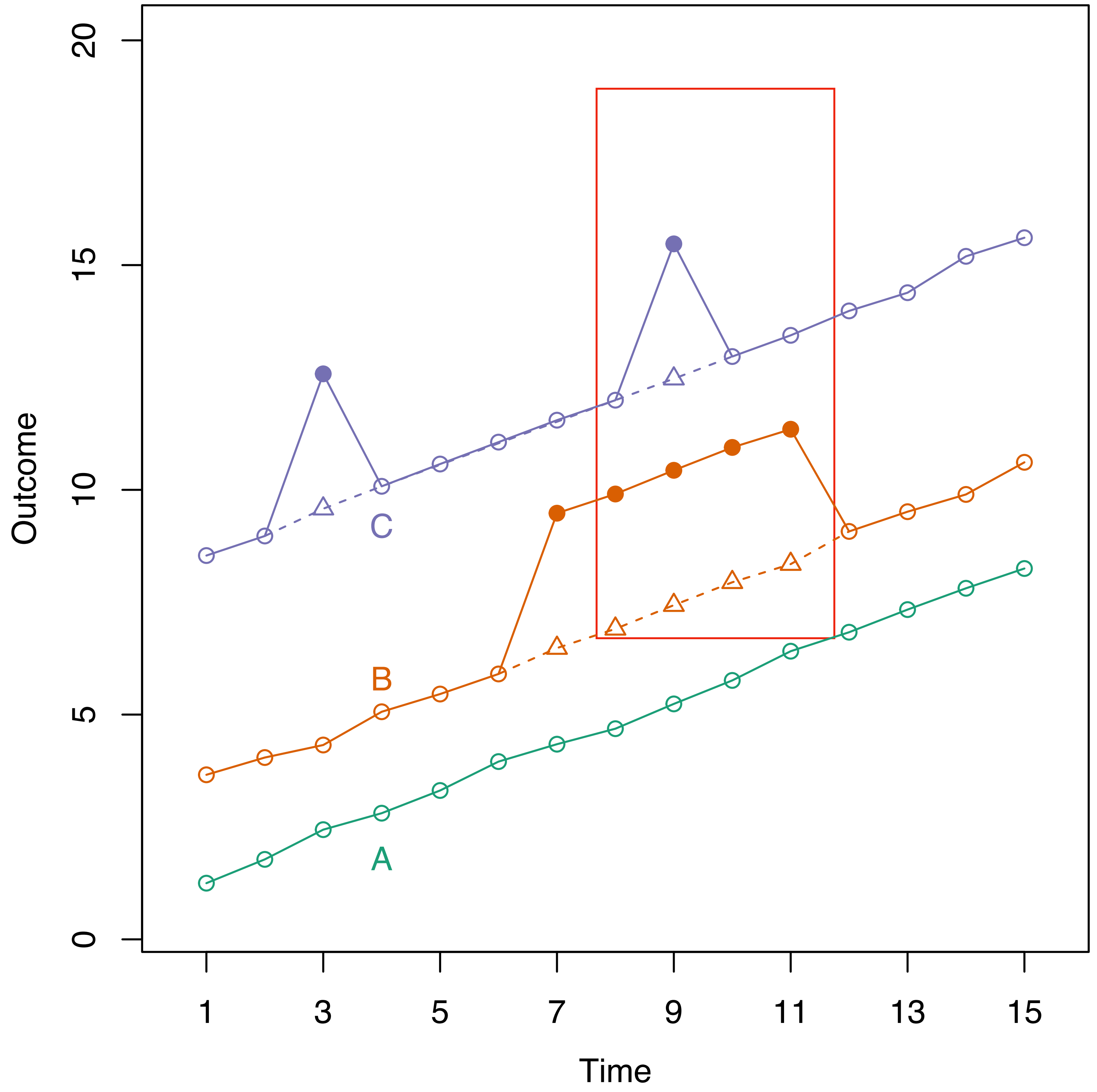
# HTE Compounded with Other Issues...

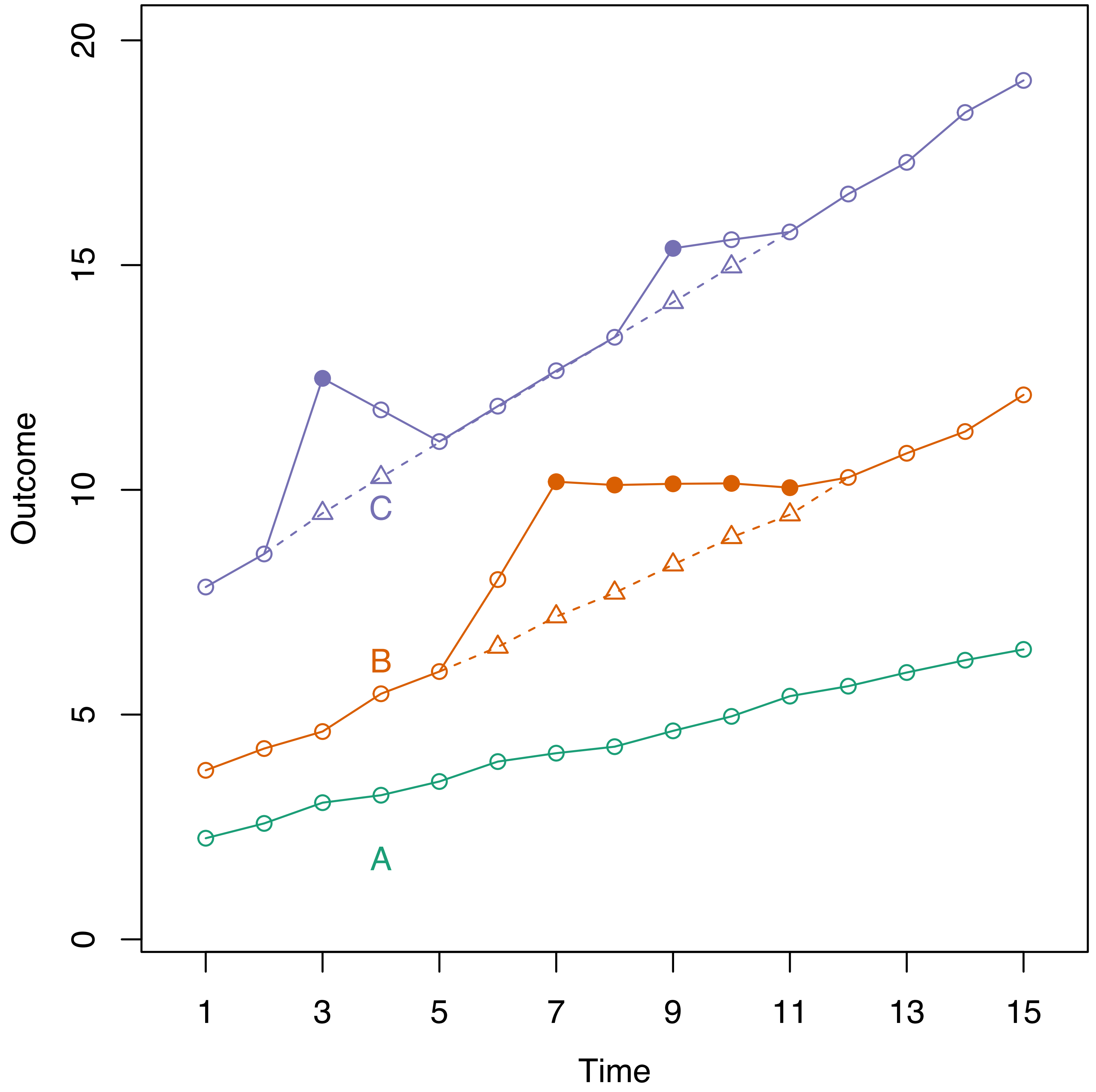
- Treatment reversal (majority of PoliSci studies)
- PT violations
- Anticipation and carryover effects



# HTE Compounded with Other Issues...

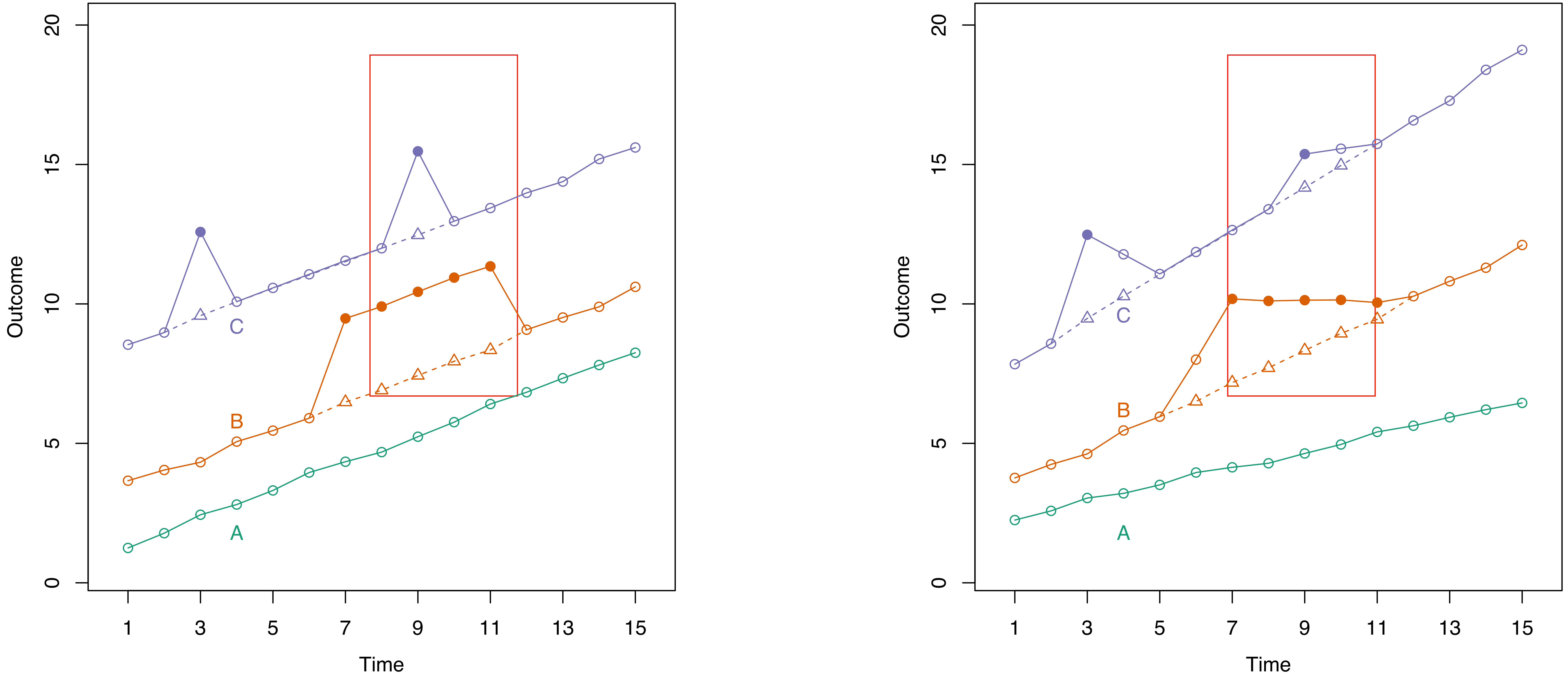
- Treatment reversal (majority of PoliSci studies)
- PT violations
- Anticipation and carryover effects



# HTE Compounded with Other Issues...

- Treatment reversal (majority of PoliSci studies)
- PT violations
- Anticipation and carryover effects



# This Project

# This Project

- Widespread confusion

# This Project

- Widespread confusion
  - Are existing results based on TWFE regressions reliable?

# This Project

- Widespread confusion
  - Are existing results based on TWFE regressions reliable?
  - With so many options, what's the current best practice?

# This Project

- Widespread confusion
  - Are existing results based on TWFE regressions reliable?
  - With so many options, what's the current best practice?
  - What are the main challenges of conducting causal panel analysis under parallel trends?

# This Project

- Widespread confusion
  - Are existing results based on TWFE regressions reliable?
  - With so many options, what's the current best practice?
  - What are the main challenges of conducting causal panel analysis under parallel trends?
- What we do

# This Project

- Widespread confusion
  - Are existing results based on TWFE regressions reliable?
  - With so many options, what's the current best practice?
  - What are the main challenges of conducting causal panel analysis under parallel trends?
- What we do
  - Replicated a main result of **49** top publications in a **seven-year** span (2017-2023)

# This Project

- Widespread confusion
  - Are existing results based on TWFE regressions reliable?
  - With so many options, what's the current best practice?
  - What are the main challenges of conducting causal panel analysis under parallel trends?

- What we do
  - Replicated a main result of **49** top publications in a **seven-year** span (2017-2023)
  - Standardize tools and reanalyze these findings using a large set of new methods

# This Project

- Widespread confusion
  - Are existing results based on TWFE regressions reliable?
  - With so many options, what's the current best practice?
  - What are the main challenges of conducting causal panel analysis under parallel trends?

- What we do
  - Replicated a main result of **49** top publications in a **seven-year** span (2017-2023)
  - Standardize tools and reanalyze these findings using a large set of new methods
  - Provide recommendations to improve practice

# This Project

- Widespread confusion
  - Are existing results based on TWFE regressions reliable?
  - With so many options, what's the current best practice?
  - What are the main challenges of conducting causal panel analysis under parallel trends?
- What we do
  - Replicated a main result of **49** top publications in a **seven-year** span (2017-2023)
  - Standardize tools and reanalyze these findings using a large set of new methods
  - Provide recommendations to improve practice
- Why large scale replication/reanalysis?

# This Project

- Widespread confusion
  - Are existing results based on TWFE regressions reliable?
  - With so many options, what's the current best practice?
  - What are the main challenges of conducting causal panel analysis under parallel trends?
- What we do
  - Replicated a main result of **49** top publications in a **seven-year** span (2017-2023)
  - Standardize tools and reanalyze these findings using a large set of new methods
  - Provide recommendations to improve practice
- Why large scale replication/reanalysis?
  - To understand the relevance of theoretical findings and the challenges in implementing changes

# This Project

- Widespread confusion
  - Are existing results based on TWFE regressions reliable?
  - With so many options, what's the current best practice?
  - What are the main challenges of conducting causal panel analysis under parallel trends?

- What we do
  - Replicated a main result of **49** top publications in a **seven-year** span (2017-2023)
  - Standardize tools and reanalyze these findings using a large set of new methods
  - Provide recommendations to improve practice

- Why large scale replication/reanalysis?
  - To understand the relevance of theoretical findings and the challenges in implementing changes
  - To identify researchers' needs and improve scientific practices

# Preview of Findings

# Preview of Findings

Common practice?

# Preview of Findings

## Common practice?

- FE (77%), including TWFE (58%)

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping

# Preview of Findings

## Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

# Preview of Findings

## Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

### Do results hold up?

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

### Do results hold up?

Yes and No

# Preview of Findings

## Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

## Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs

# Preview of Findings

## Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

## Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

### Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common
- No — Insufficient power when HTE-robust estimators used

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

### Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common
- No — Insufficient power when HTE-robust estimators used
- No — Few studies survive mild sensitivity analyses

# Preview of Findings

## Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

## Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common
- No — Insufficient power when HTE-robust estimators used
- No — Few studies survive mild sensitivity analyses

Strong empirical support for $<1/3$ of the findings

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

### Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common
- No — Insufficient power when HTE-robust estimators used
- No — Few studies survive mild sensitivity analyses

Strong empirical support for <1/3 of the findings

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

### Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common
- No — Insufficient power when HTE-robust estimators used
- No — Few studies survive mild sensitivity analyses

Strong empirical support for <1/3 of the findings

### Takeaways

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

### Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common
- No — Insufficient power when HTE-robust estimators used
- No — Few studies survive mild sensitivity analyses

Strong empirical support for <1/3 of the findings

### Takeaways

- PT (& research design) is a first-order issue

# Preview of Findings

## Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

## Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common
- No — Insufficient power when HTE-robust estimators used
- No — Few studies survive mild sensitivity analyses

Strong empirical support for $<1/3$ of the findings

## Takeaways

- PT (& research design) is a first-order issue
- Concerns over HTE is valid but seems second-order

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

### Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common
- No — Insufficient power when HTE-robust estimators used
- No — Few studies survive mild sensitivity analyses

Strong empirical support for <1/3 of the findings

### Takeaways

- PT (& research design) is a first-order issue
- Concerns over HTE is valid but seems second-order
- Validation is the key:

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

### Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common
- No — Insufficient power when HTE-robust estimators used
- No — Few studies survive mild sensitivity analyses

Strong empirical support for <1/3 of the findings

### Takeaways

- PT (& research design) is a first-order issue
- Concerns over HTE is valid but seems second-order
- Validation is the key:
  - Event-study plots are a minimal requirement

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

### Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common
- No — Insufficient power when HTE-robust estimators used
- No — Few studies survive mild sensitivity analyses

Strong empirical support for <1/3 of the findings

### Takeaways

- PT (& research design) is a first-order issue
- Concerns over HTE is valid but seems second-order
- Validation is the key:
  - Event-study plots are a minimal requirement
  - Sensitivity analysis is helpful

# Preview of Findings

### Common practice?

- FE (77%), including TWFE (58%)
- Cluster-robust SE (98%); few use bootstrapping
- 59% with some graphic inspections

### Do results hold up?

Yes and No

- Yes — HTE-robust estimators rarely flip signs
- No — PT violations still common
- No — Insufficient power when HTE-robust estimators used
- No — Few studies survive mild sensitivity analyses

Strong empirical support for <1/3 of the findings

### Takeaways

- PT (& research design) is a first-order issue
- Concerns over HTE is valid but seems second-order
- Validation is the key:
  - Event-study plots are a minimal requirement
  - Sensitivity analysis is helpful
- "Robust" DID requires a strong design and a lot of power

# Related Literature

- Review articles: Roth et al. (2023), Xu (2023), Arkhangelsky and Imbens (2023)
  - New diagnostic and estimation strategies not applied to data
  - Difficult to assess their relevance to empirical research

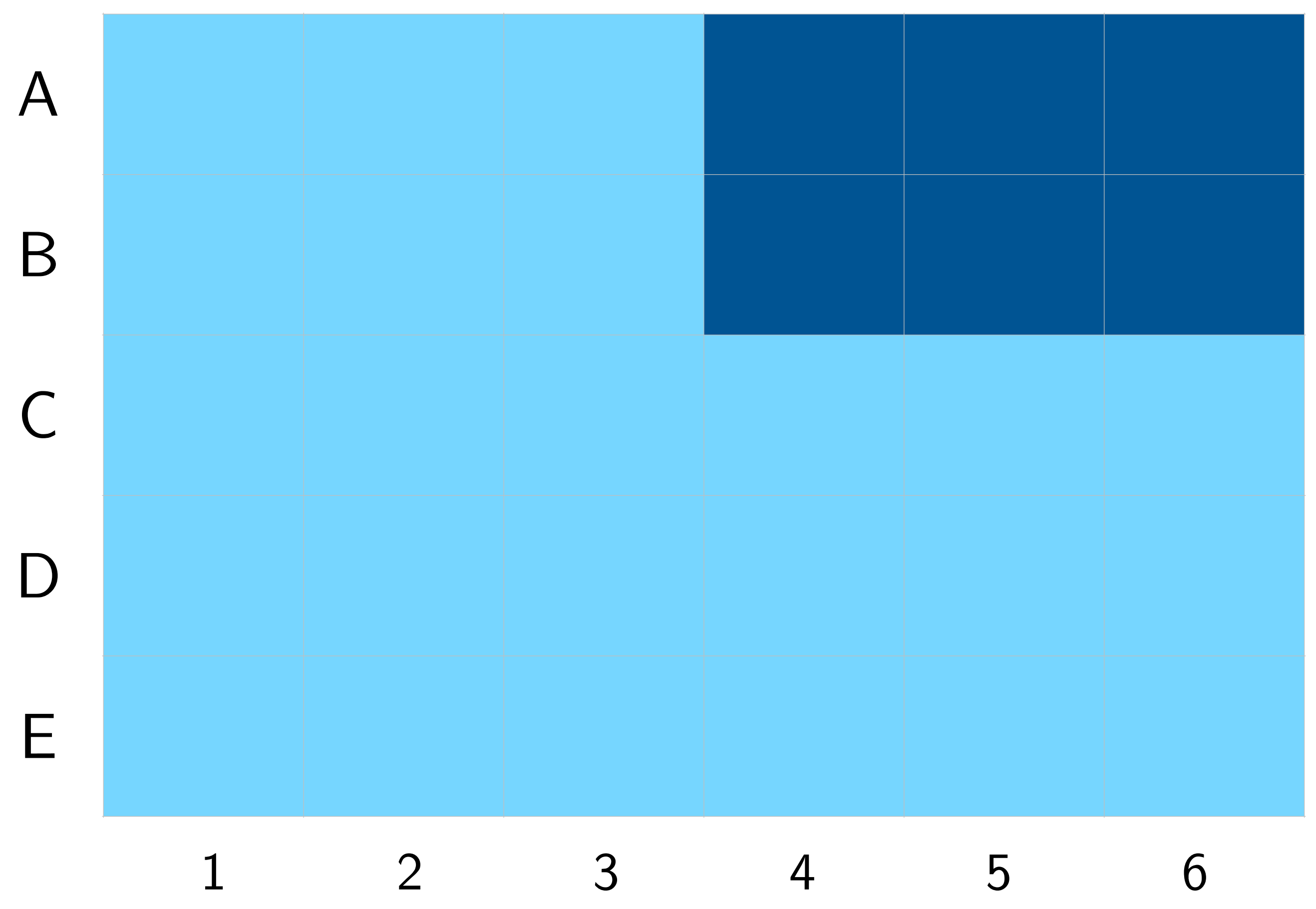
- Replication studies: Baker et al. (2022)
  - Replicated five economics and finance studies with staggered treatments
  - Focused on the consequence of HTE

# [roadmap]

- Estimators
  - Review 6 HTE-robust estimators
  - Typology & comparison
- Data and Procedure
  - Sample
  - Procedure
- Findings
  - Three examples
  - Overall assessment
- Recommendations

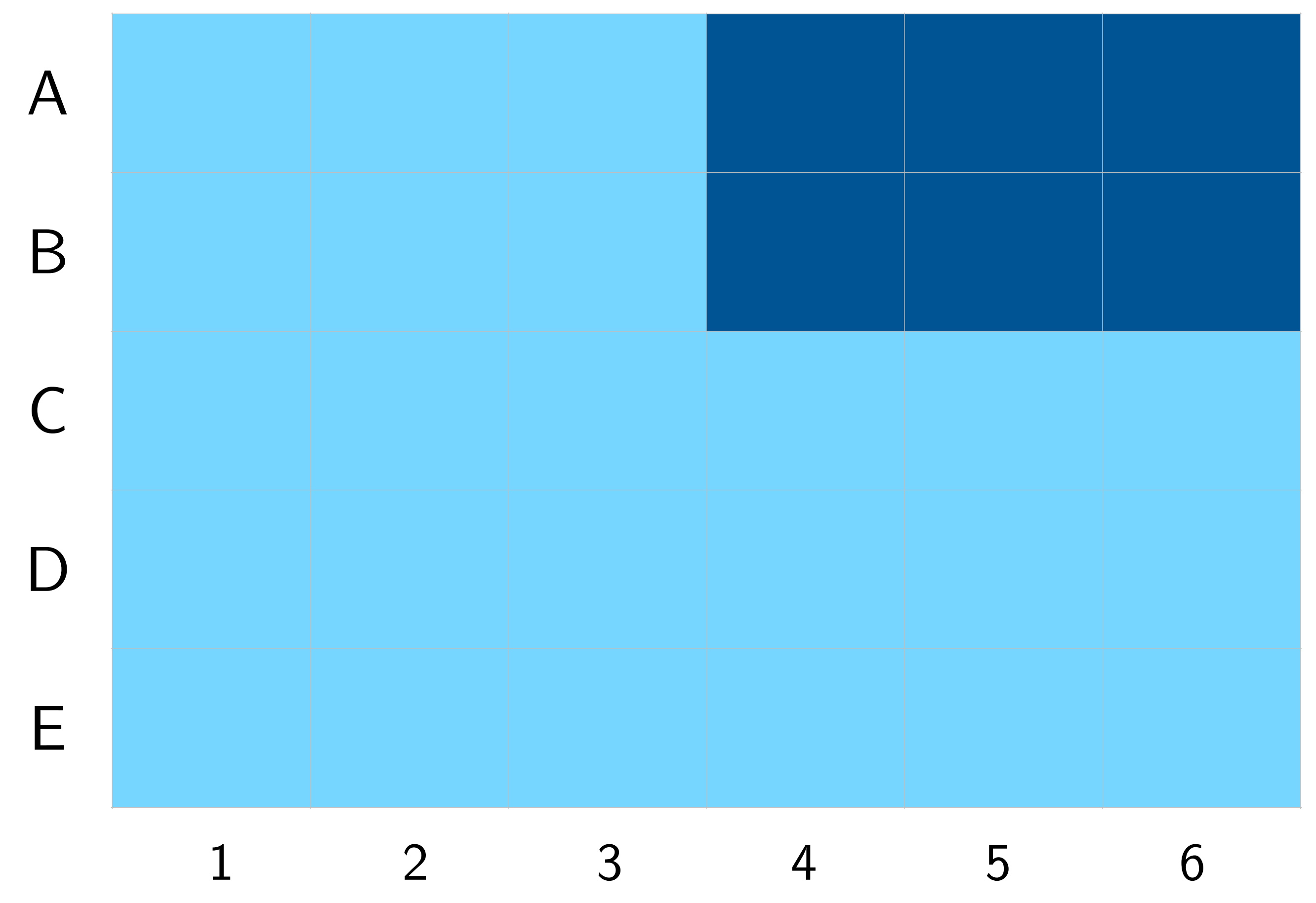
# Methods

# Settings

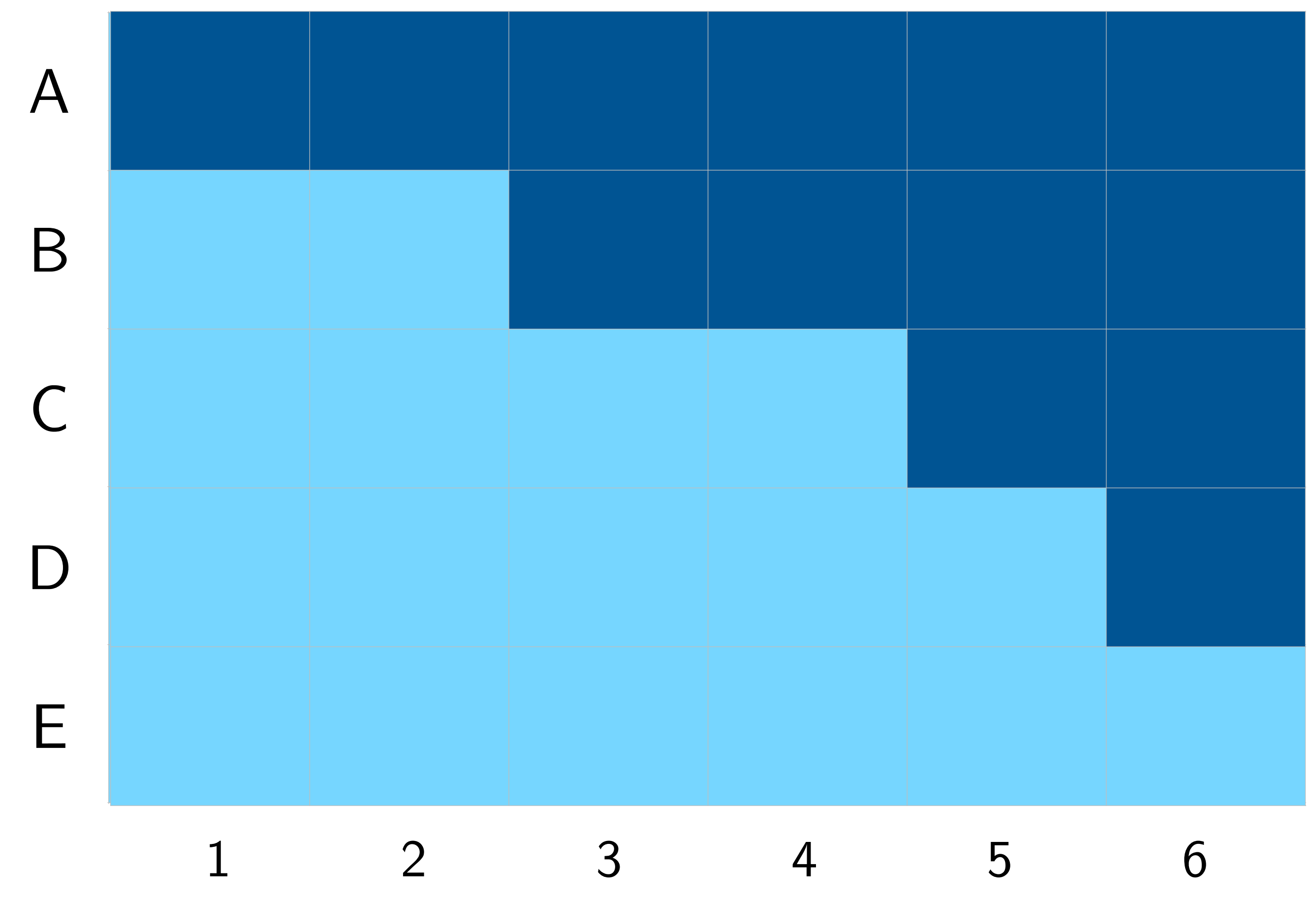



(Multi-Period) Block DID Setting

# Settings

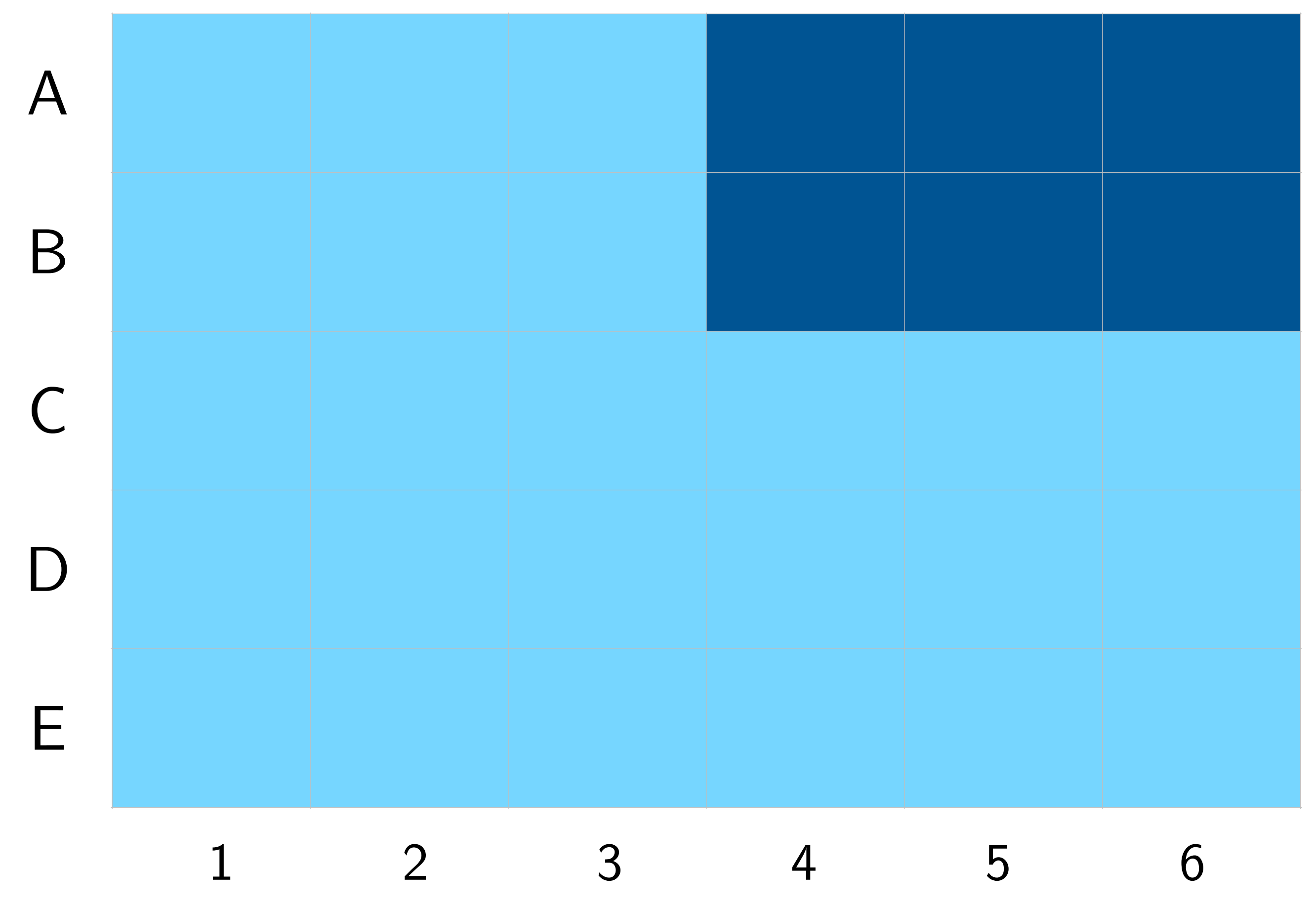



(Multi-Period) Block DID Setting

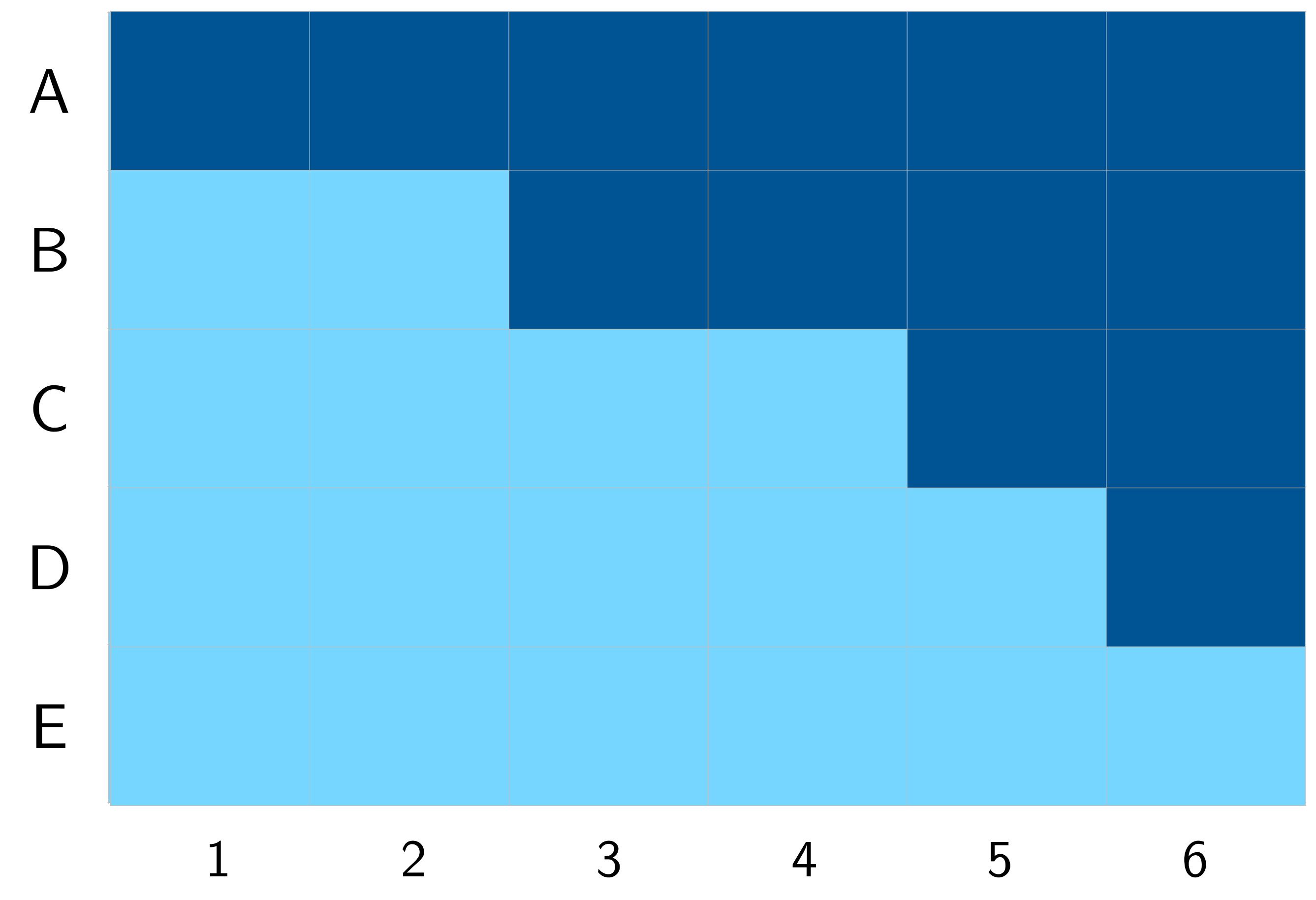



Staggered DID Setting

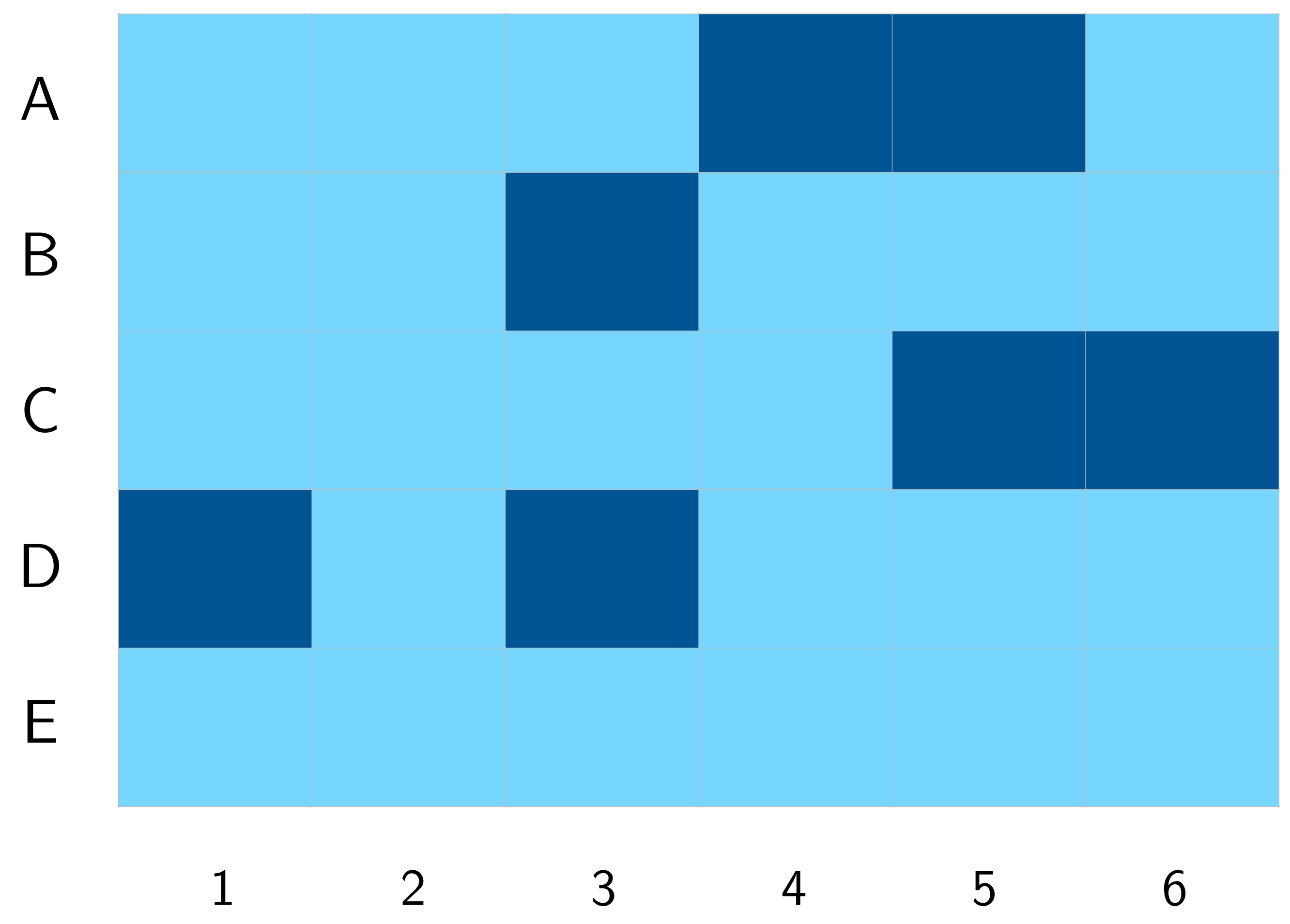
# Settings



(Multi-Period) Block DID Setting

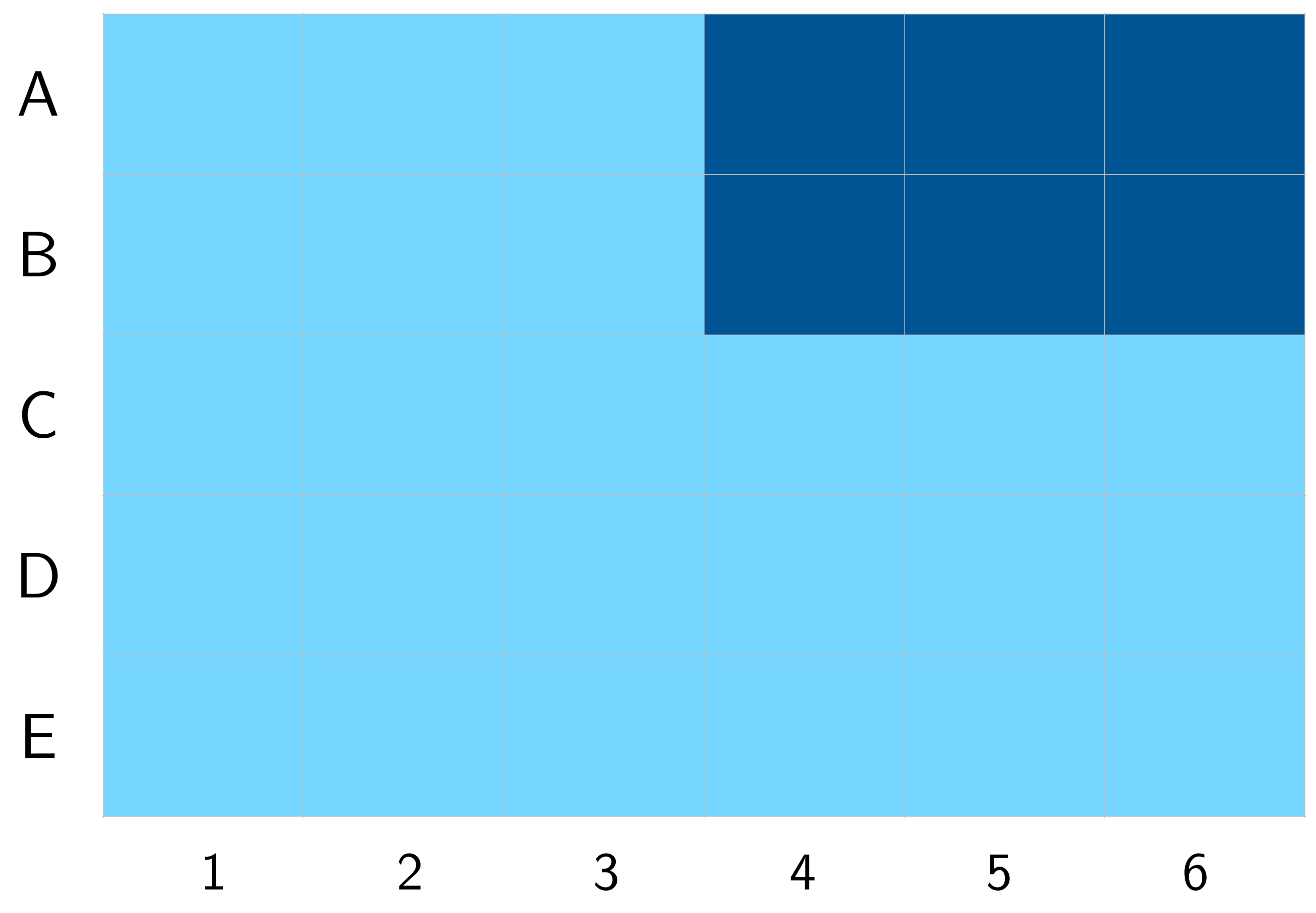


Staggered DID Setting

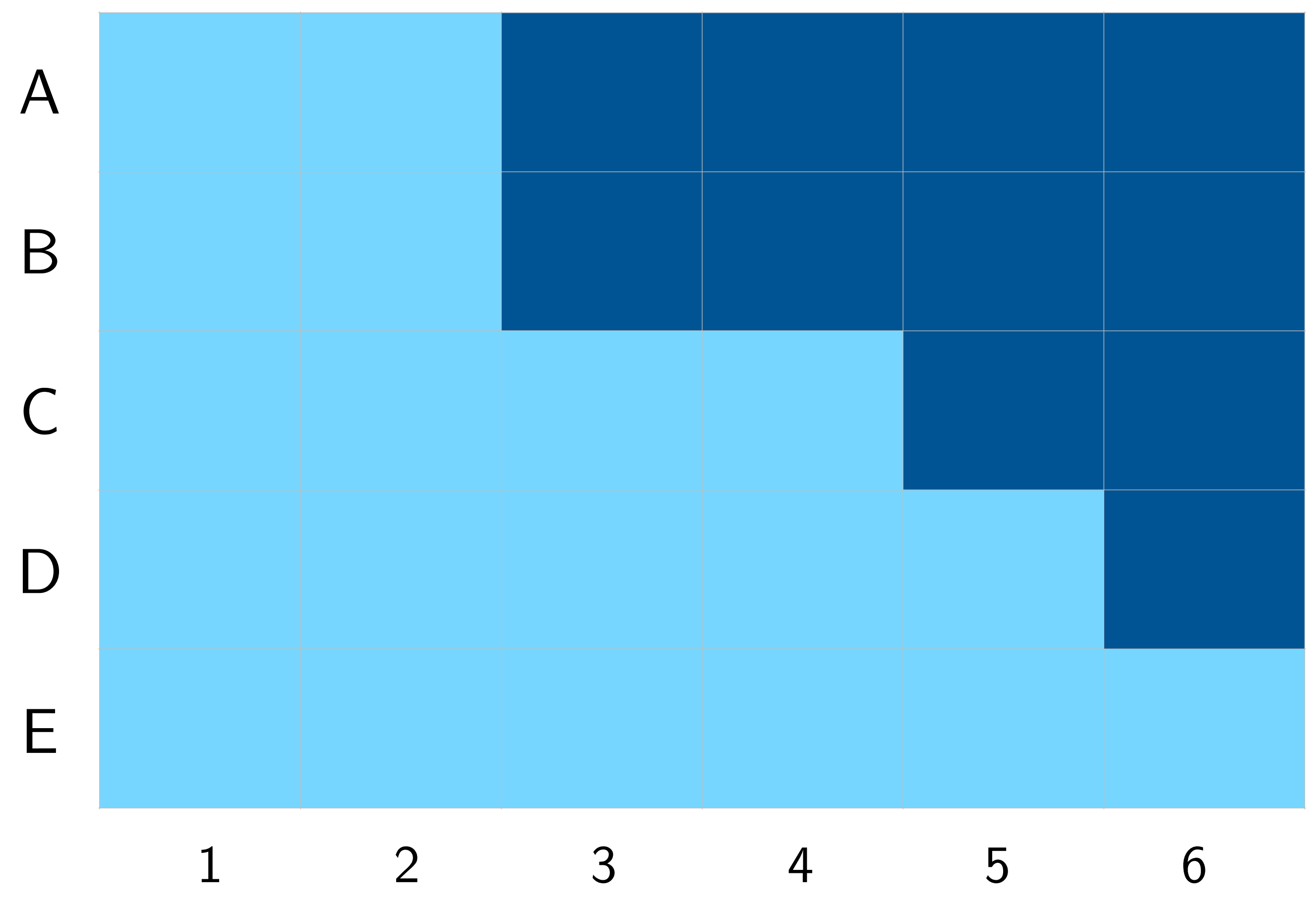


General Setting
(w/ Treatment Reversal)

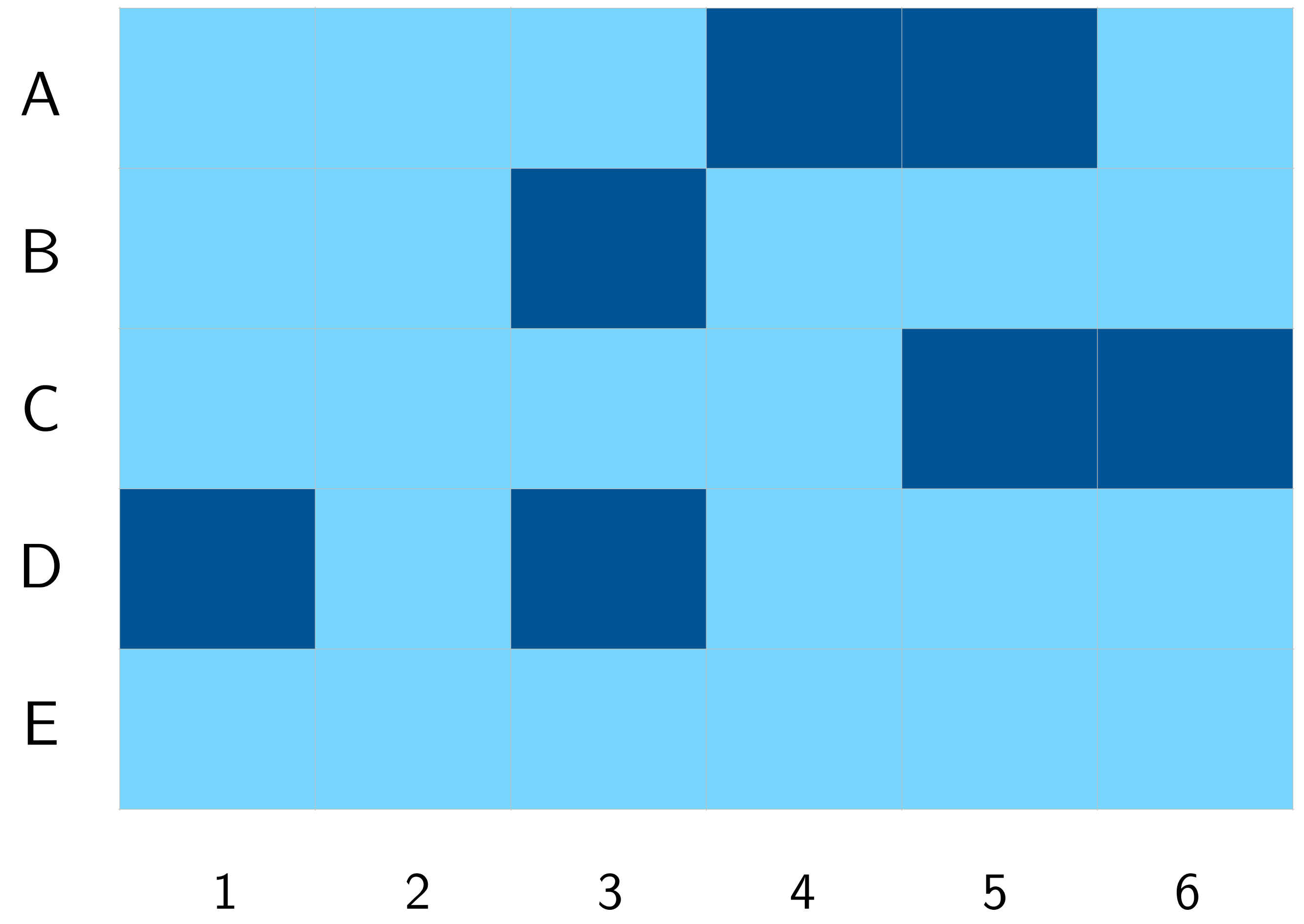
# Settings




(Multi-Period) Block DID Setting

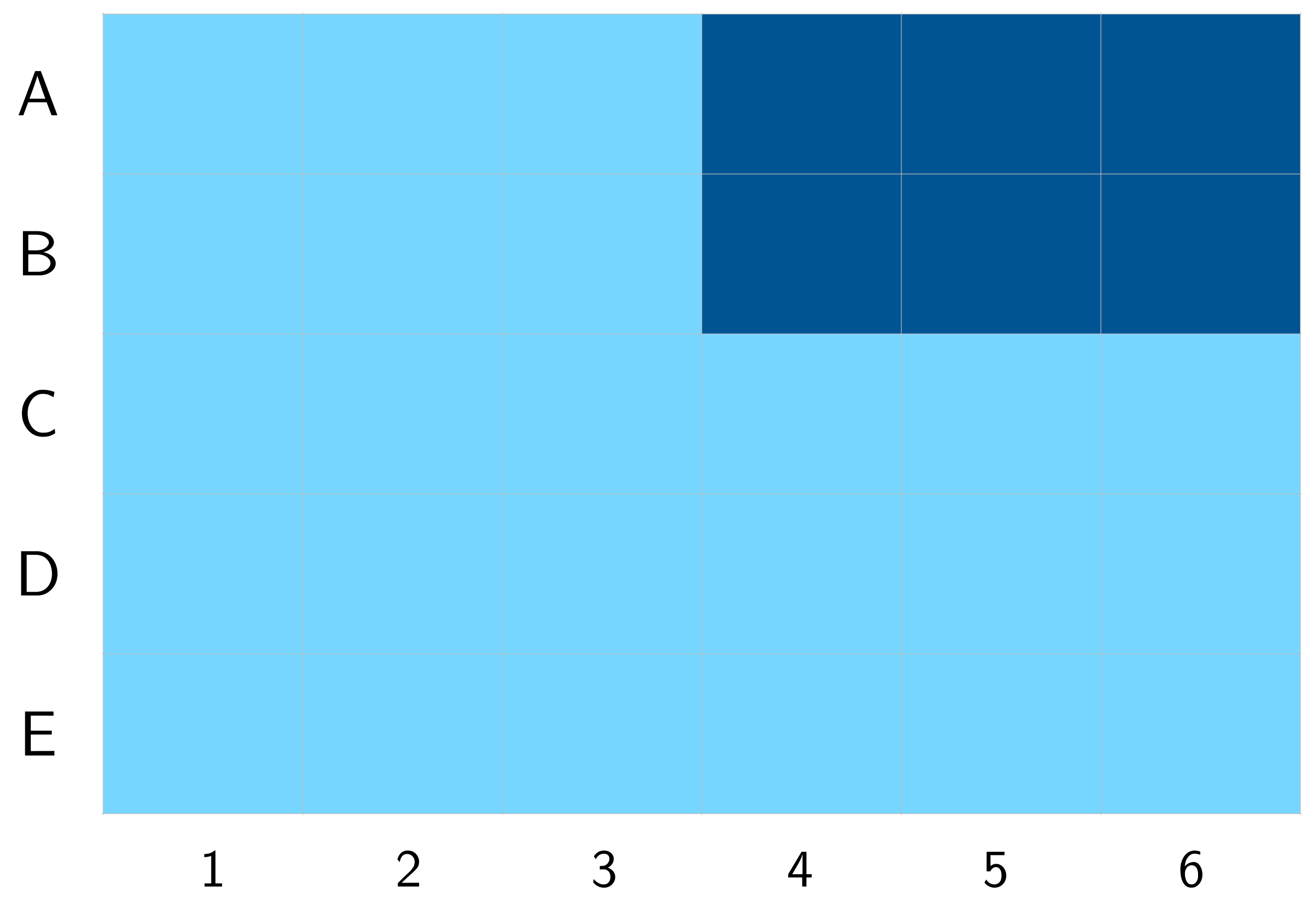



Staggered DID Setting

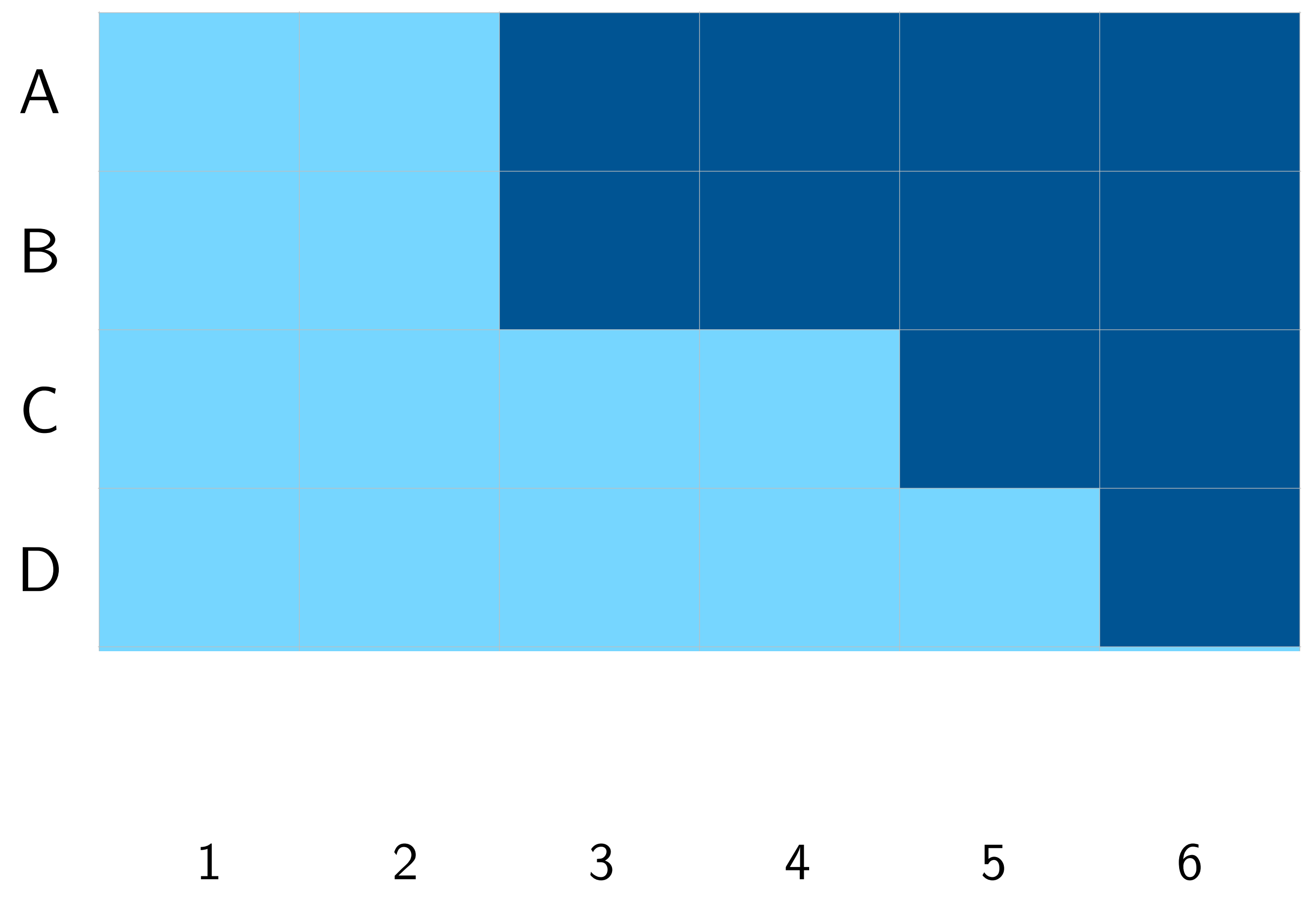



General Setting
(w/ Treatment Reversal)

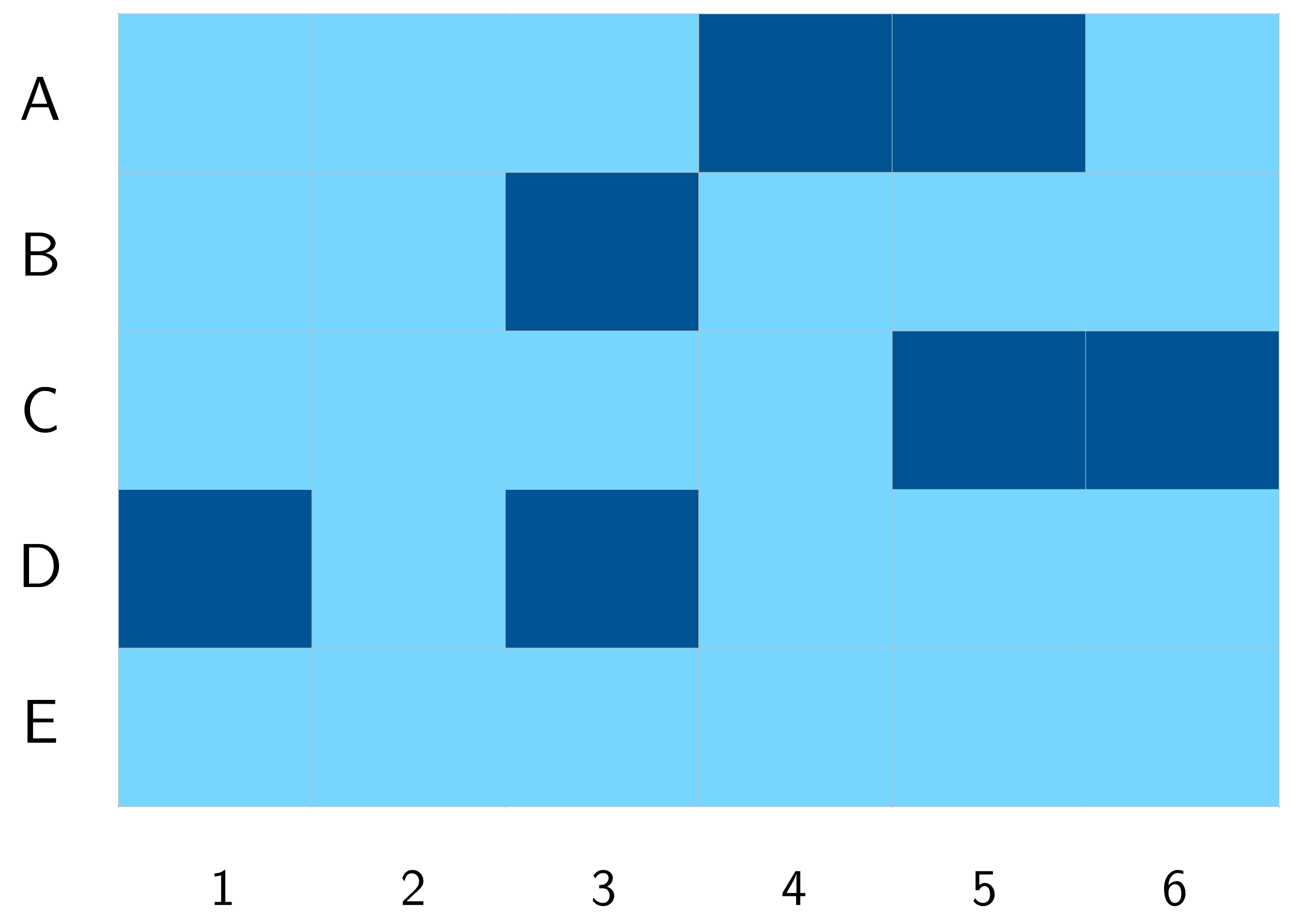
# Settings




(Multi-Period) Block DID Setting

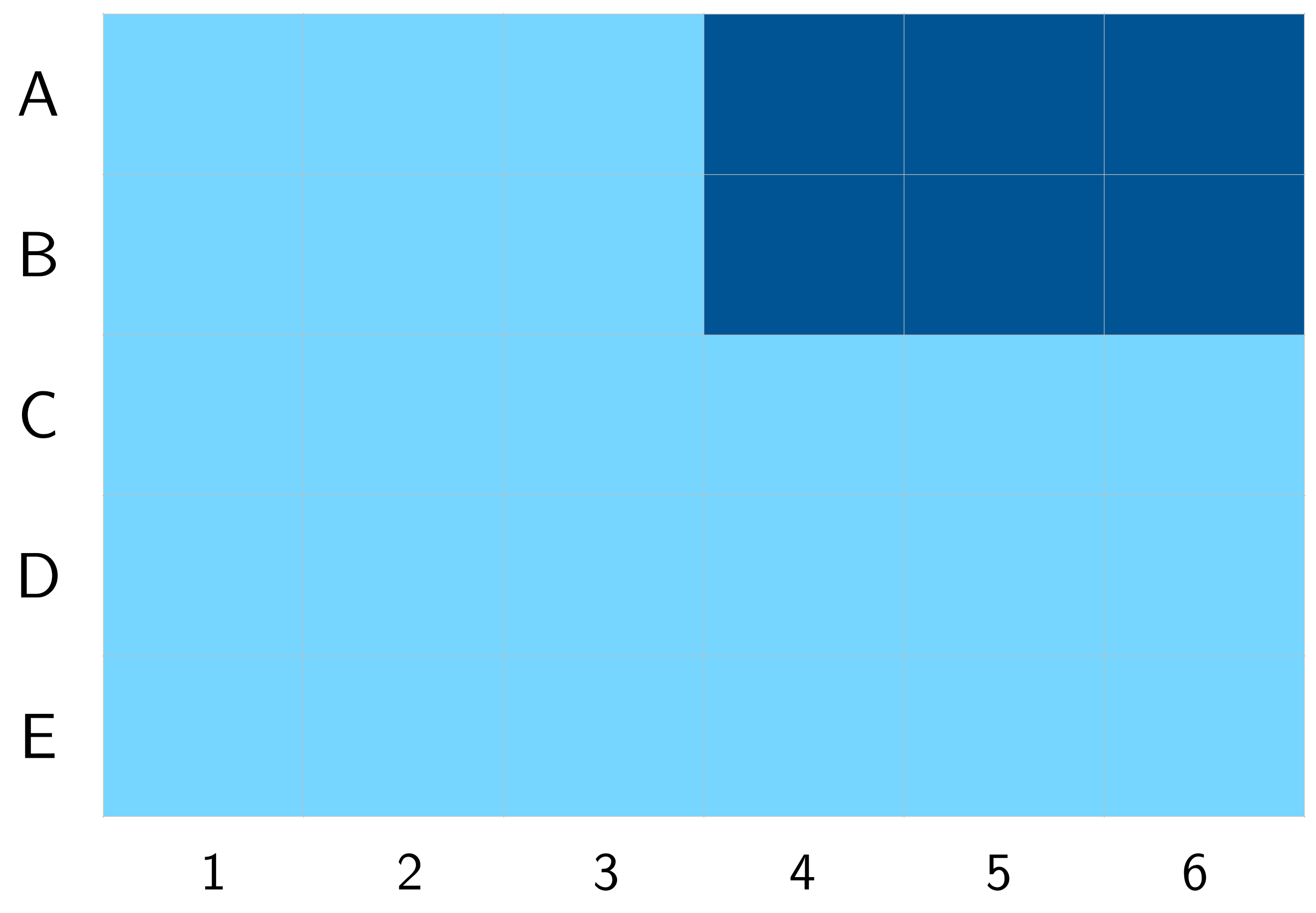



Staggered DID Setting

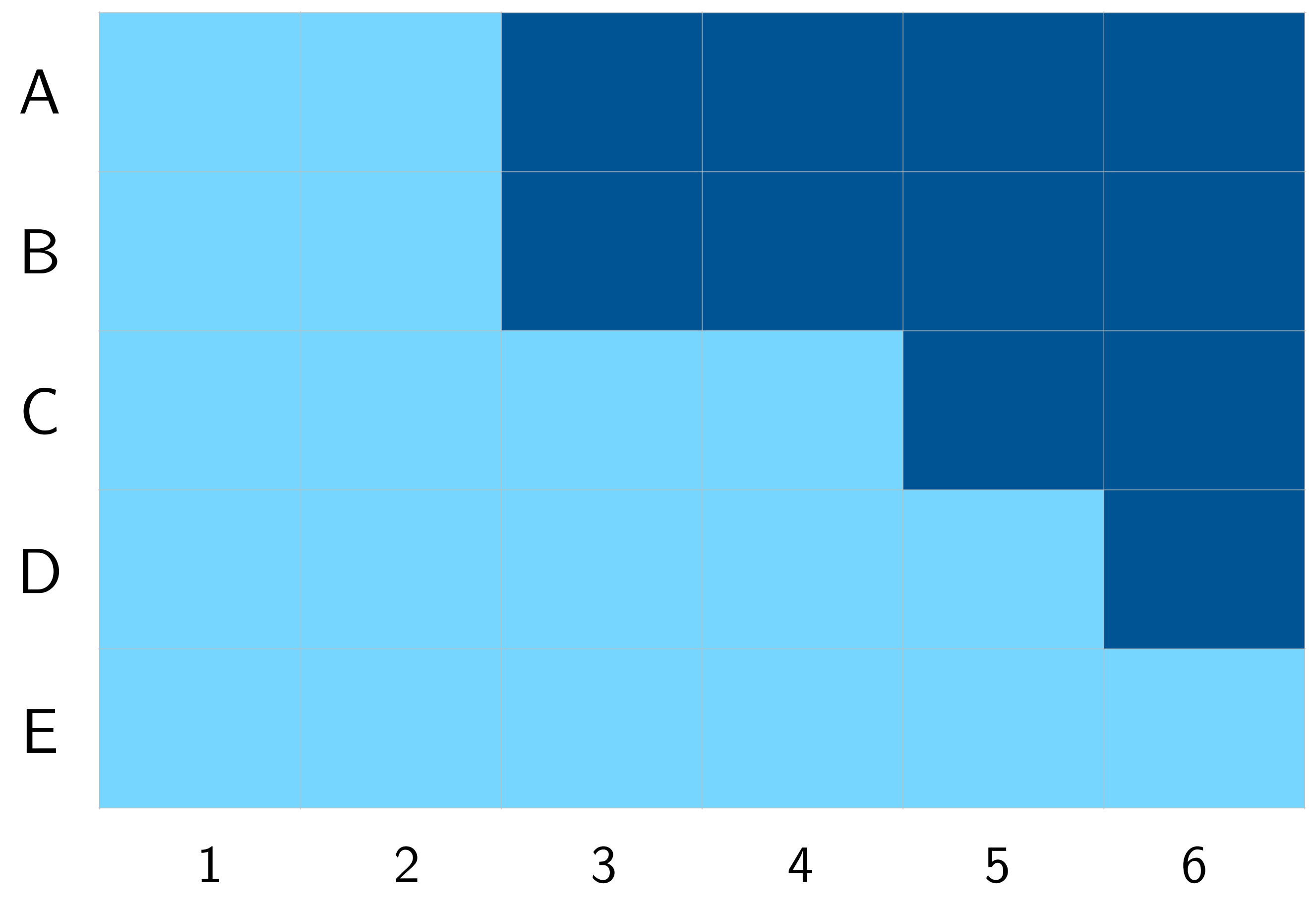



General Setting
(w/ Treatment Reversal)

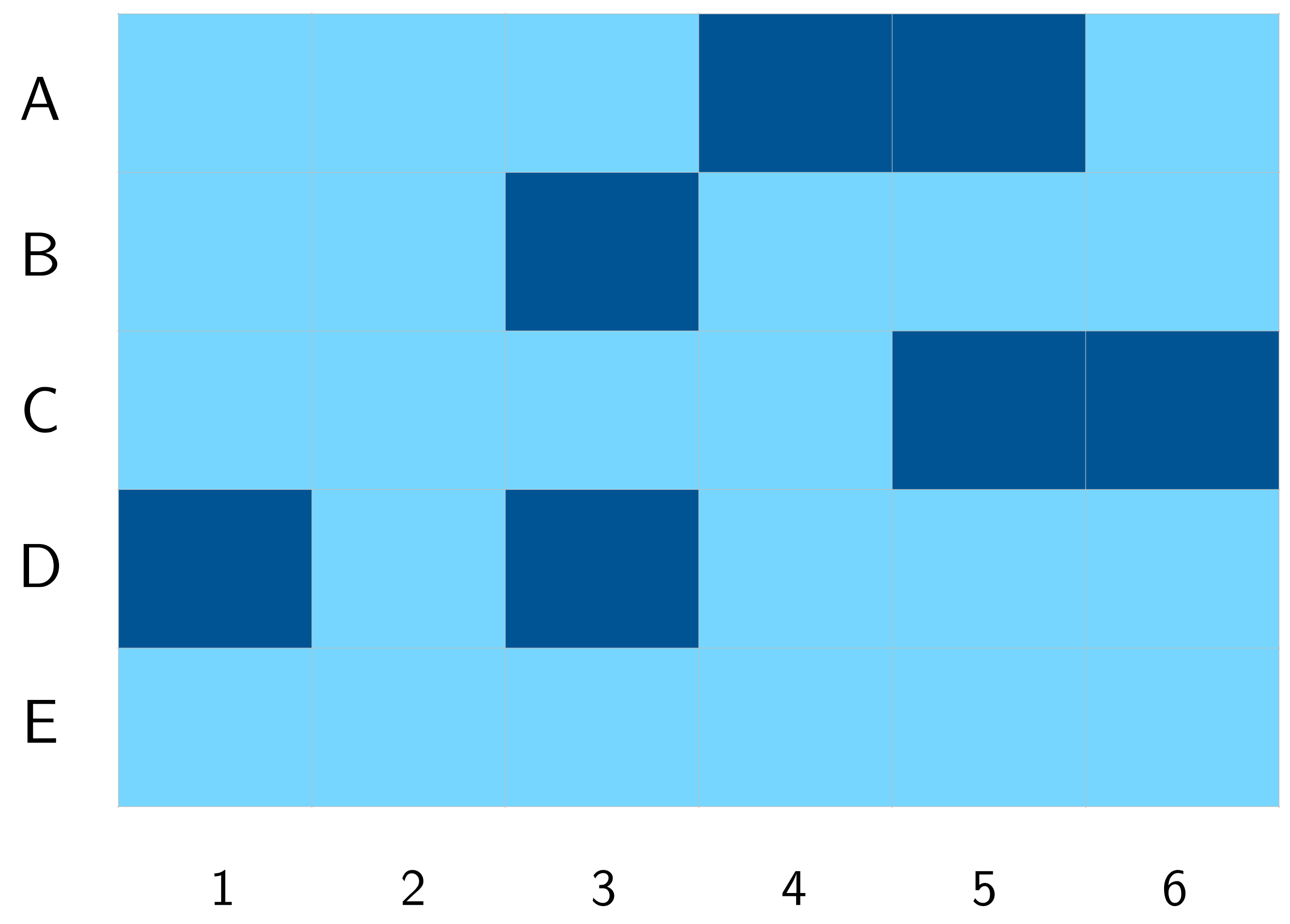
# Settings




(Multi-Period) Block DID Setting




Staggered DID Setting

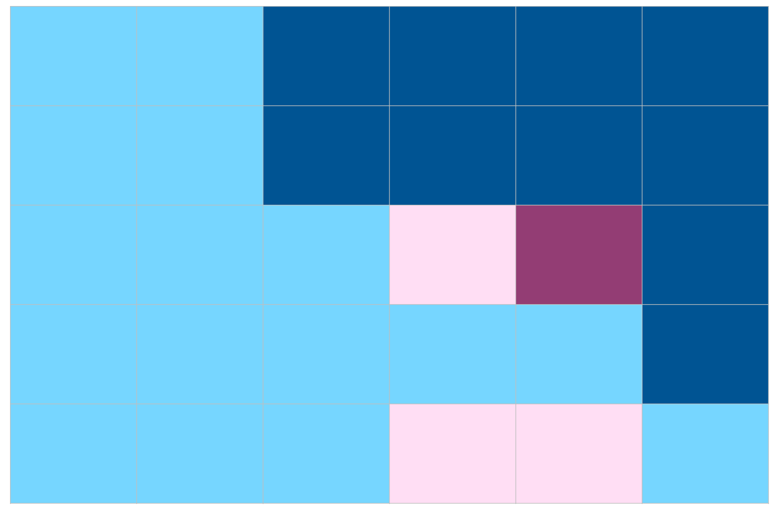



General Setting
(w/ Treatment Reversal)

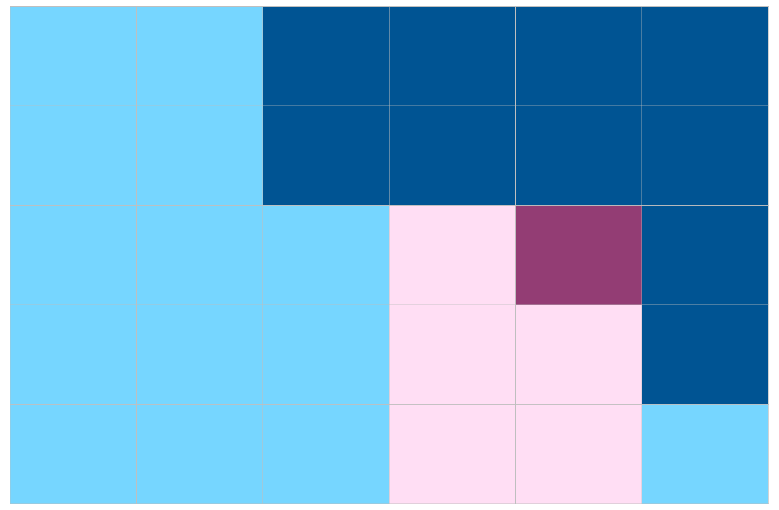
# Different Estimators Use Different Comparison Groups

DID Extension

Staggered



Interaction Weighted
& Stacked DID

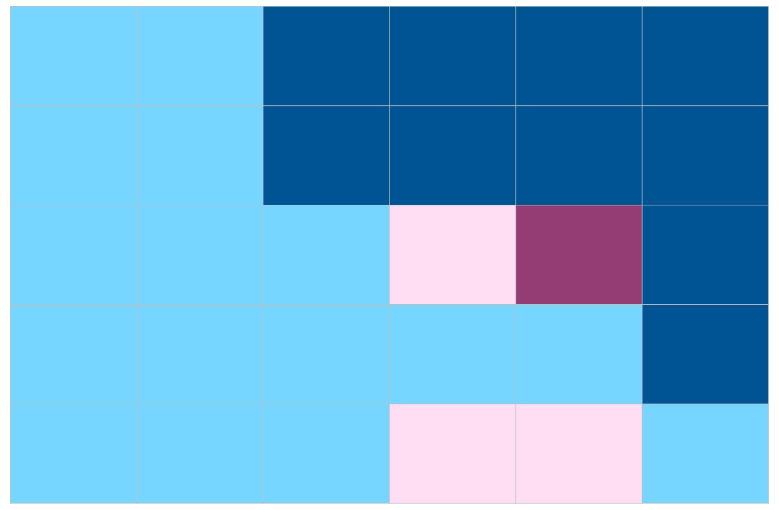


CSDID

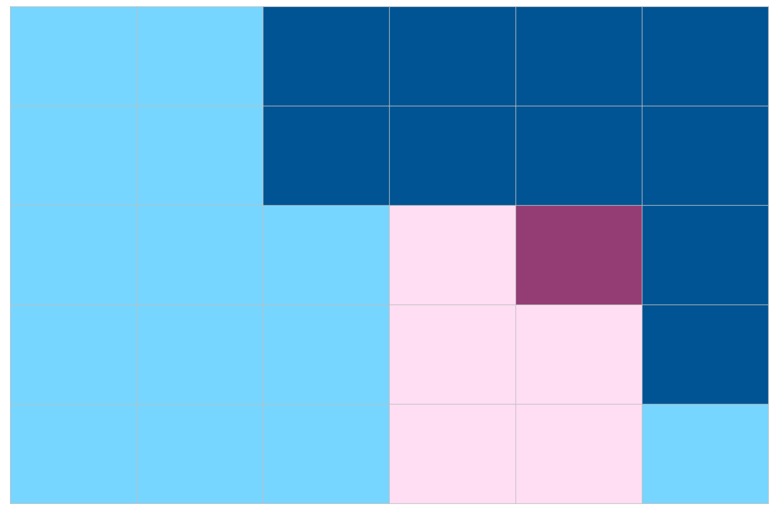
# Different Estimators Use Different Comparison Groups

DID Extension

Staggered

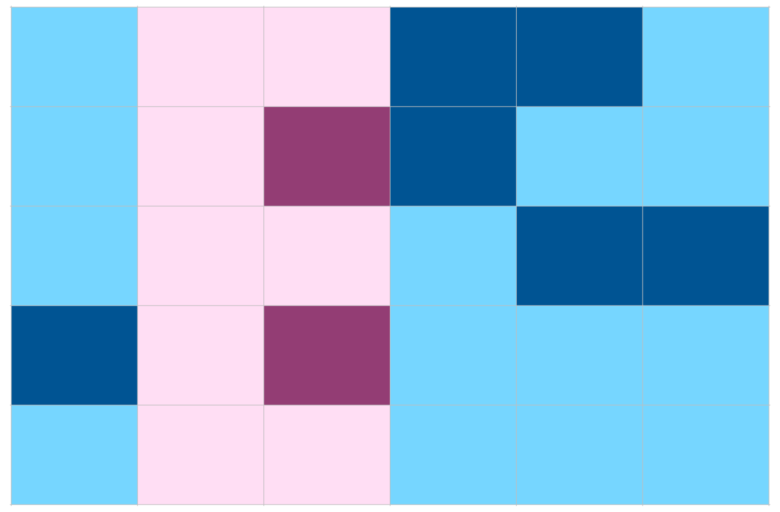


Interaction Weighted
& Stacked DID



CSDID

General



DID multiple/PanelMatch

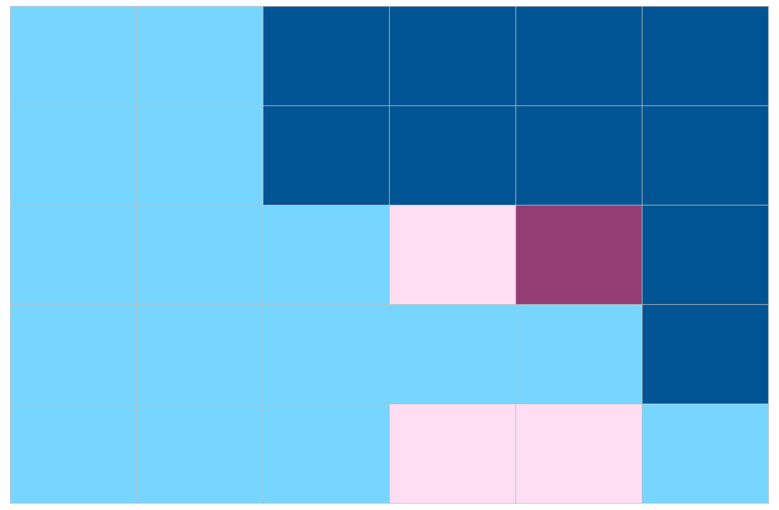
# Different Estimators Use Different Comparison Groups

DID Extension

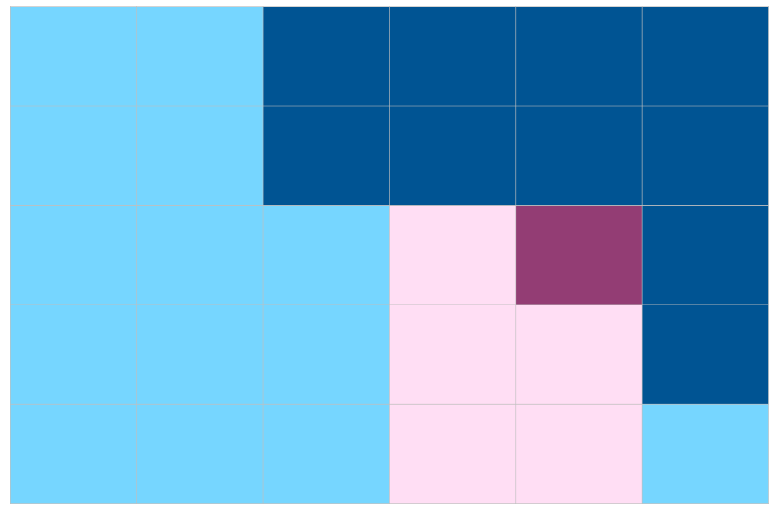
Imputation

Staggered

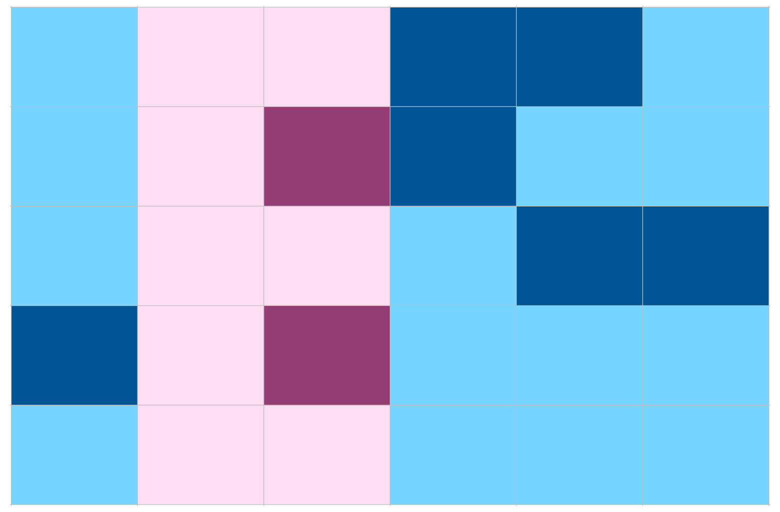
General



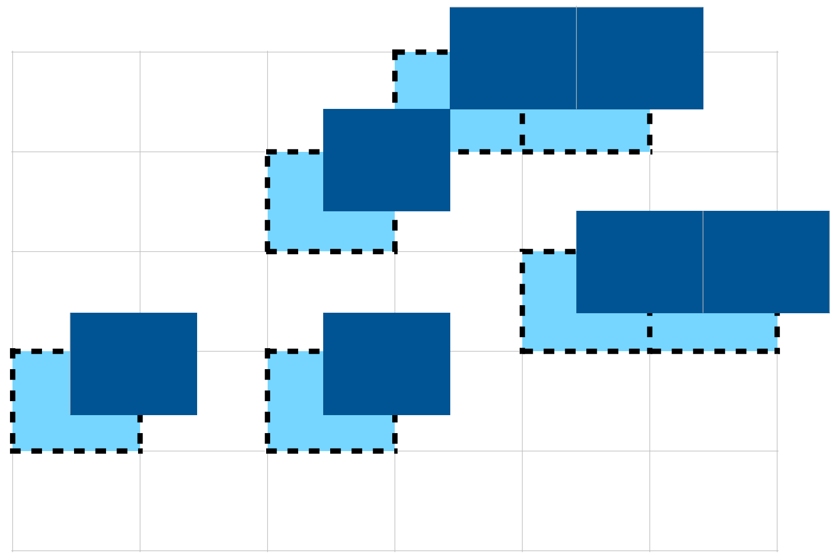
Interaction Weighted
& Stacked DID



CSDID



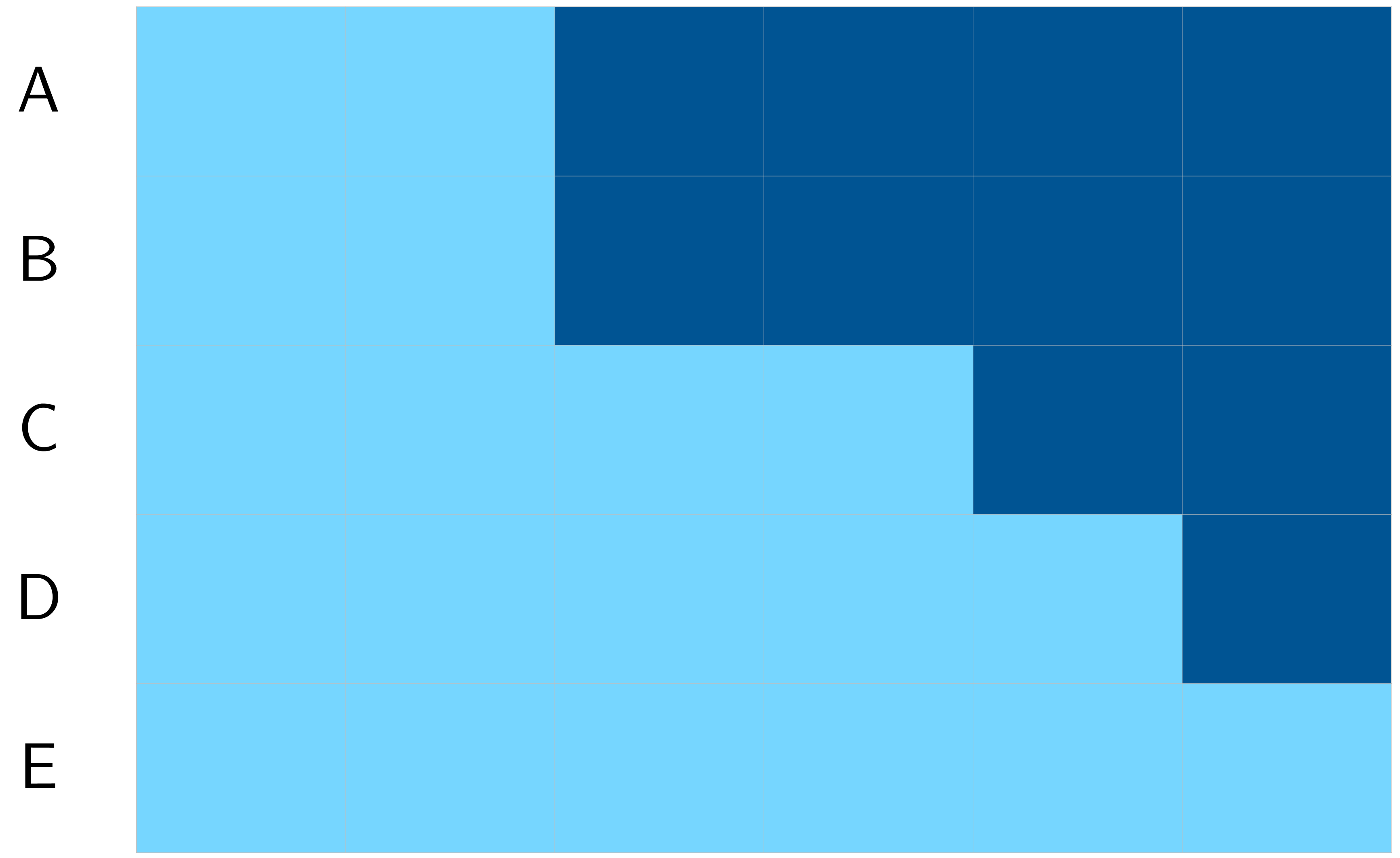
DID multiple/PanelMatch



Imputation Method
$\text{DID}_{\text{impute}}$, FEct

# Sun & Abraham (2021)

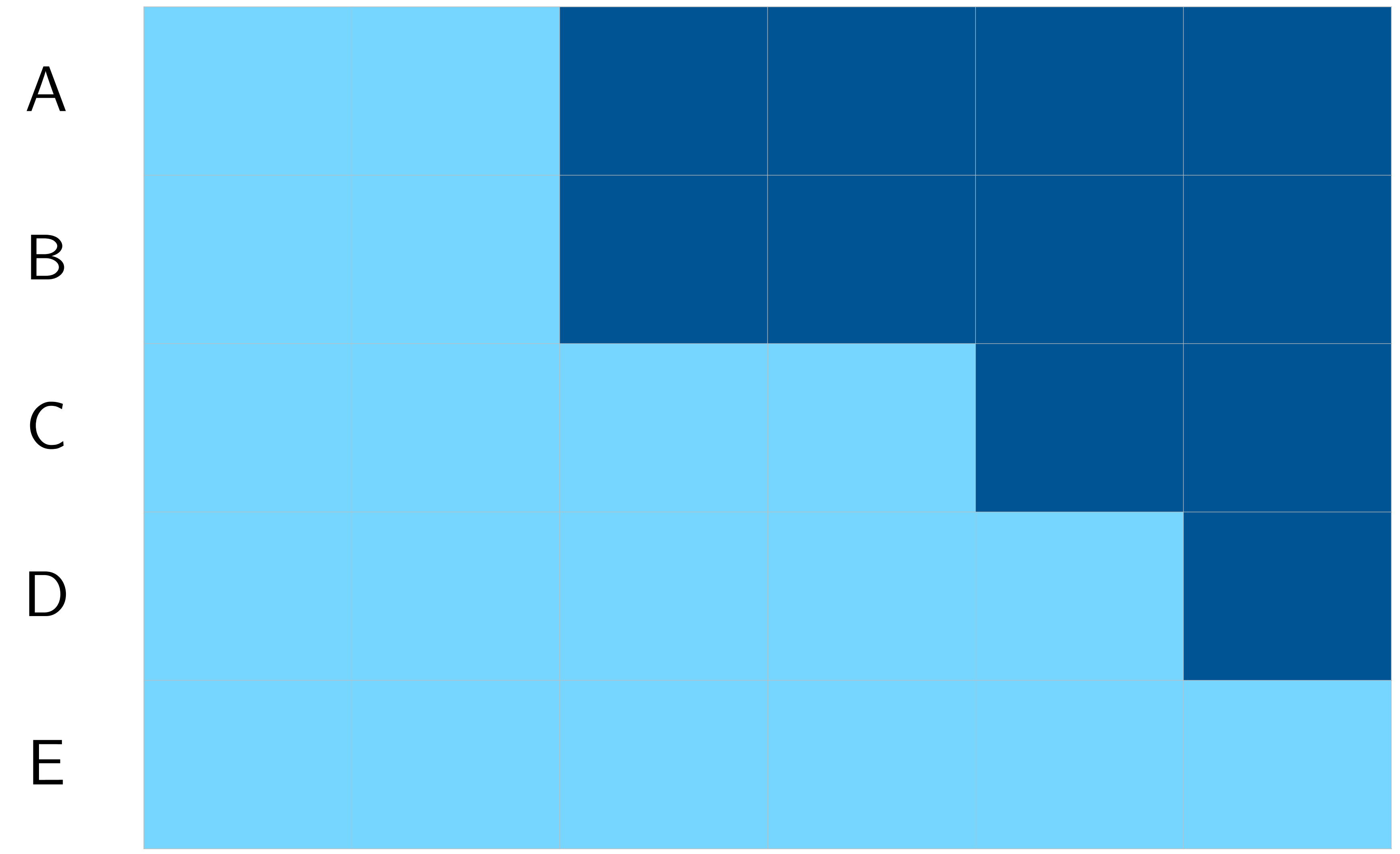
Interaction Weighted (IW)



# Sun & Abraham (2021)

Interaction Weighted (IW)

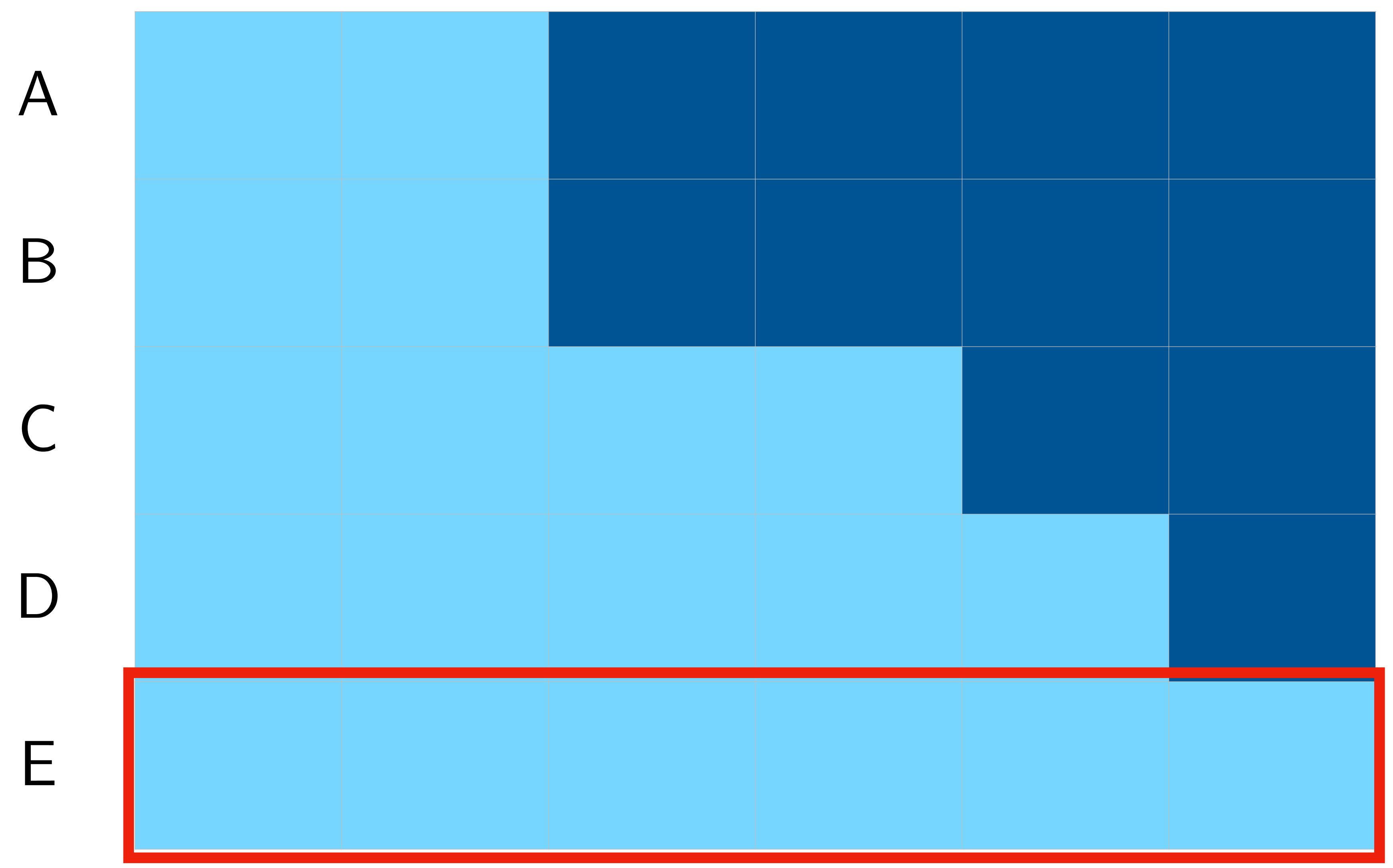
- Comparison group: never-treated



# Sun & Abraham (2021)

Interaction Weighted (IW)

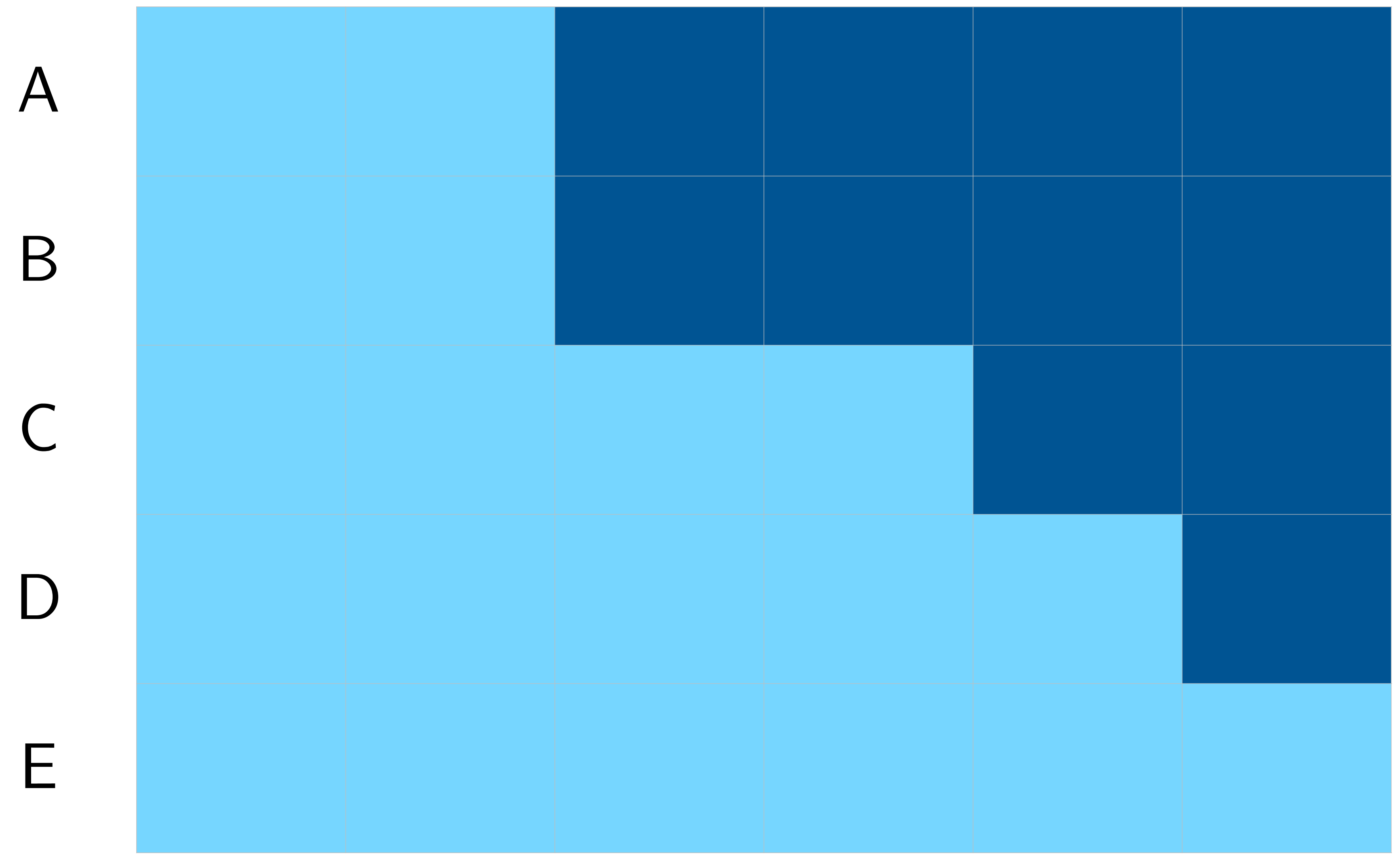
- Comparison group: never-treated



# Sun & Abraham (2021)

Interaction Weighted (IW)

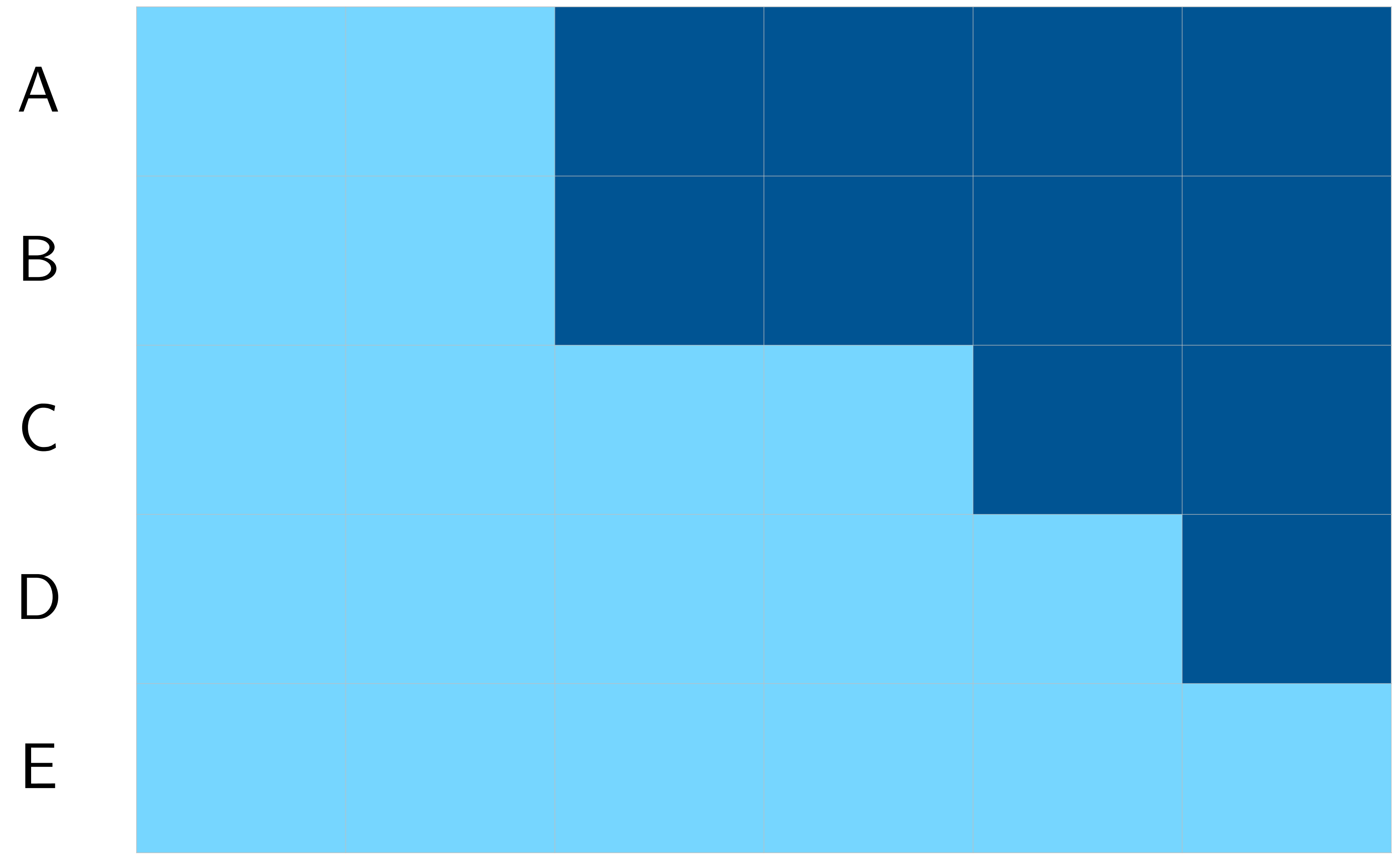
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using $2 \times 2$ DID for each cohort $g$ and period since treatment $l$



# Sun & Abraham (2021)

Interaction Weighted (IW)

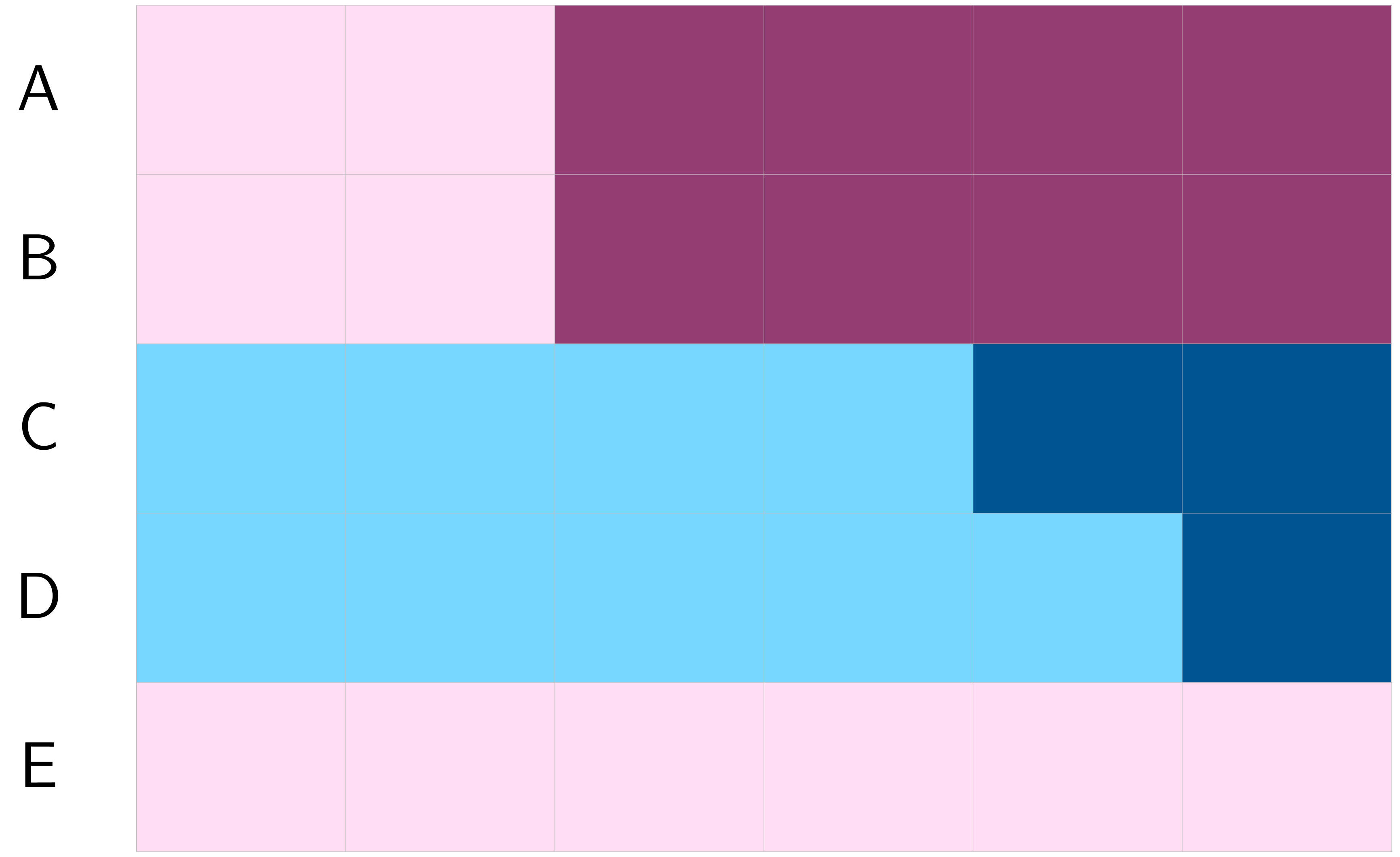
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using $2 \times 2$ DID for each cohort $g$ and period since treatment $l$
- ATT $=$ average CATT, weighted by cohort size



# Sun & Abraham (2021)

Interaction Weighted (IW)

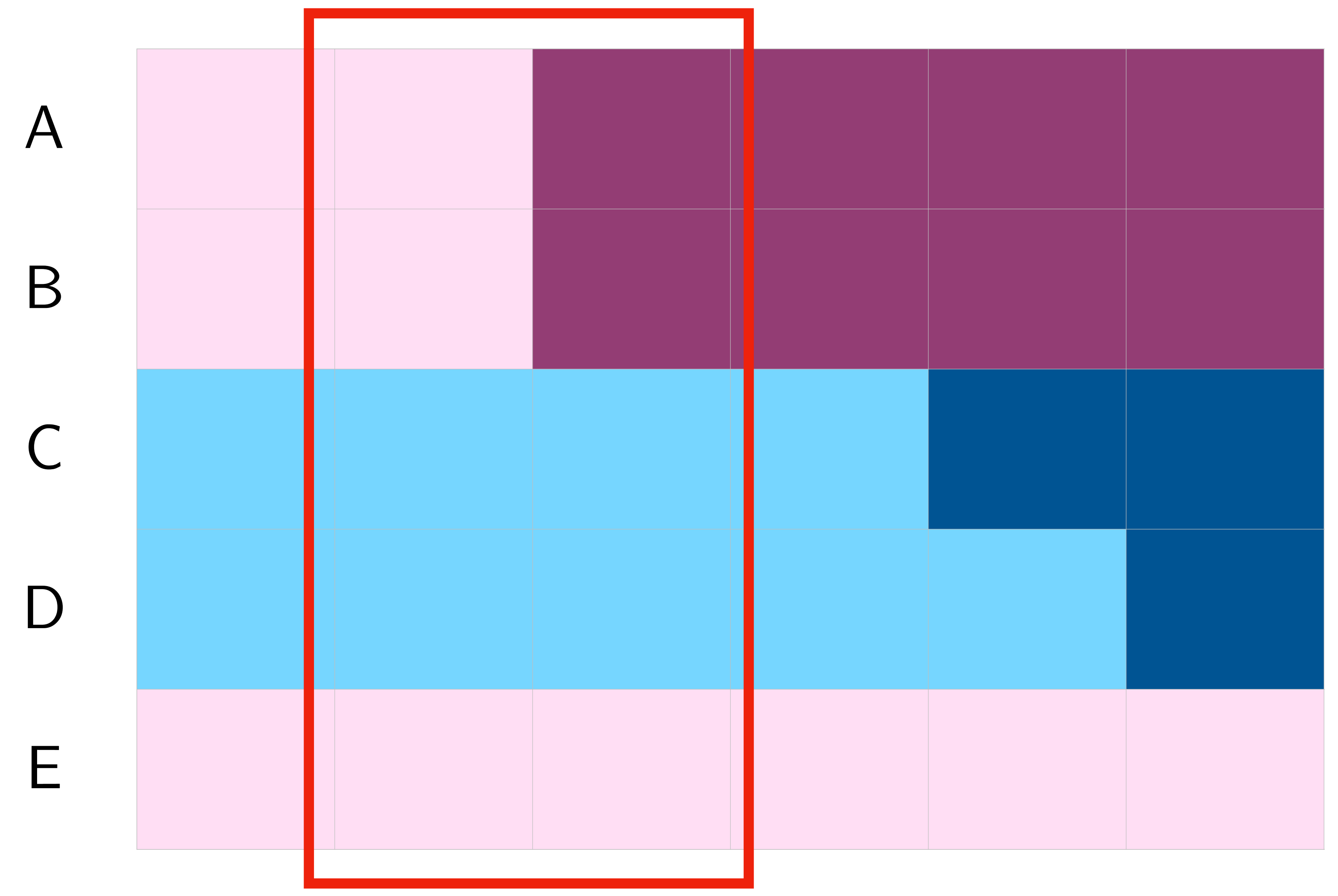
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using $2 \times 2$ DID for each cohort $g$ and period since treatment $l$
- ATT $=$ average CATT, weighted by cohort size



# Sun & Abraham (2021)

Interaction Weighted (IW)

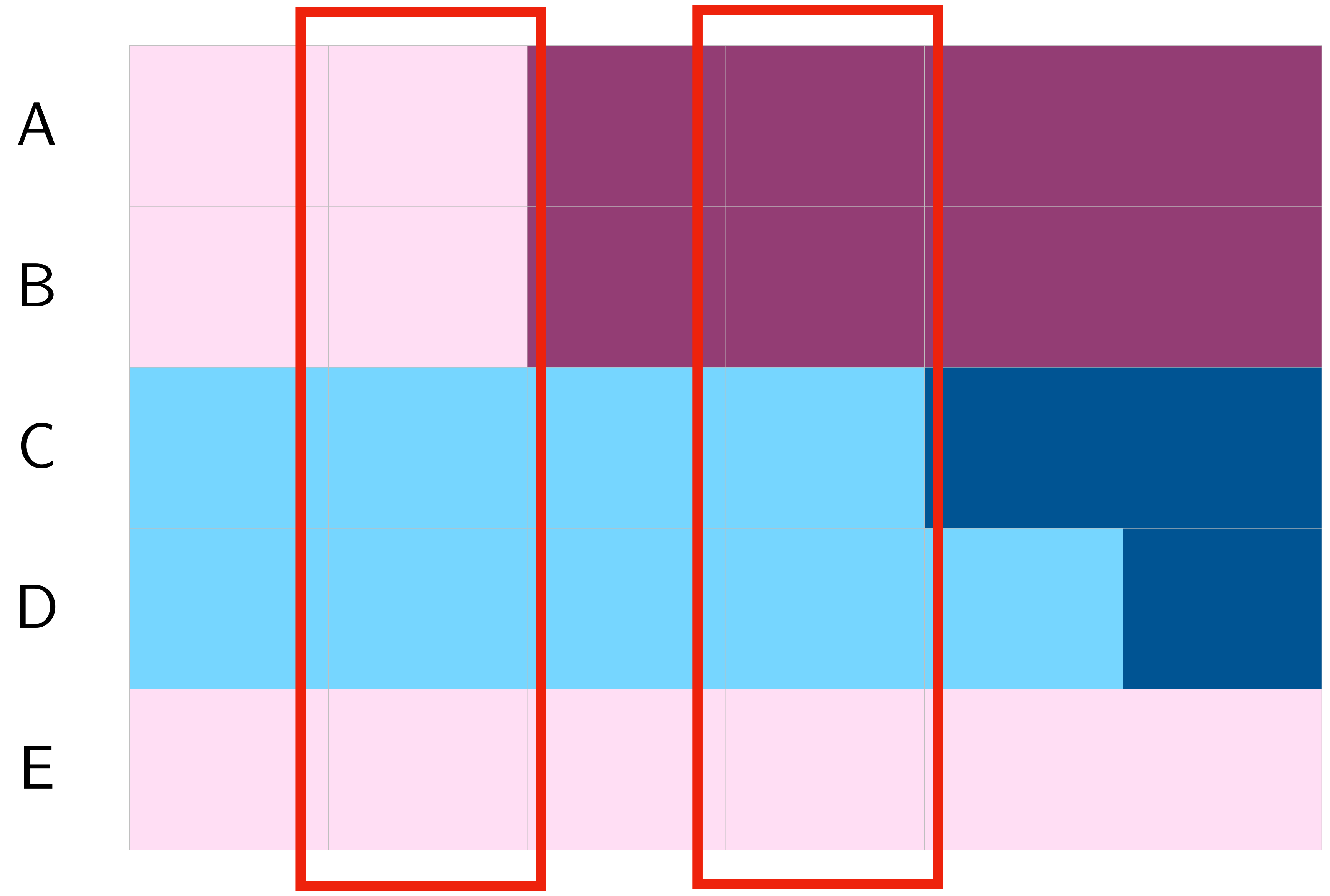
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using $2 \times 2$ DID for each cohort $g$ and period since treatment $l$
- ATT $=$ average CATT, weighted by cohort size



# Sun & Abraham (2021)

Interaction Weighted (IW)

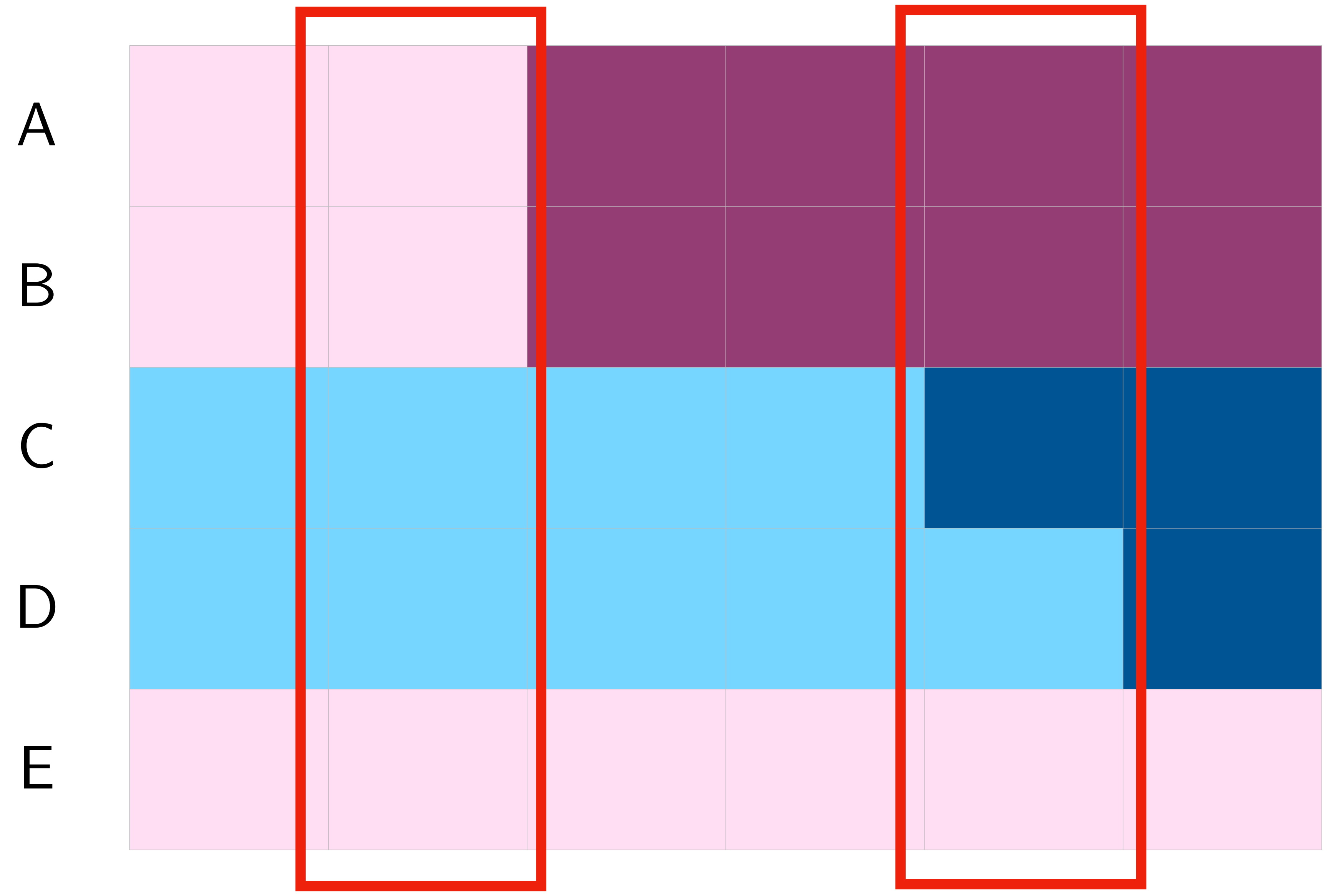
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using $2 \times 2$ DID for each cohort $g$ and period since treatment $l$
- ATT $=$ average CATT, weighted by cohort size



# Sun & Abraham (2021)

Interaction Weighted (IW)

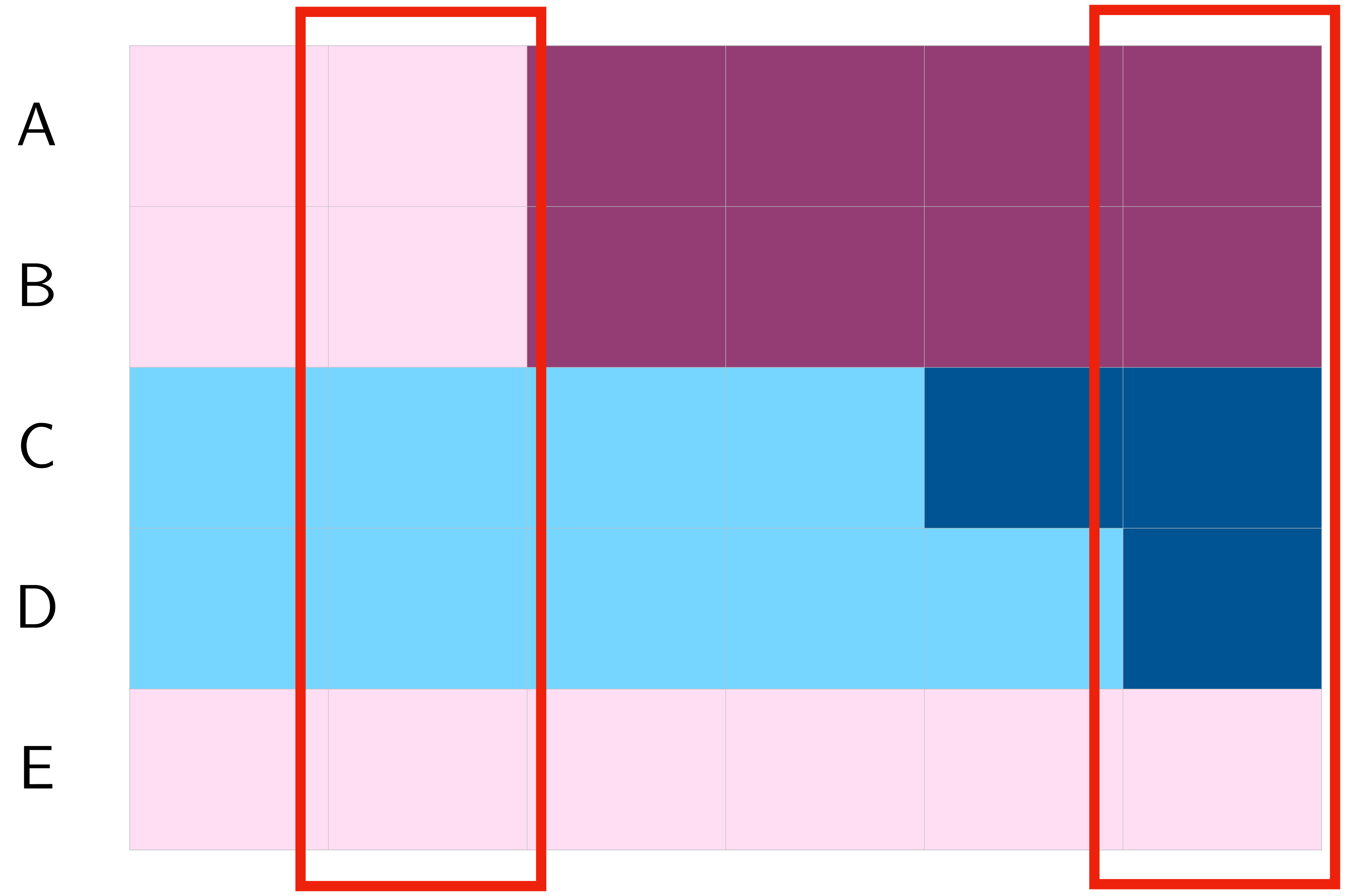
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using $2 \times 2$ DID for each cohort $g$ and period since treatment $l$
- ATT $=$ average CATT, weighted by cohort size



# Sun & Abraham (2021)

Interaction Weighted (IW)

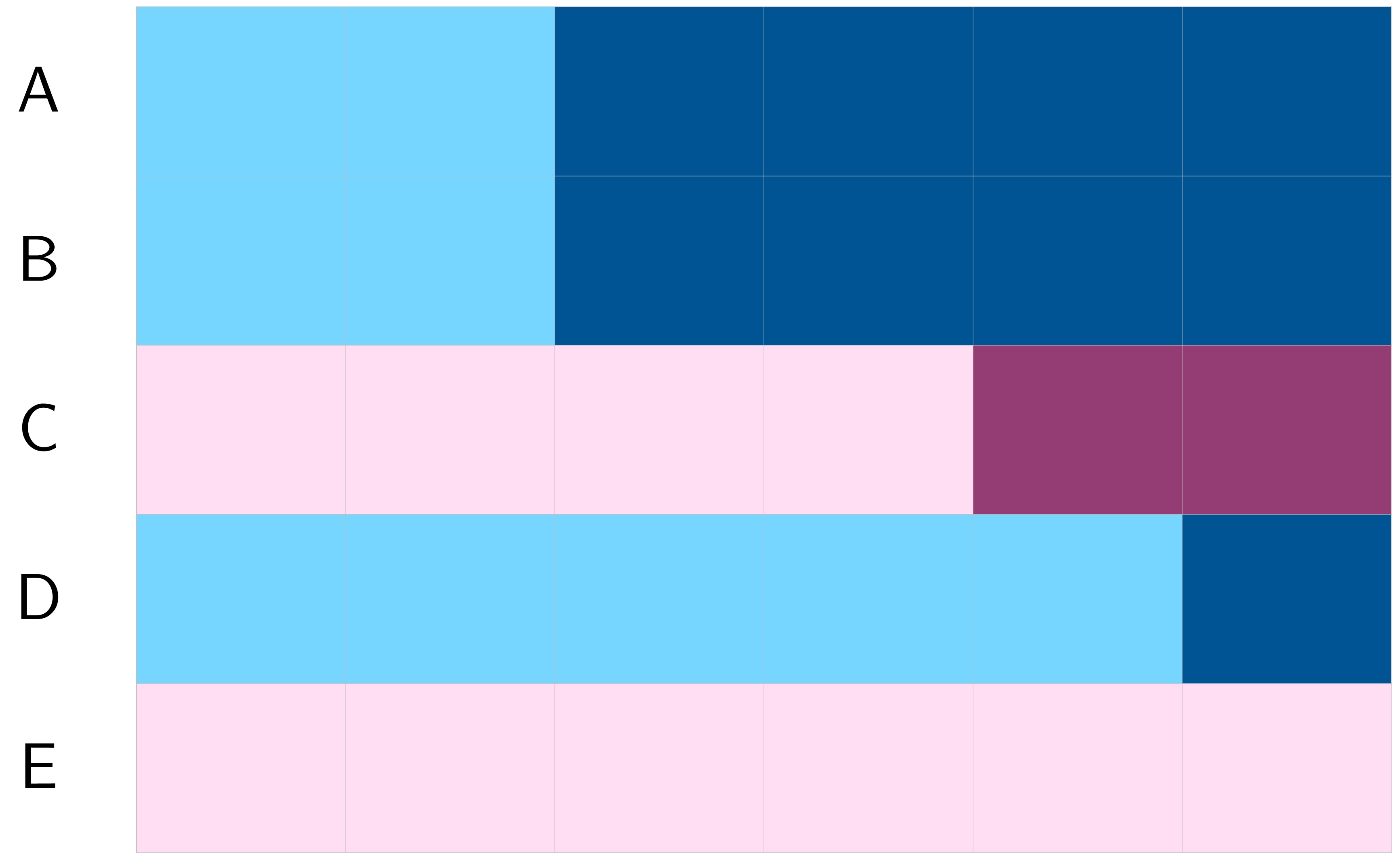
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using $2 \times 2$ DID for each cohort $g$ and period since treatment $l$
- ATT $=$ average CATT, weighted by cohort size



# Sun & Abraham (2021)

Interaction Weighted (IW)

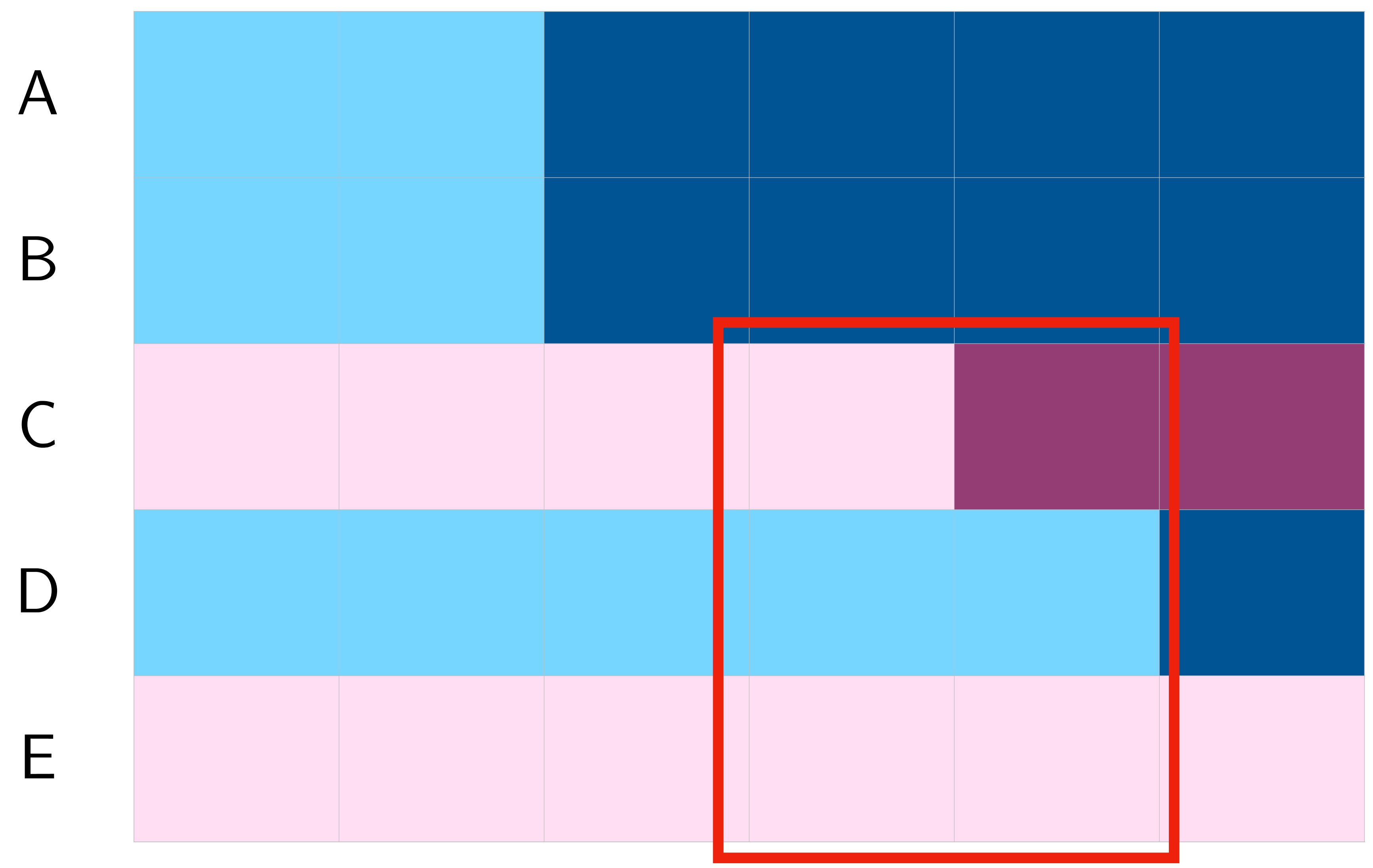
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using $2 \times 2$ DID for each cohort $g$ and period since treatment $l$
- ATT $=$ average CATT, weighted by cohort size



# Sun & Abraham (2021)

Interaction Weighted (IW)

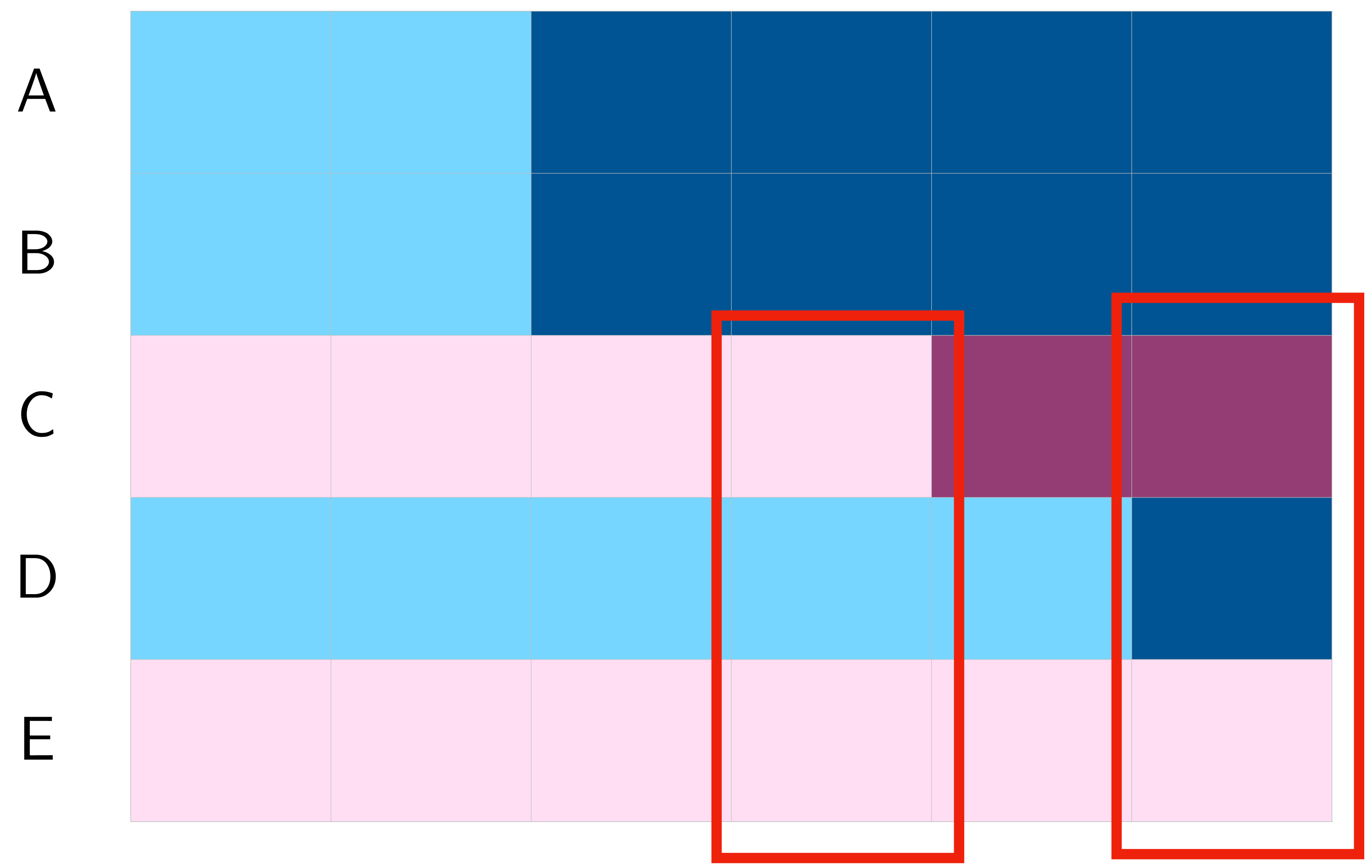
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using $2 \times 2$ DID for each cohort $g$ and period since treatment $l$
- ATT $=$ average CATT, weighted by cohort size



# Sun & Abraham (2021)

Interaction Weighted (IW)

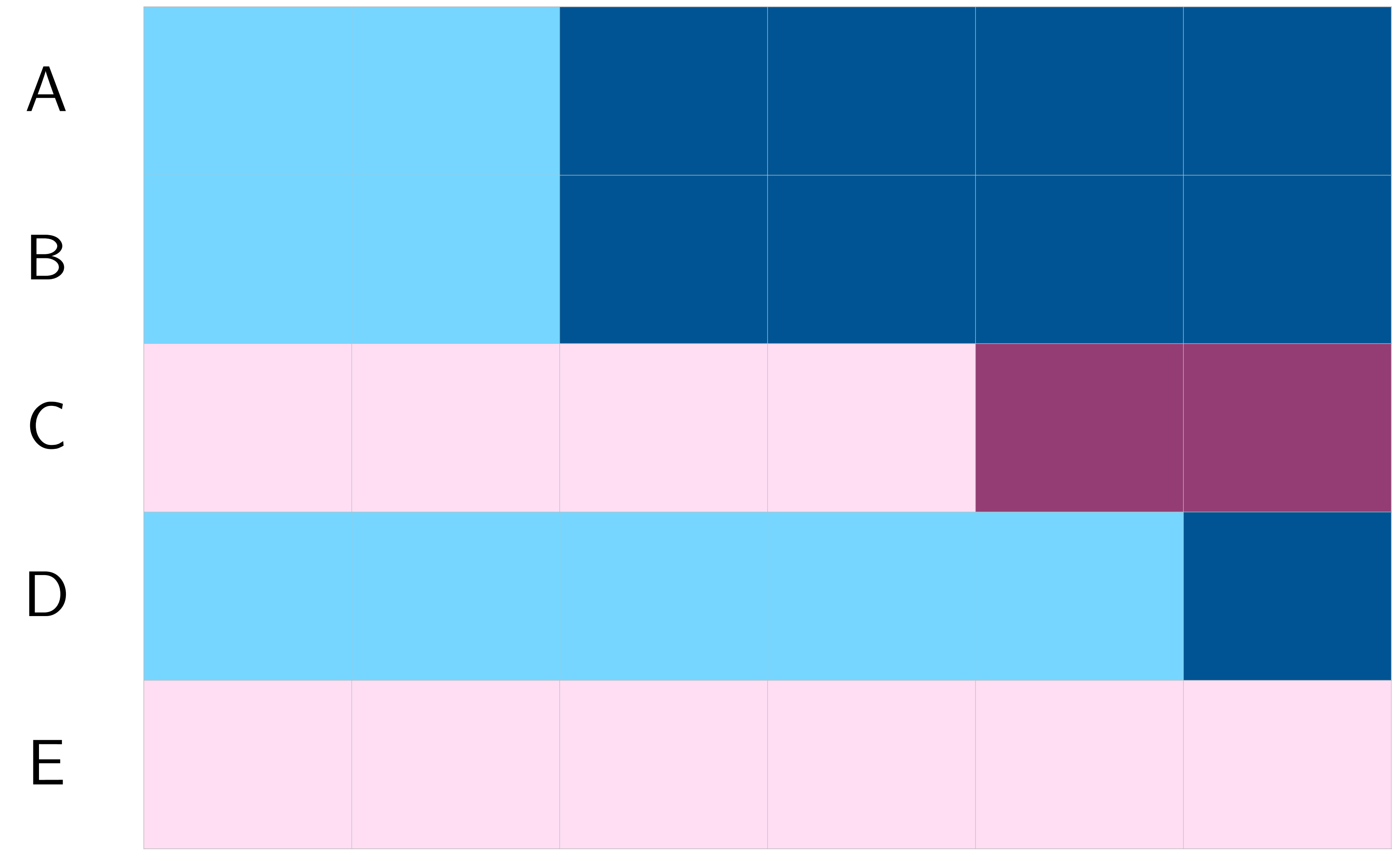
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using $2 \times 2$ DID for each cohort $g$ and period since treatment $l$
- ATT $=$ average CATT, weighted by cohort size



# Sun & Abraham (2021)

Interaction Weighted (IW)

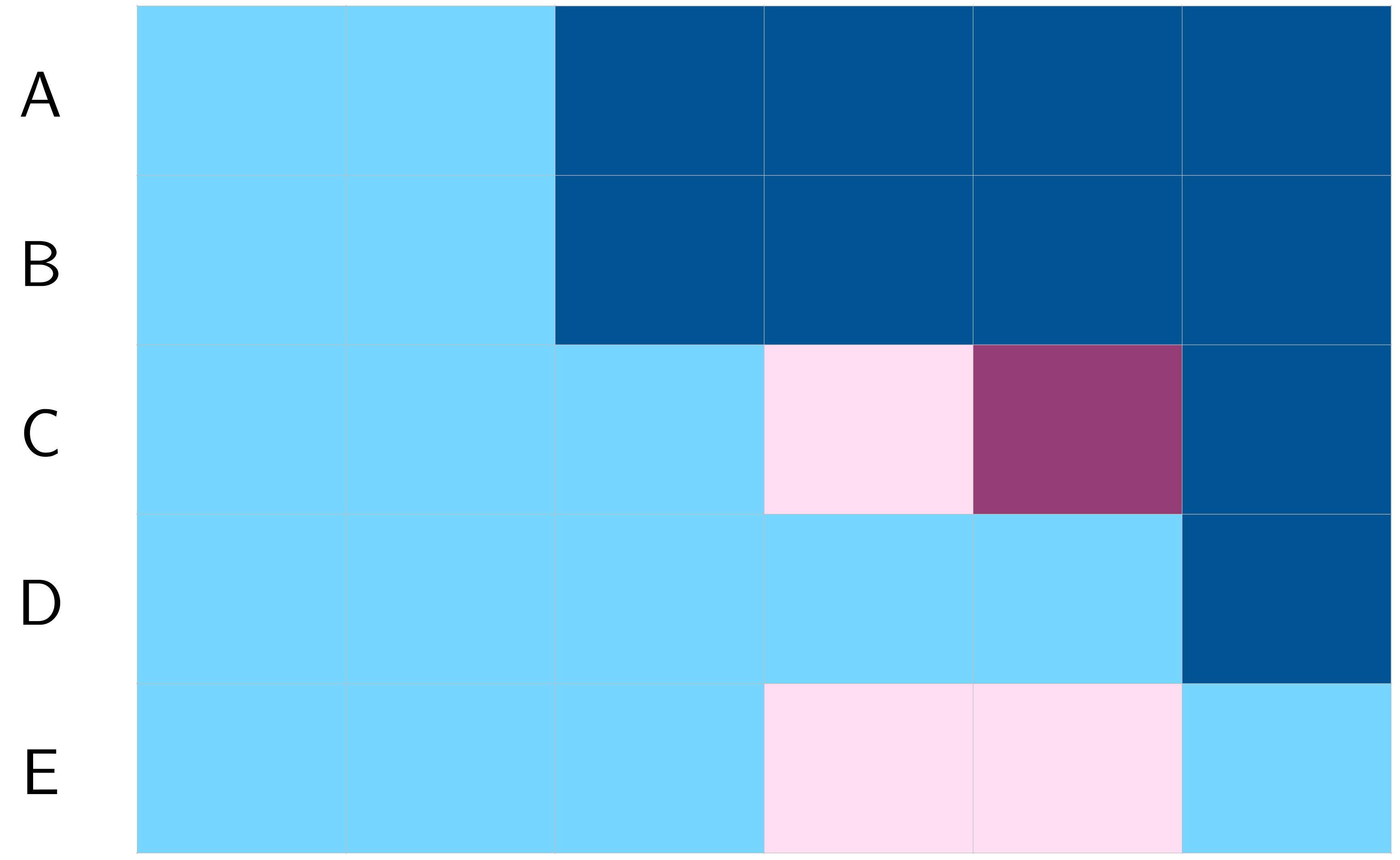
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using $2 \times 2$ DID for each cohort $g$ and period since treatment $l$
- ATT $=$ average CATT, weighted by cohort size



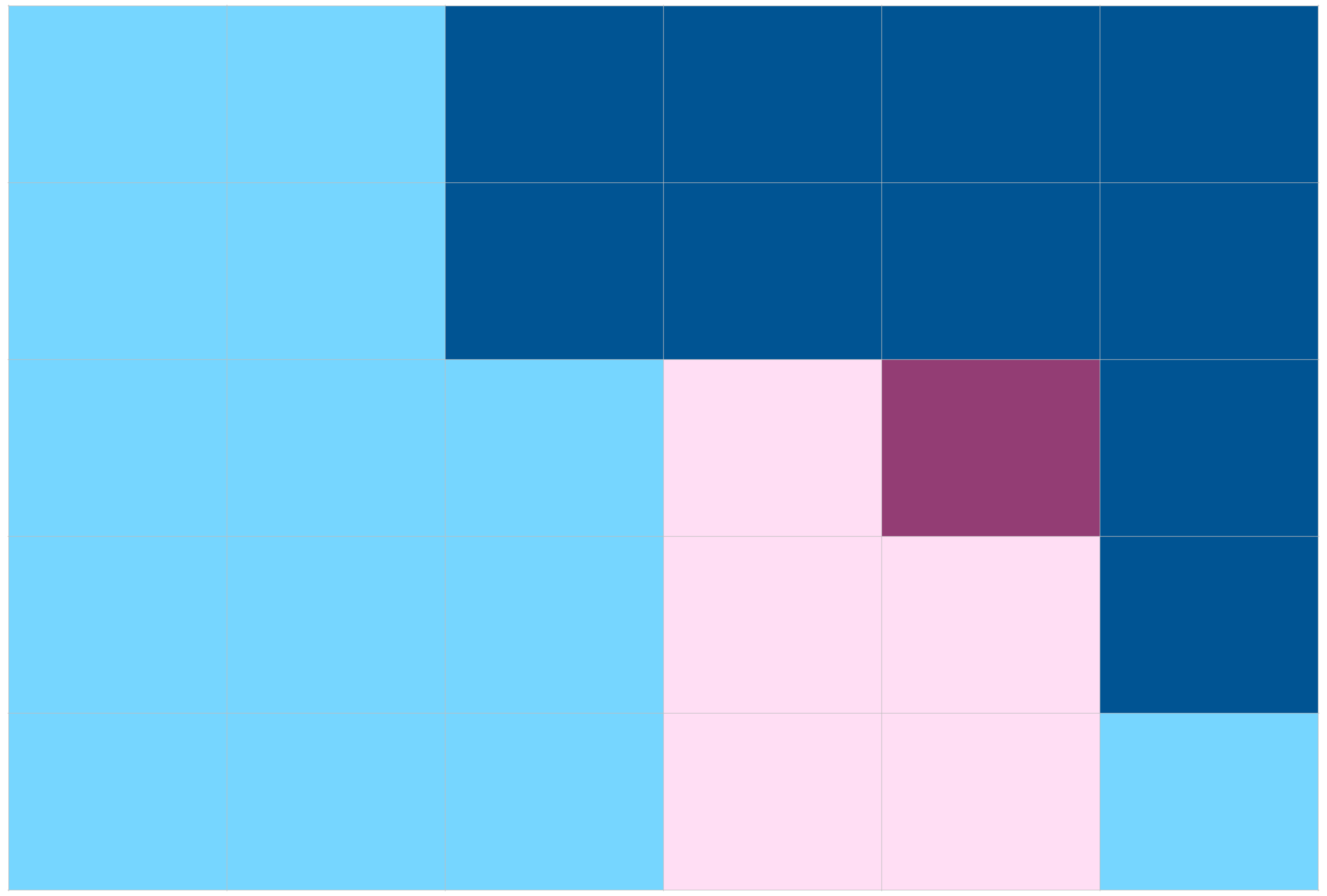
# Callaway & Sant'Anna (2021)

- Comparison group: <u>not-yet-treated</u> (in additional to never treated)
- "Doubly robust" with covariates

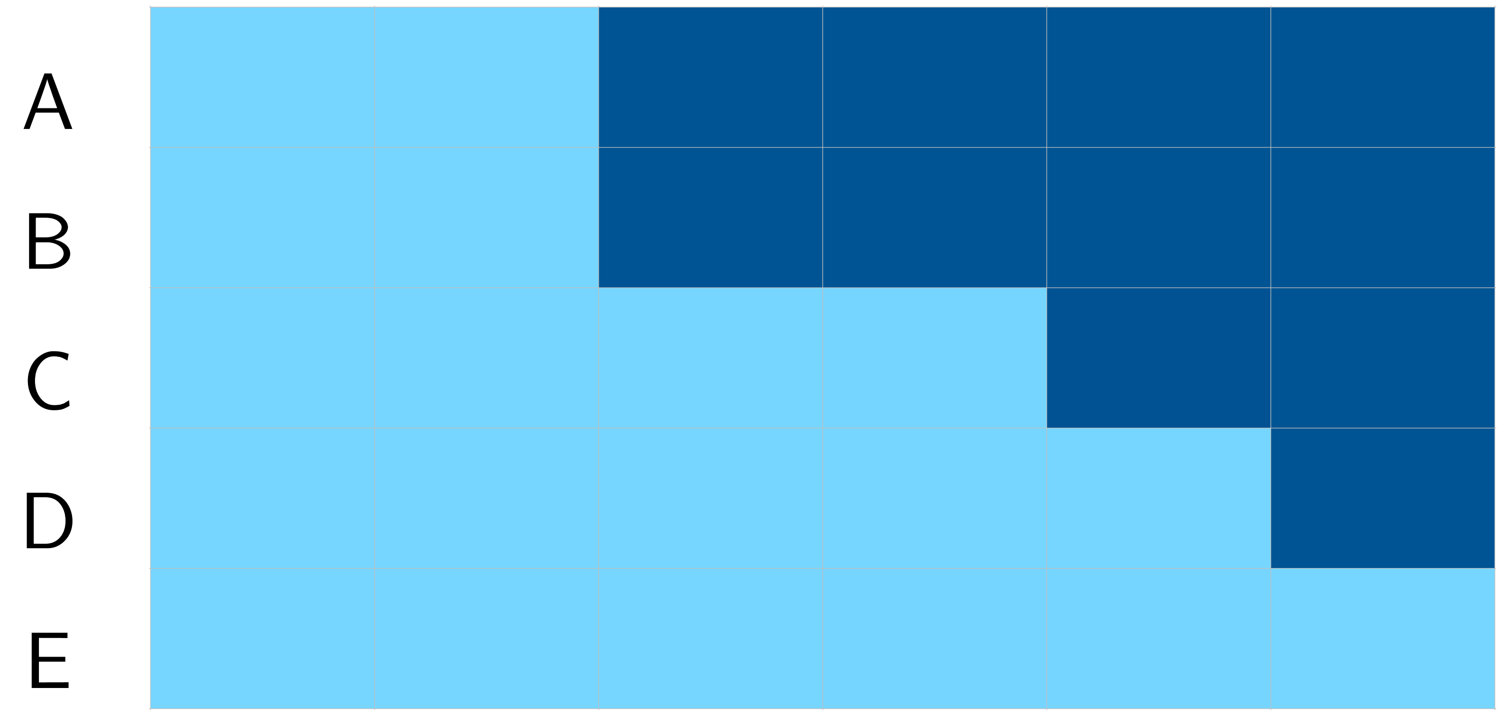
Sun & Abraham (2021)




Callaway & Sant'Anna (2021)

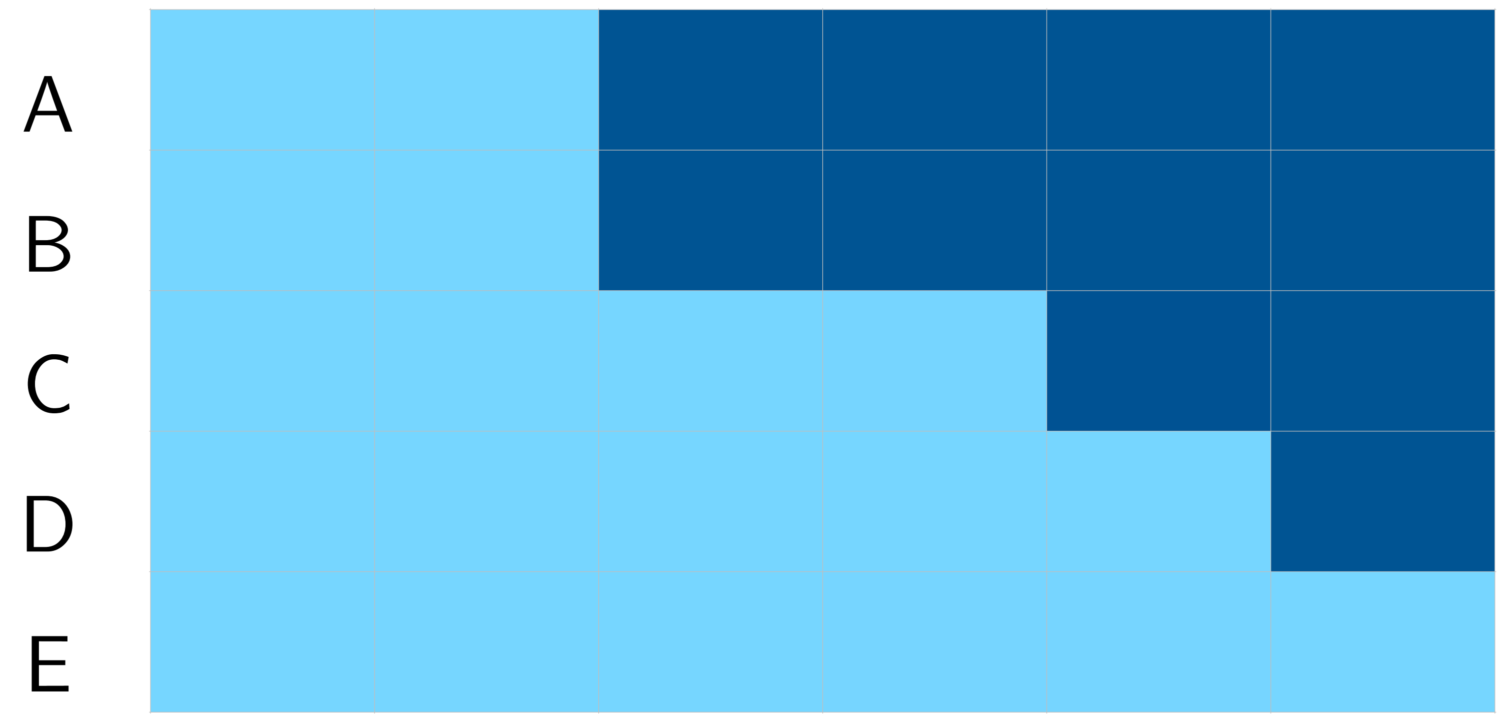


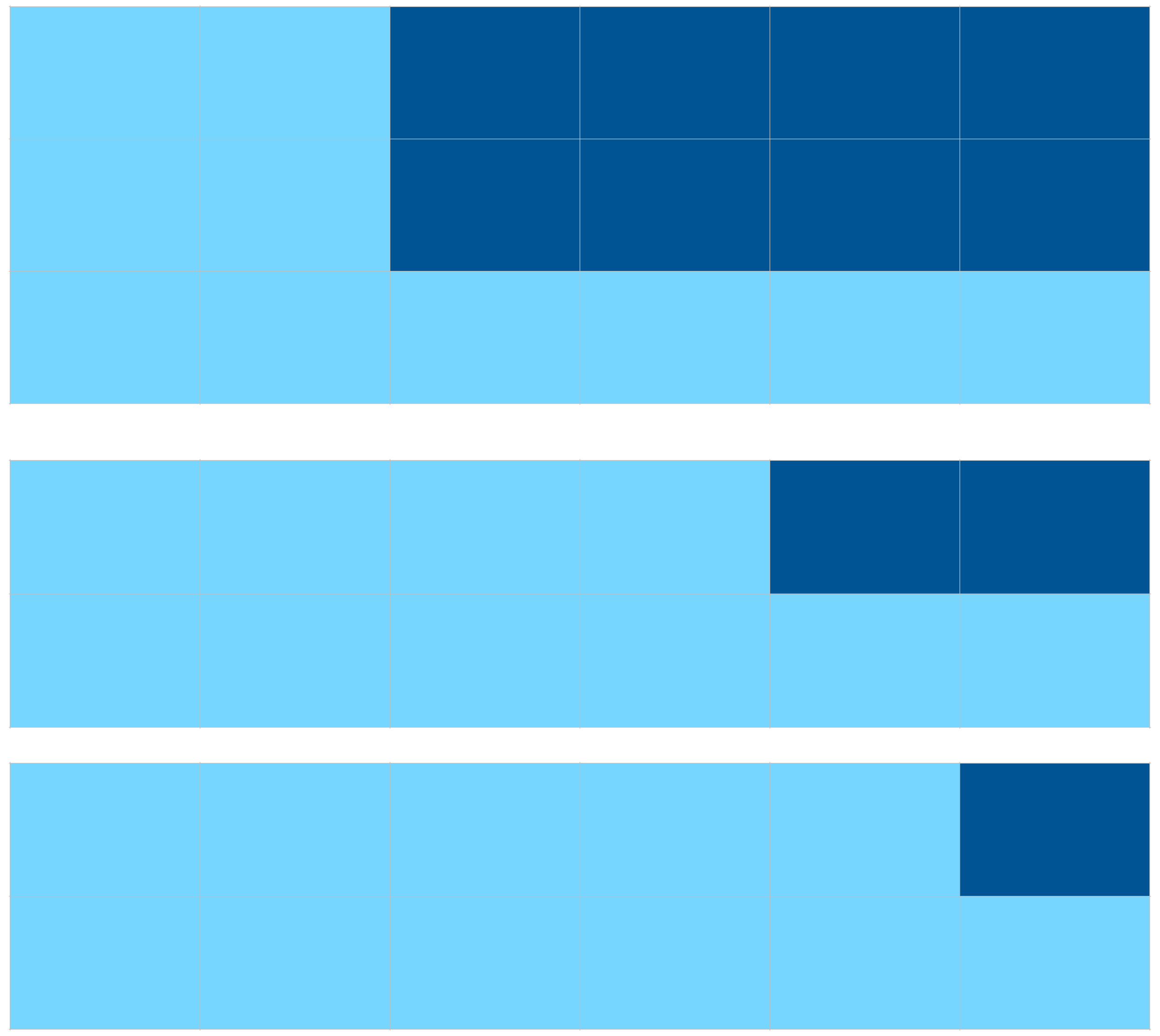
# Stacked DID: Cengiz et al. (2019)



# Stacked DID: Cengiz et al. (2019)

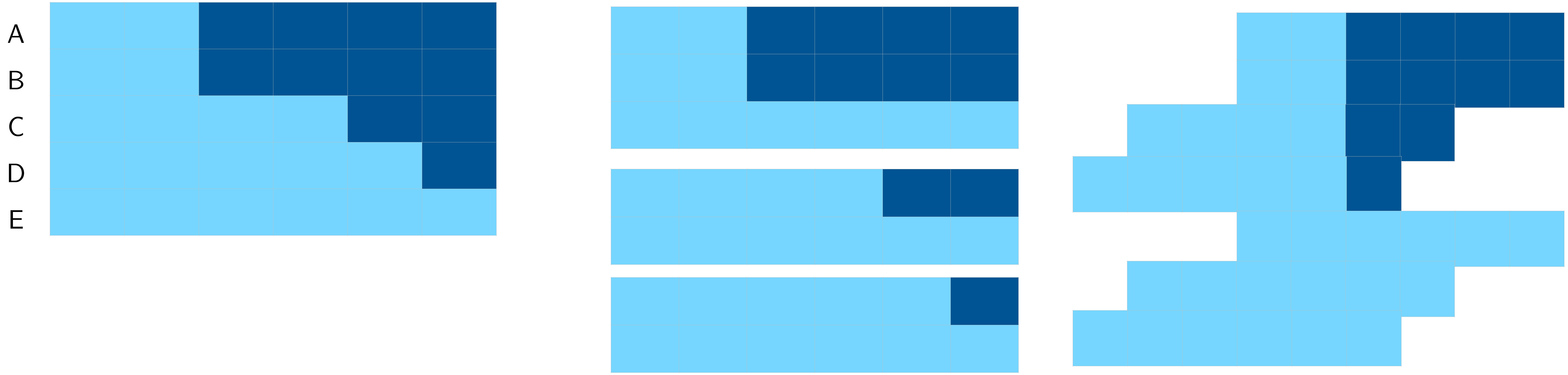
- Duplicate the pure control group for each cohort



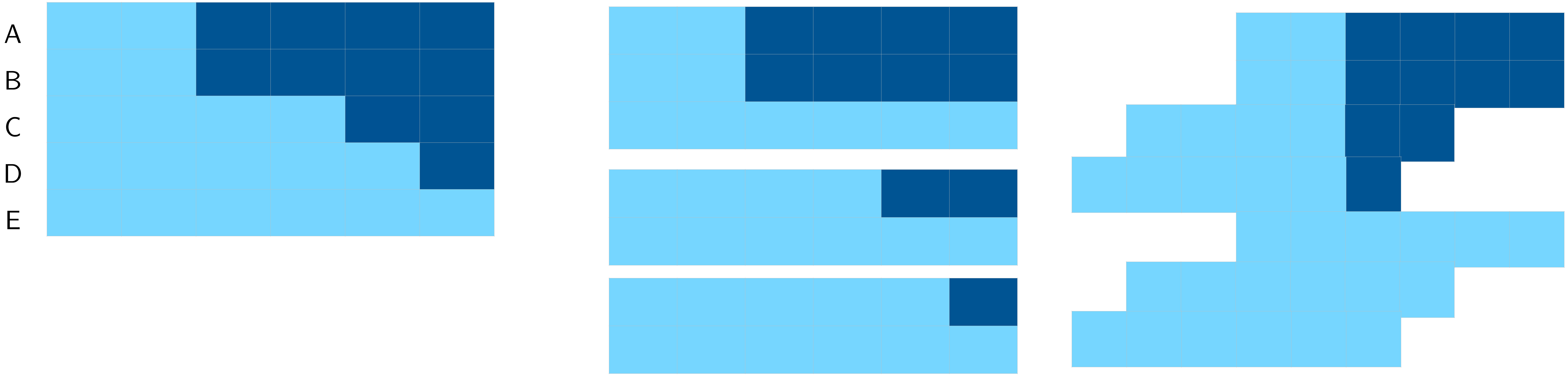
# Stacked DID: Cengiz et al. (2019)

- Duplicate the pure control group for each cohort
- "Stack" on top of each other, align by relative time to treatment onset



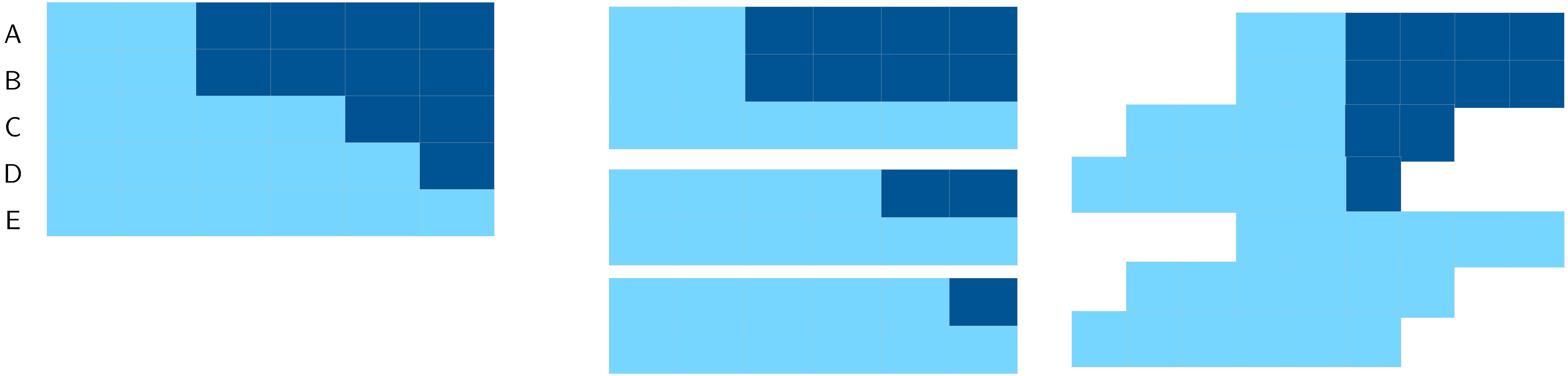
# Stacked DID: Cengiz et al. (2019)

- Duplicate the pure control group for each cohort
- "Stack" on top of each other, align by relative time to treatment onset
- Run saturated regression

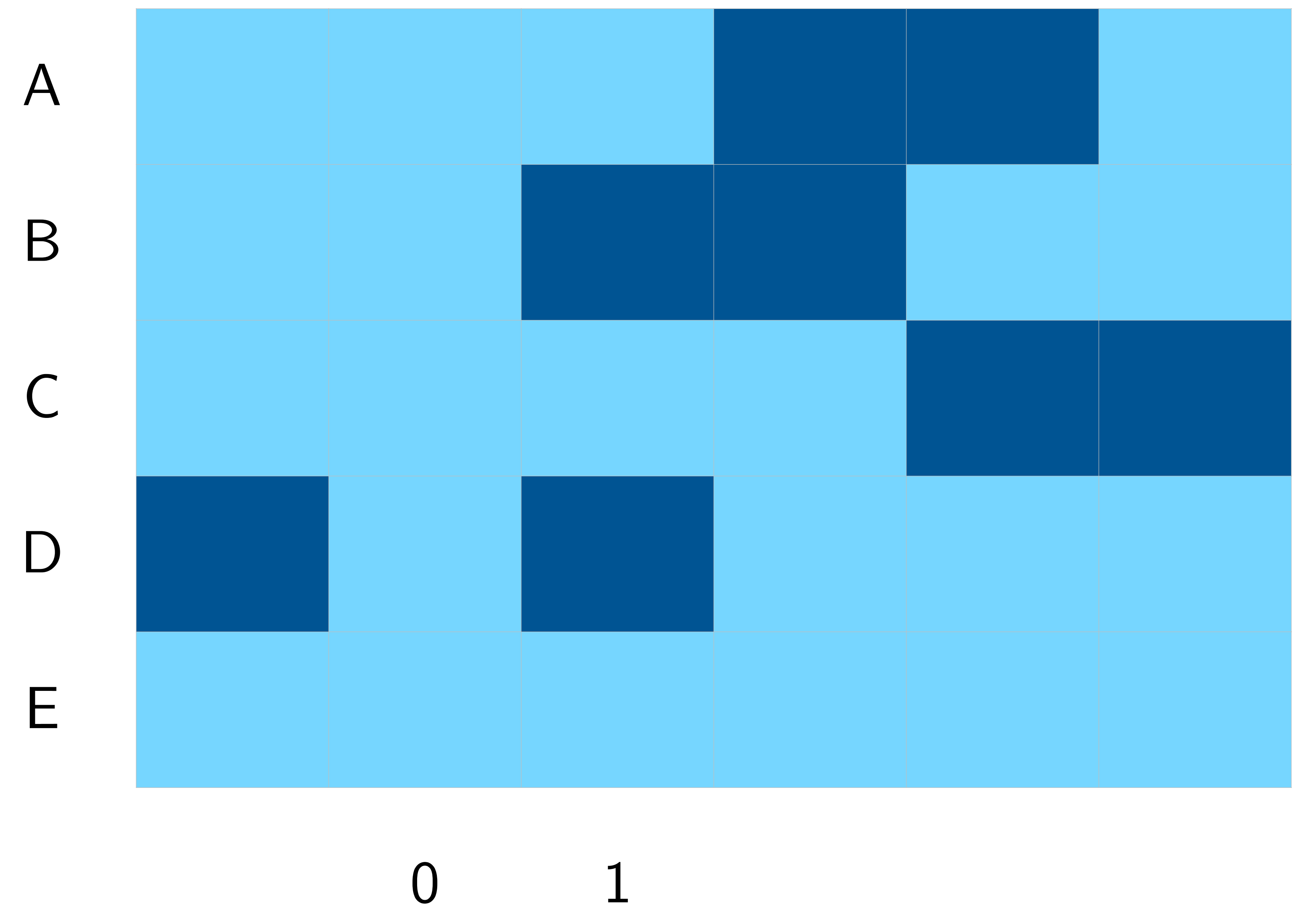


# Stacked DID: Cengiz et al. (2019)

- Duplicate the pure control group for each cohort
- "Stack" on top of each other, align by relative time to treatment onset
- Run saturated regression
- Similar to IW with disproportionate weights

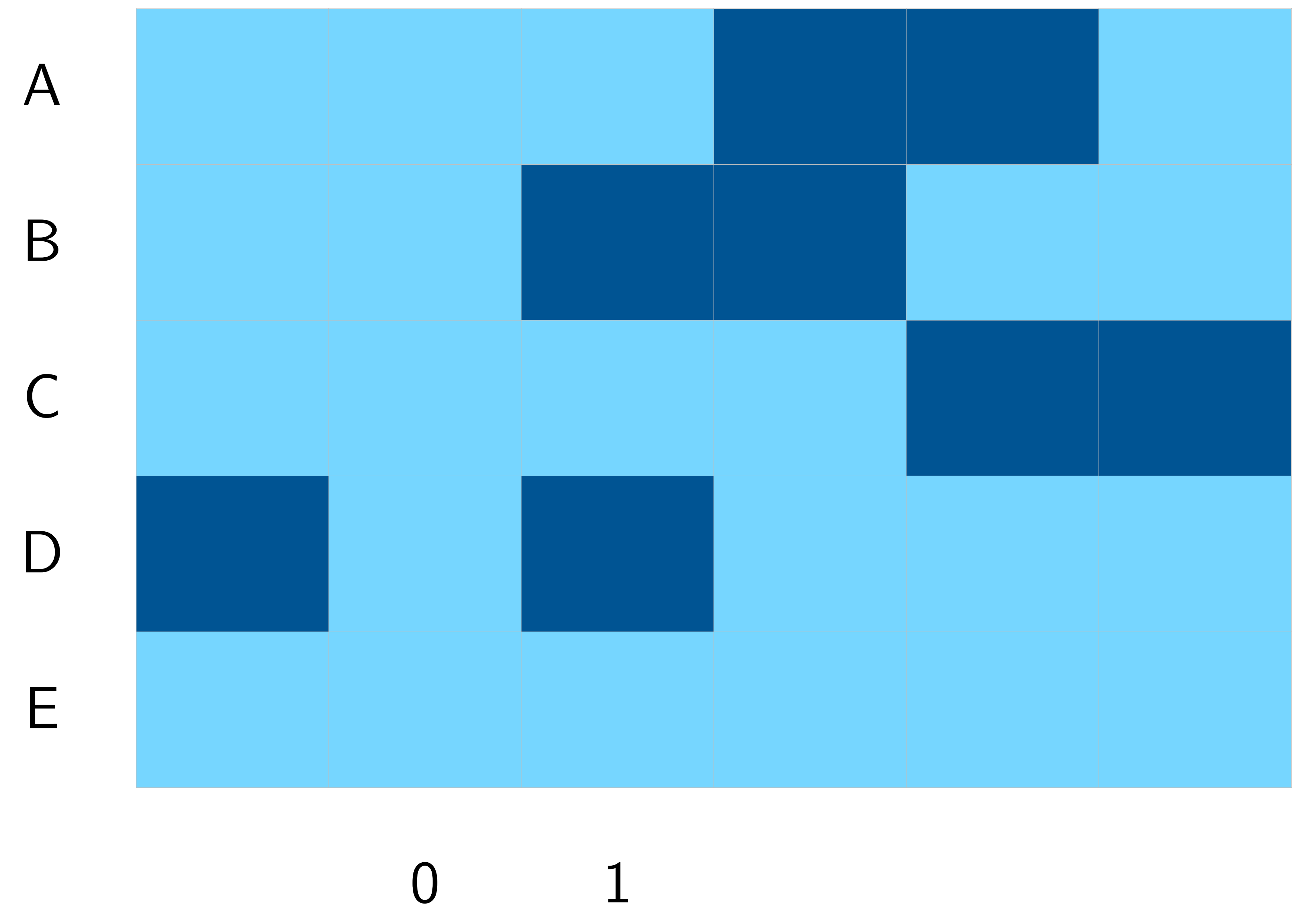


# De Chaisemartin and D'Haultfœuille (2020)



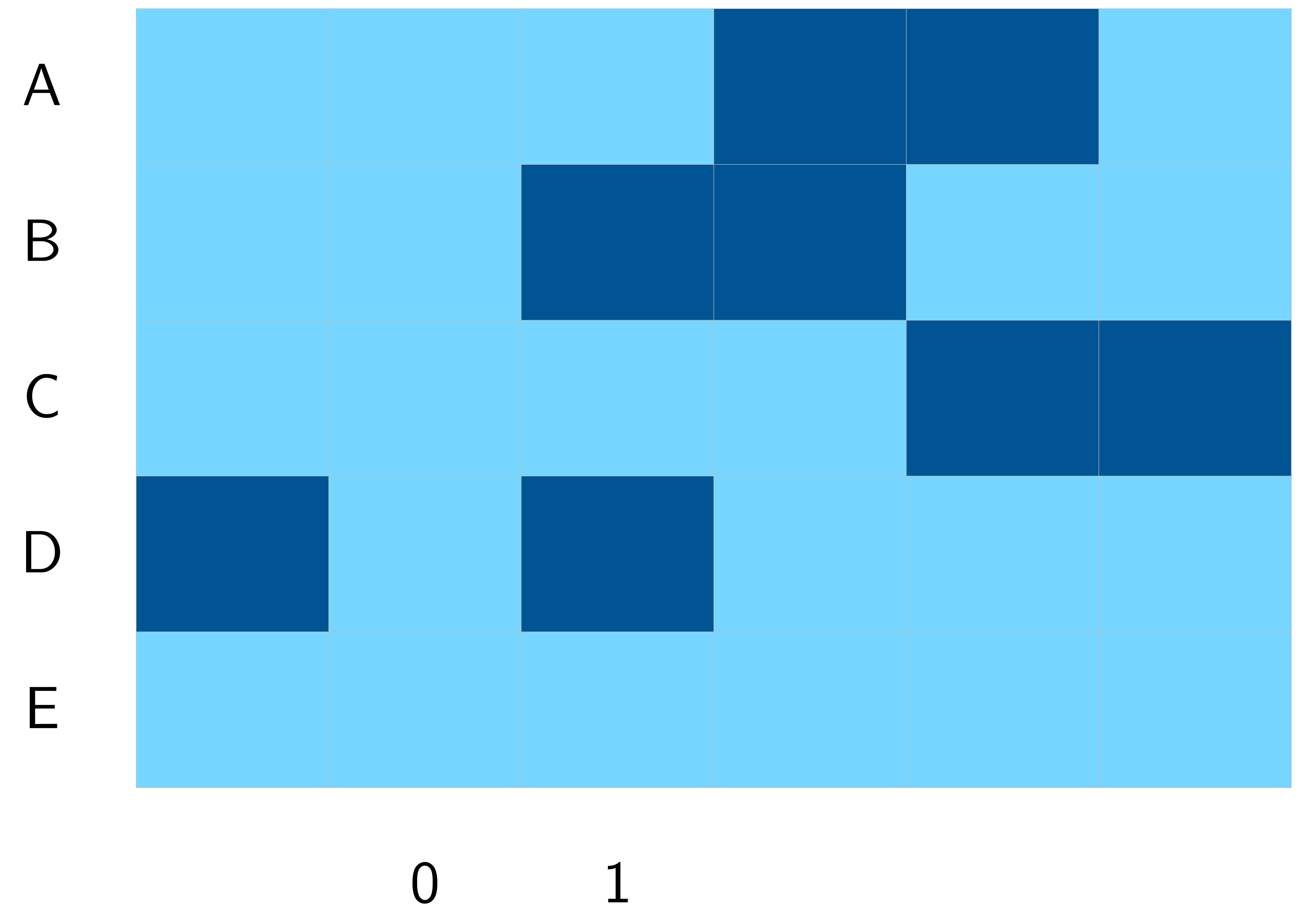
# De Chaisemartin and D'Haultfœuille (2020)

- No cohorts — estimates a single average effect



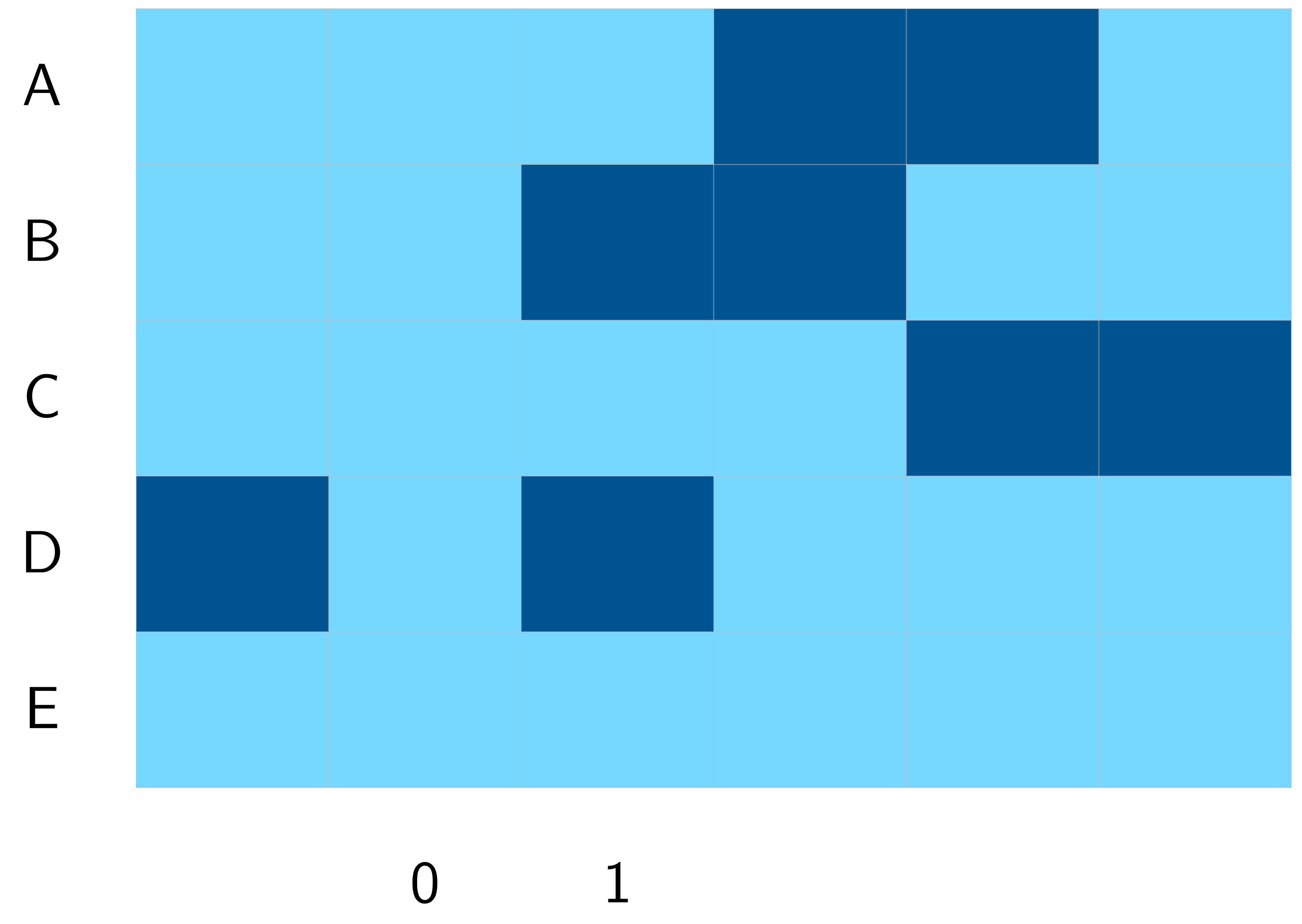
# De Chaisemartin and D'Haultfœuille (2020)

- No cohorts — estimates a single average effect
- Effect for switchers (not ATT)



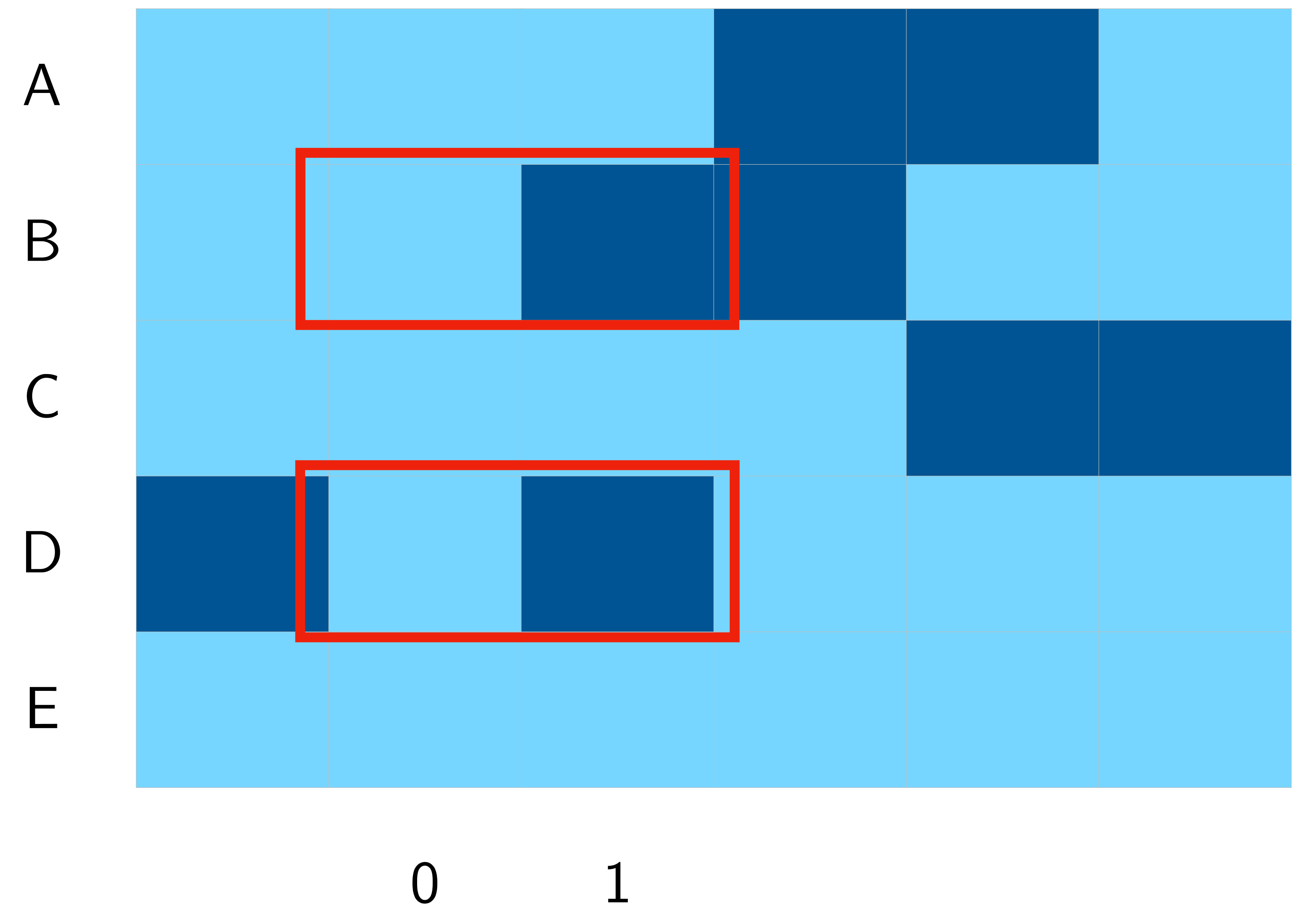
# De Chaisemartin and D'Haultfœuille (2020)

- No cohorts — estimates a single average effect
- Effect for switchers (not ATT)
- Match treated to control with shared treatment status in previous period



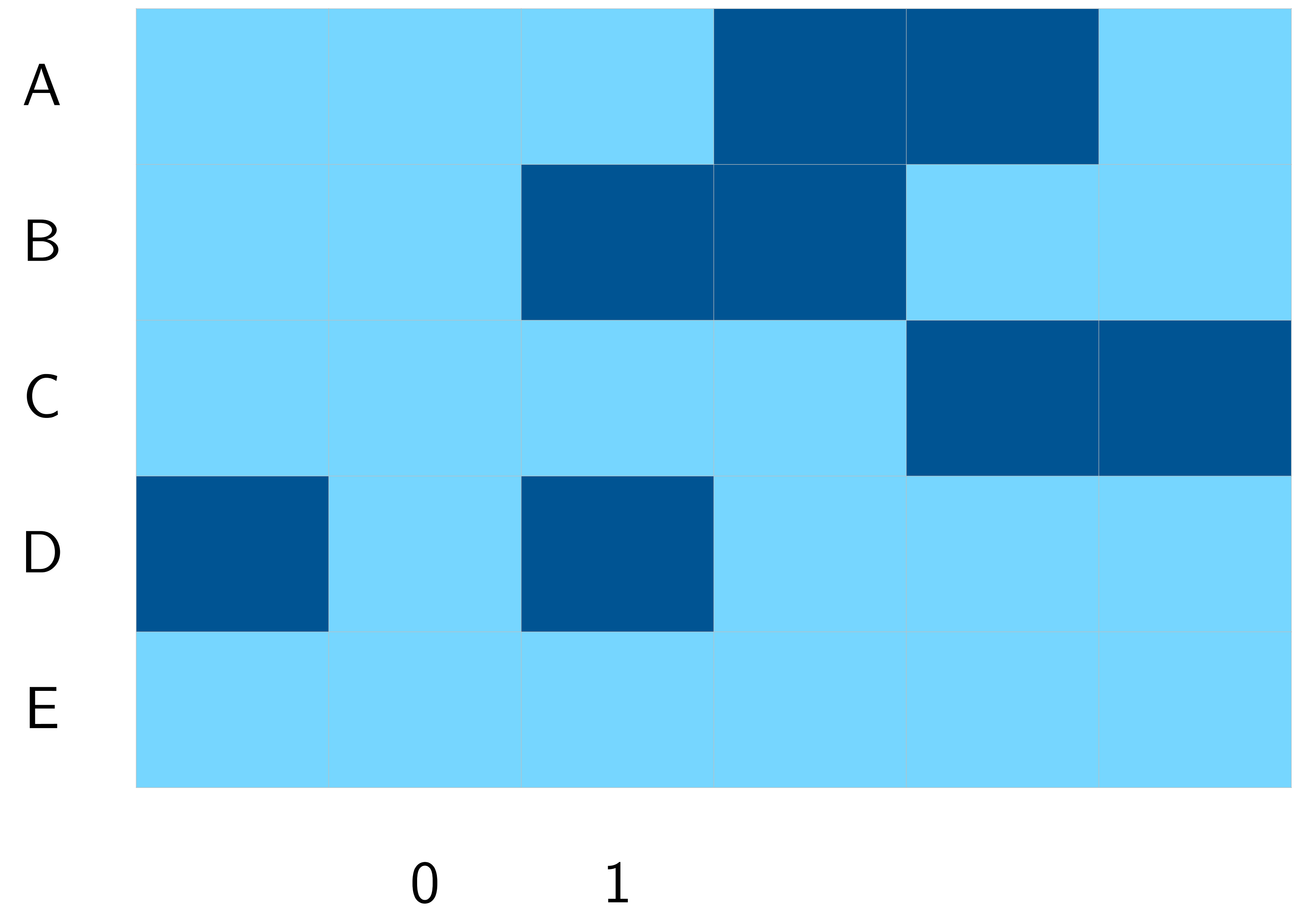
# De Chaisemartin and D'Haultfœuille (2020)

- No cohorts — estimates a single average effect
- Effect for switchers (not ATT)
- Match treated to control with shared treatment status in previous period
  - Switchers $(i, t) : D_{it} \neq D_{it-1}$



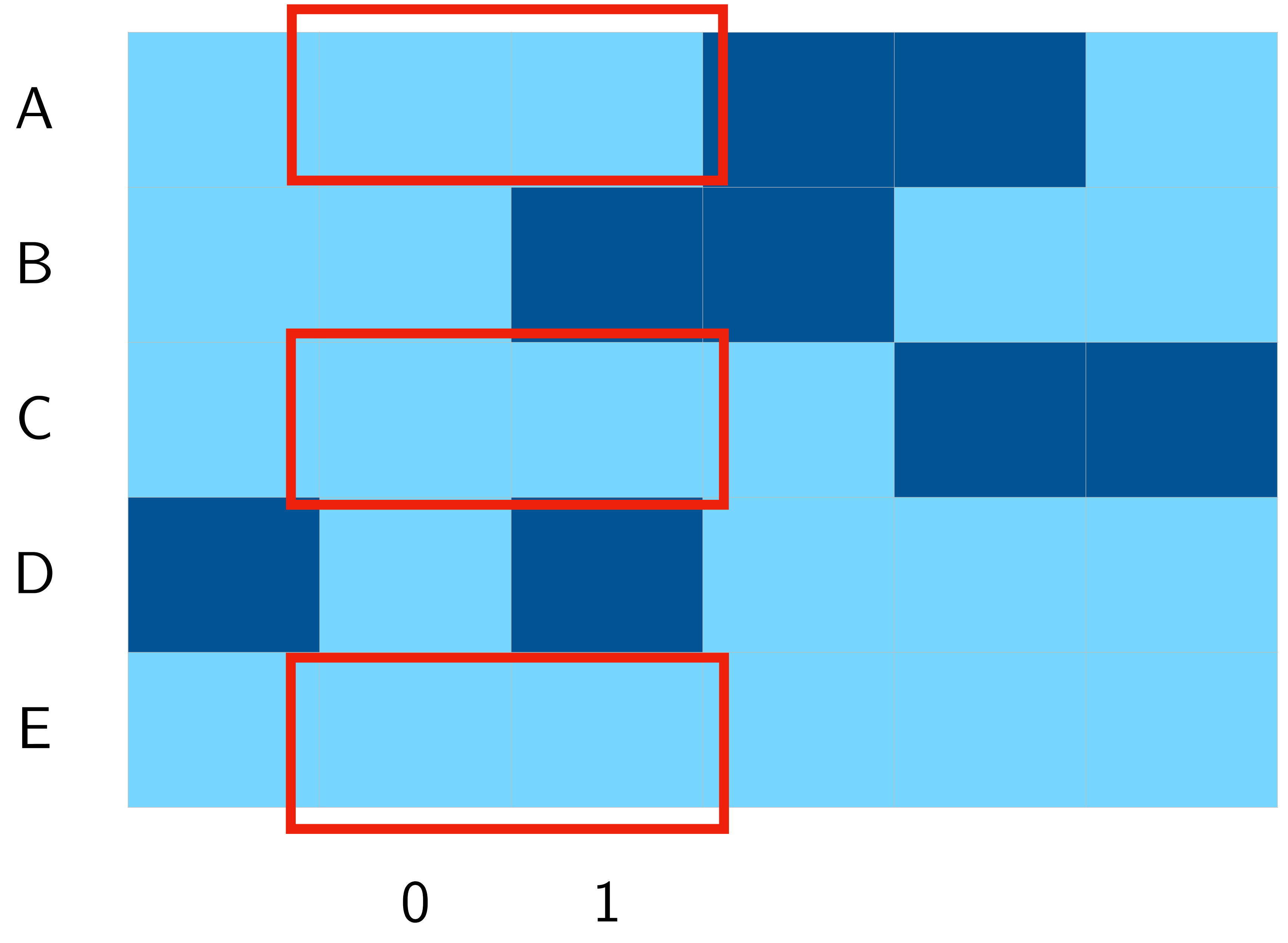
# De Chaisemartin and D'Haultfœuille (2020)

- No cohorts — estimates a single average effect
- Effect for switchers (not ATT)
- Match treated to control with shared treatment status in previous period
  - Switchers $(i,t) : D_{it} \neq D_{it-1}$



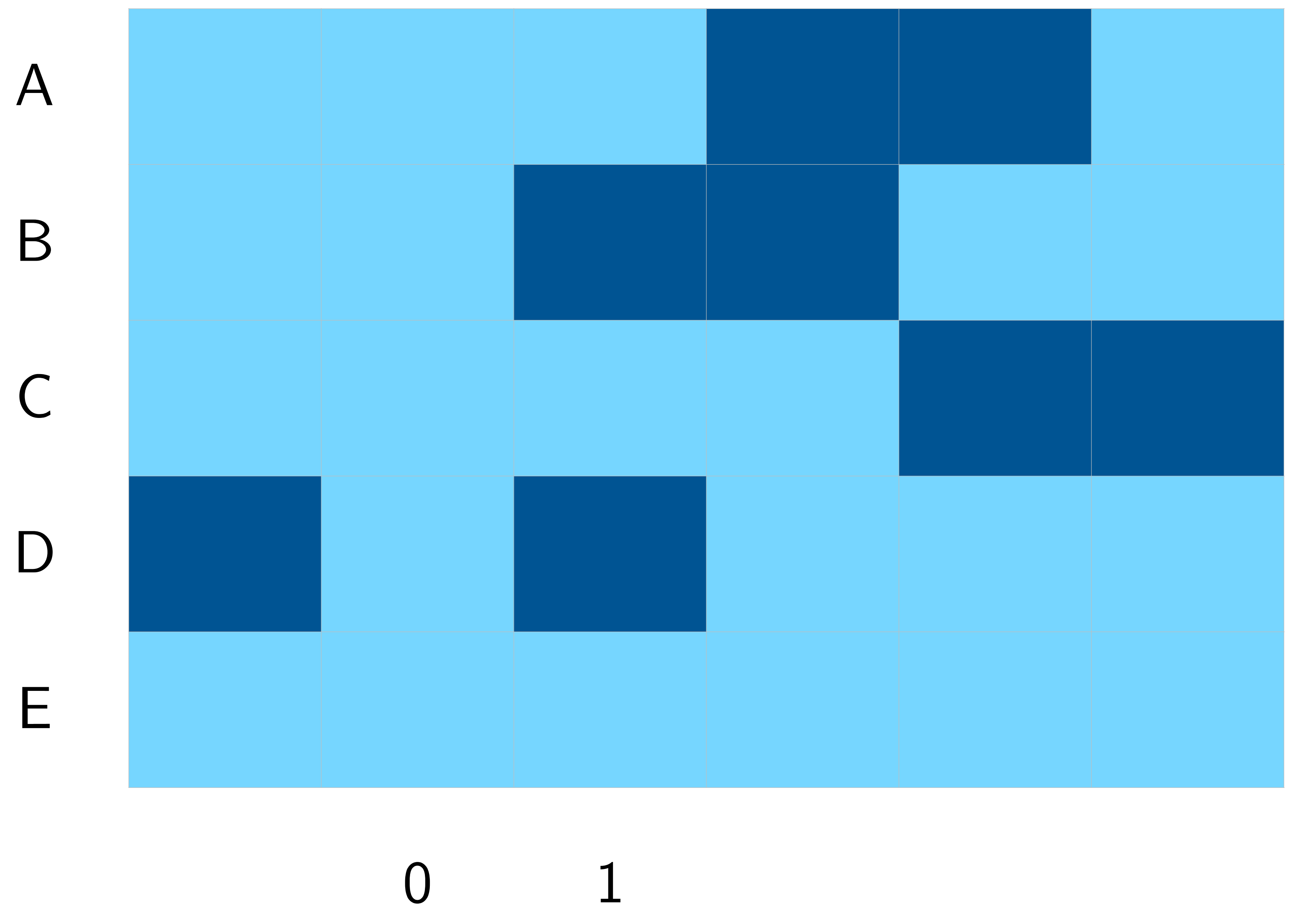
# De Chaisemartin and D'Haultfœuille (2020)

- No cohorts — estimates a single average effect
- Effect for switchers (not ATT)
- Match treated to control with shared treatment status in previous period
  - Switchers $(i, t) : D_{it} \neq D_{it-1}$
  - Stable group $(i, t) : D_{it} = D_{it}$



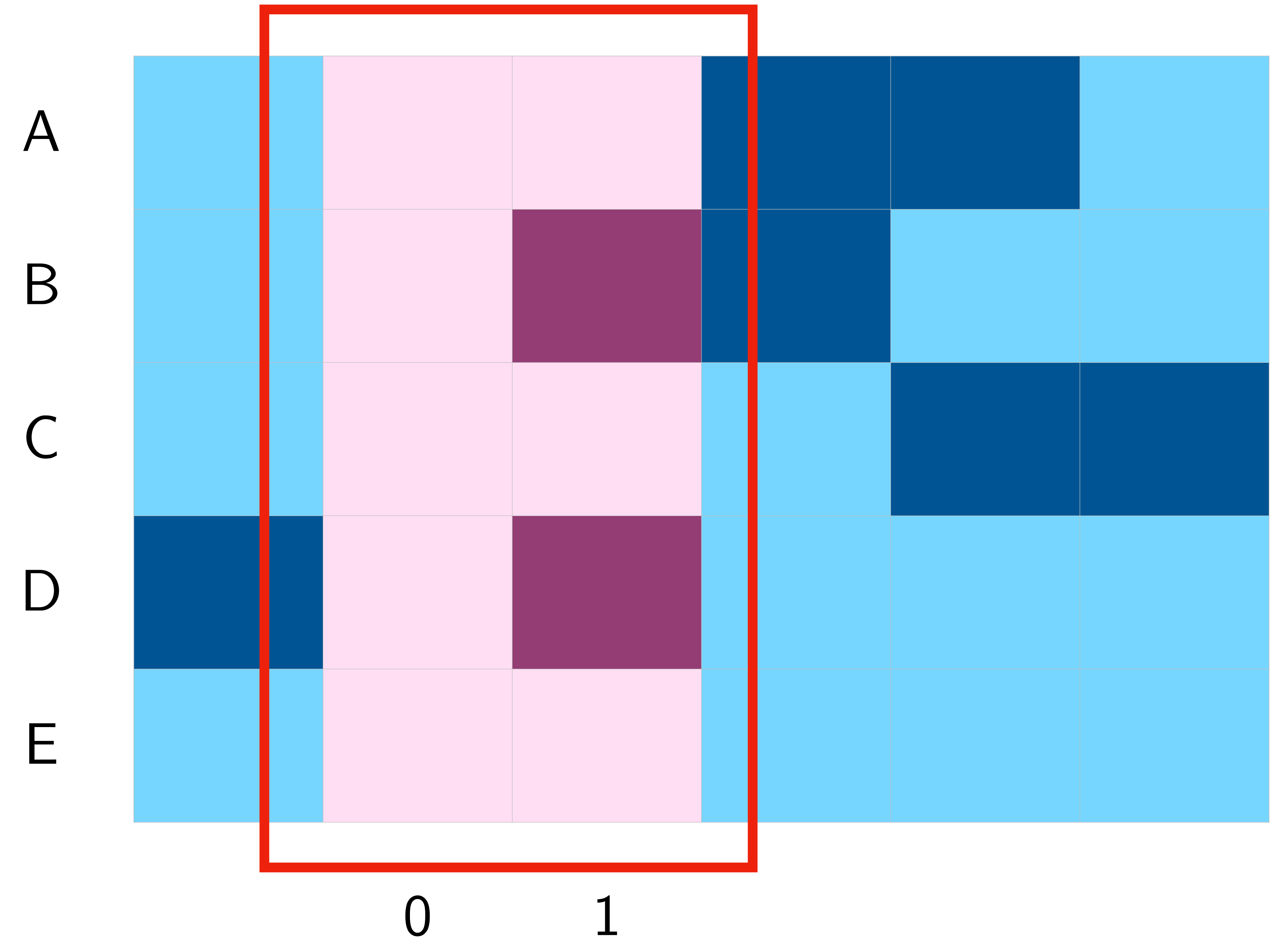
# De Chaisemartin and D'Haultfœuille (2020)

- No cohorts — estimates a single average effect
- Effect for switchers (not ATT)
- Match treated to control with shared treatment status in previous period
  - Switchers $(i, t) : D_{it} \neq D_{it-1}$
  - Stable group $(i, t) : D_{it} = D_{it}$



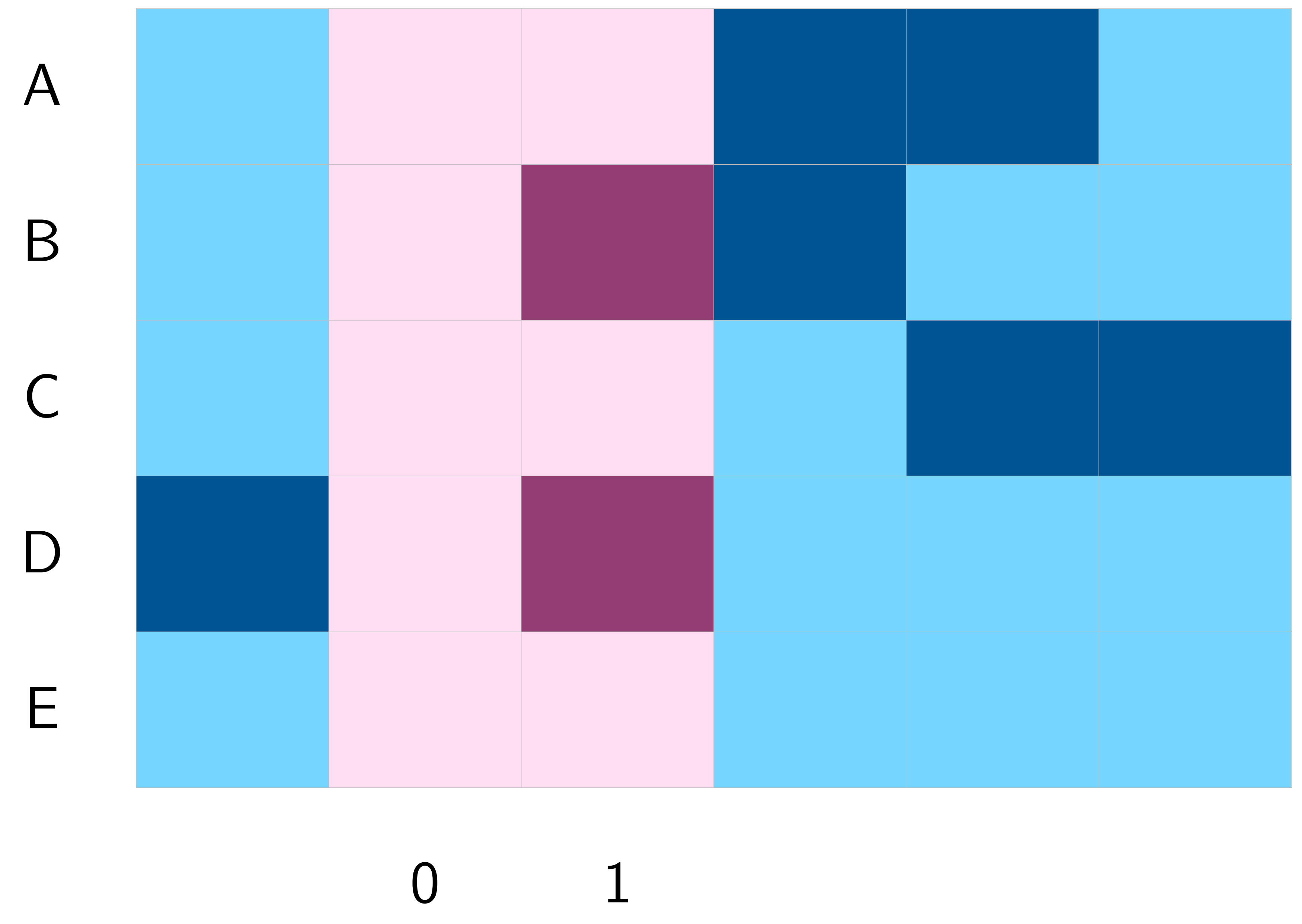
# De Chaisemartin and D'Haultfœuille (2020)

- No cohorts — estimates a single average effect
- Effect for switchers (not ATT)
- Match treated to control with shared treatment status in previous period
  - Switchers $(i, t) : D_{it} \neq D_{it-1}$
  - Stable group $(i, t) : D_{it} = D_{it}$
- $DID_M$: DID to estimate contemporaneous effect at period of switch

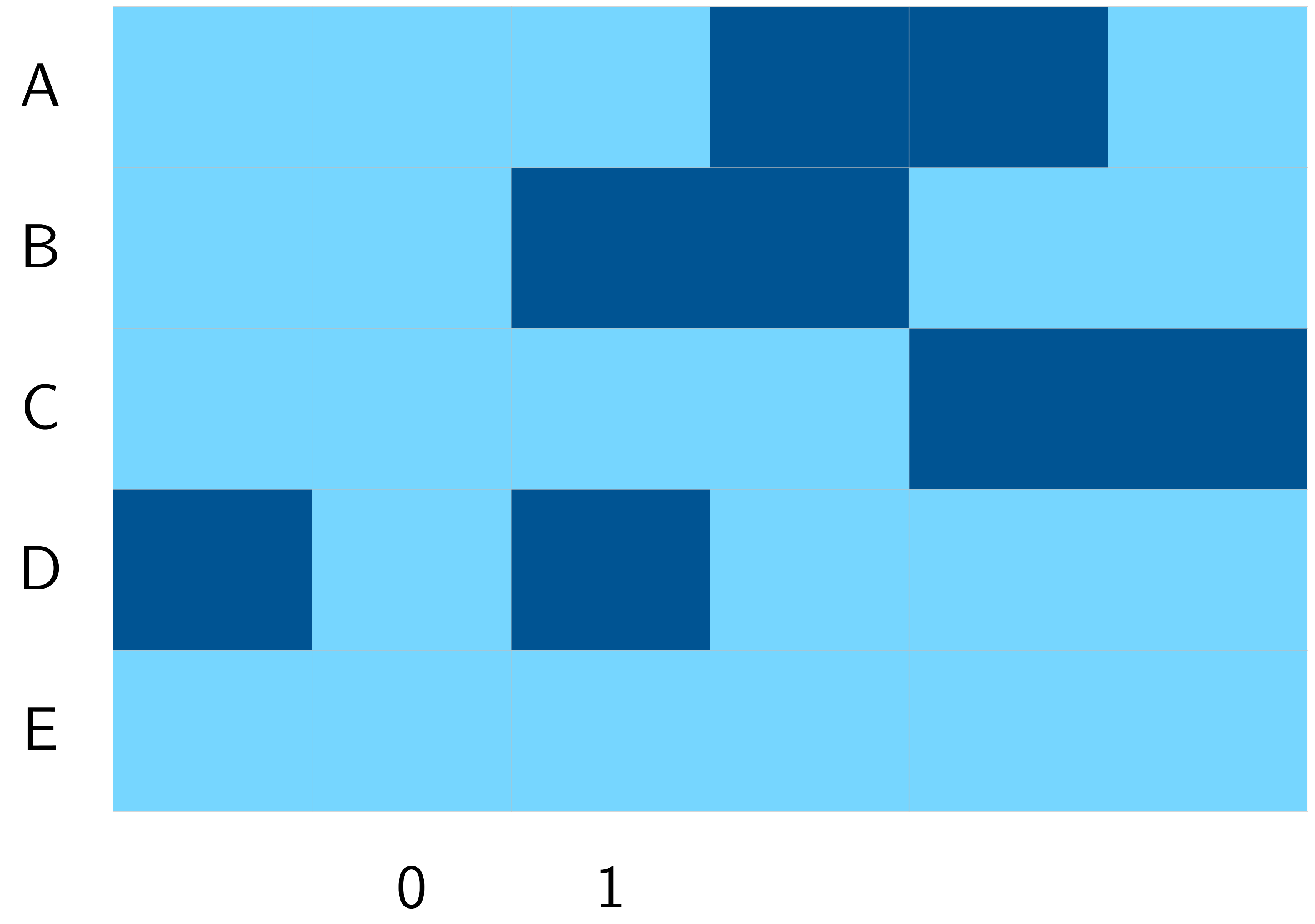


# De Chaisemartin and D'Haultfœuille (2020)

- No cohorts — estimates a single average effect
- Effect for switchers (not ATT)
- Match treated to control with shared treatment status in previous period
  - Switchers $(i, t) : D_{it} \neq D_{it-1}$
  - Stable group $(i, t) : D_{it} = D_{it}$
- $DID_M$: DID to estimate contemporaneous effect at period of switch

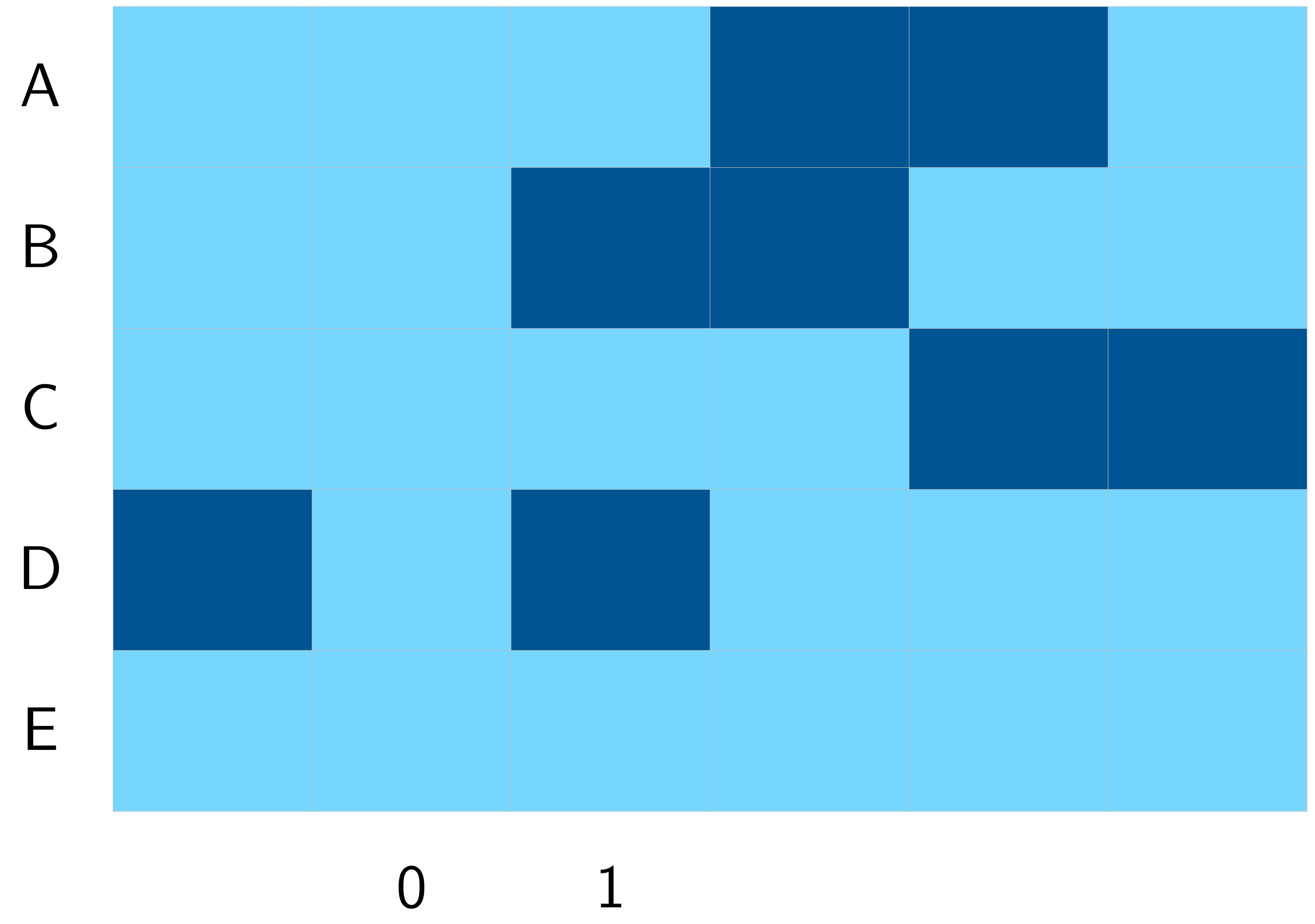


# Imai, Kim & Wang (2021) “PanelMatch”



# Imai, Kim & Wang (2021) "PanelMatch"

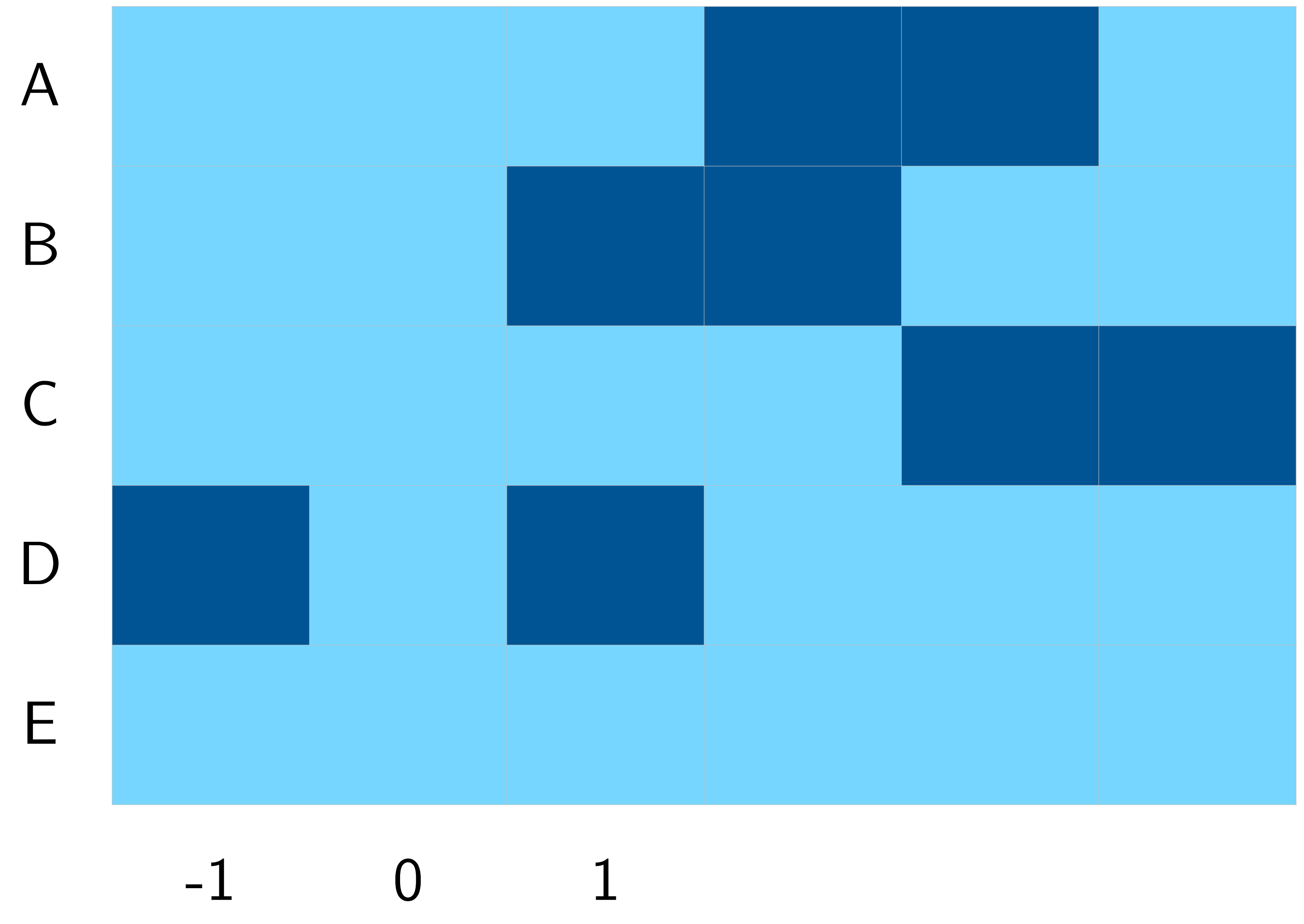
- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i,t)$ with $\left\{j : D_{is} = D_{js} \text{ for all } s \in \{t-1, t-2, \ldots, t-a\}\right\}$



# Imai, Kim & Wang (2021) "PanelMatch"

- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i, t)$ with $\{j : D_{is} = D_{js}$ for all $s \in \{t-1, t-2, \ldots, t-a\}\}$

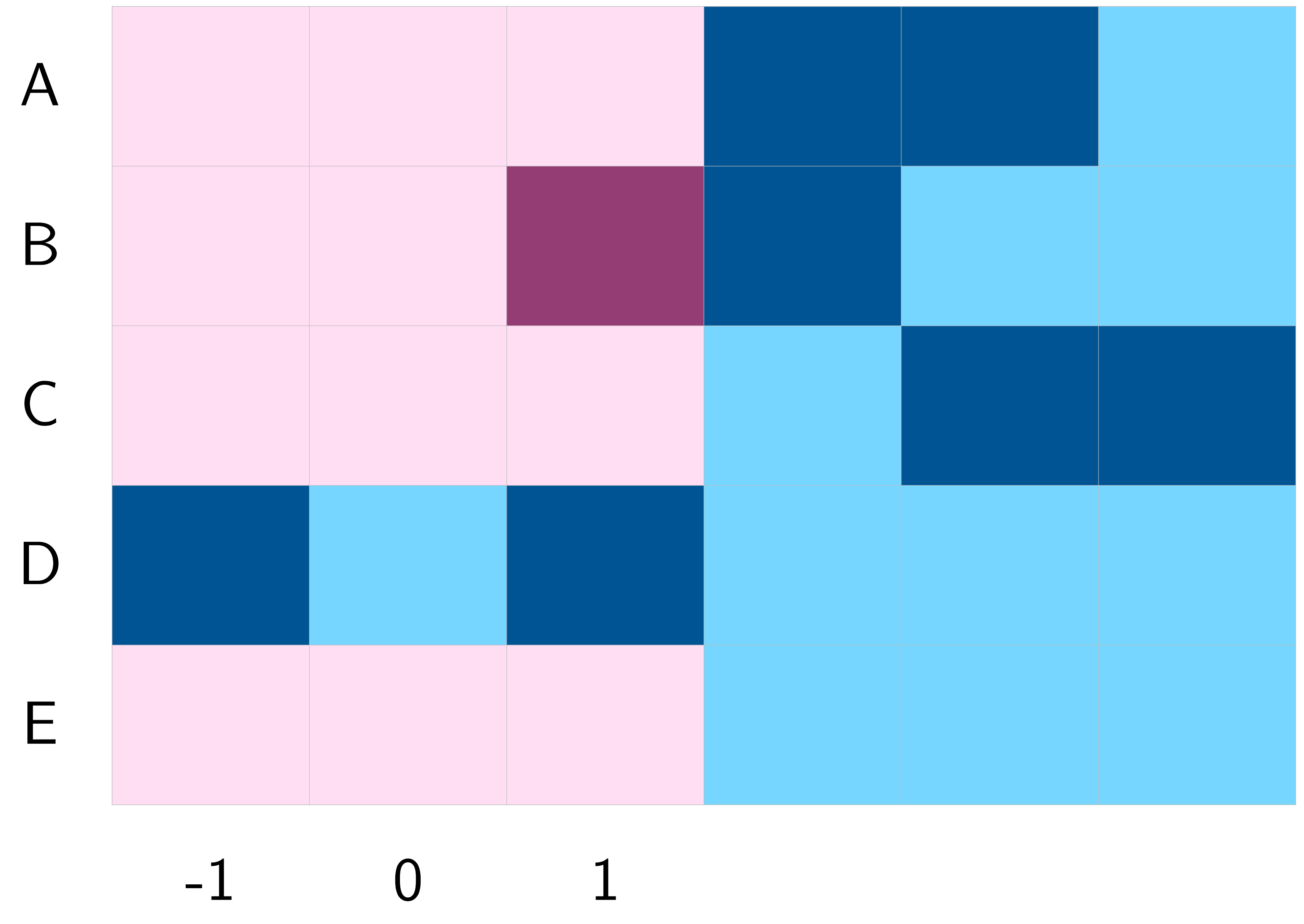
e.g., $a = 2$



# Imai, Kim & Wang (2021) "PanelMatch"

- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i, t)$ with $\left\{j : D_{is} = D_{js} \text{ for all } s \in \{t-1, t-2, \ldots, t-a\}\right\}$

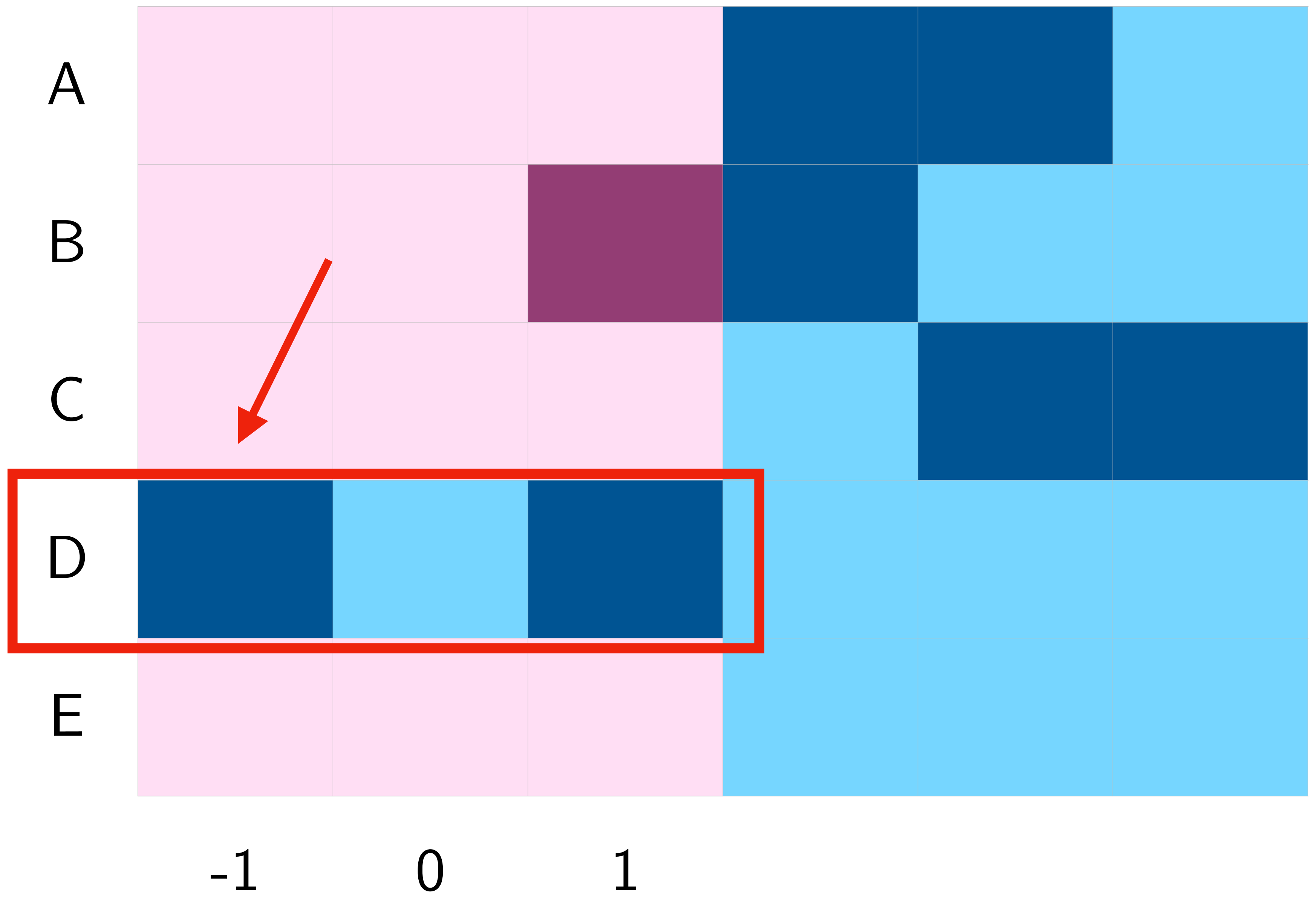
e.g., $a = 2$



# Imai, Kim & Wang (2021) "PanelMatch"

- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i, t)$ with $\left\{j : D_{is} = D_{js} \text{ for all } s \in \{t-1, t-2, \ldots, t-a\}\right\}$

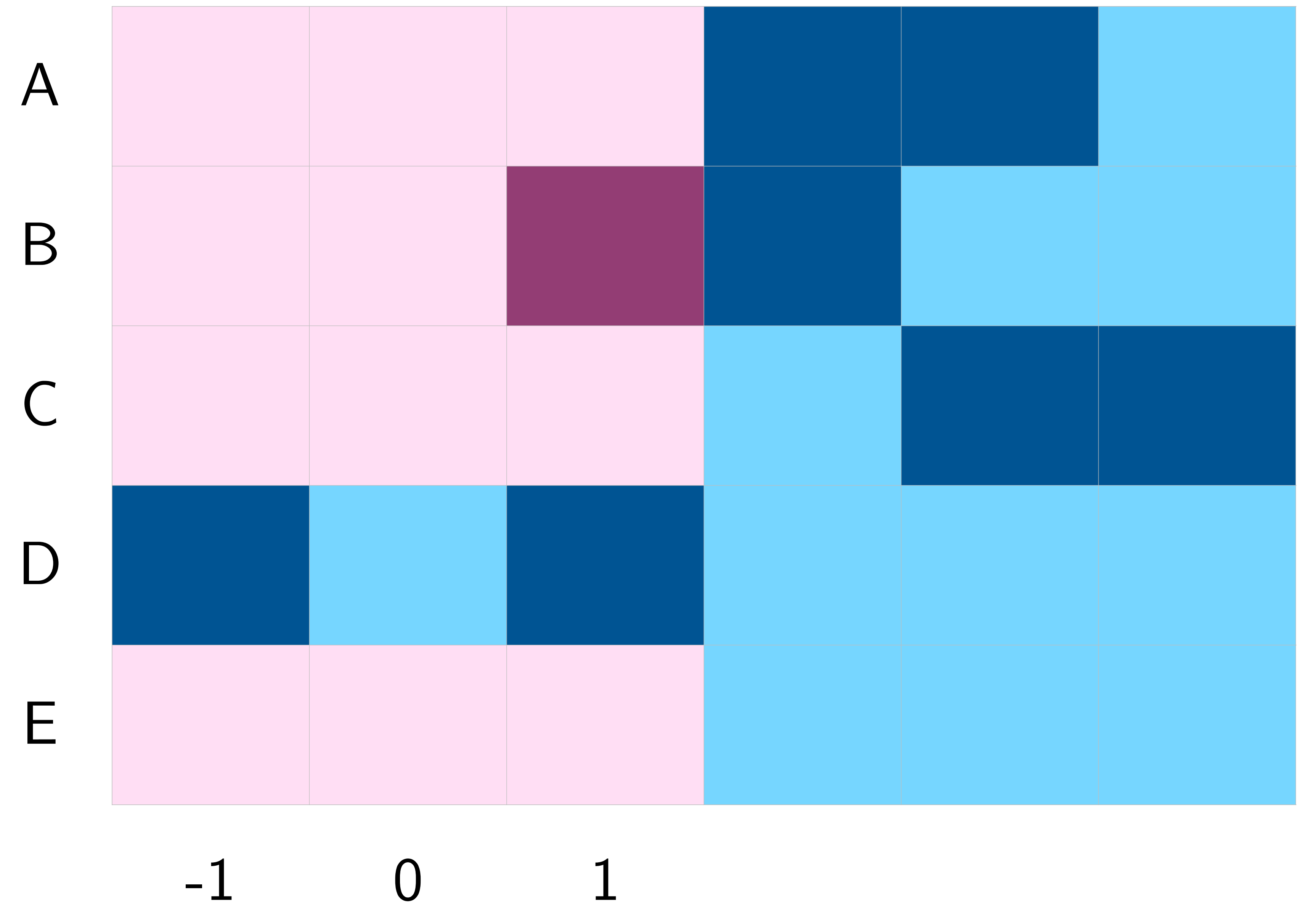
e.g., $a = 2$



# Imai, Kim & Wang (2021) "PanelMatch"

- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i,t)$ with $\left\{j : D_{is} = D_{js} \text{ for all } s \in \{t-1, t-2, \ldots, t-a\}\right\}$

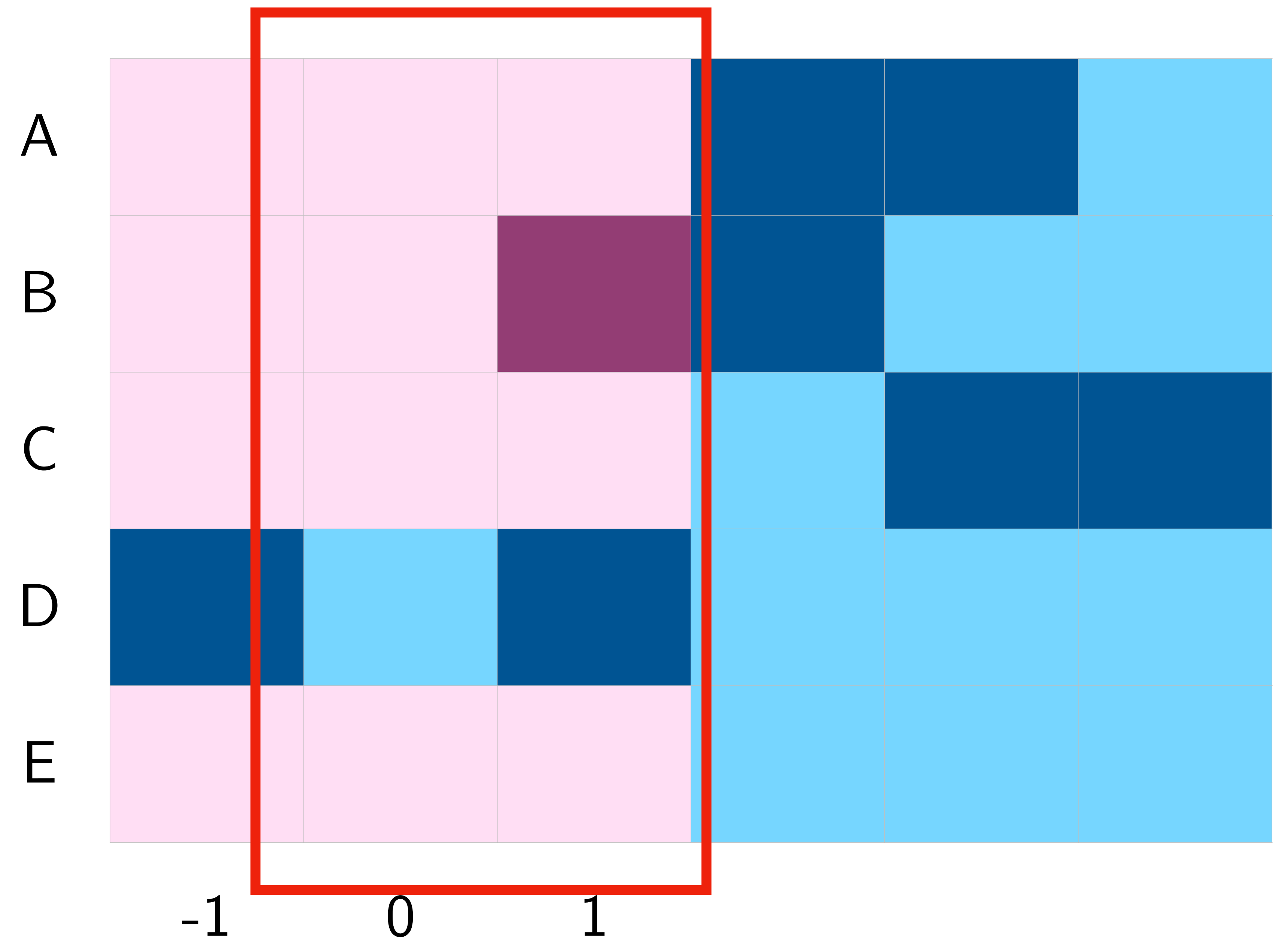
e.g., $a = 2$



# Imai, Kim & Wang (2021) "PanelMatch"

- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i,t)$ with $\{j : D_{is} = D_{js}$ for all $s \in \{t-1, t-2, \ldots, t-a\}\}$

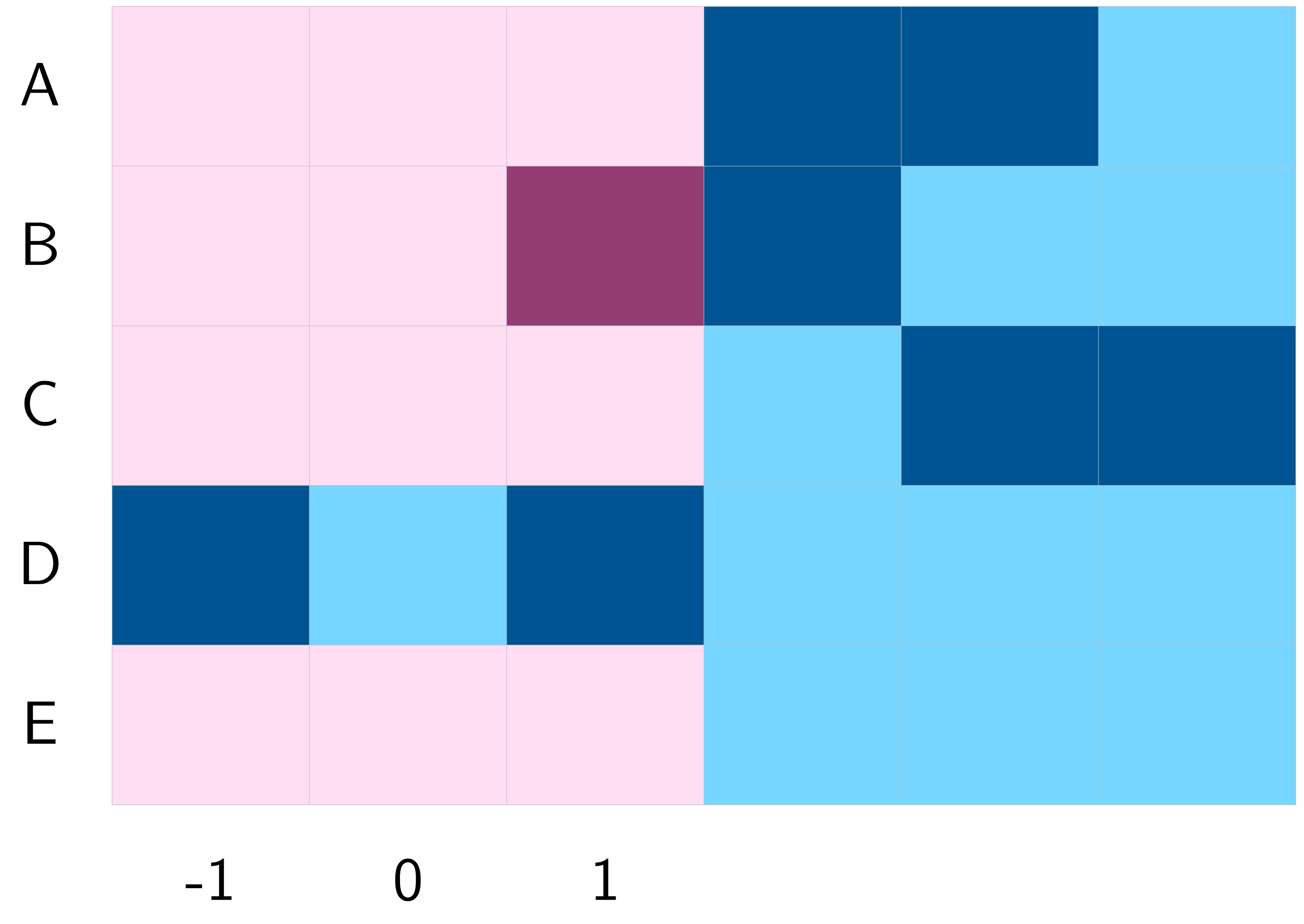
e.g., $a = 2$



# Imai, Kim & Wang (2021) "PanelMatch"

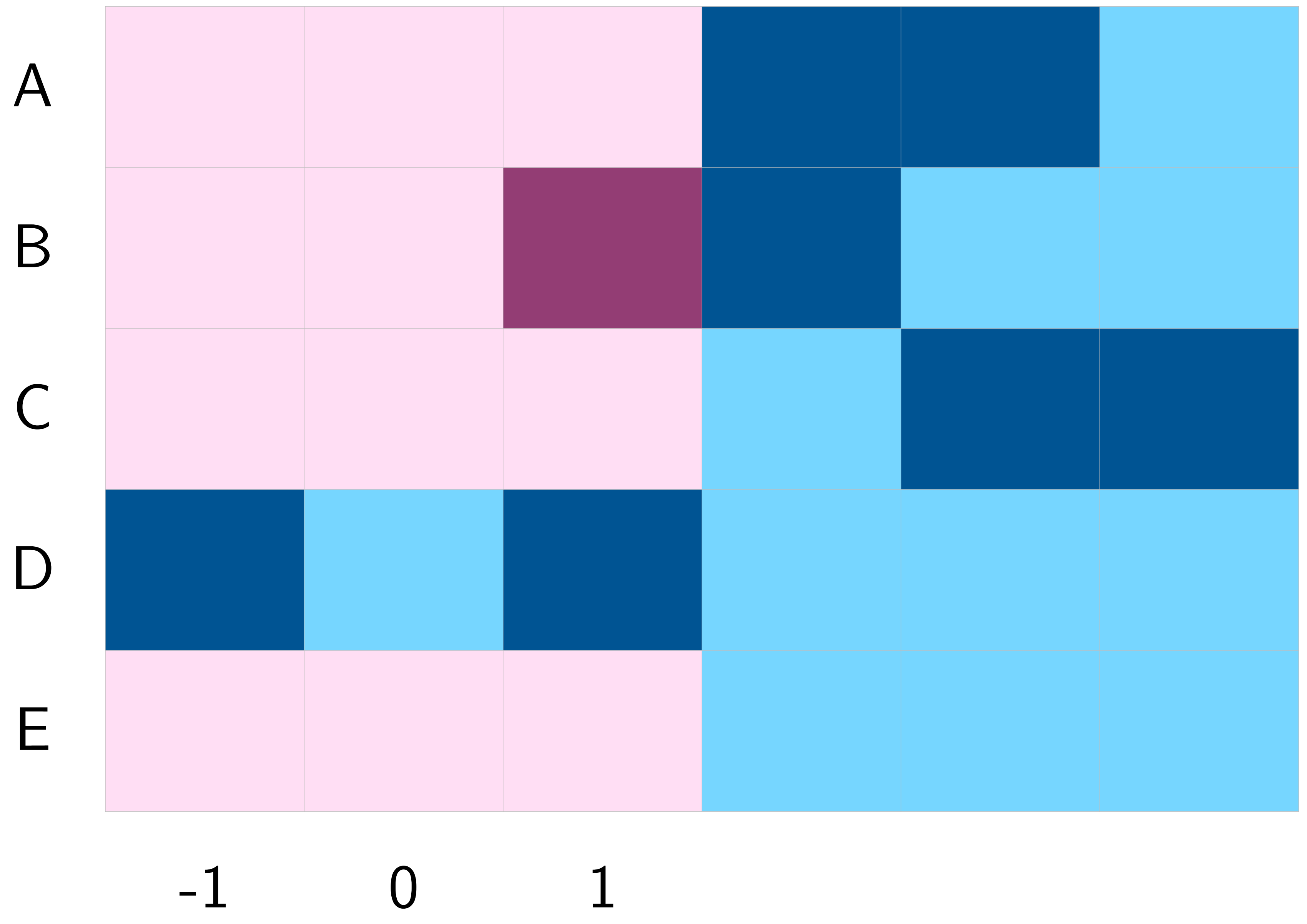
- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i, t)$ with $\{j : D_{is} = D_{js}$ for all $s \in \{t-1, t-2, \ldots, t-a\}\}$

e.g., $a = 2$



# Imai, Kim & Wang (2021) “PanelMatch”

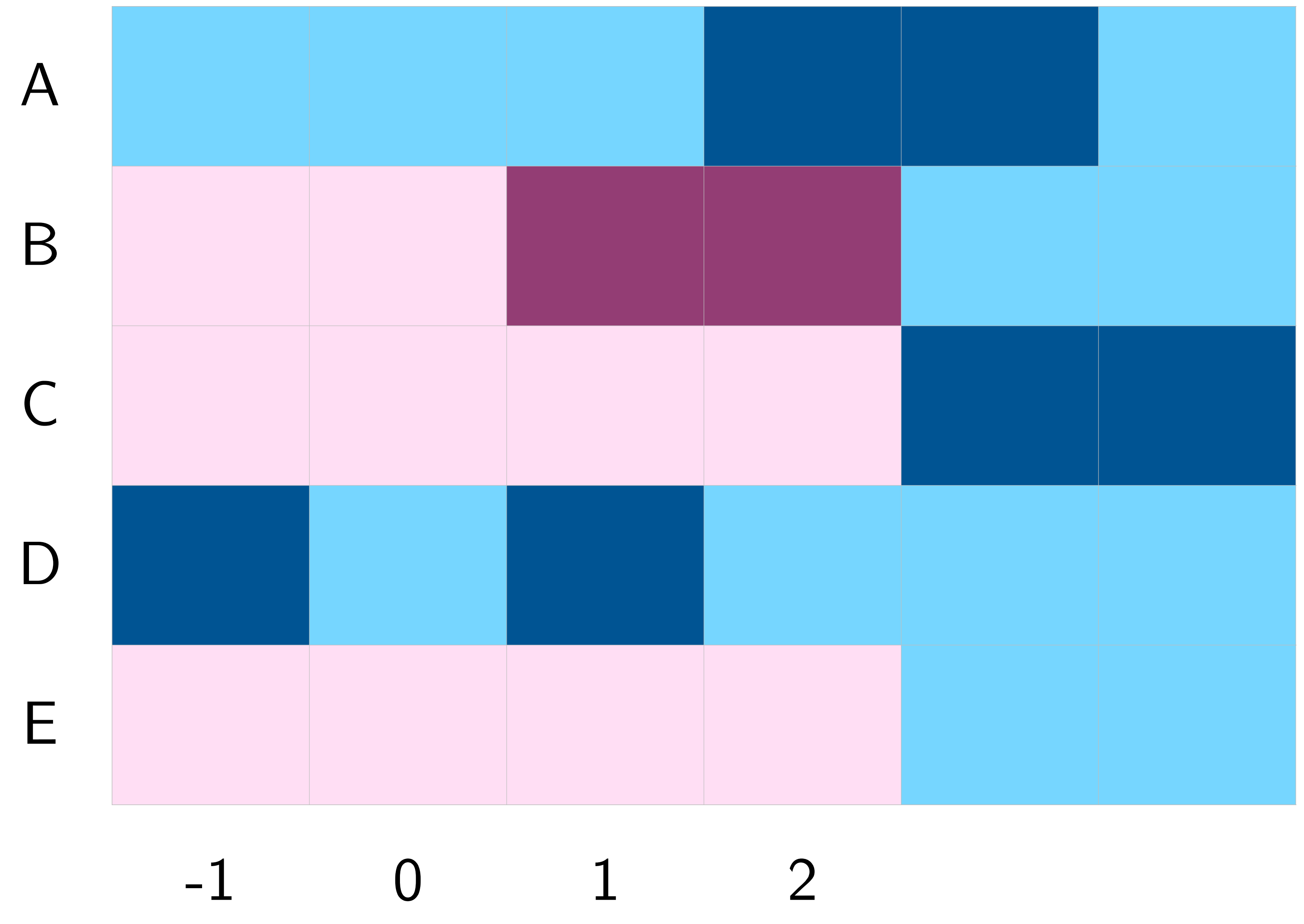
- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i, t)$ with $\{j : D_{is} = D_{js} \text{ for all } s \in \{t-1, t-2, \ldots, t-a\}\}$
- DID to estimate dynamic effects for future periods $l = 1, 2, \ldots$ (up to reversal)



# Imai, Kim & Wang (2021) "PanelMatch"

- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i, t)$ with $\{j : D_{is} = D_{js}$ for all $s \in \{t-1, t-2, \ldots, t-a\}\}$
- DID to estimate dynamic effects for future periods $l = 1,2,\ldots$ (up to reversal)

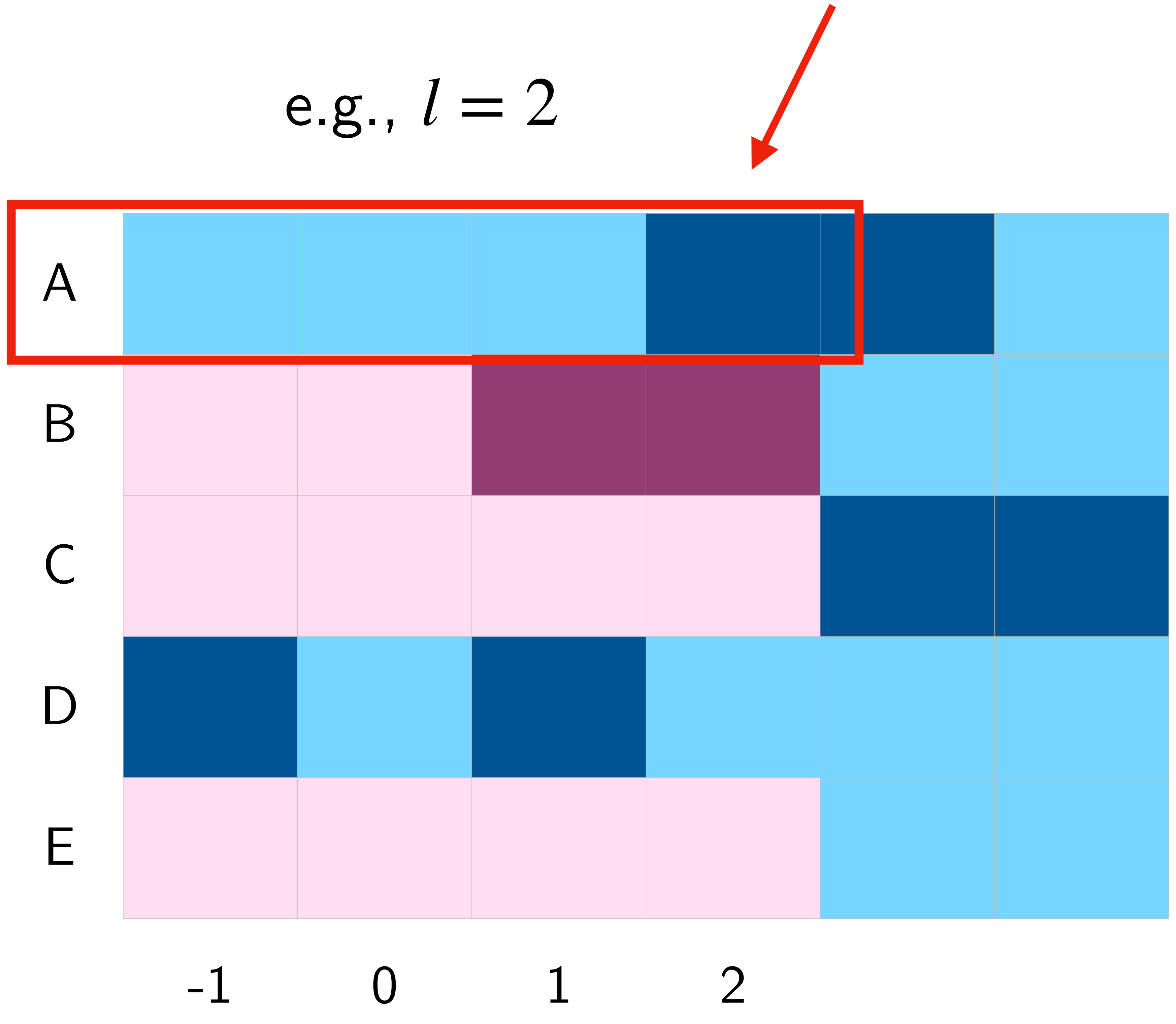
e.g., $l = 2$



# Imai, Kim & Wang (2021) "PanelMatch"

- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i, t)$ with $\{j : D_{is} = D_{js} \text{ for all } s \in \{t-1, t-2, \ldots, t-a\}\}$
- DID to estimate dynamic effects for future periods $l = 1,2,\ldots$ (up to reversal)

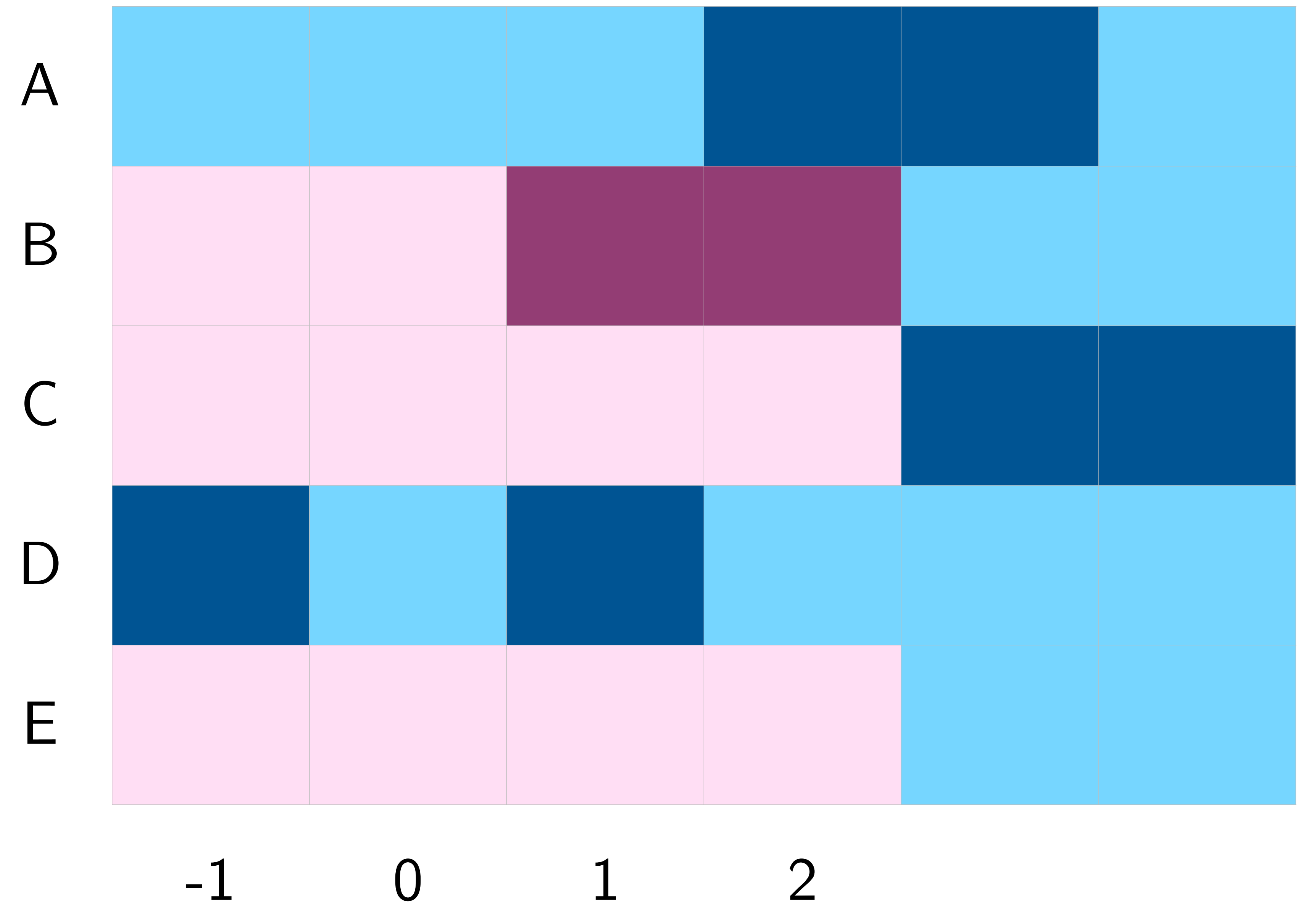
e.g., $l = 2$



# Imai, Kim & Wang (2021) "PanelMatch"

- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i,t)$ with $\{j : D_{is} = D_{js}$ for all $s \in \{t-1, t-2, \ldots, t-a\}\}$
- DID to estimate dynamic effects for future periods $l = 1,2,\ldots$ (up to reversal)

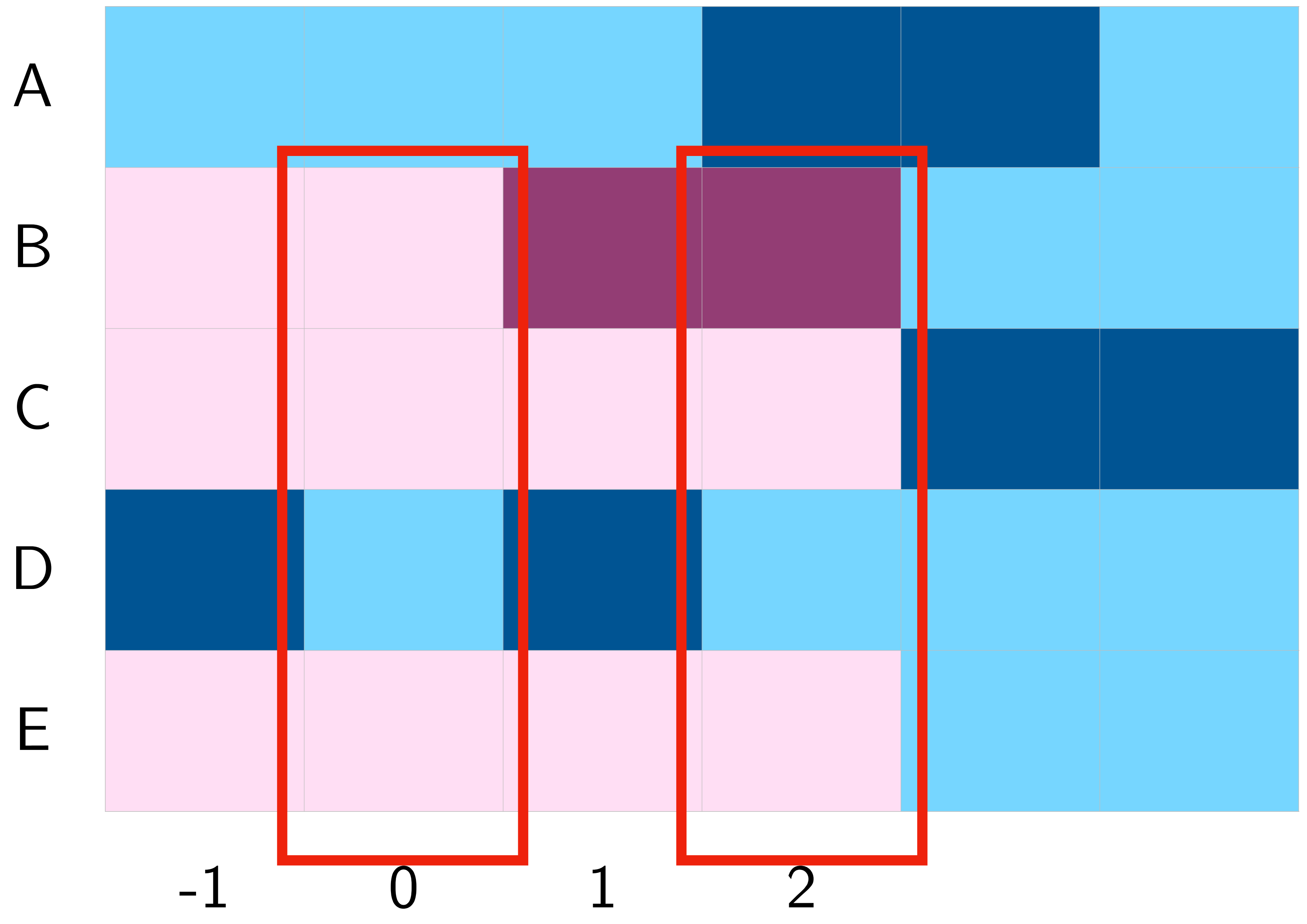
e.g., $l = 2$



# Imai, Kim & Wang (2021) "PanelMatch"

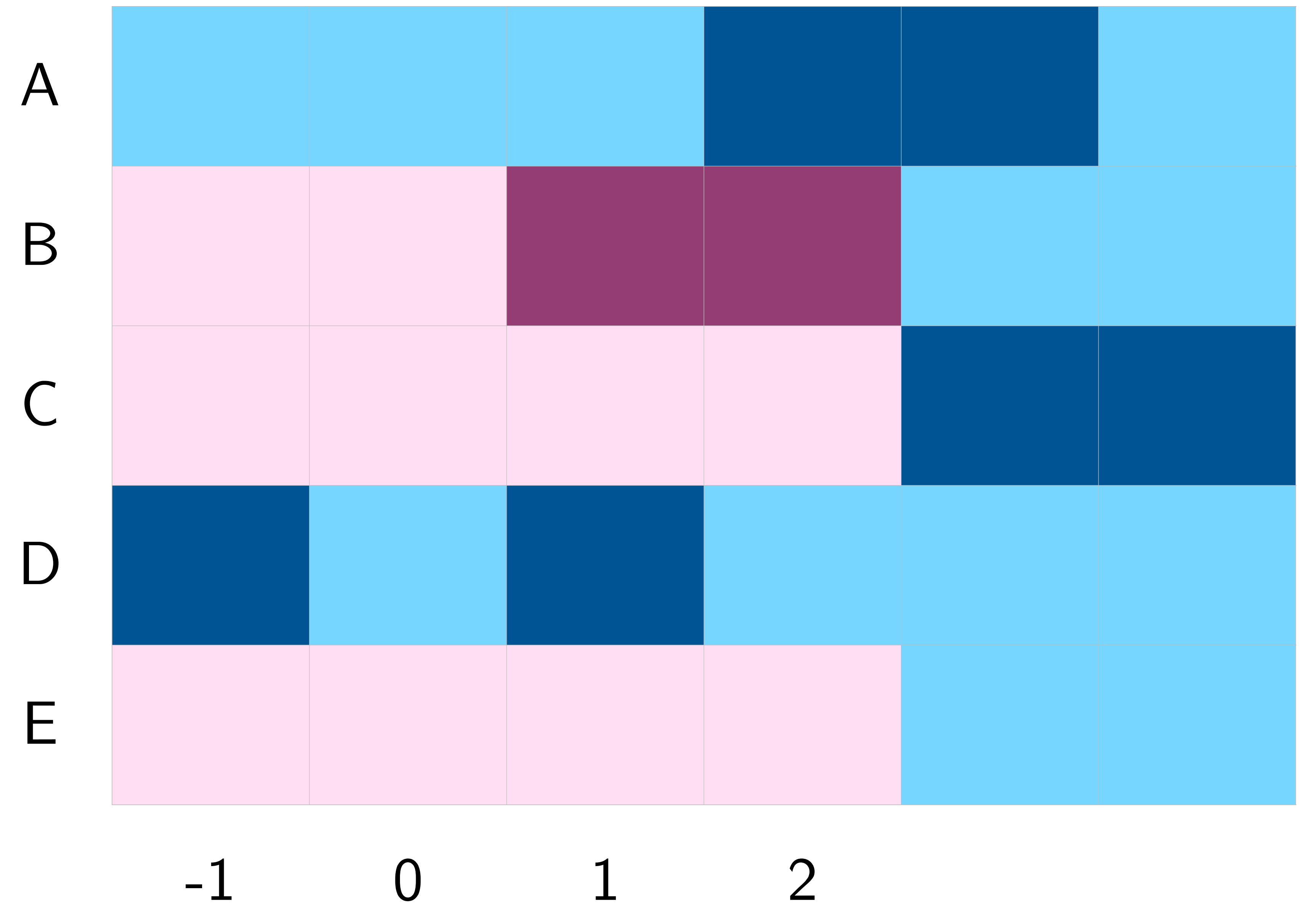
- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i,t)$ with $\{j : D_{is} = D_{js}$ for all $s \in \{t-1, t-2, \ldots, t-a\}\}$
- DID to estimate dynamic effects for future periods $l = 1, 2, \ldots$ (up to reversal)

e.g., $l = 2$



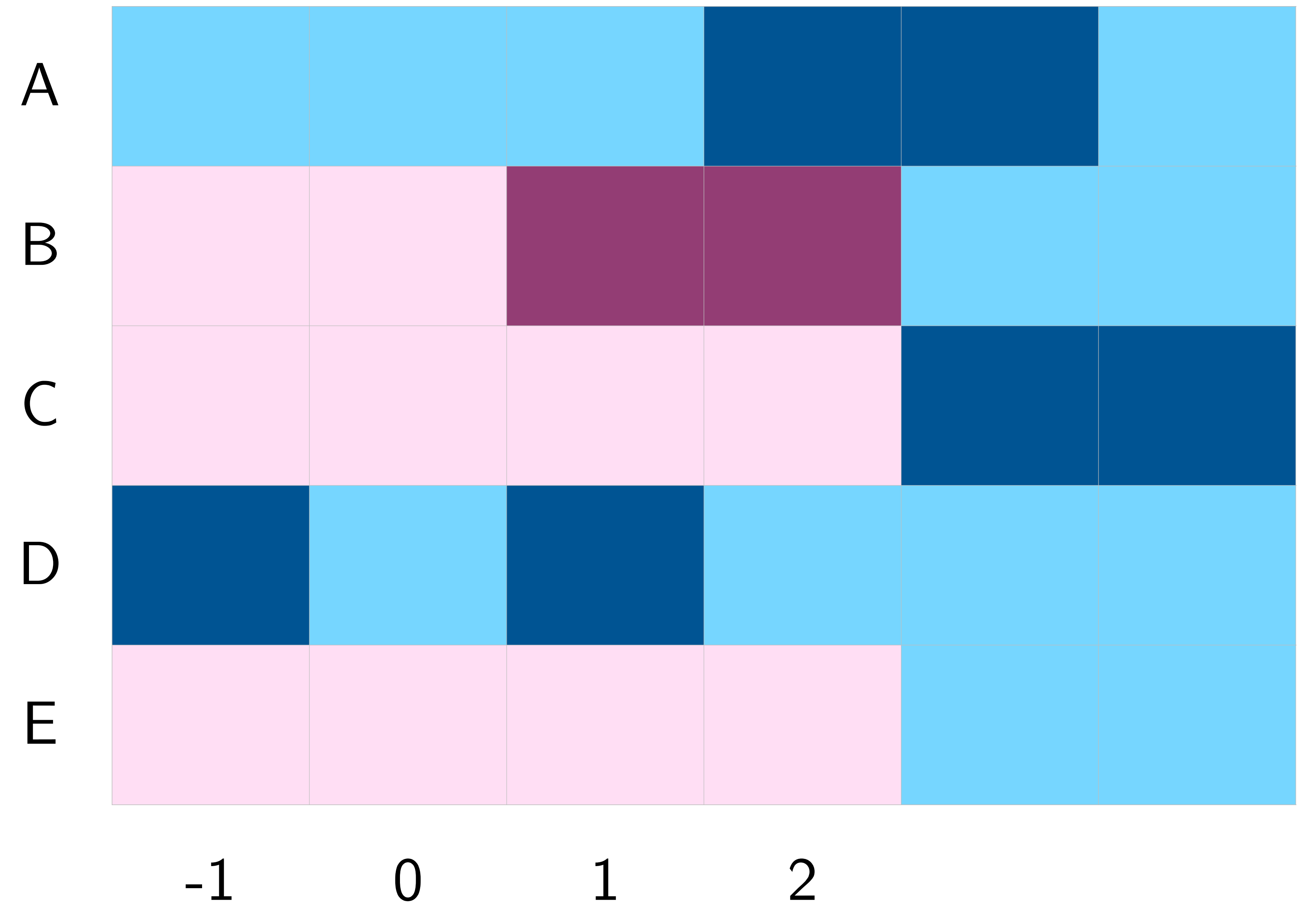
# Imai, Kim & Wang (2021) "PanelMatch"

- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i, t)$ with $\{j : D_{is} = D_{js}$ for all $s \in \{t-1, t-2, \ldots, t-a\}\}$
- DID to estimate dynamic effects for future periods $l = 1, 2, \ldots$ (up to reversal)



# Imai, Kim & Wang (2021) "PanelMatch"

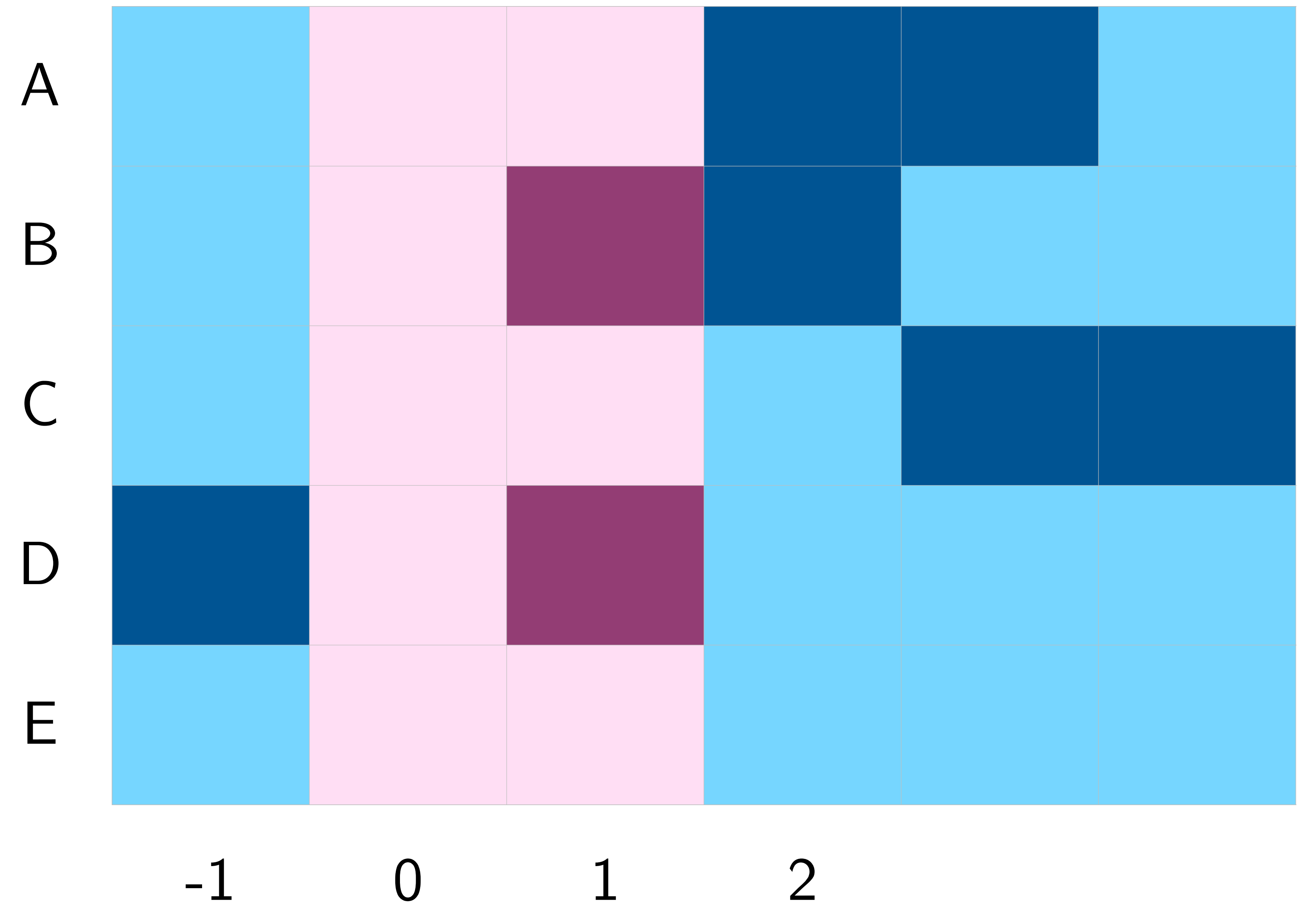
- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i, t)$ with $\{j : D_{is} = D_{js}$ for all $s \in \{t-1, t-2, \ldots, t-a\}\}$
- DID to estimate dynamic effects for future periods $l = 1, 2, \ldots$ (up to reversal)
- Refine matched set based on covariates $X_{it}$



# Imai, Kim & Wang (2021) "PanelMatch"

- Match up to $a$ periods before joining (or leaving)
  - Match treated $(i,t)$ with $\{j : D_{is} = D_{js} \text{ for all } s \in \{t-1, t-2, \ldots, t-a\}\}$
- DID to estimate dynamic effects for future periods $l = 1, 2, \ldots$ (up to reversal)

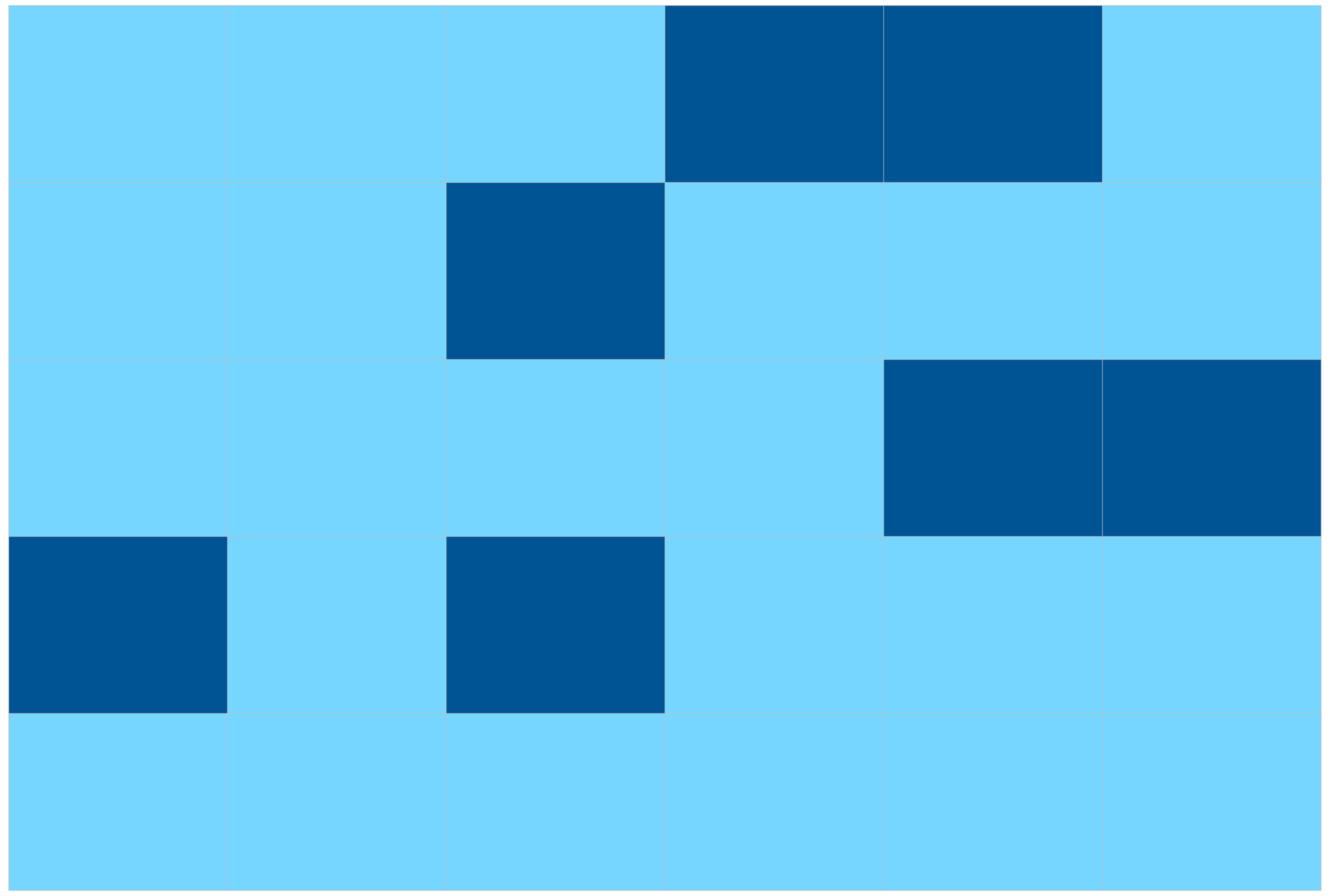
- Refine matched set based on covariates $X_{it}$
- $DID_M$ (de Chaisemartin and D'Haultfœuille, 2020) is weighted sum of PanelMatch estimators for joiners + leavers, $a = l = 1$ (without refinement)



# Imputation Methods

Borusyak, Jaravel & Spiess (2023); Liu, Wang & Xu (2022)

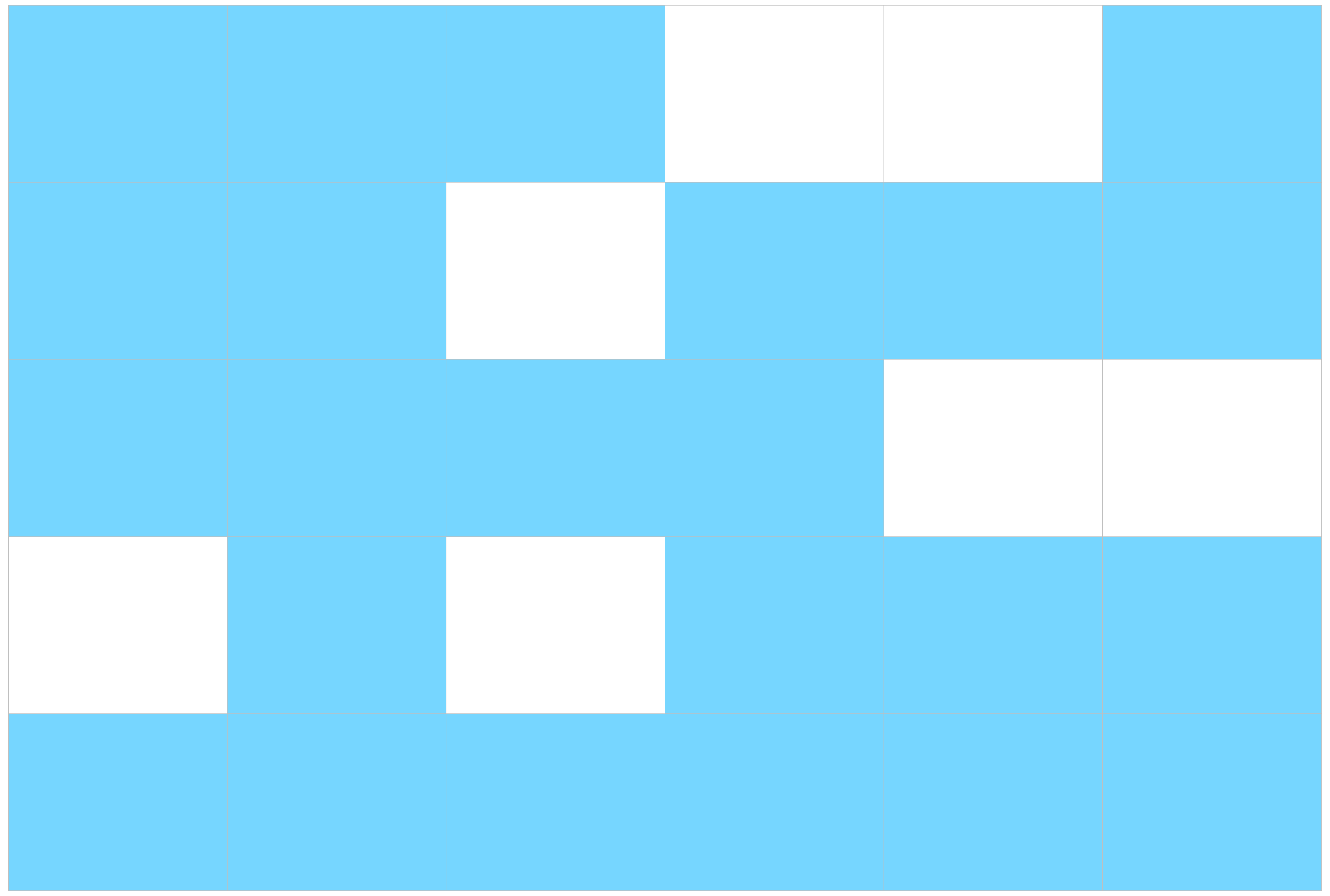
- Fit model for $Y_{it}(0)$ on controls
  - TWFE: Fixed effects counterfactual estimator
- Impute $\hat{Y}_{it}(0)$ for treated
- Estimate individual treatment effects $\hat{\delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$ for treated
- Summarize based on $\hat{\delta}_{it}$
- Efficient under homoskedasticity (BJS 2023)



# Imputation Methods

Borusyak, Jaravel & Spiess (2023); Liu, Wang & Xu (2022)

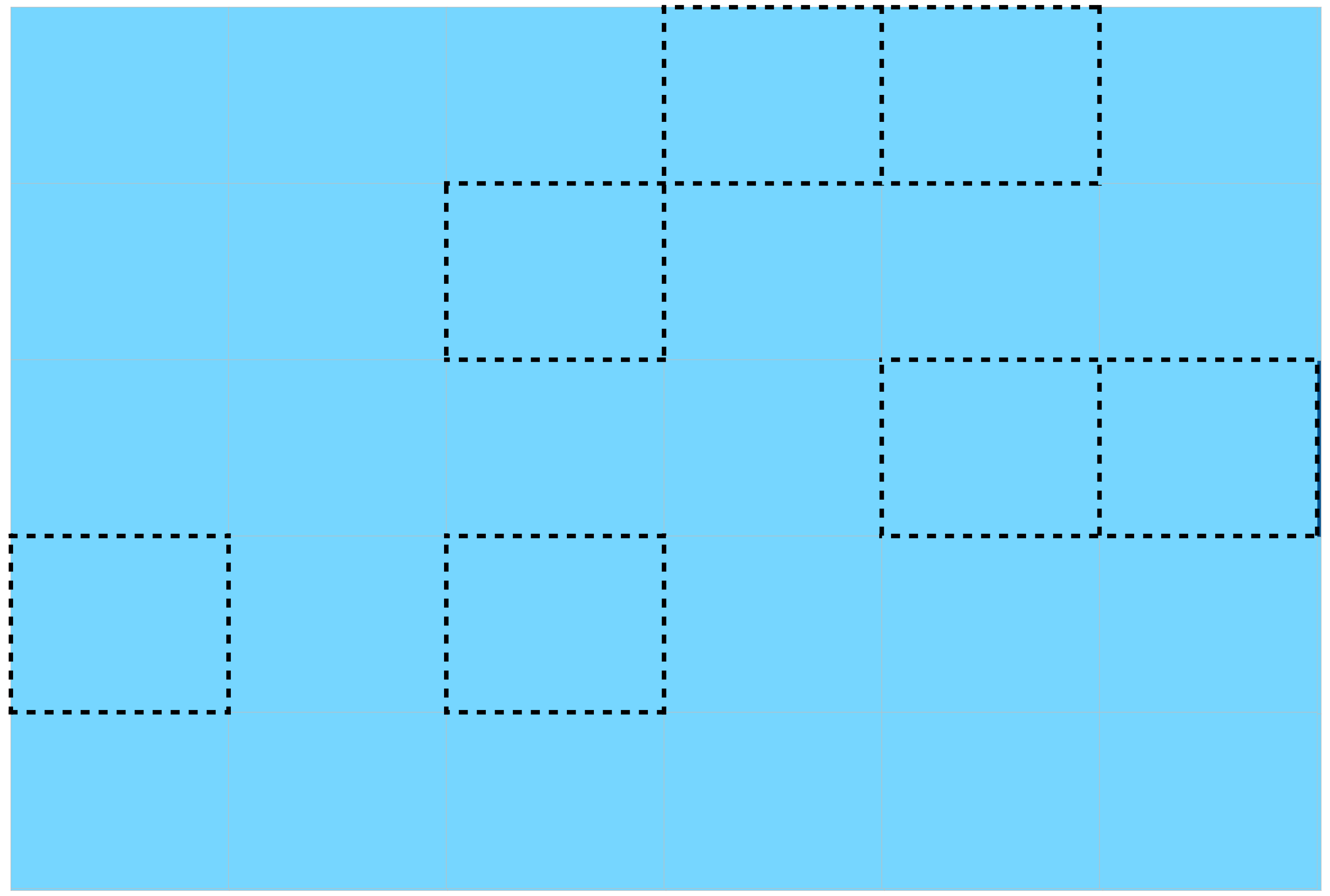
- Fit model for $Y_{it}(0)$ on controls
  - TWFE: Fixed effects counterfactual estimator
- Impute $\hat{Y}_{it}(0)$ for treated
- Estimate individual treatment effects $\hat{\delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$ for treated
- Summarize based on $\hat{\delta}_{it}$
- Efficient under homoskedasticity (BJS 2023)



# Imputation Methods

Borusyak, Jaravel & Spiess (2023); Liu, Wang & Xu (2022)

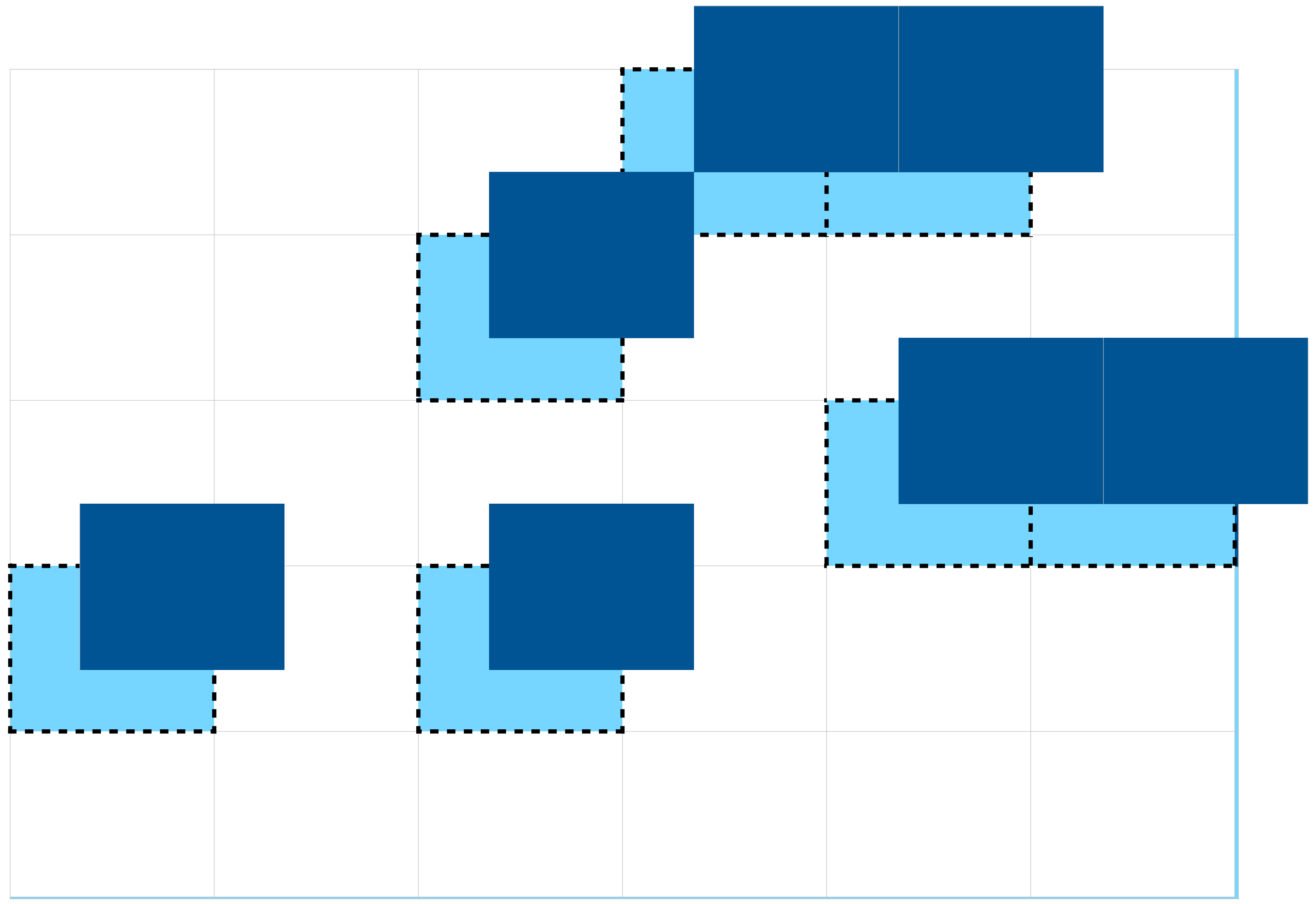
- Fit model for $Y_{it}(0)$ on controls
  - TWFE: Fixed effects counterfactual estimator
- Impute $\hat{Y}_{it}(0)$ for treated
- Estimate individual treatment effects $\hat{\delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$ for treated
- Summarize based on $\hat{\delta}_{it}$
- Efficient under homoskedasticity (BJS 2023)



# Imputation Methods

Borusyak, Jaravel & Spiess (2023); Liu, Wang & Xu (2022)

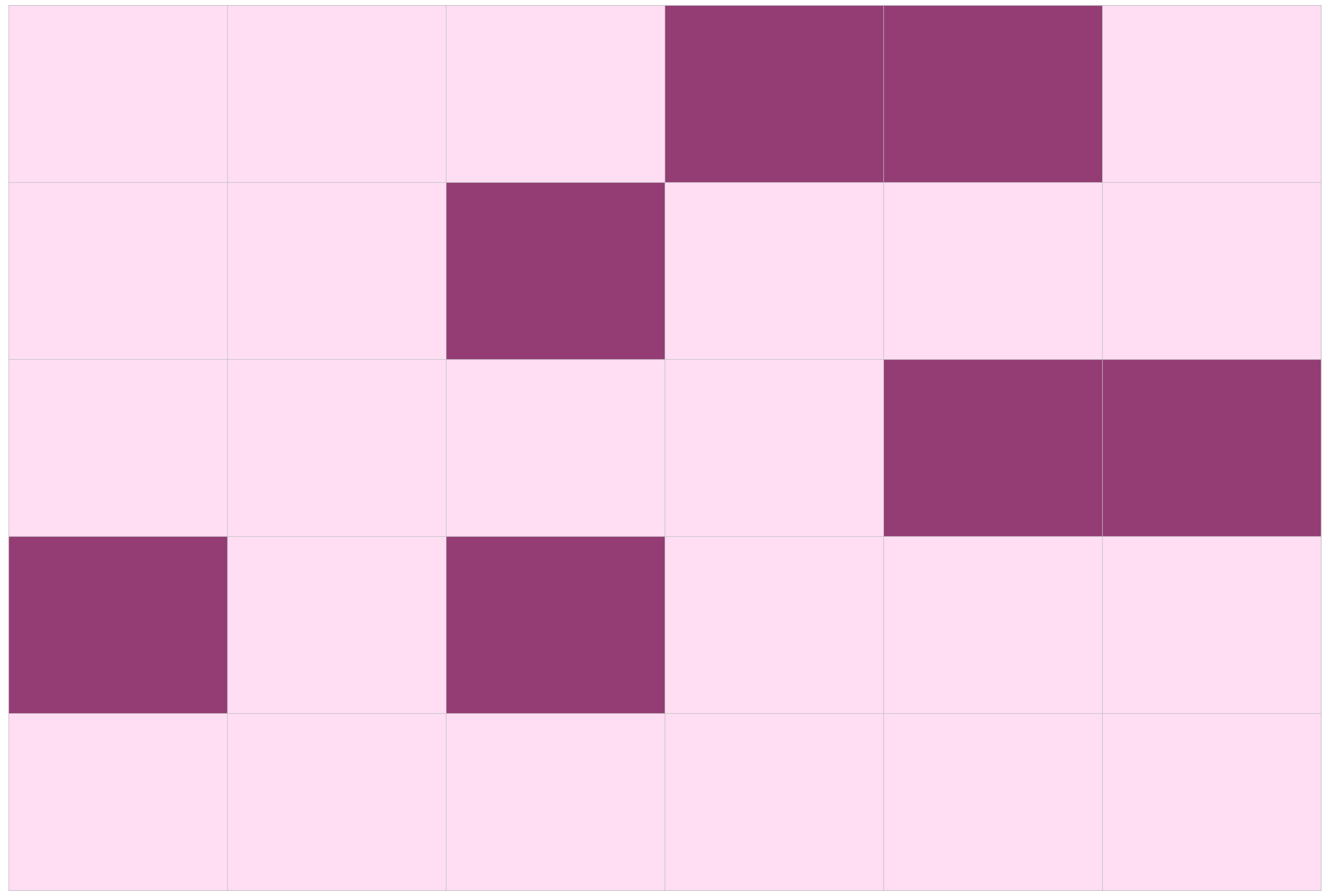
- Fit model for $Y_{it}(0)$ on controls
  - TWFE: Fixed effects counterfactual estimator
- Impute $\hat{Y}_{it}(0)$ for treated
- Estimate individual treatment effects $\hat{\delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$ for treated
- Summarize based on $\hat{\delta}_{it}$
- Efficient under homoskedasticity (BJS 2023)



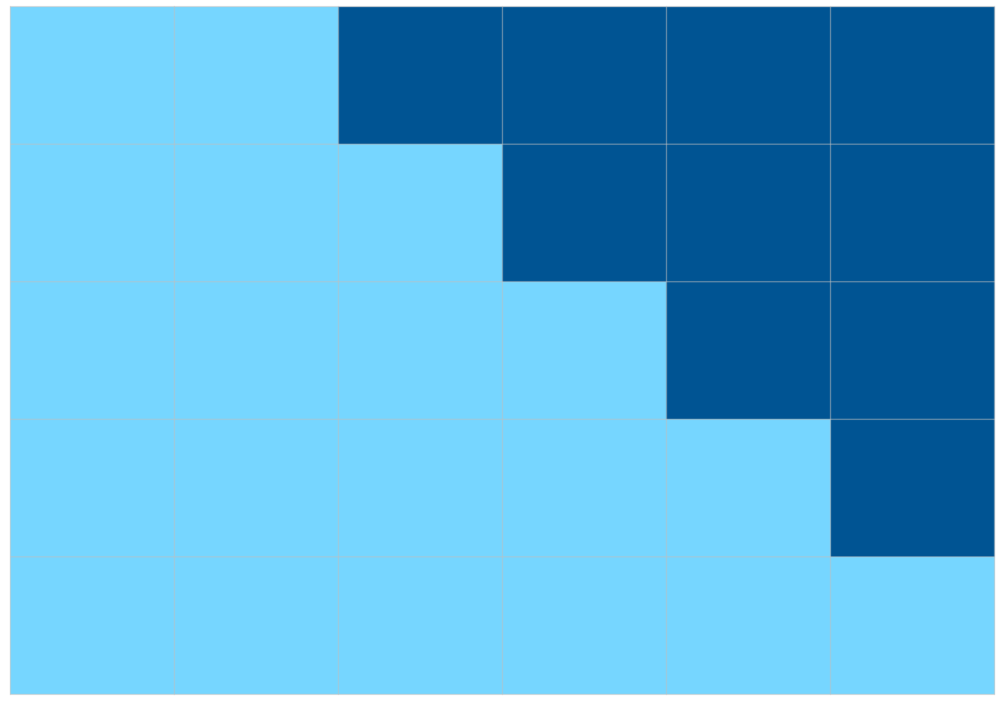
# Imputation Methods

Borusyak, Jaravel & Spiess (2023); Liu, Wang & Xu (2022)

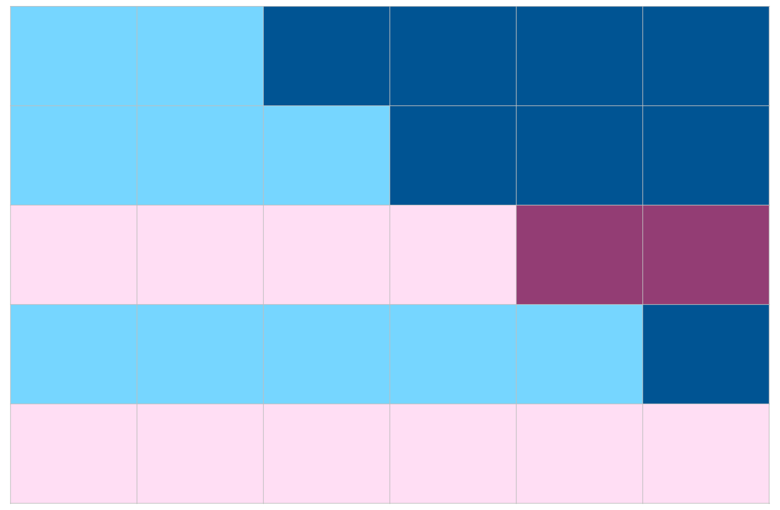
- Fit model for $Y_{it}(0)$ on controls
  - TWFE: Fixed effects counterfactual estimator
- Impute $\hat{Y}_{it}(0)$ for treated
- Estimate individual treatment effects $\hat{\delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$ for treated
- Summarize based on $\hat{\delta}_{it}$
- Efficient under homoskedasticity (BJS 2023)



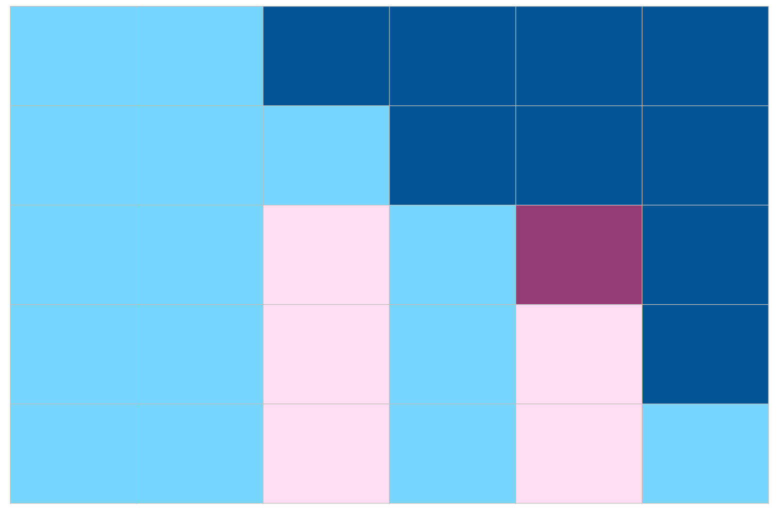
# Comparison — Staggered



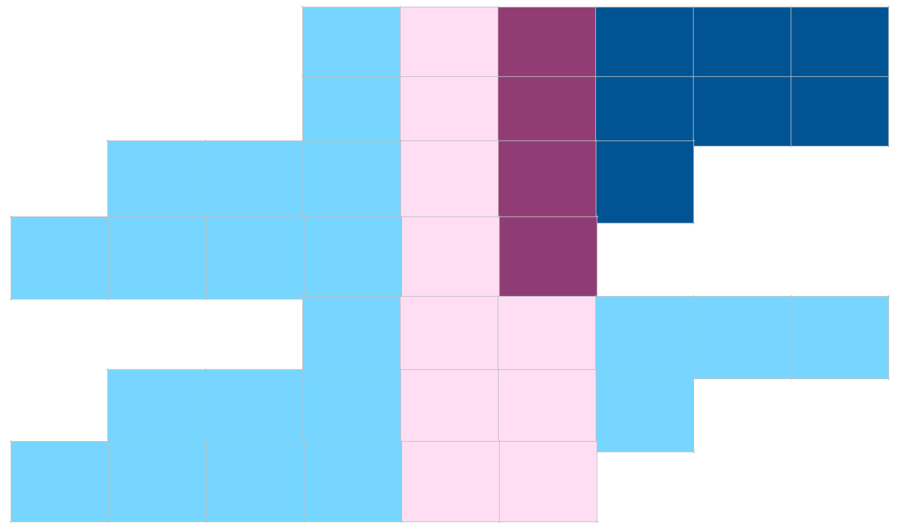
Original Data



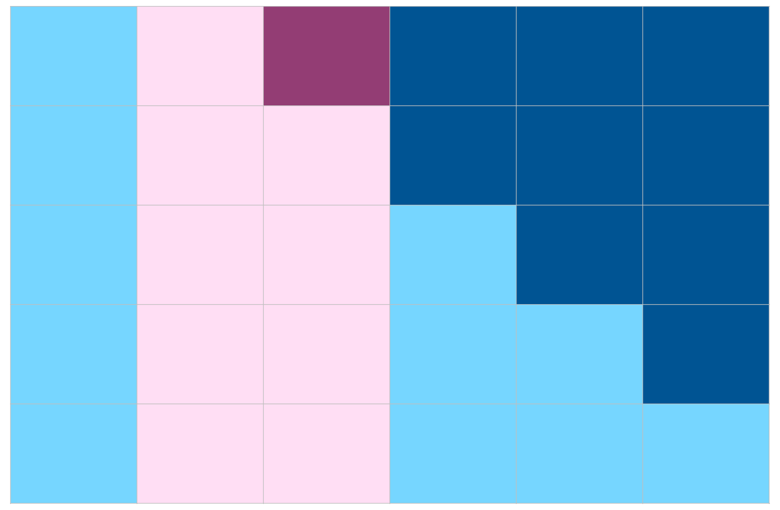
Interaction Weighted
& Stacked DID



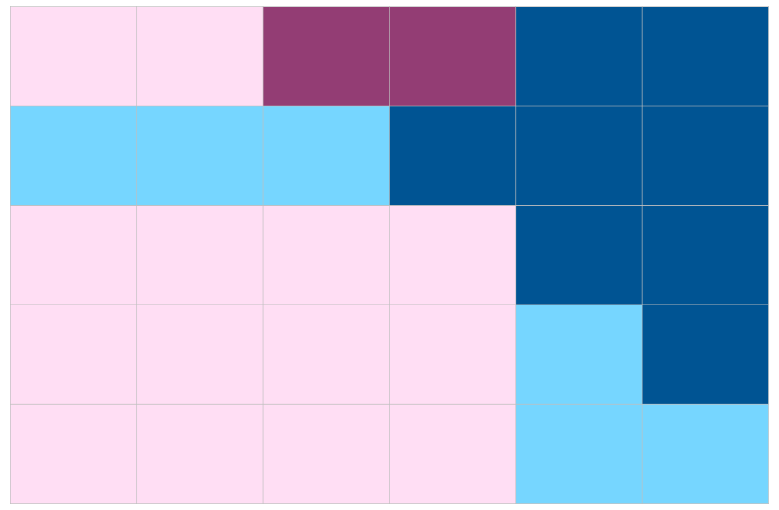
CSDID



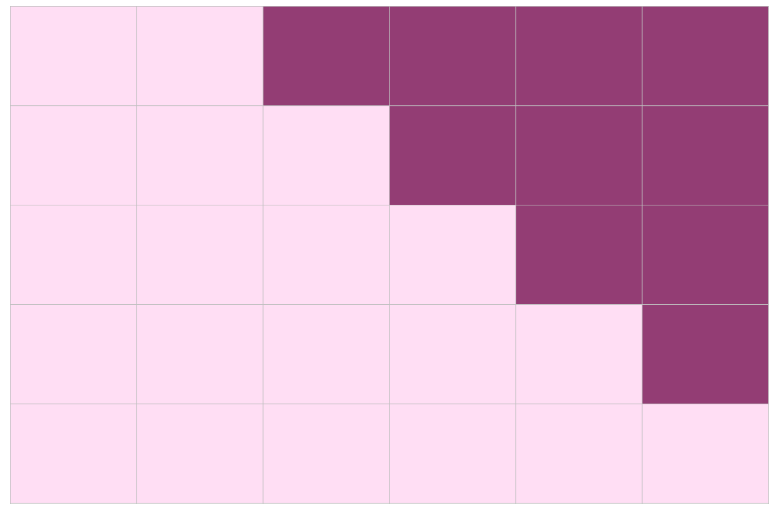
Stacked DID



DID multiple



PanelMatch



Imputation Method
e.g., FEct

# HTE-Robust Estimators

| | DID Extensions (2x2 DID as building blocks) | | Imputation Methods (outcome model w/ FE) |
|---|---|---|---|
| *Setting* | Staggered | General | General |
| *Estimand* | ATT | ATT for Switchers | ATT |
| *Estimator* | IW, CSDID, Stacked DID | PanelMatch, $\text{DID}_{\text{M}}$ | $\text{DID}_{\text{impute}}$, FEct |
| *Comparison Group* | Never/last/not-yet-treated | Matched set | Imputed counterfactual |
| *Key assumption* | Parallel Trends | Parallel Trends | Zero Conditional Mean or Parallel Trends |

# Replication & Reanalysis

Data

Procedure

# The Replication Sample (2017-2023)

# The Replication Sample (2017-2023)

Case selection:

- Use panel data analysis as a critical piece of evidence to support a causal argument
- Binary treatment
- A "proper" TWFE (DID) research design
- Use a DID or TWFE estimator
- Focus on the authors' preferred specification

# The Replication Sample (2017-2023)

## Case selection:

- Use panel data analysis as a critical piece of evidence to support a causal argument
- Binary treatment
- A "proper" TWFE (DID) research design
- Use a DID or TWFE estimator
- Focus on the authors' preferred specification

| Journal | All Linear Panel |
|---|---|
| APSR | 22 |
| AJPS | 31 |
| JOP | 49 |
| **Total** | **102** |

# The Replication Sample (2017-2023)

## Case selection:

- Use panel data analysis as a critical piece of evidence to support a causal argument
- Binary treatment
- A "proper" TWFE (DID) research design
- Use a DID or TWFE estimator
- Focus on the authors' preferred specification

| Journal | All Linear Panel | "Proper" TWFE |
|---|---|---|
| APSR | 22 | 13 |
| AJPS | 31 | 21 |
| JOP | 49 | 30 |
| **Total** | **102** | **64** |

# The Replication Sample (2017-2023)

## Case selection:

- Use panel data analysis as a critical piece of evidence to support a causal argument
- Binary treatment
- A "proper" TWFE (DID) research design
- Use a DID or TWFE estimator
- Focus on the authors' preferred specification

| Journal | All Linear Panel | "Proper" TWFE | Incomplete Data |
|---|---|---|---|
| APSR | 22 | 13 | 2 |
| AJPS | 31 | 21 | 3 |
| JOP | 49 | 30 | 6 |
| **Total** | **102** | **64** | **11 (17.2%)** |

# The Replication Sample (2017-2023)

## Case selection:

- Use panel data analysis as a critical piece of evidence to support a causal argument
- Binary treatment
- A "proper" TWFE (DID) research design
- Use a DID or TWFE estimator
- Focus on the authors' preferred specification

| Journal | All Linear Panel | "Proper" TWFE | Incomplete Data | Error in Code |
|---|---|---|---|---|
| **APSR** | 22 | 13 | 2 | 1 |
| **AJPS** | 31 | 21 | 3 | 3 |
| **JOP** | 49 | 30 | 6 | 0 |
| **Total** | **102** | **64** | **11 (17.2%)** | **4 (6.3%)** |

# The Replication Sample (2017-2023)

## Case selection:

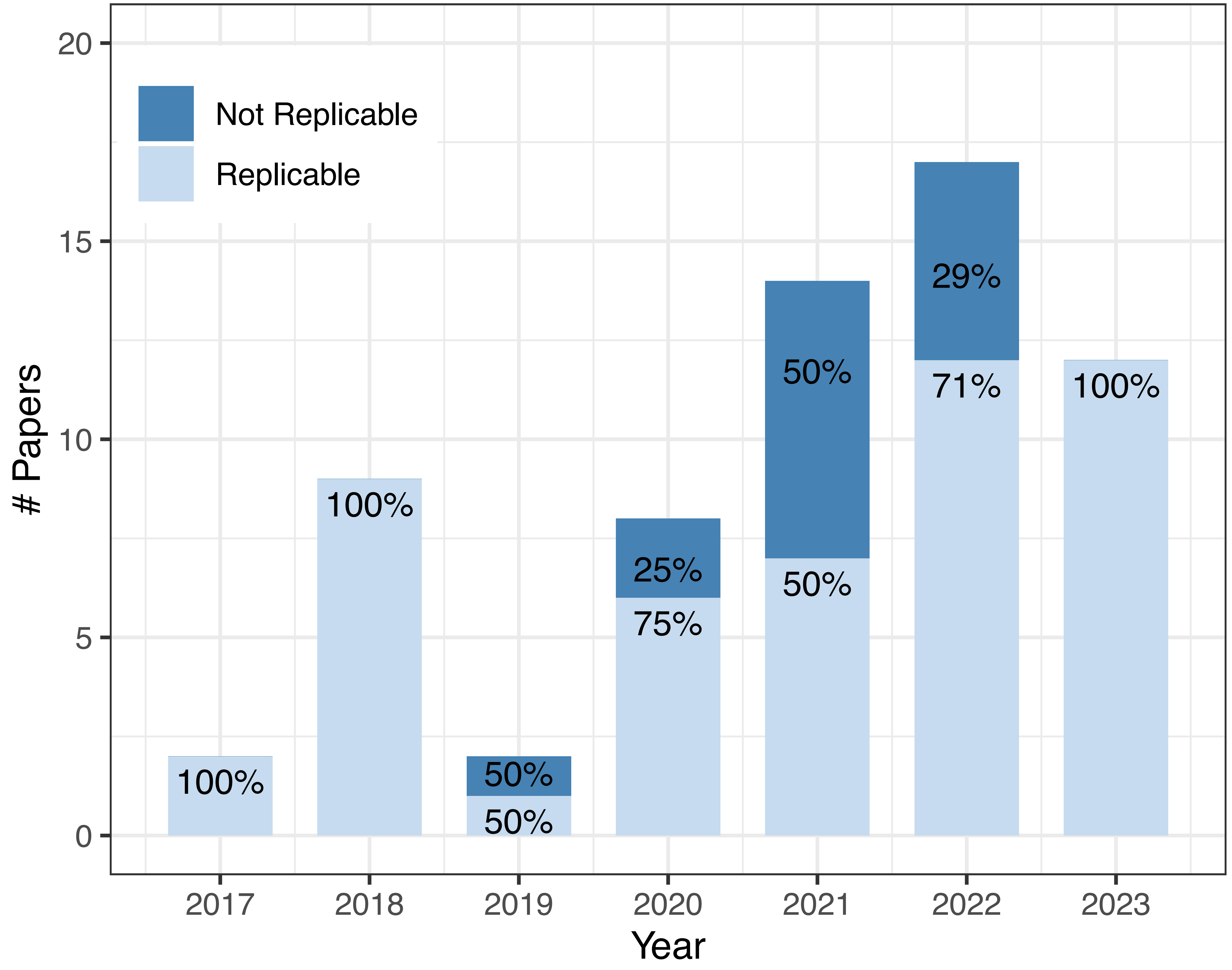
- Use panel data analysis as a critical piece of evidence to support a causal argument
- Binary treatment
- A "proper" TWFE (DID) research design
- Use a DID or TWFE estimator
- Focus on the authors' preferred specification

| Journal | All Linear Panel | "Proper" TWFE | Incomplete Data | Error in Code | Replicable |
|---|---|---|---|---|---|
| **APSR** | 22 | 13 | 2 | 1 | 10 (76.9%) |
| **AJPS** | 31 | 21 | 3 | 3 | 15 (71.4%) |
| **JOP** | 49 | 30 | 6 | 0 | 24 (80%) |
| **Total** | **102** | **64** | **11 (17.2%)** | **4 (6.3%)** | **49 (76.6%)** |

# The Replication Sample (2017-2023)

## Case selection:

- Use panel data analysis as a critical piece of evidence to support a causal argument
- Binary treatment
- A "proper" TWFE (DID) research design
- Use a DID or TWFE estimator
- Focus on the authors' preferred specification



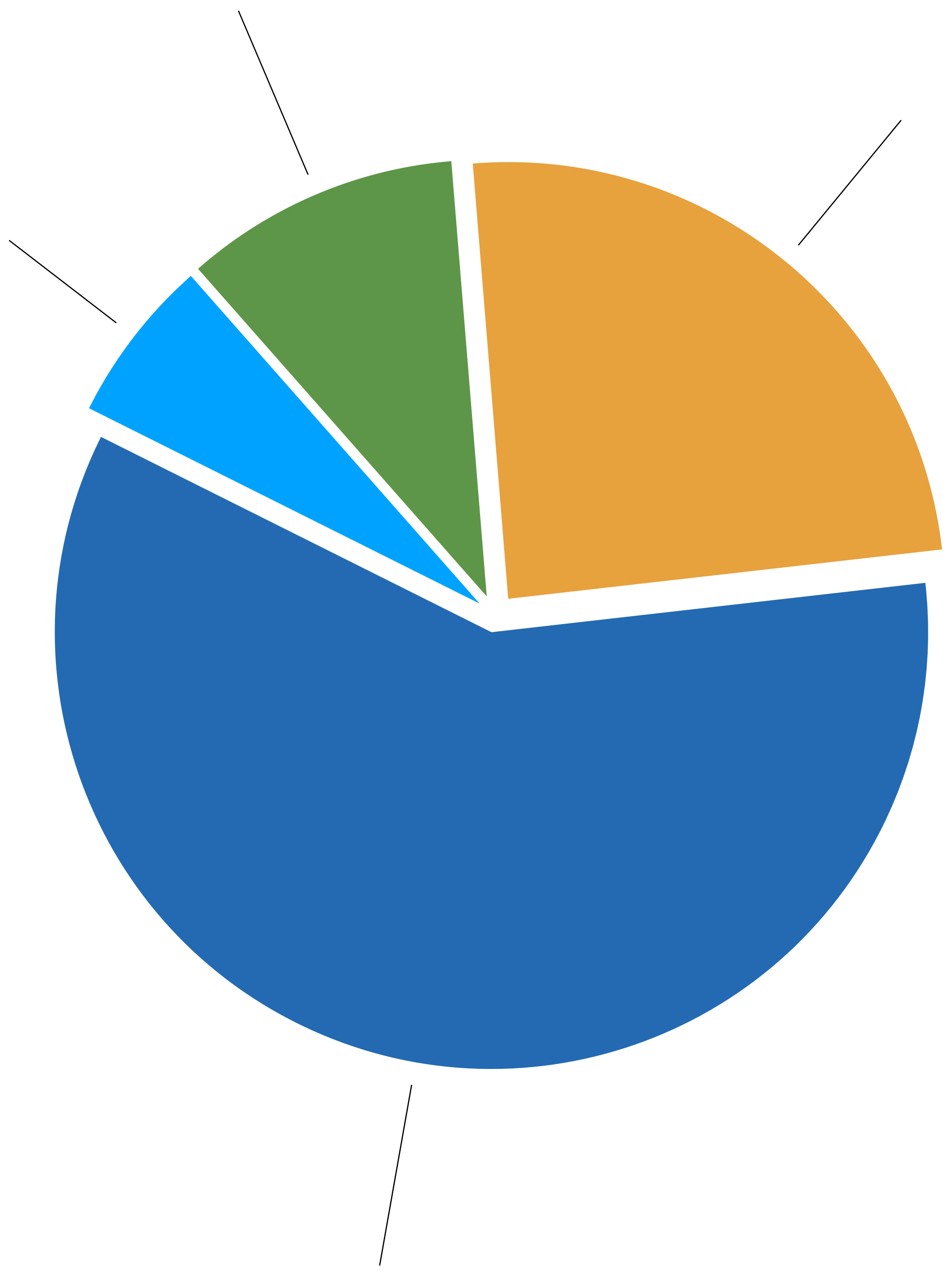
# Common Settings and Practice

# Common Settings and Practice

Among 49 Replicable Studies

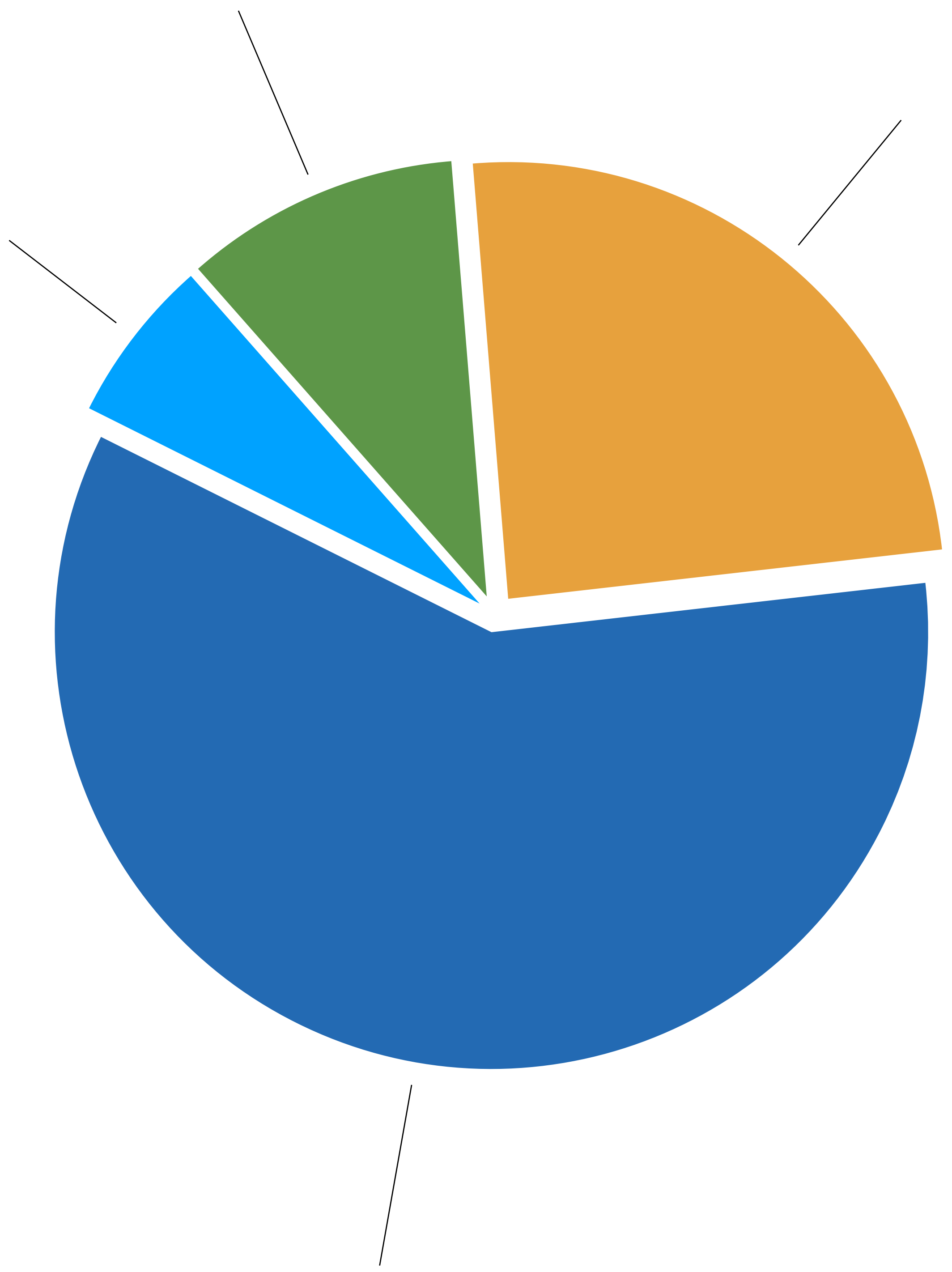
# Common Settings and Practice

Among 49 Replicable Studies



# Common Settings and Practice

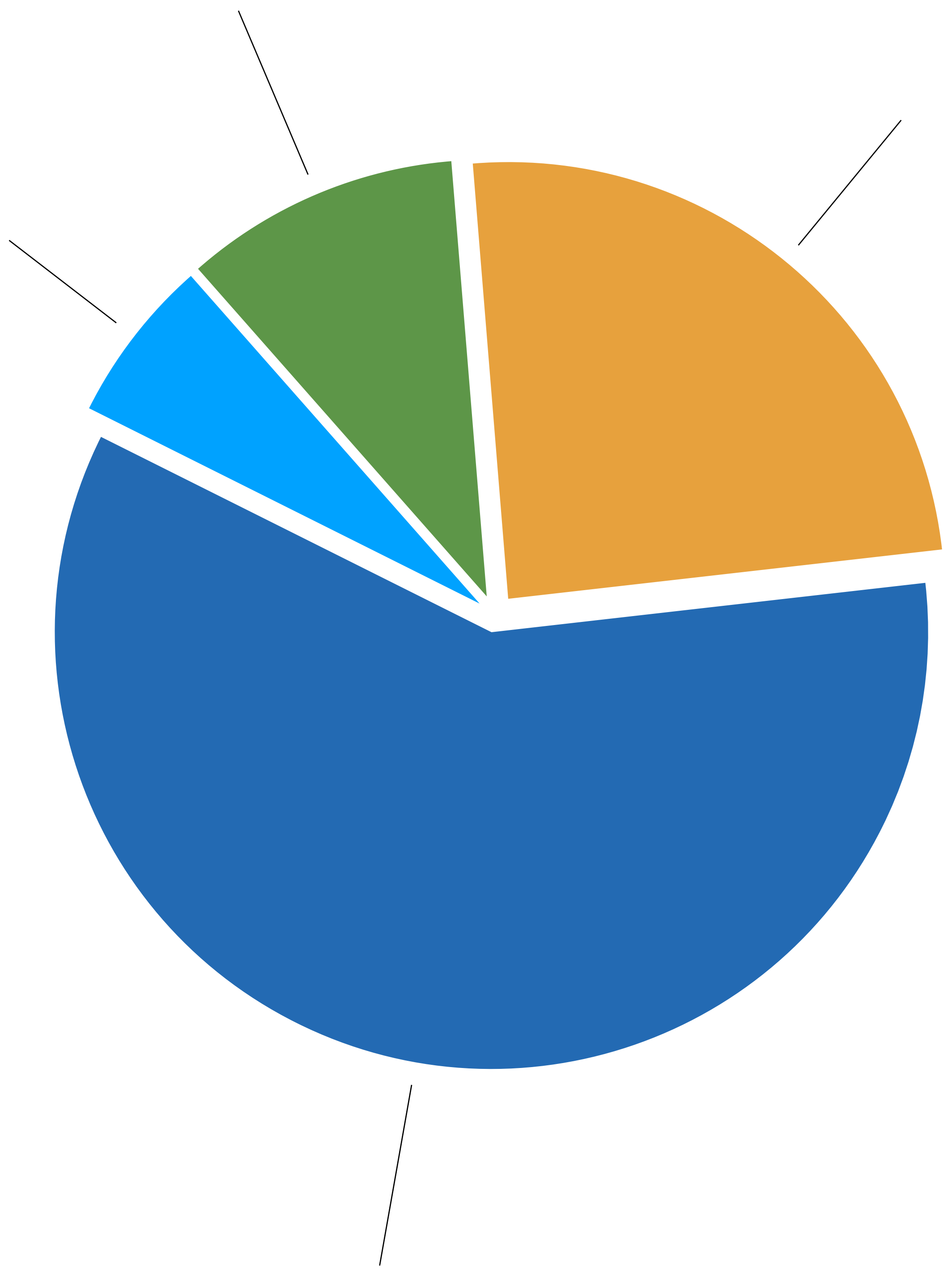
Among 49 Replicable Studies



| Variance Estimator | | |
|---|---|---|
| Cluster-robust SE or PCSE | 48 | 98% |
| Clustered bootstrapping | 8 | 16% |

# Common Settings and Practice

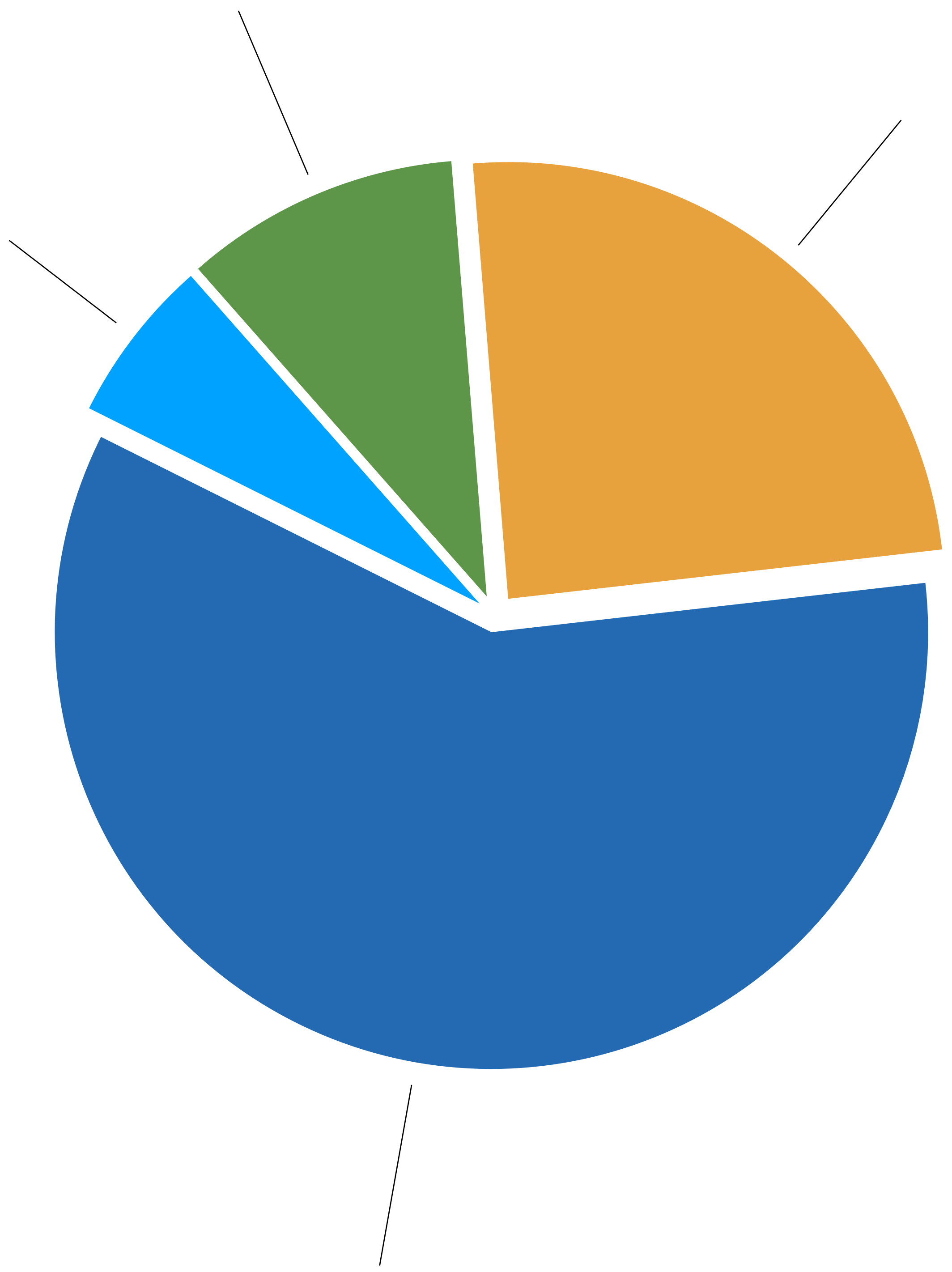
Among 49 Replicable Studies



| Variance Estimator | | |
|---|---|---|
| Cluster-robust SE or PCSE | 48 | 98% |
| Clustered bootstrapping | 8 | 16% |
| | | |
| **Variants in TWFE Specifications** | | |
| w/ lagged outcomes | 8 | 16% |
| w/ higher-than-unit-level time trends | 5 | 10% |
| w/ unit-level time trends | 15 | 30% |

# Common Settings and Practice

Among 49 Replicable Studies



| Variance Estimator | | |
|---|---|---|
| Cluster-robust SE or PCSE | 48 | 98% |
| Clustered bootstrapping | 8 | 16% |
| | | |
| **Variants in TWFE Specifications** | | |
| w/ lagged outcomes | 8 | 16% |
| w/ higher-than-unit-level time trends | 5 | 10% |
| w/ unit-level time trends | 15 | 30% |
| | | |
| **Visual Inspection** | | |
| Group average outcome trajectories | 19 | 39% |
| Event-study plots | 23 | 47% |
| Neither | 19 | 39% |

# Procedure

# Procedure

- Step 1. Understand the context, setting, and data structure
  - Plot raw data
  - Record key information

# Procedure

- Step 1. Understand the context, setting, and data structure
  - Plot raw data
  - Record key information
- Step 2. Replicate a main result
  - Original variance estimator & cluster-bootstrap procedure

# Procedure

- Step 1. Understand the context, setting, and data structure
  - Plot raw data
  - Record key information
- Step 2. Replicate a main result
  - Original variance estimator & cluster-bootstrap procedure
- Step 3. Re-estimate ATT and the event study plot using TWFE and several HTE-robust estimators, including
  - IW (Sun & Abraham 2021) — If staggered DID
  - CSDID (Callaway and Sant'Anna 2021) — If staggered DID
  - Stacked DID (Cengiz et al. 2019) — If staggered DID
  - PanelMatch/DID multiple (Imai, Kim & Wang 2021; De Chaisemartin and D'Haultfœuille)
  - Imputation (Borusyak, Jaravel and Spiess 2021; Liu, Wang & Xu 2022)

# Procedure

- Step 1. Understand the context, setting, and data structure
  - Plot raw data
  - Record key information

- Step 2. Replicate a main result
  - Original variance estimator & cluster-bootstrap procedure

- Step 3. Re-estimate ATT and the event study plot using TWFE and several HTE-robust estimators, including
  - IW (Sun & Abraham 2021) — If staggered DID
  - CSDID (Callaway and Sant'Anna 2021) — If staggered DID
  - Stacked DID (Cengiz et al. 2019) — If staggered DID
  - PanelMatch/DID multiple (Imai, Kim & Wang 2021; De Chaisemartin and D'Haultfœuille)
  - Imputation (Borusyak, Jaravel and Spiess 2021; Liu, Wang & Xu 2022)

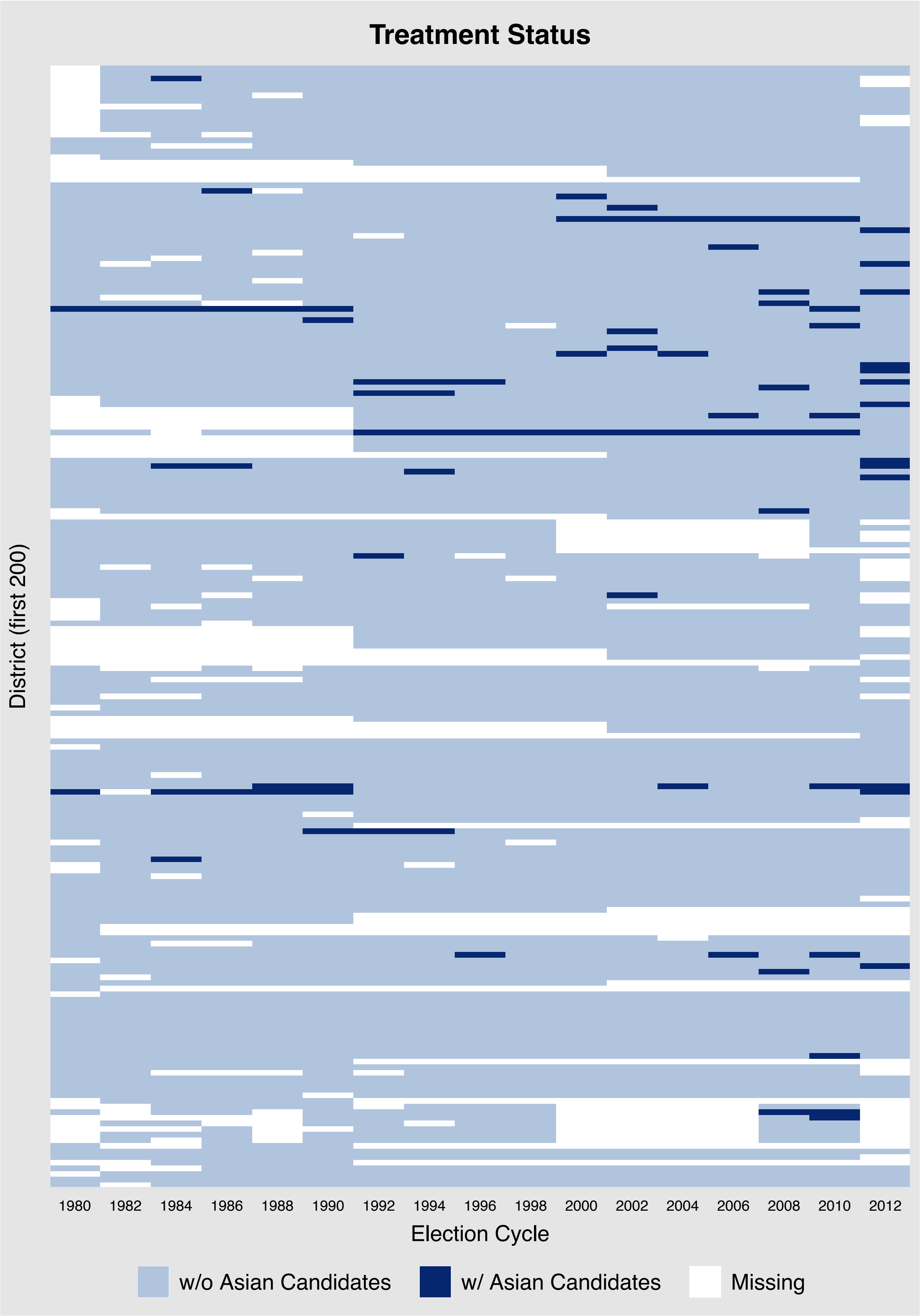
- Step 4. Conduct diagnostic test based on the imputation estimator (Liu, Wang & Xu 2022)
  - Tests for pretrend & carryover effects
  - Sensitivity analysis (Rambachan & Roth 2023)

# Findings

Three examples

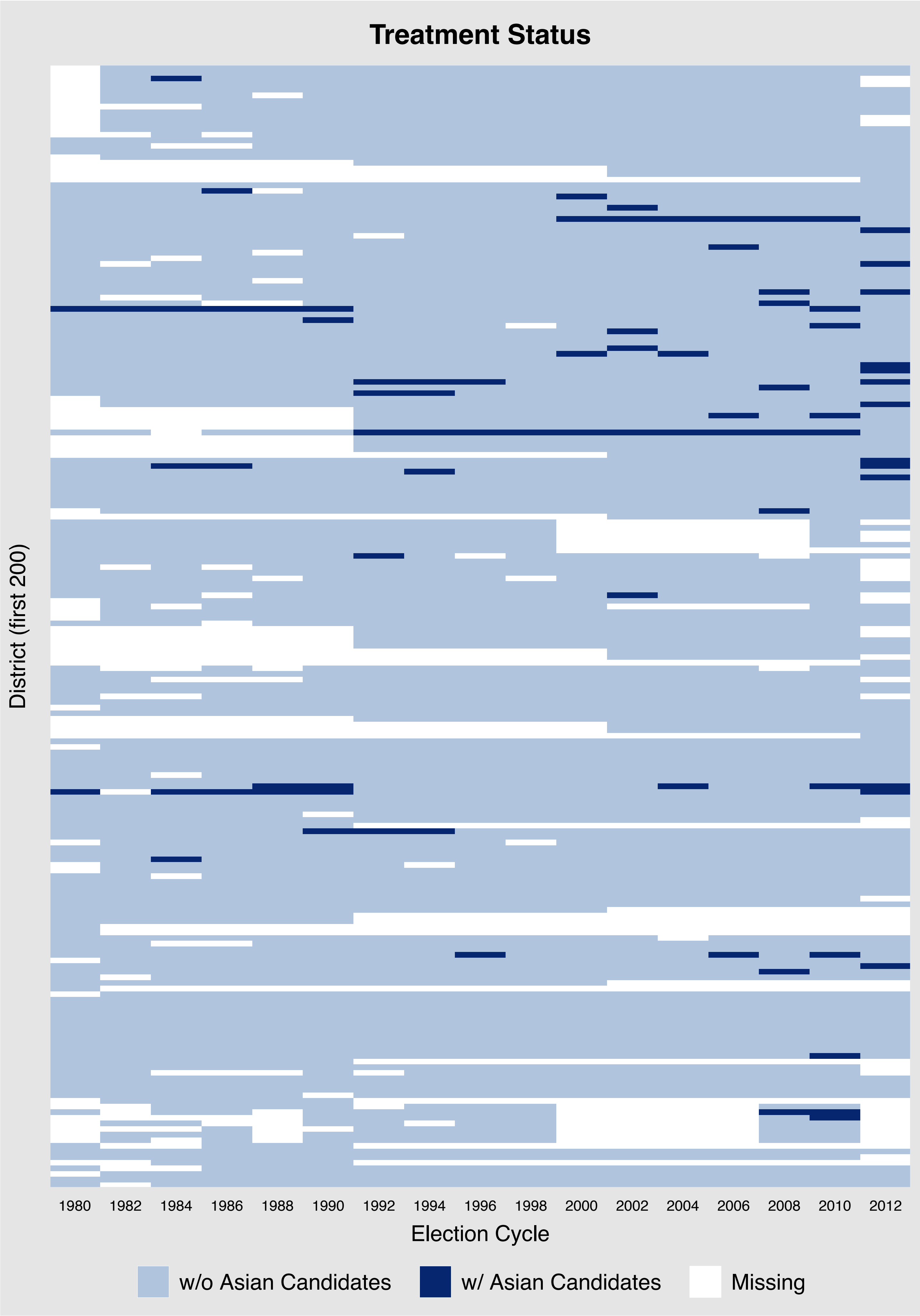
Overall assessment

# Example 1: Coethnic Mobilization (APSR 2020)



# Example 1: Coethnic Mobilization (APSR 2020)

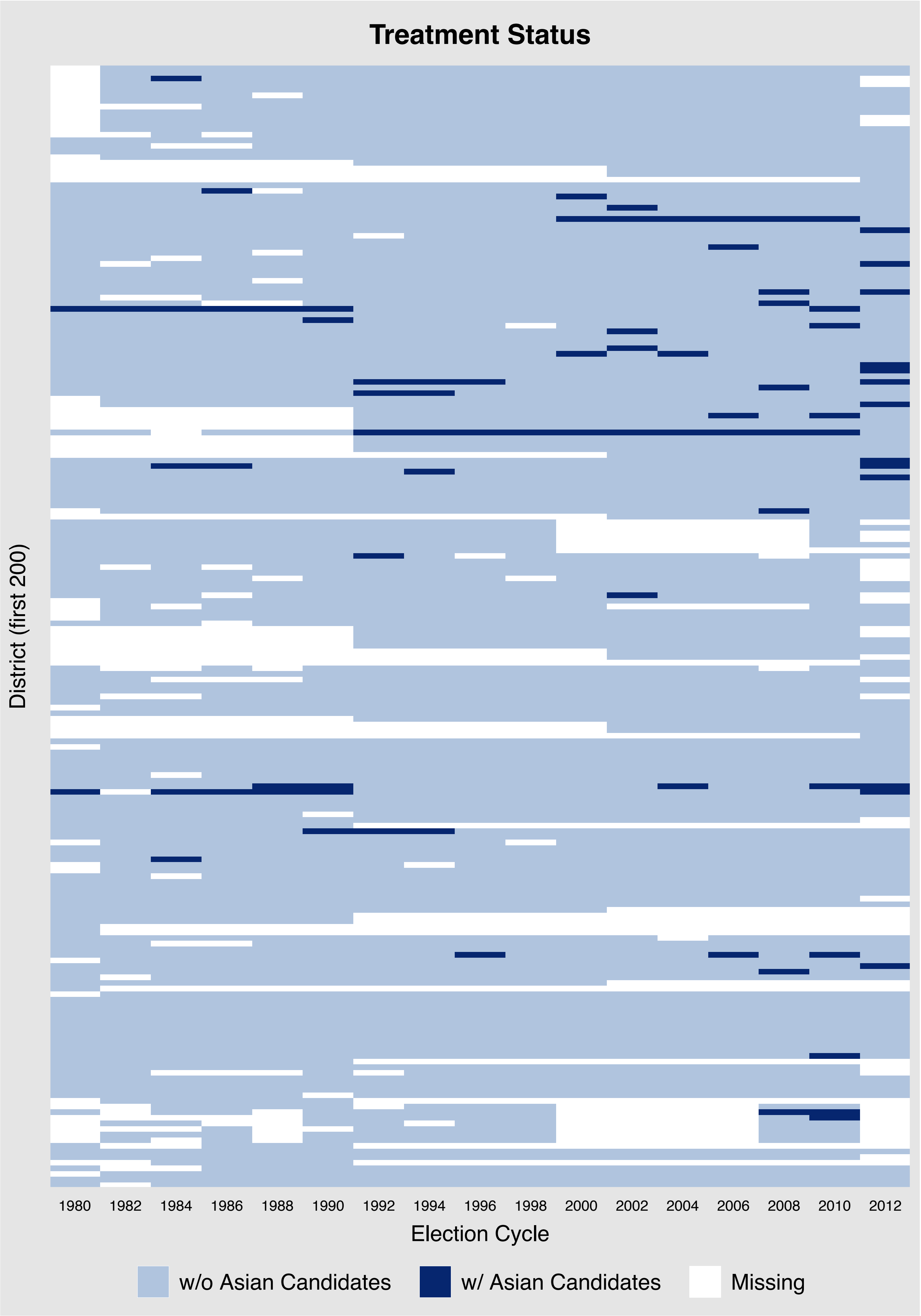
- Grumbach & Sahn (2020): Do minority candidates in US congressional elections mobilize coethnic donators?



# Example 1: Coethnic Mobilization (APSR 2020)

- Grumbach & Sahn (2020): Do minority candidates in US congressional elections mobilize coethnic donators?

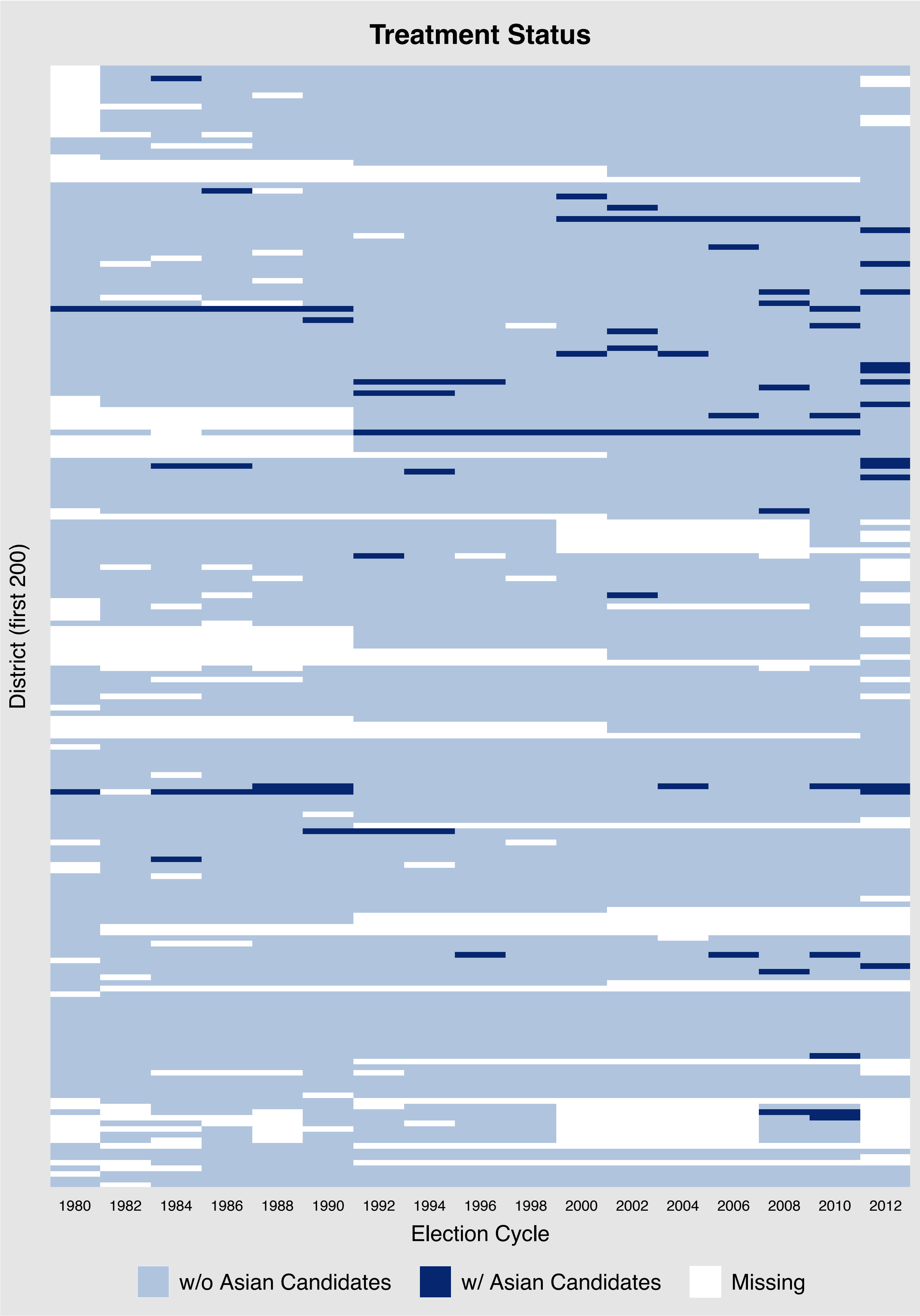
  - Treatment: Asian candidates



# Example 1: Coethnic Mobilization (APSR 2020)

- Grumbach & Sahn (2020): Do minority candidates in US congressional elections mobilize coethnic donators?

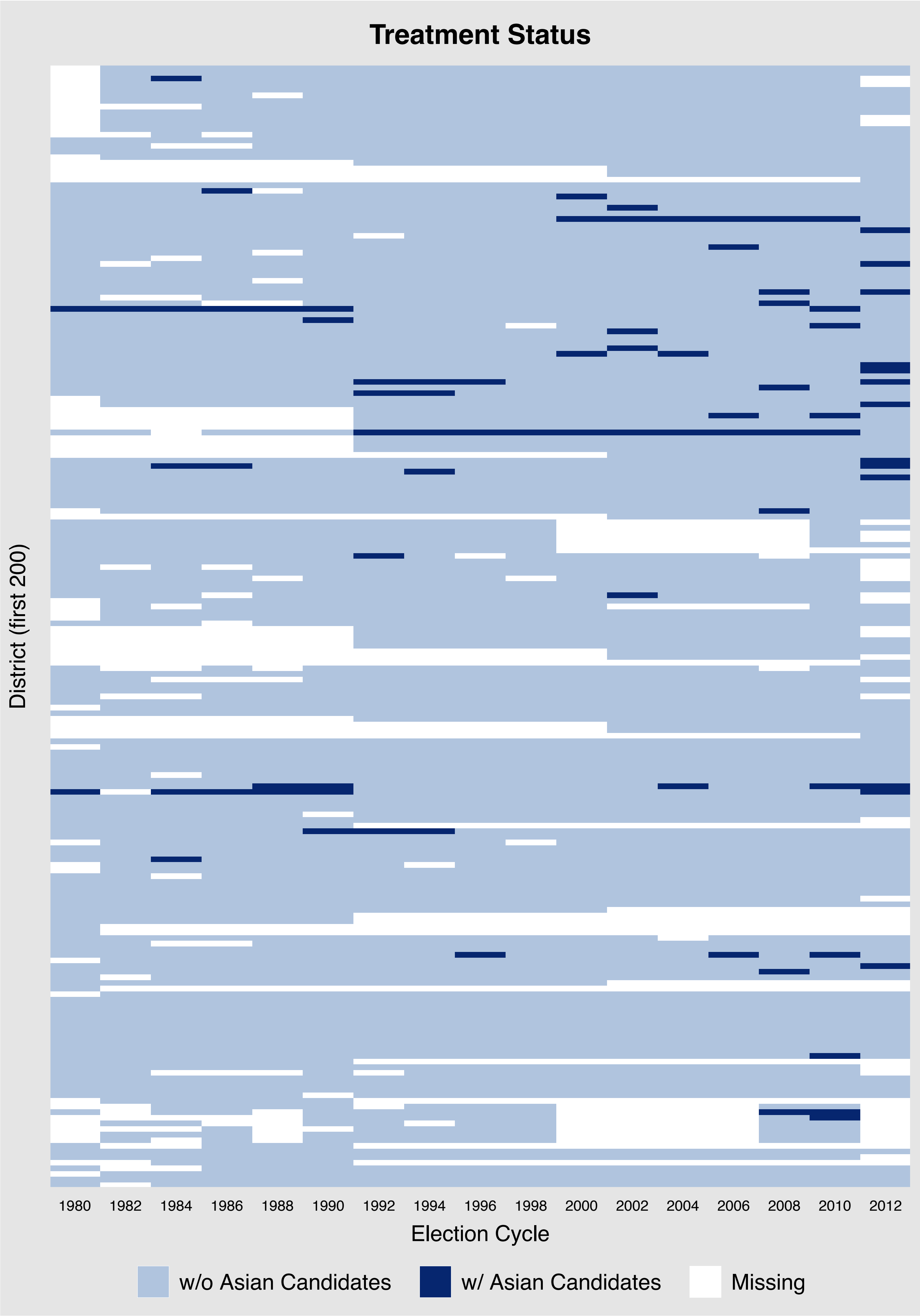
  - Treatment: Asian candidates
  - Outcome: share of Asian donations



# Example 1: Coethnic Mobilization (APSR 2020)

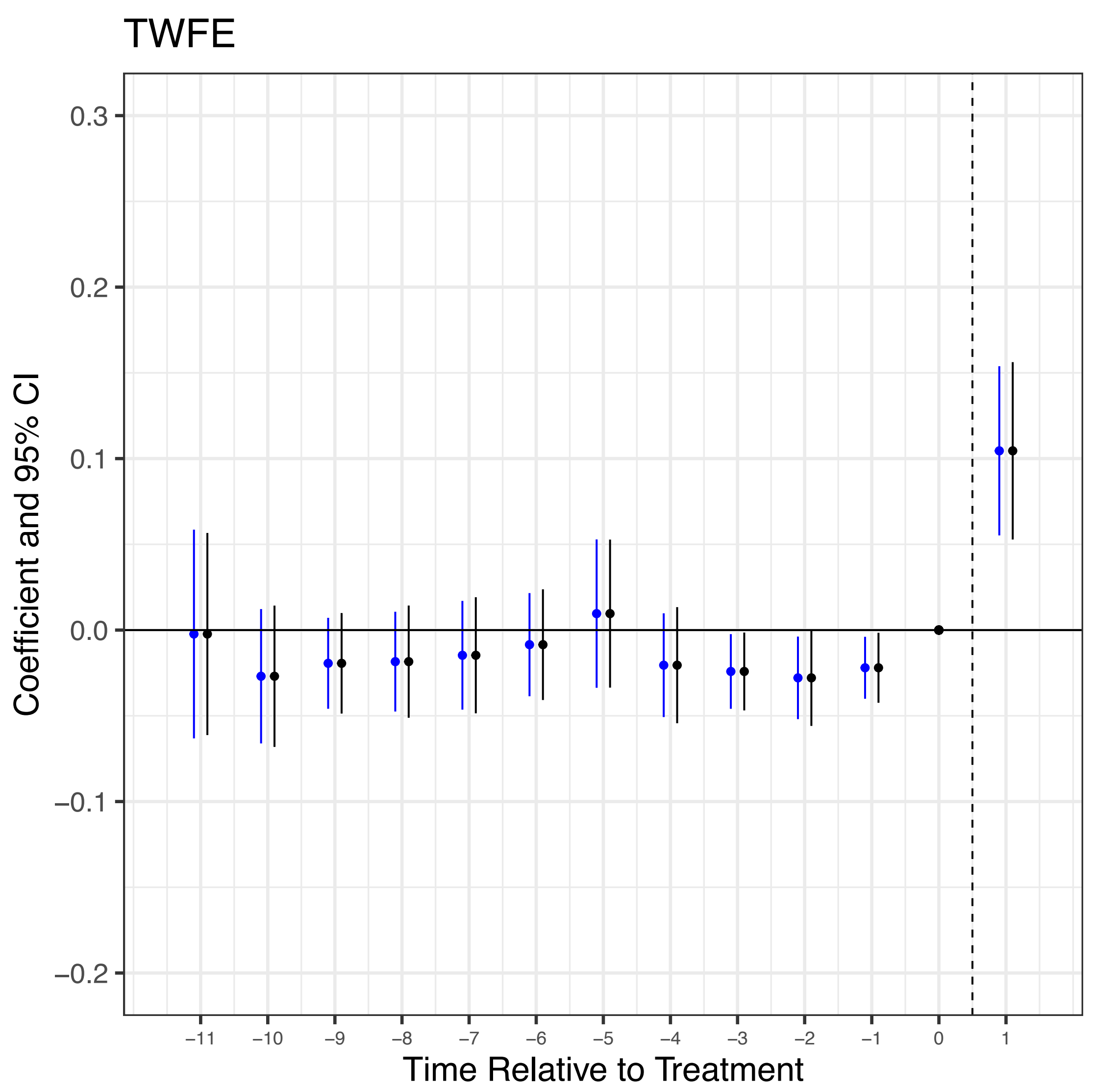
- Grumbach & Sahn (2020): Do minority candidates in US congressional elections mobilize coethnic donators?

  - Treatment: Asian candidates
  - Outcome: share of Asian donations
  - Sample size:
    - N: 489
    - T: 17 (1980-2012)
    - #obs: 7,141

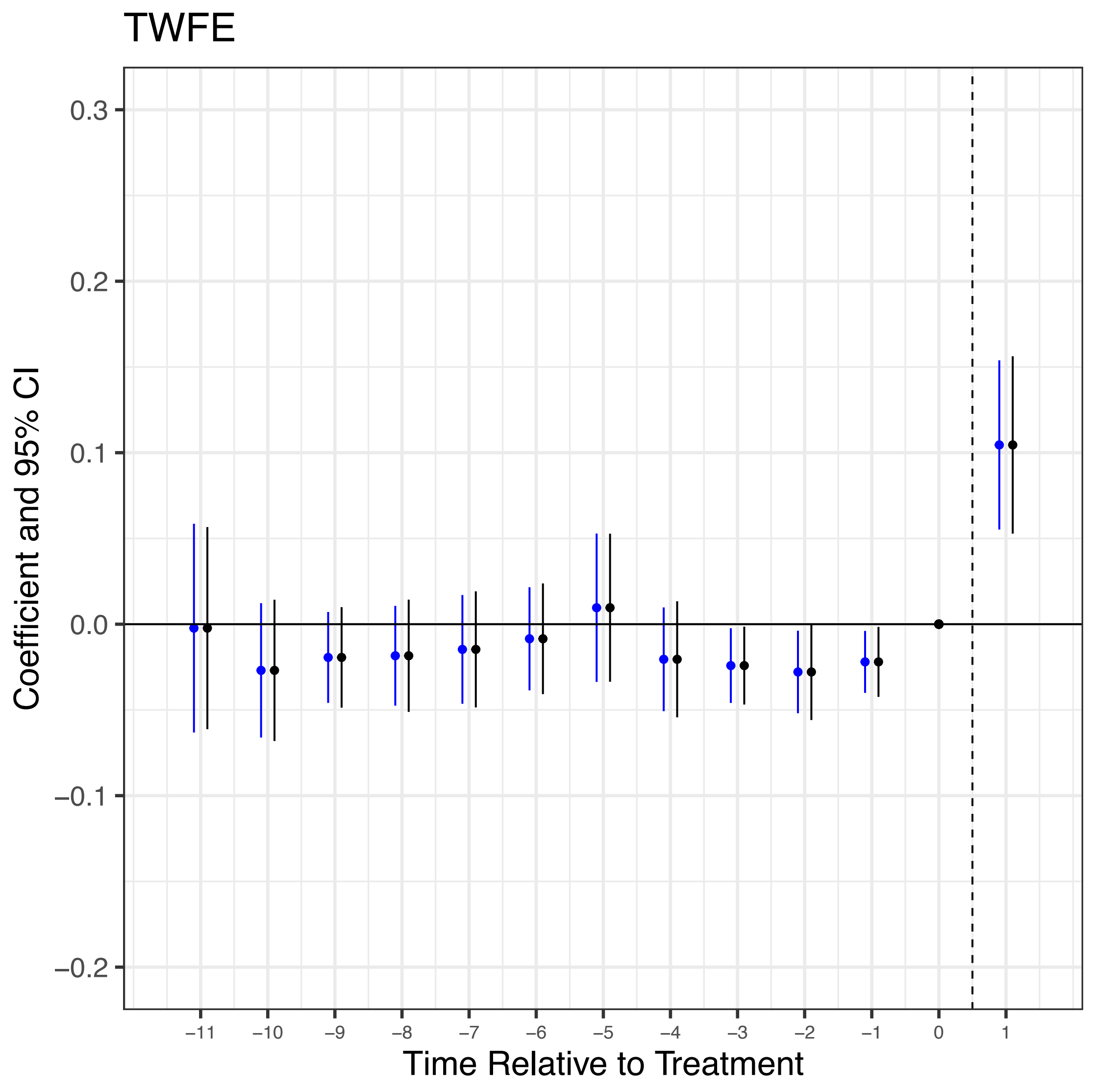


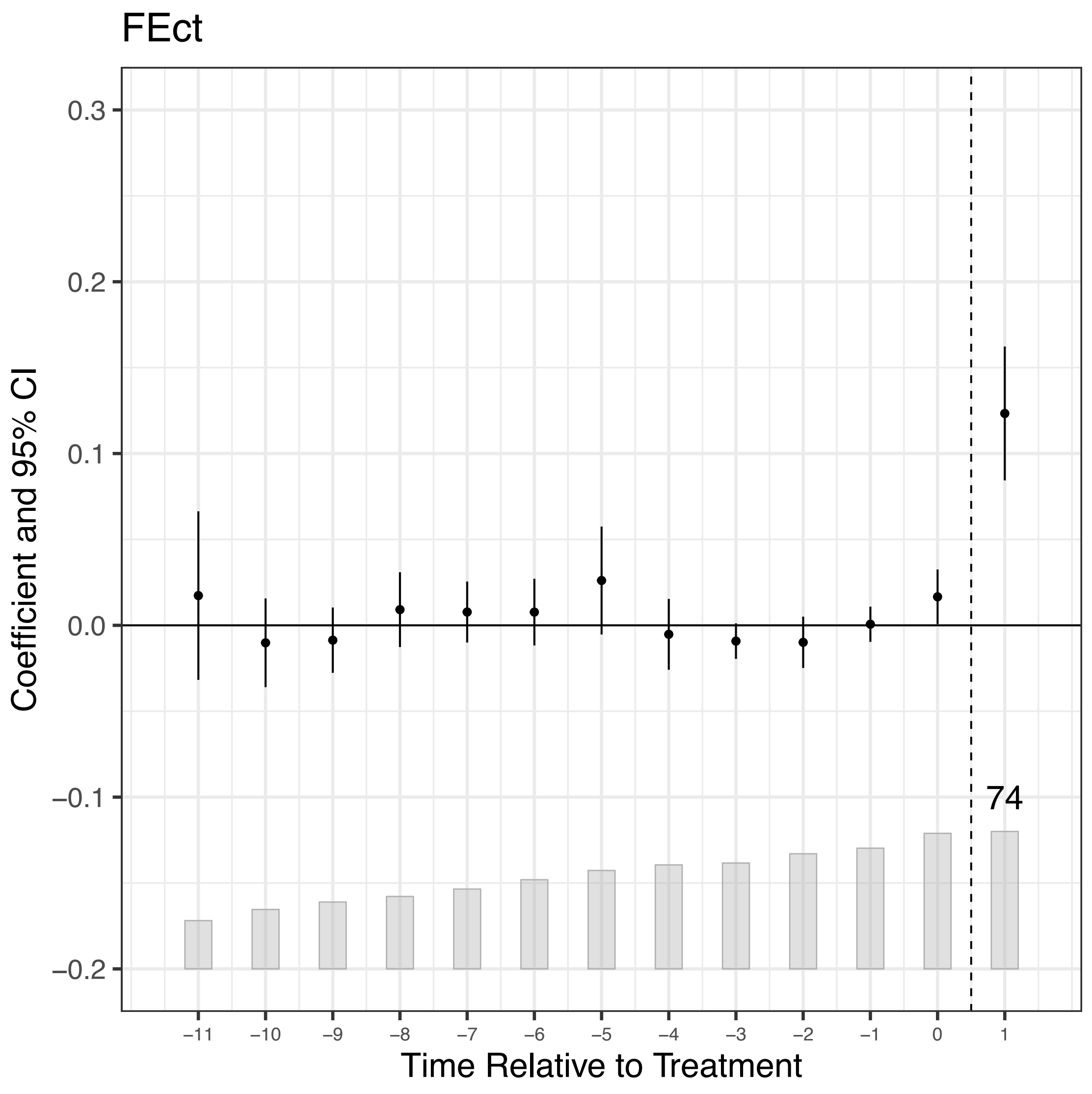
# Comparing Estimators

# Comparing Estimators

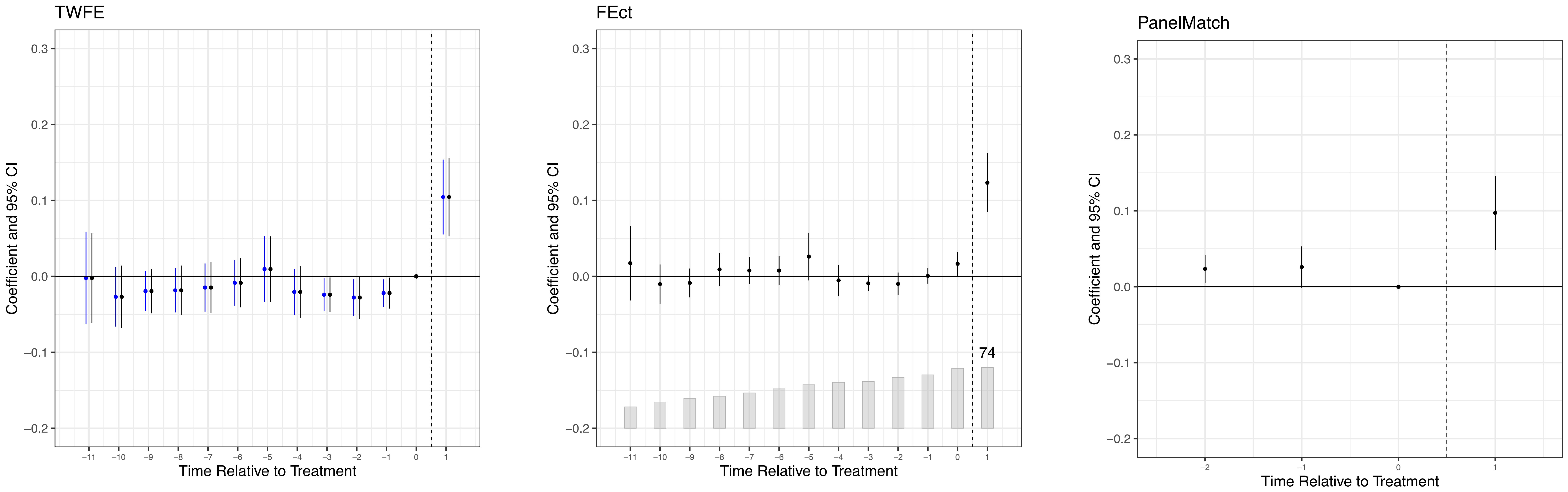


# Comparing Estimators

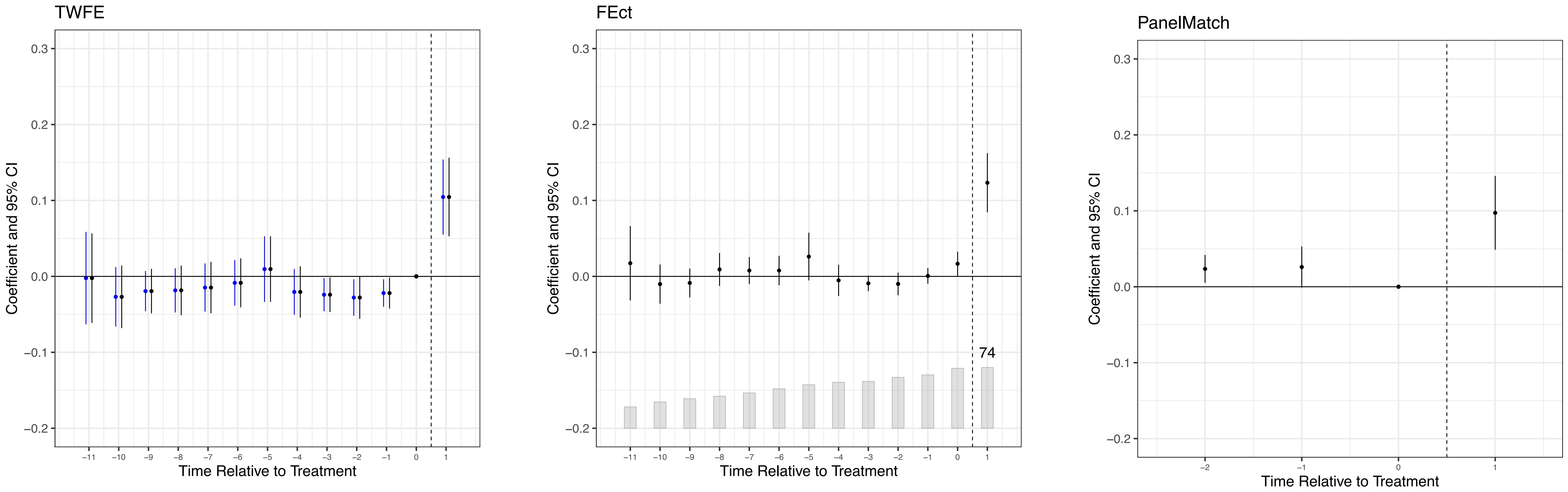


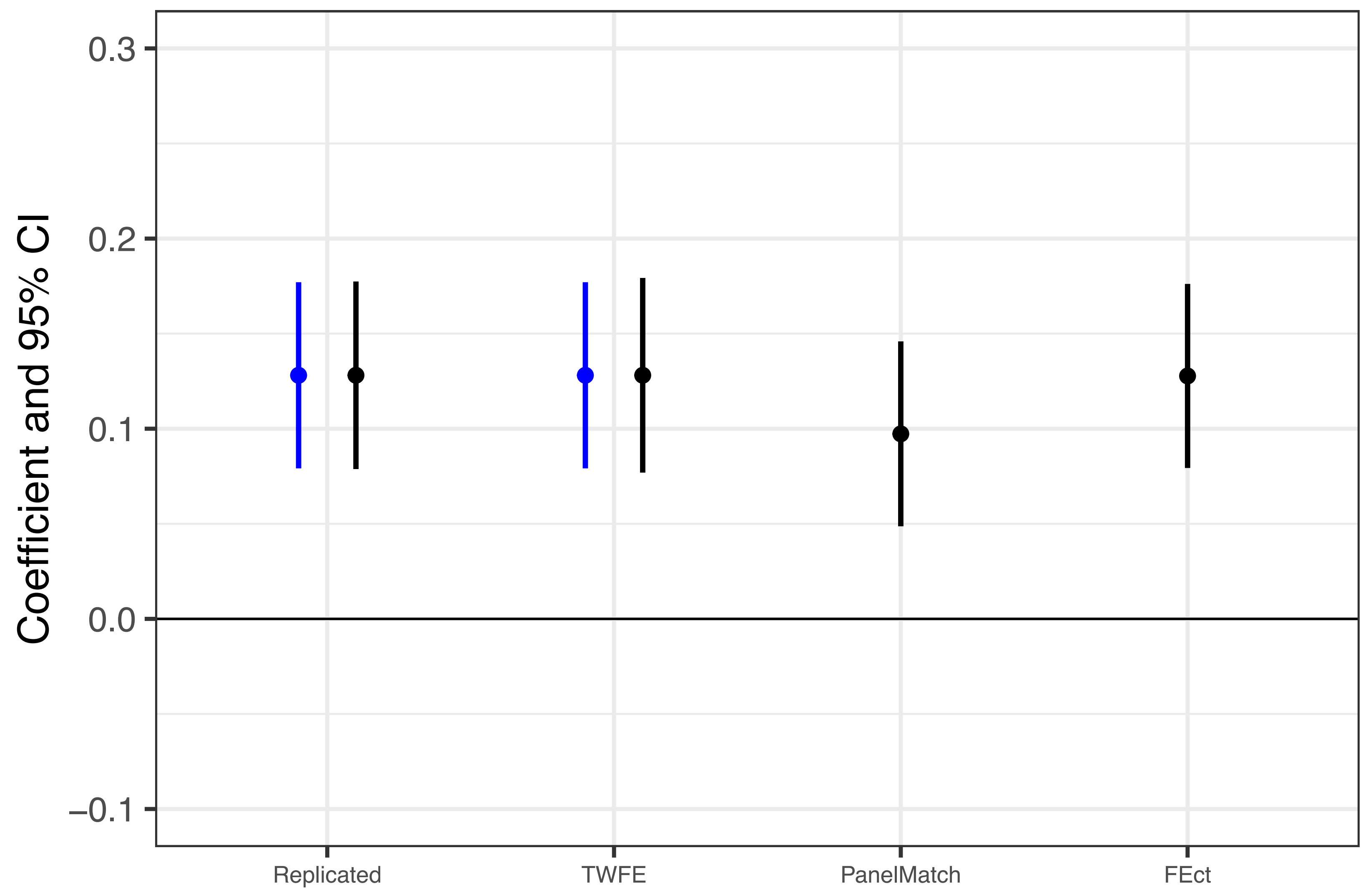
# Comparing Estimators



# Comparing Estimators

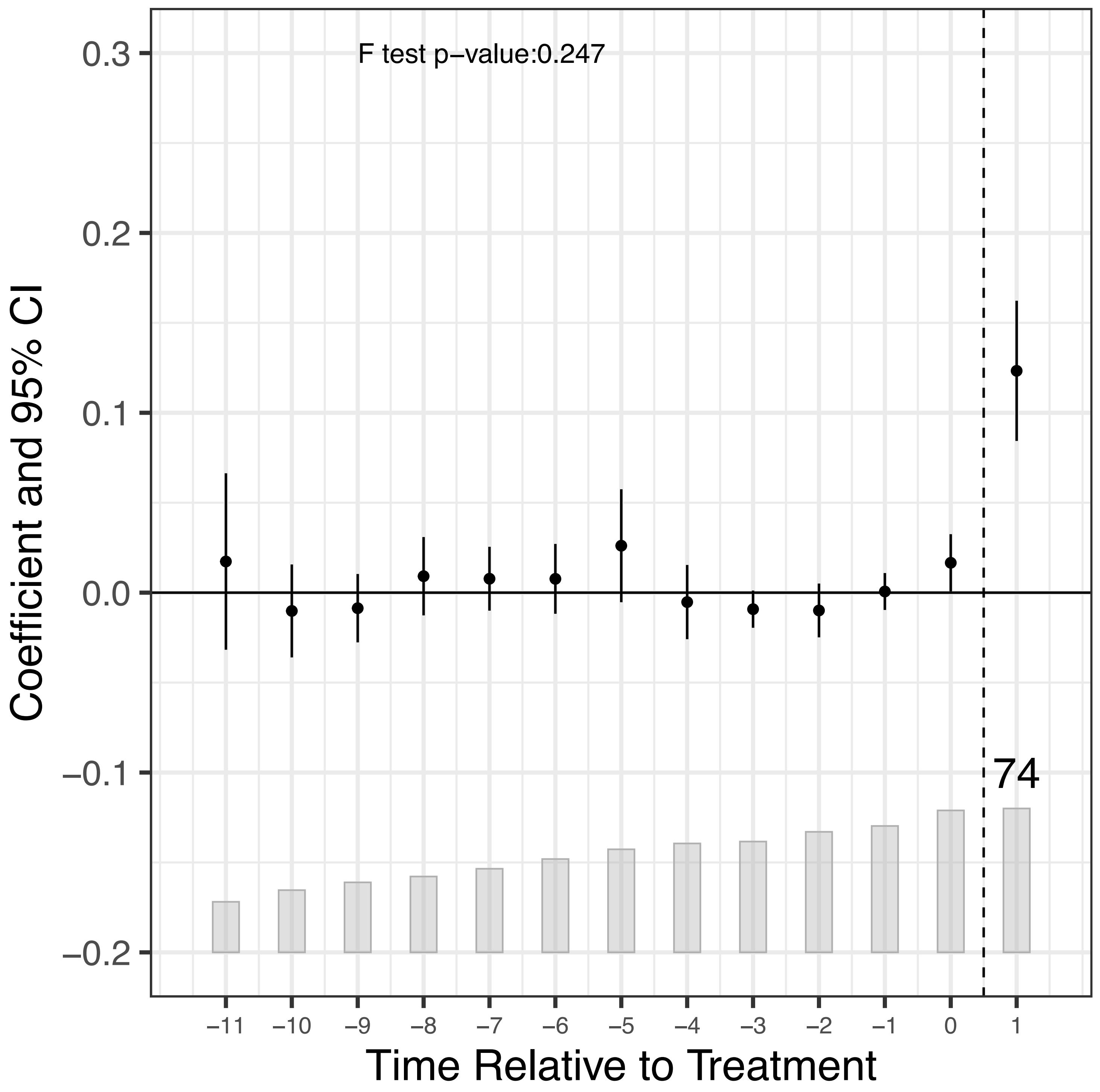



Comparison of Estimators

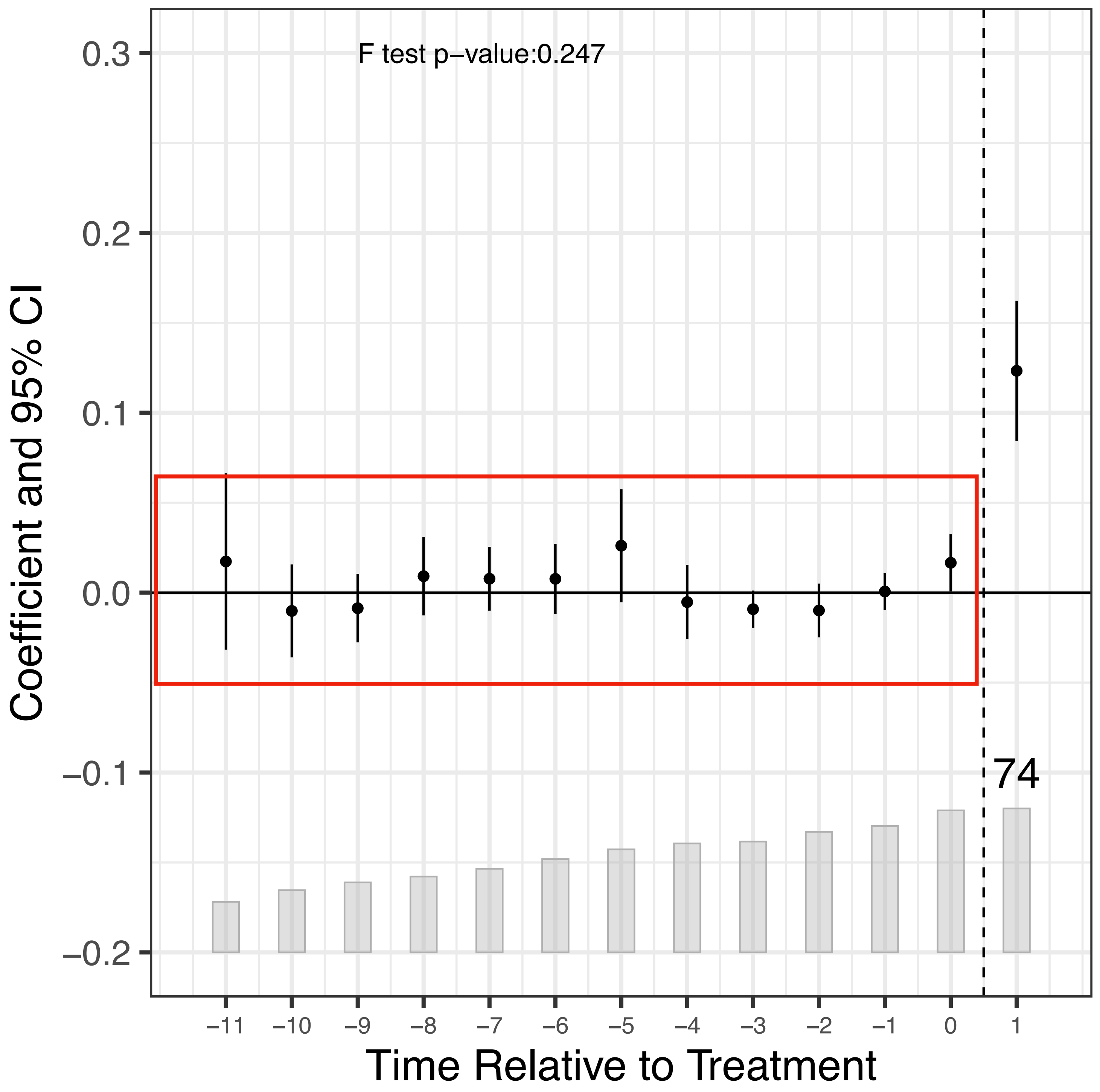


# Diagnostics

# Diagnostics

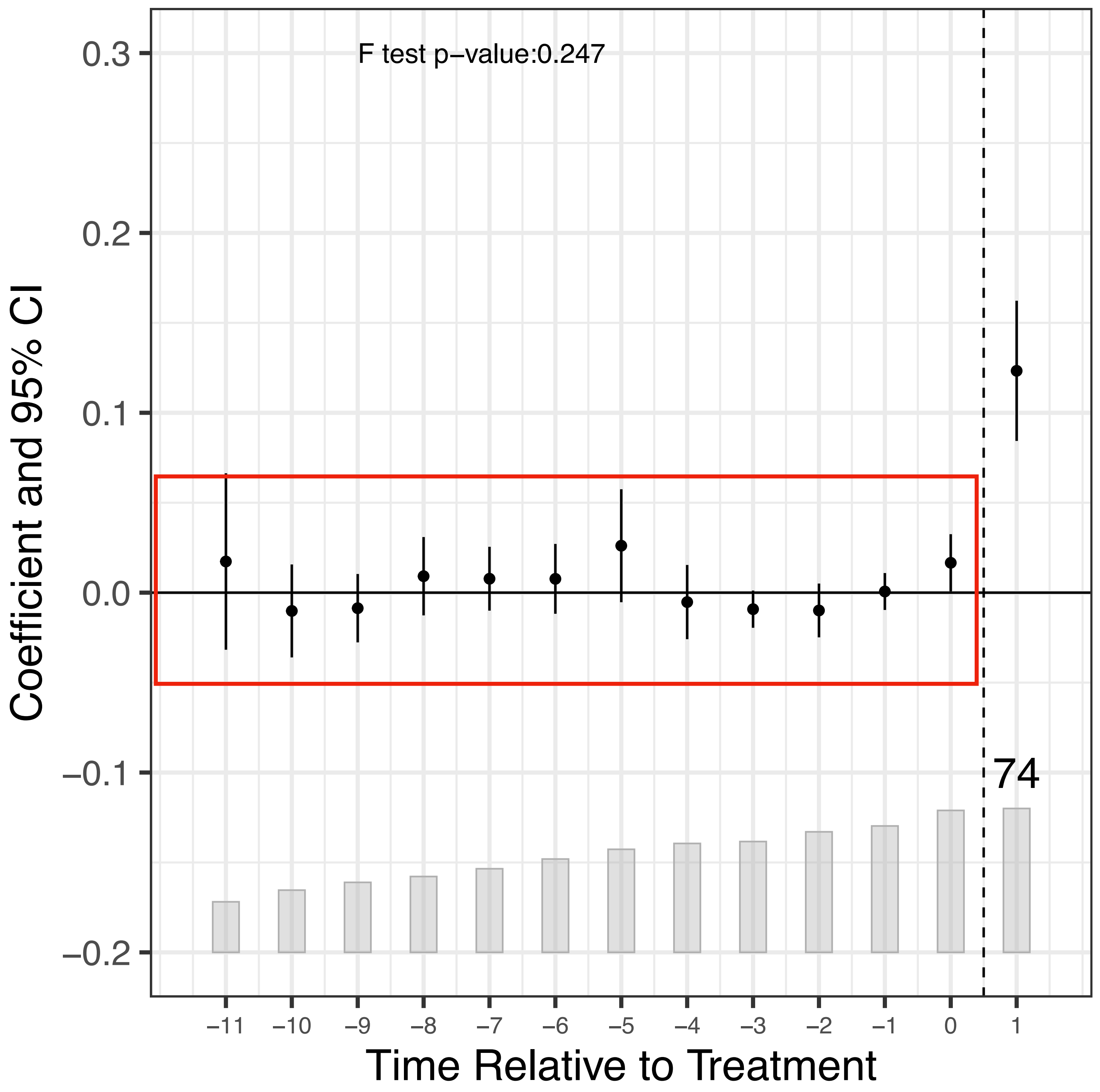


# Diagnostics

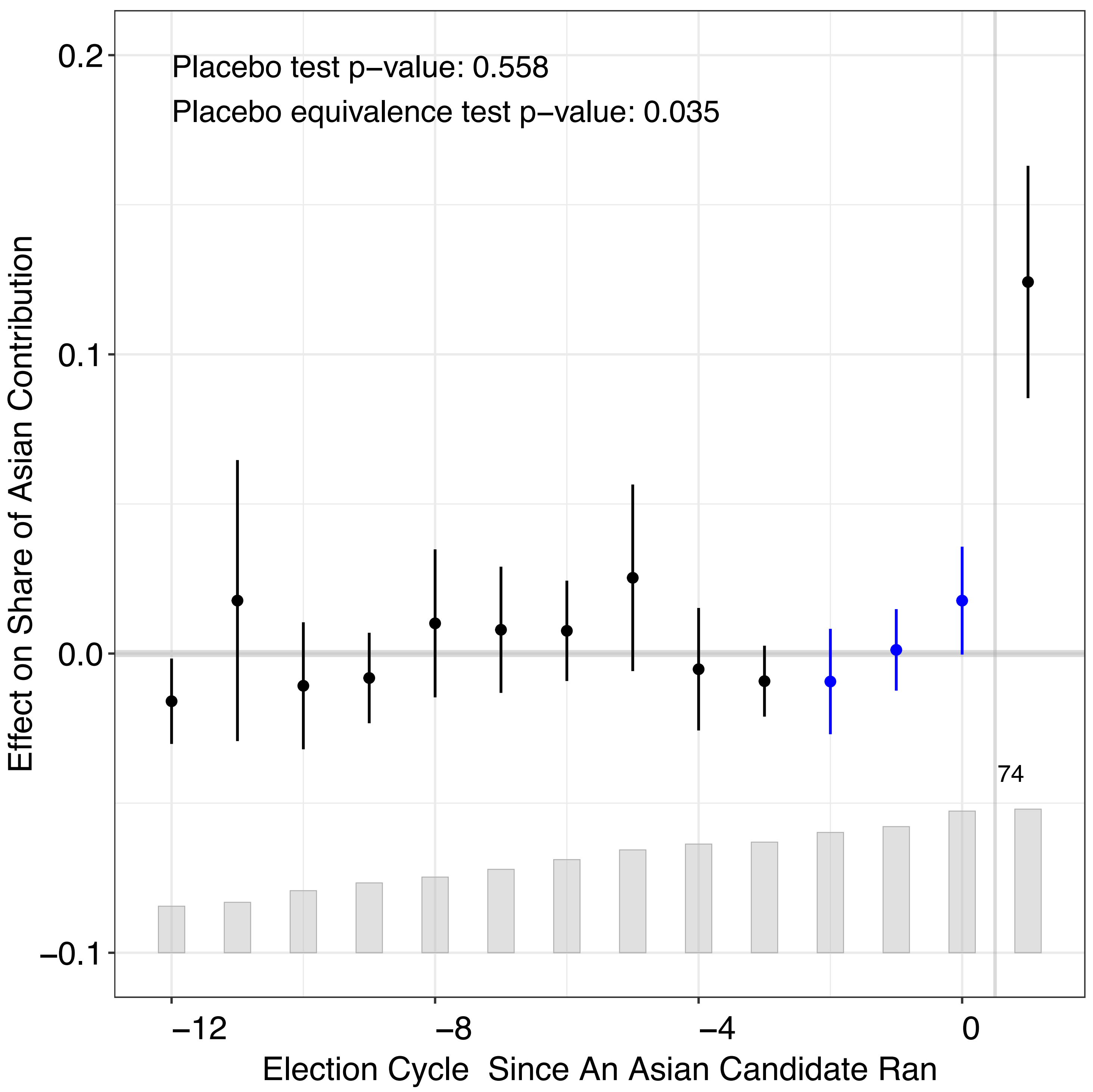



$F$ test for no pretrend
$p = 0.247$

# Diagnostics

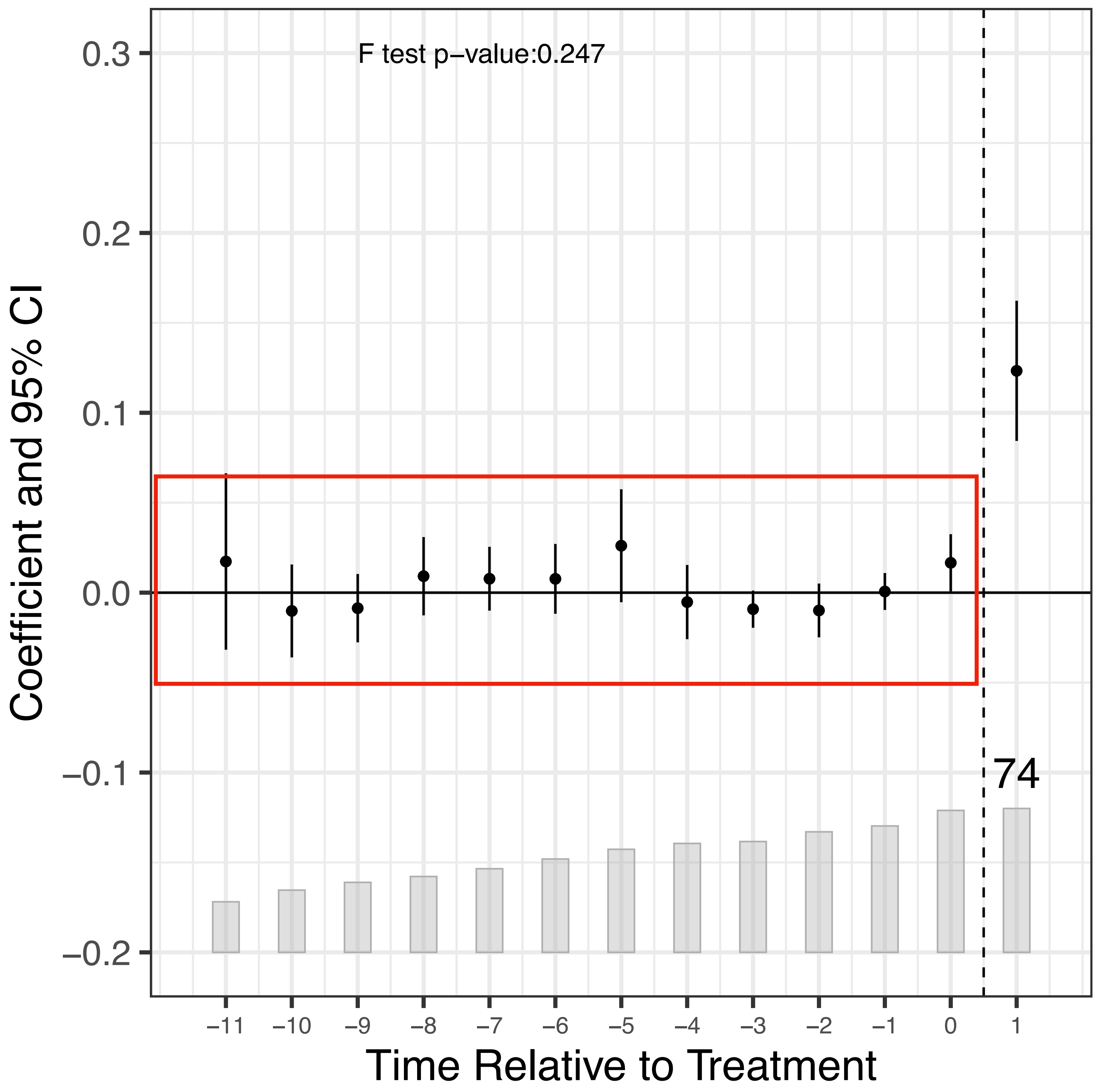


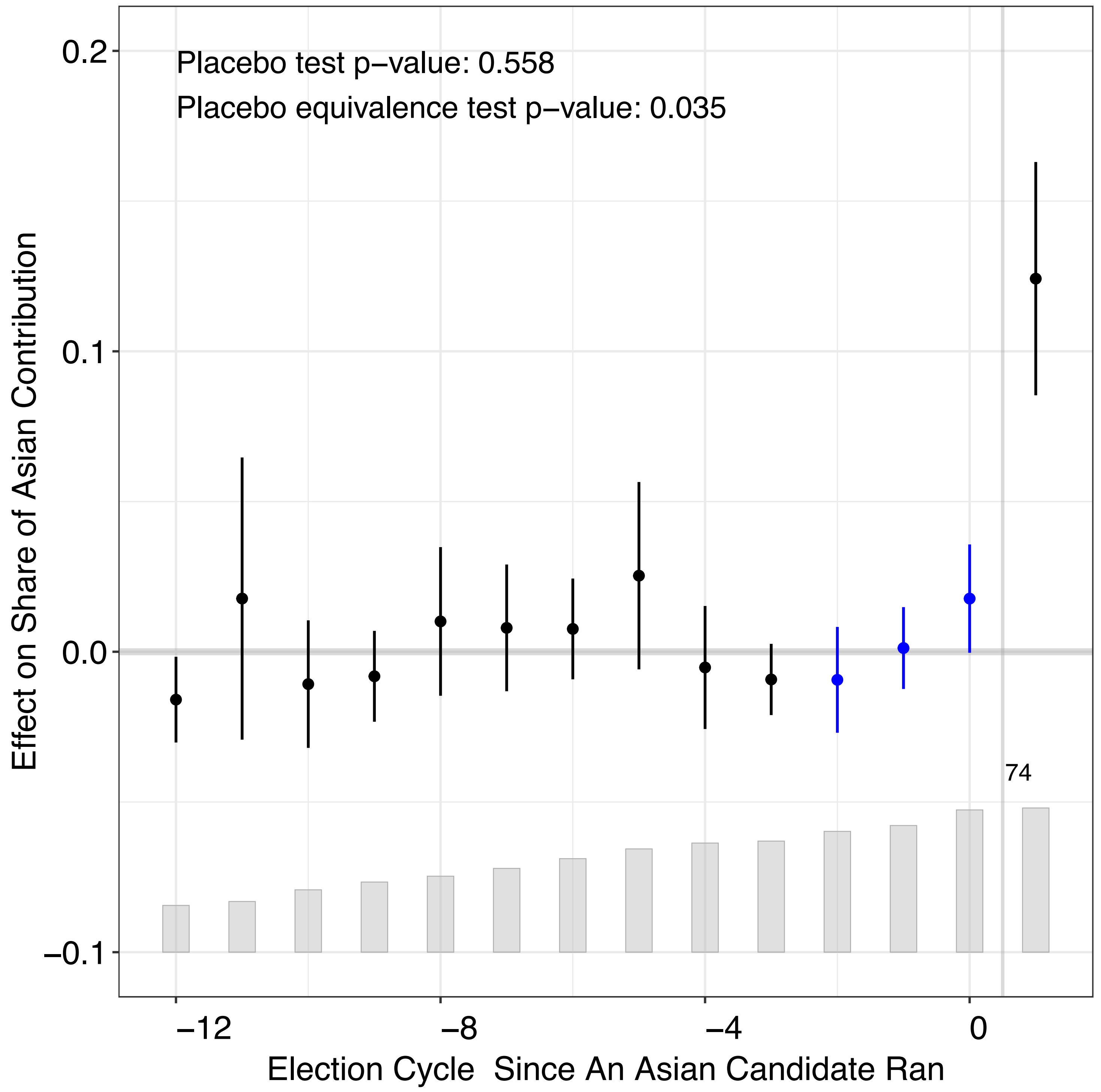

$F$ test for no pretrend
$p = 0.247$




*placebo test*
$p = 0.558$

# Diagnostics




*F* test for no pretrend

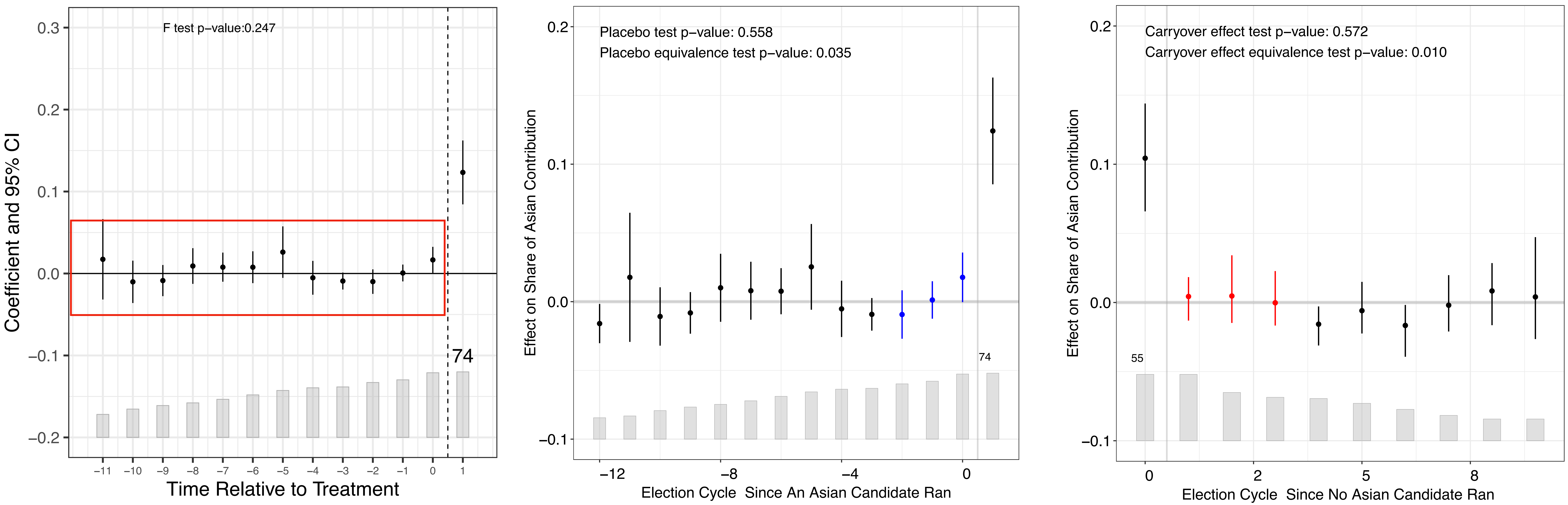
$p = 0.247$

*placebo test*

$p = 0.558$

*Assessing Pretrend*

# Diagnostics




*F* test for no pretrend
$p = 0.247$

*placebo test*
$p = 0.558$

*Test for carryover effects*
$p = 0.572$

*Assessing Pretrend*

# Sensitivity Analysis: Relaxing the PT

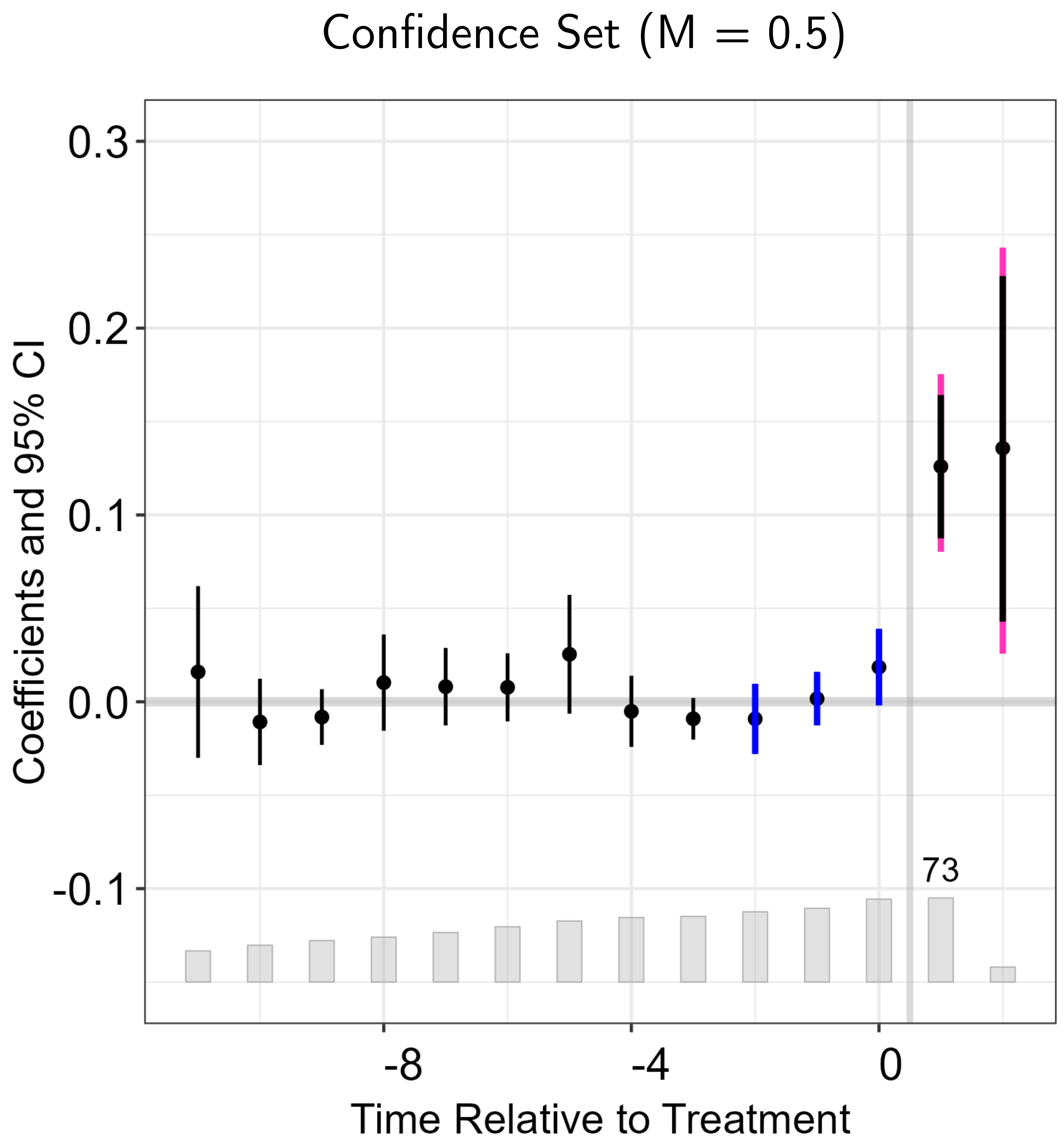
- Addapt Rambachan & Roth (2023)'s Robust Confidence Set to Imputation Estimators

# Sensitivity Analysis: Relaxing the PT

- Addapt Rambachan & Roth (2023)'s Robust Confidence Set to Imputation Estimators
- Allows for post-treatment confounding to be $M$ times the size of the maximum difference between two neighboring placebo periods (assume PT holds exactly iff $M = 0$)

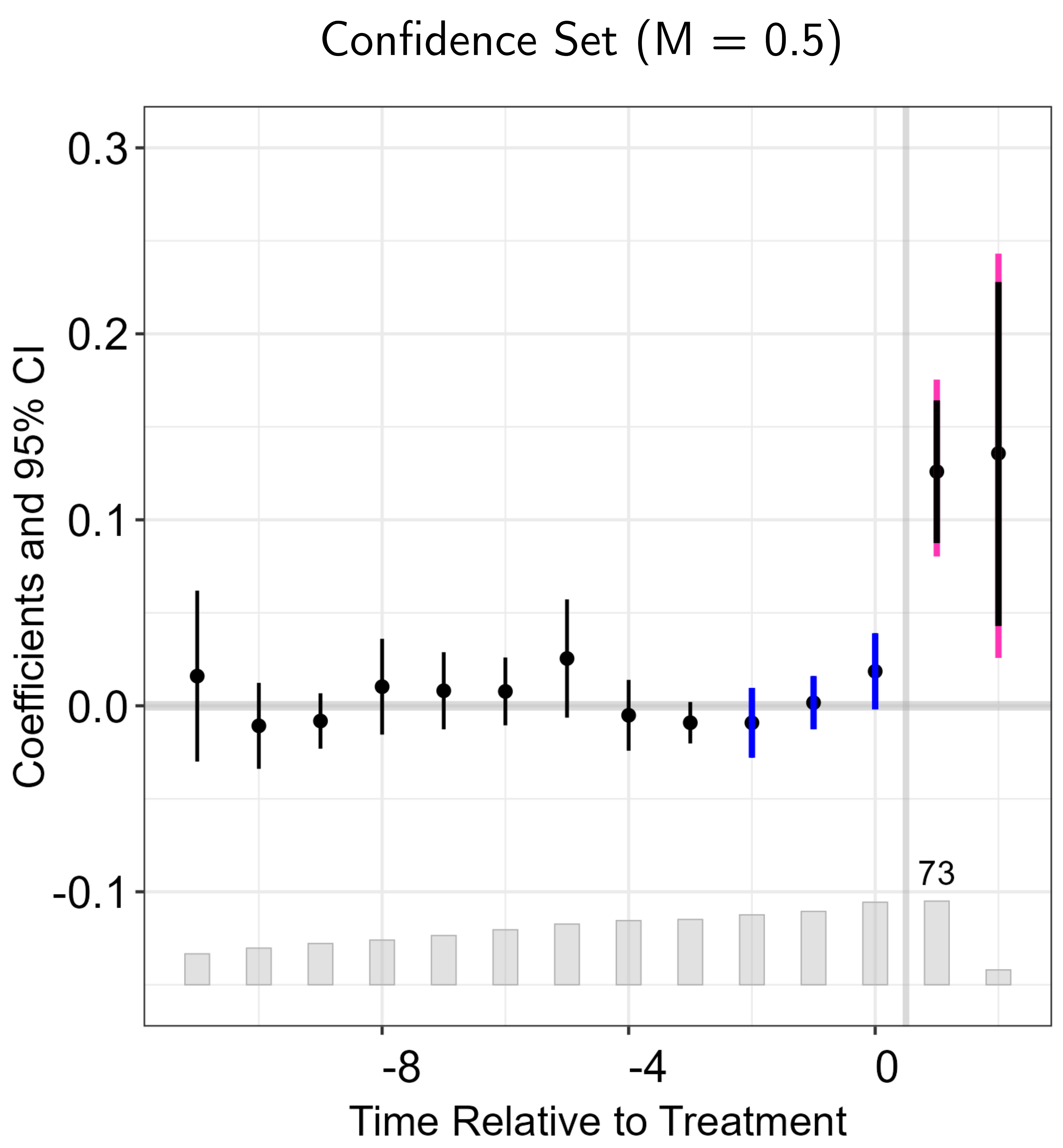
# Sensitivity Analysis: Relaxing the PT

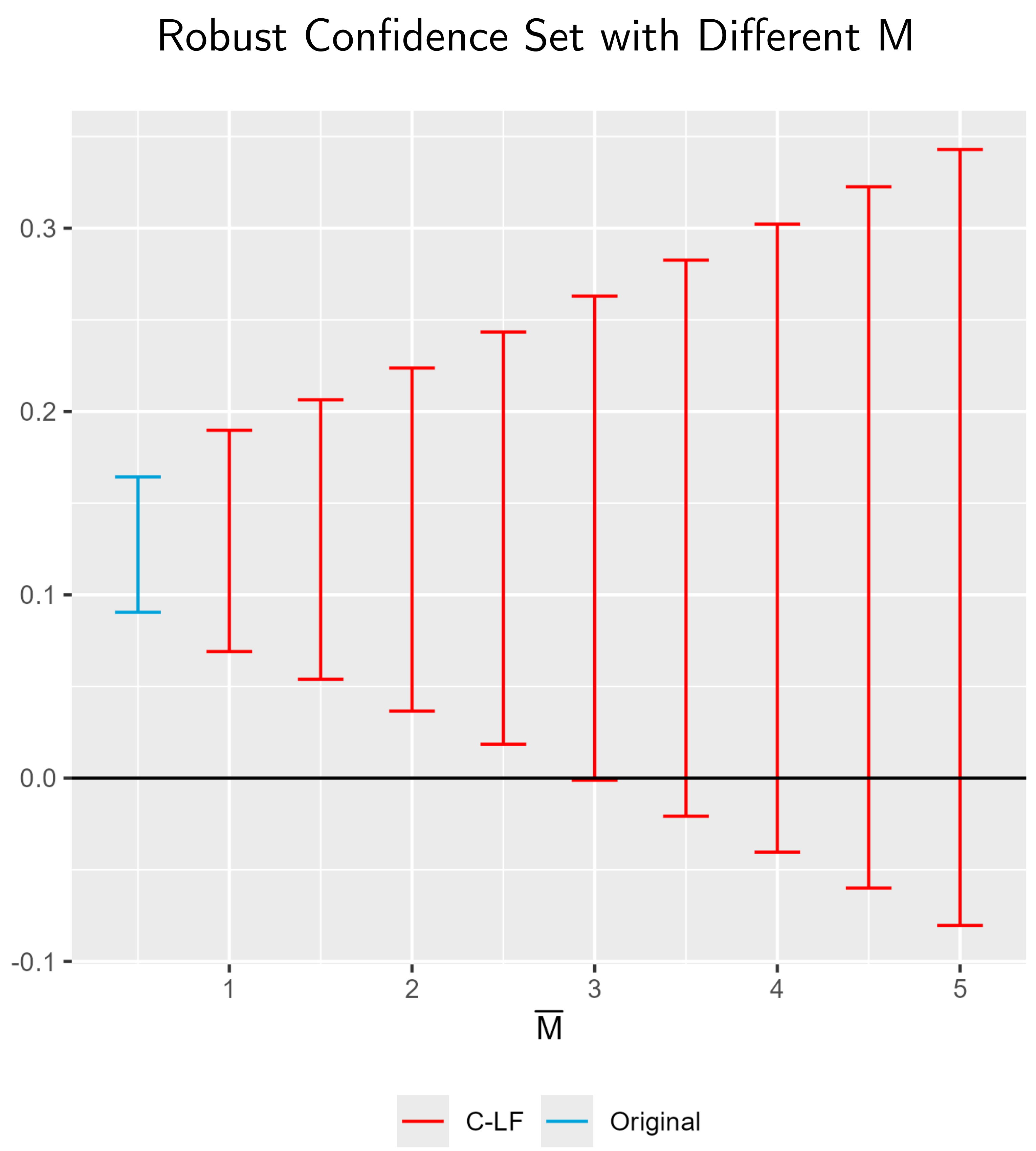
- Addapt Rambachan & Roth (2023)'s Robust Confidence Set to Imputation Estimators
- Allows for post-treatment confounding to be $M$ times the size of the maximum difference between two neighboring placebo periods (assume PT holds exactly iff $M = 0$)



# Sensitivity Analysis: Relaxing the PT

- Addapt Rambachan & Roth (2023)'s Robust Confidence Set to Imputation Estimators
- Allows for post-treatment confounding to be $M$ times the size of the maximum difference between two neighboring placebo periods (assume PT holds exactly iff $M = 0$)



# Three Examples

- **Example 1**: Coethnic Mobilization

# Three Examples

- **Example 1**: Coethnic Mobilization
  - Strong design; HTE matters marginally — estimators (including TWFE) broadly agree

# Example 2: Lawsuit against Land Use Restriction

# Example 2: Lawsuit against Land Use Restriction

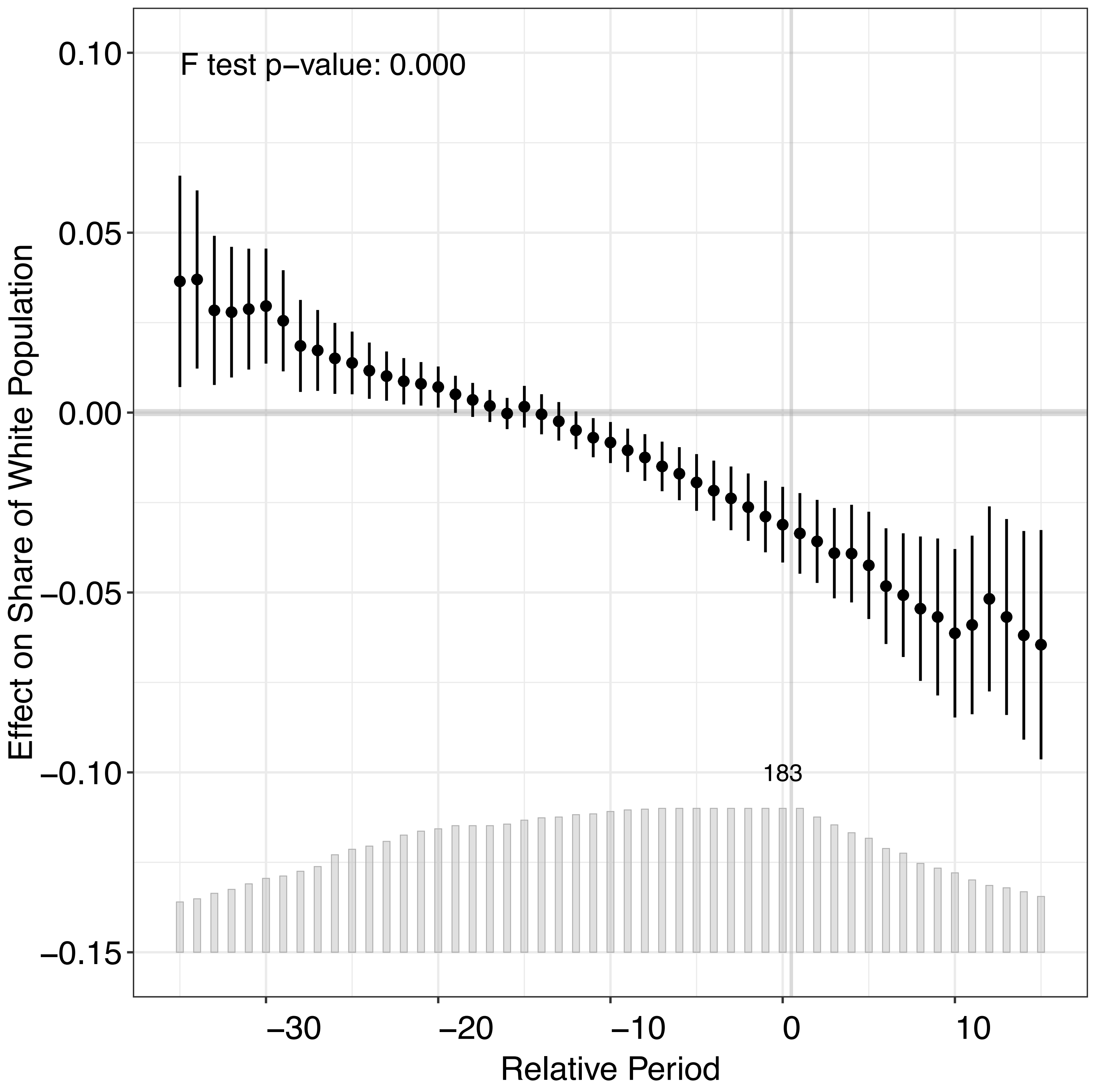
- Treatment: Fair Housing Act lawsuits against city land-use restrictions

# Example 2: Lawsuit against Land Use Restriction

- Treatment: Fair Housing Act lawsuits against city land-use restrictions
- Outcome: racial compositions of city dwellers in California

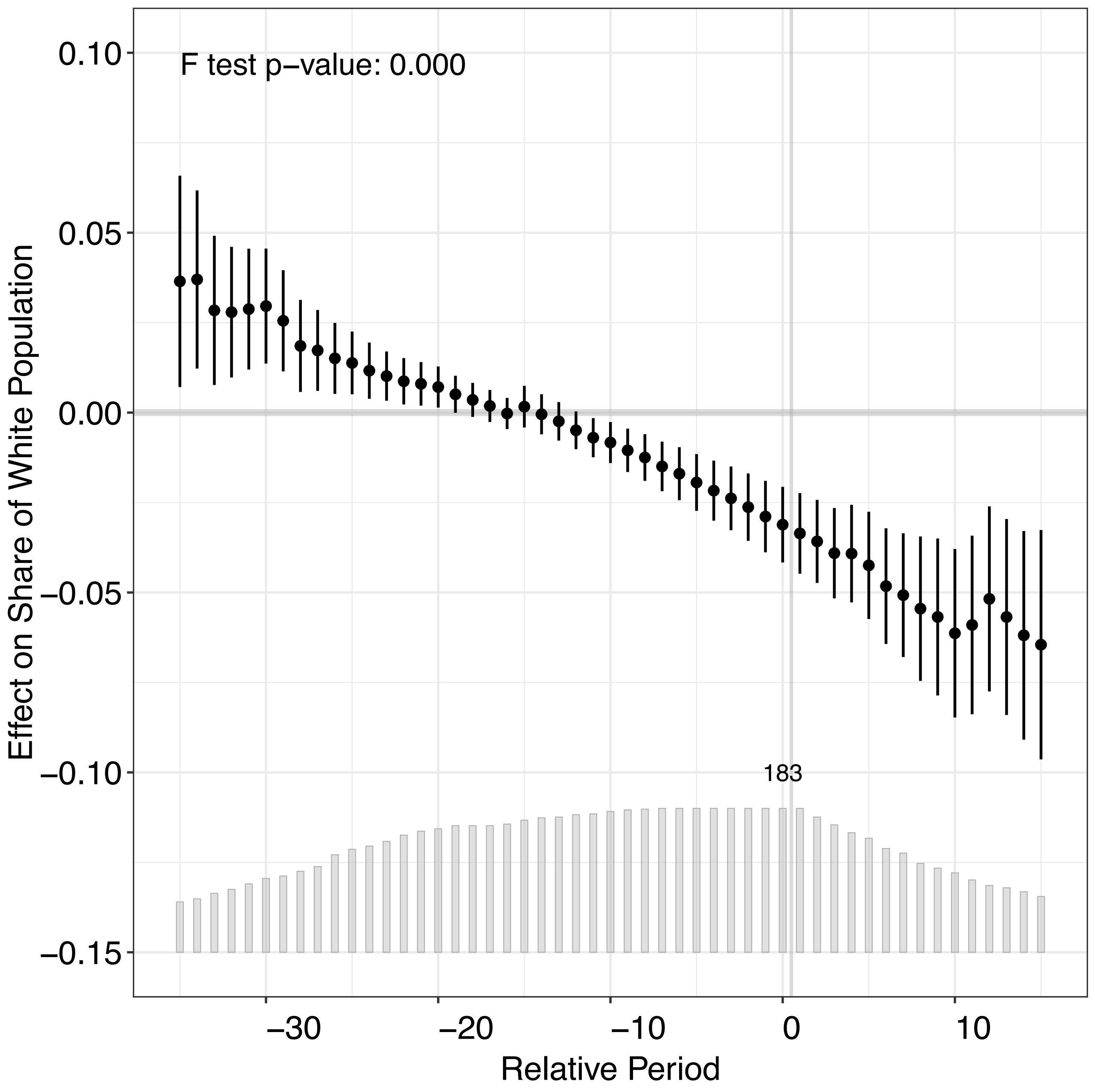
# Example 2: Lawsuit against Land Use Restriction

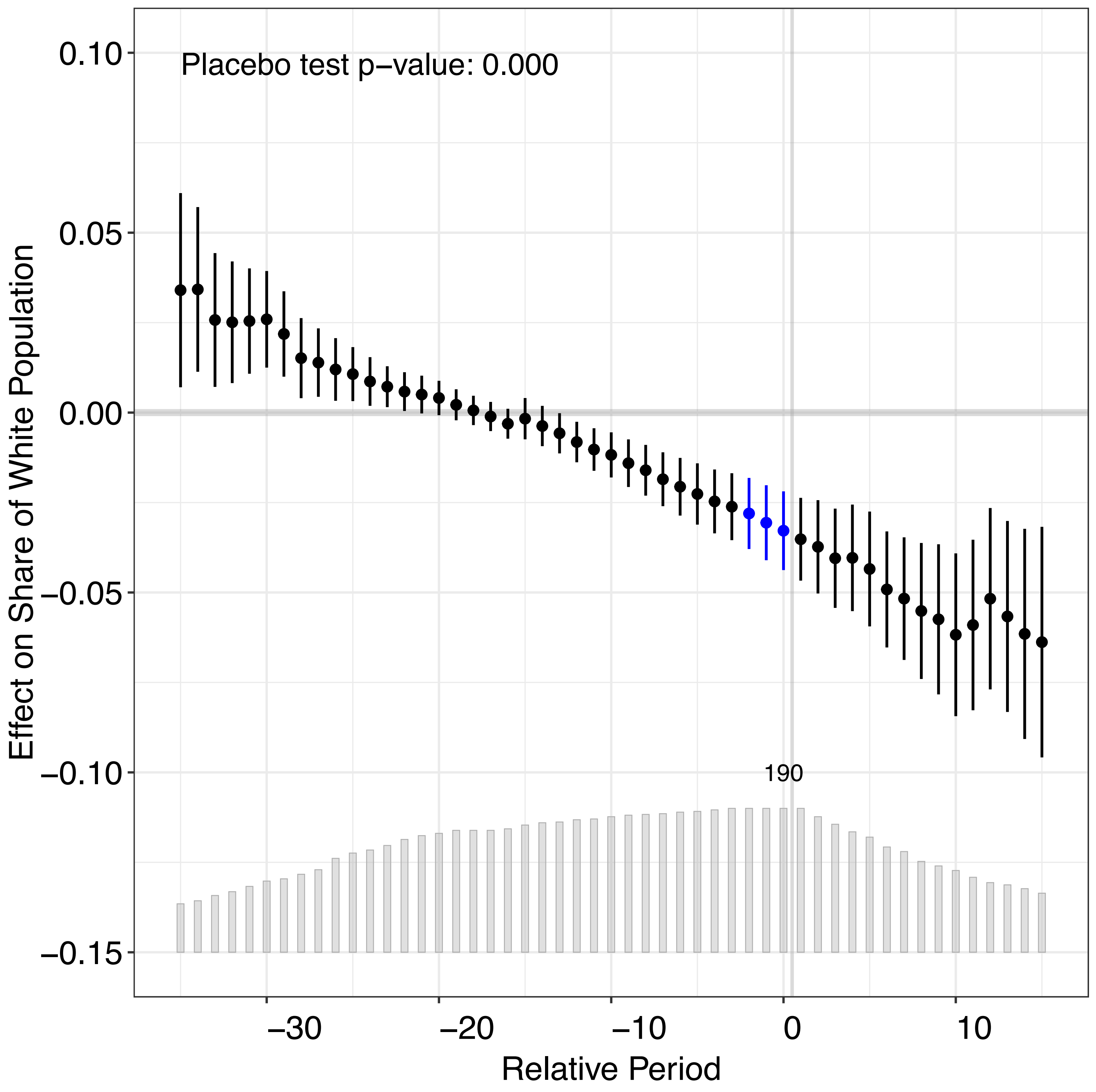
- Treatment: Fair Housing Act lawsuits against city land-use restrictions
- Outcome: racial compositions of city dwellers in California



# Example 2: Lawsuit against Land Use Restriction

- Treatment: Fair Housing Act lawsuits against city land-use restrictions
- Outcome: racial compositions of city dwellers in California



# Removing Interpolated Data and Adding Time Trends

# Removing Interpolated Data and Adding Time Trends

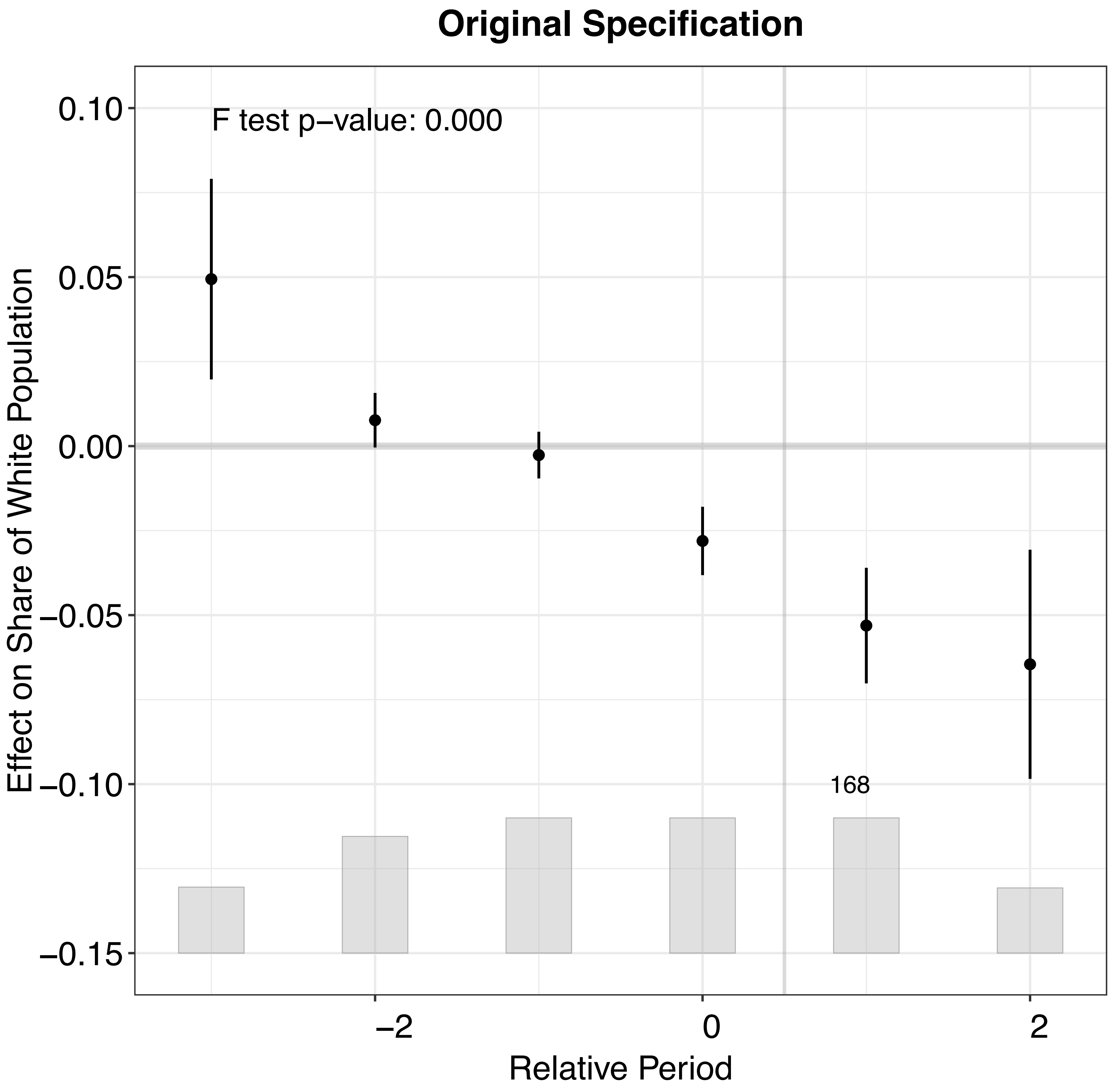
- Demographic data are mostly interpolated based on Census.

# Removing Interpolated Data and Adding Time Trends

- Demographic data are mostly interpolated based on Census.
- Findings are similar once we removed the interpolated data.

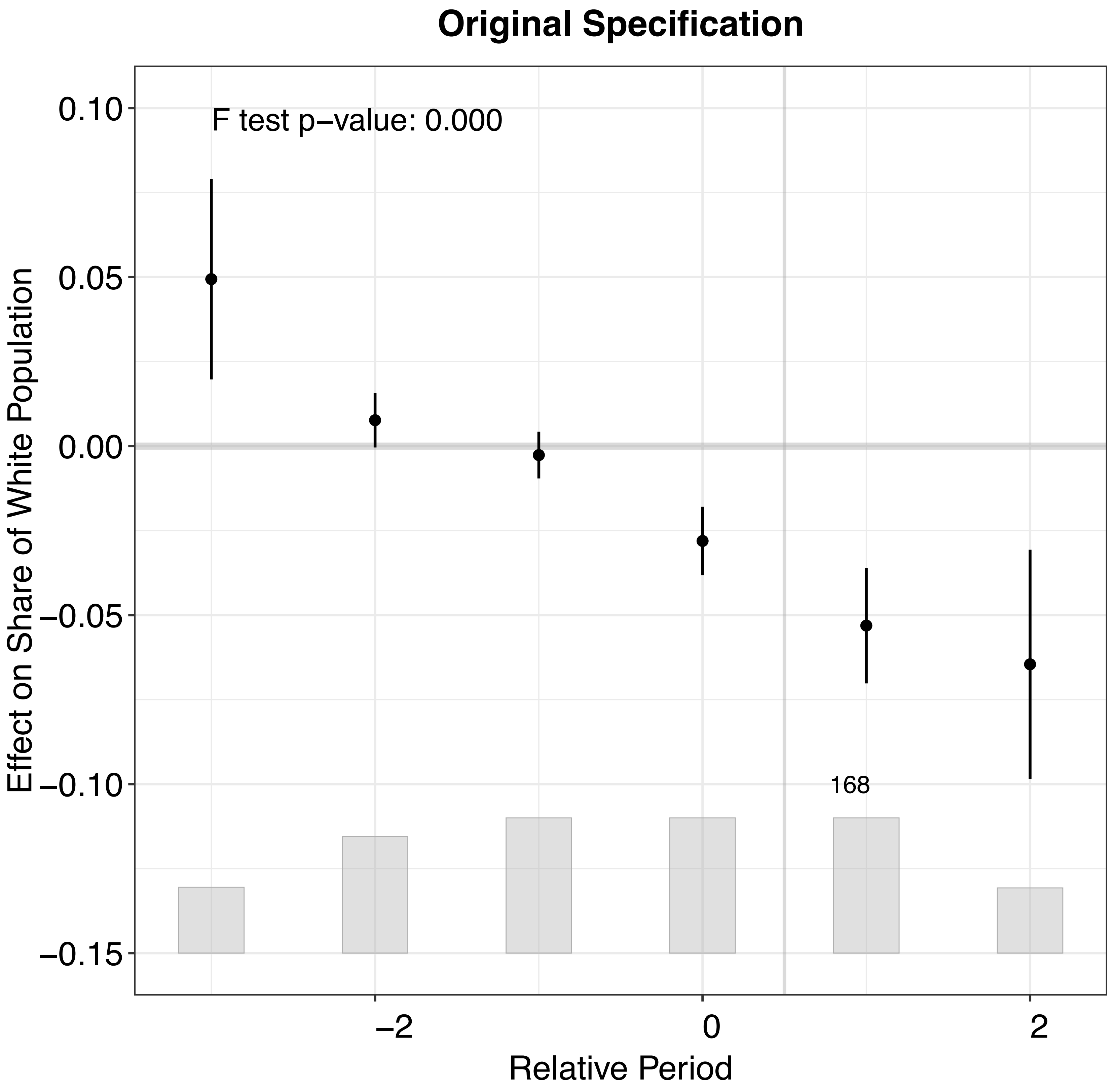
# Removing Interpolated Data and Adding Time Trends

- Demographic data are mostly interpolated based on Census.
- Findings are similar once we removed the interpolated data.



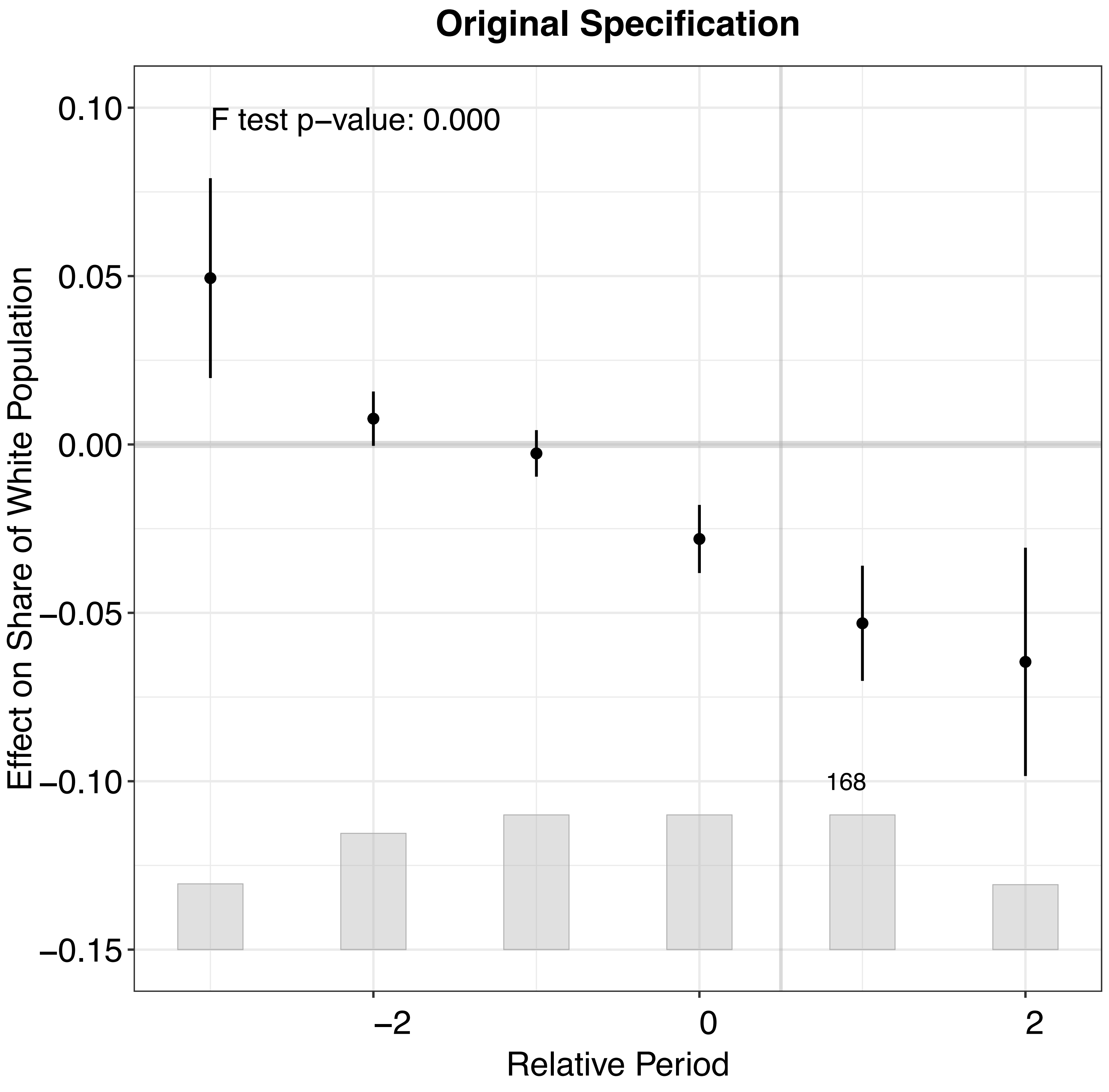
# Removing Interpolated Data and Adding Time Trends

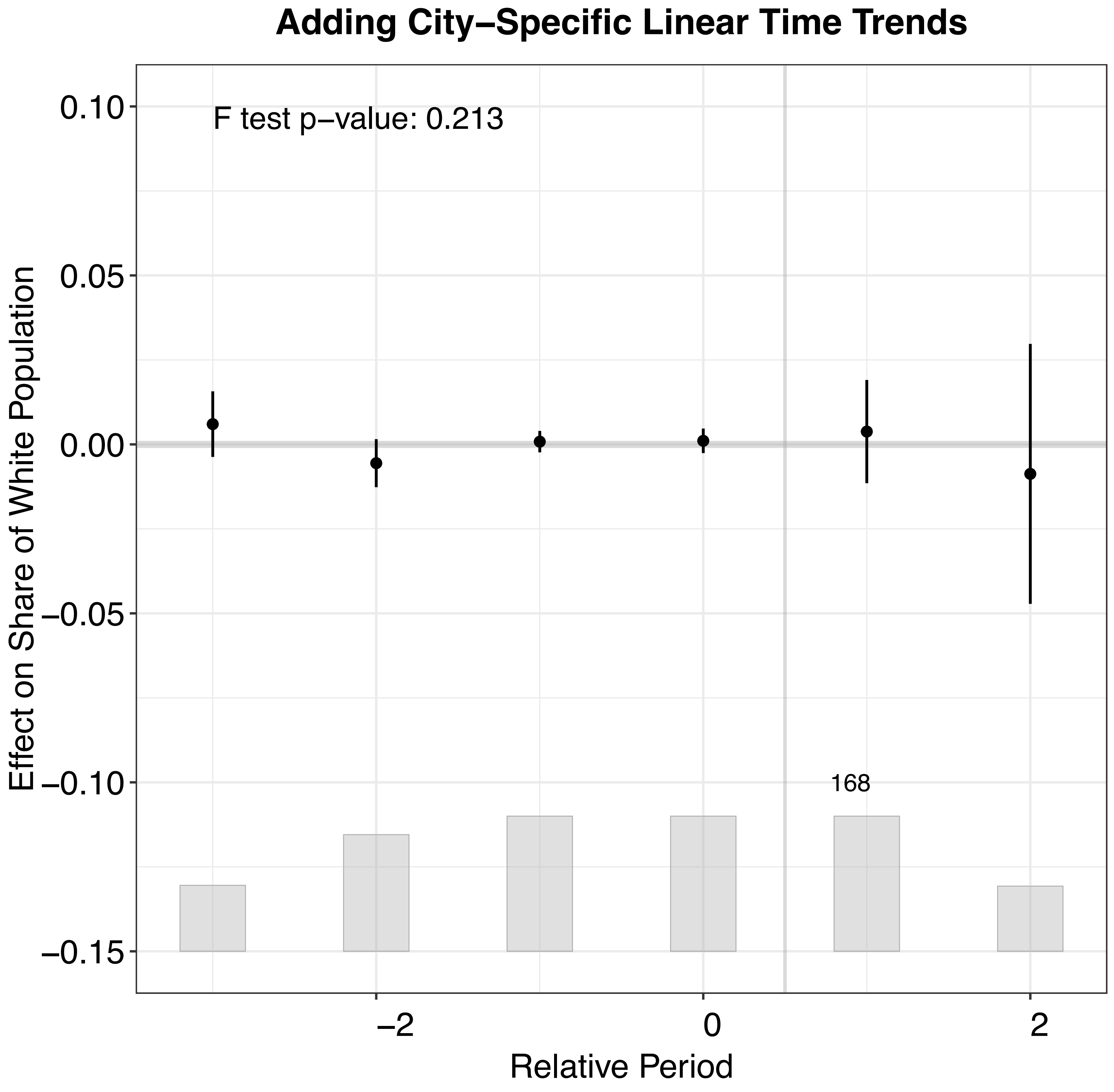
- Demographic data are mostly interpolated based on Census.
- Findings are similar once we removed the interpolated data.
- The negative result is completely gone once we added city-specific linear time-trends



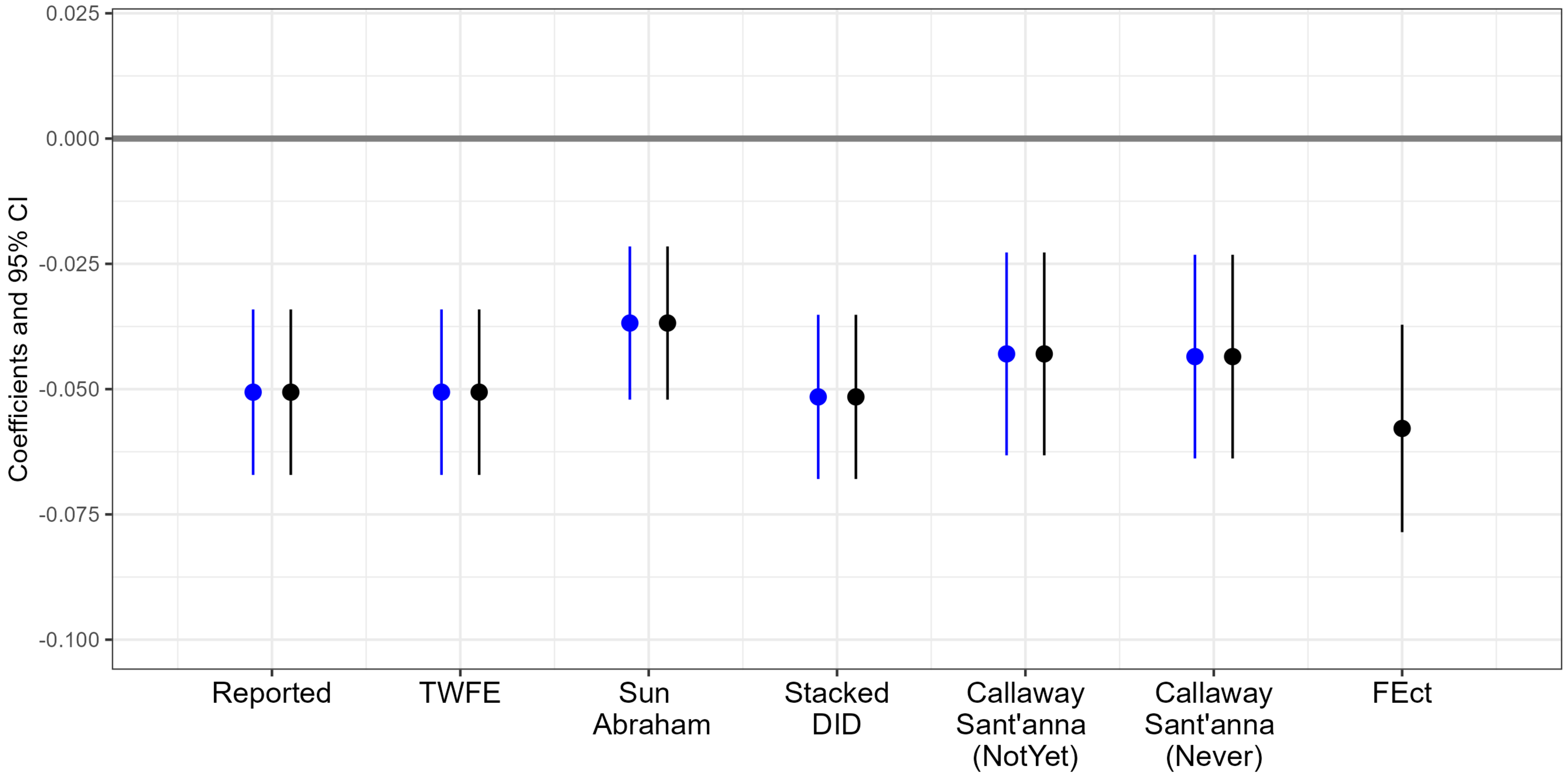
# Removing Interpolated Data and Adding Time Trends

- Demographic data are mostly interpolated based on Census.
- Findings are similar once we removed the interpolated data.
- The negative result is completely gone once we added city-specific linear time-trends

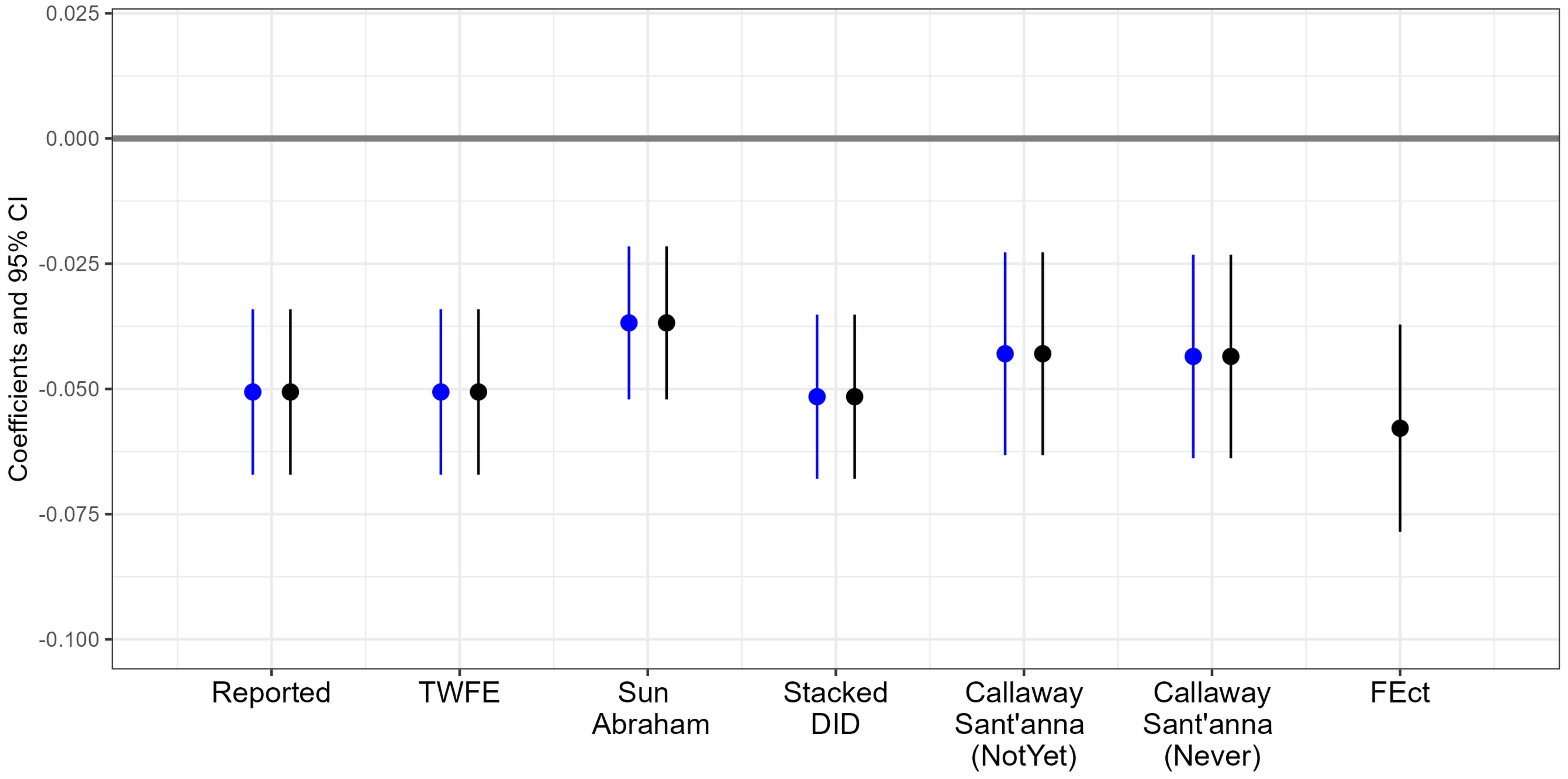


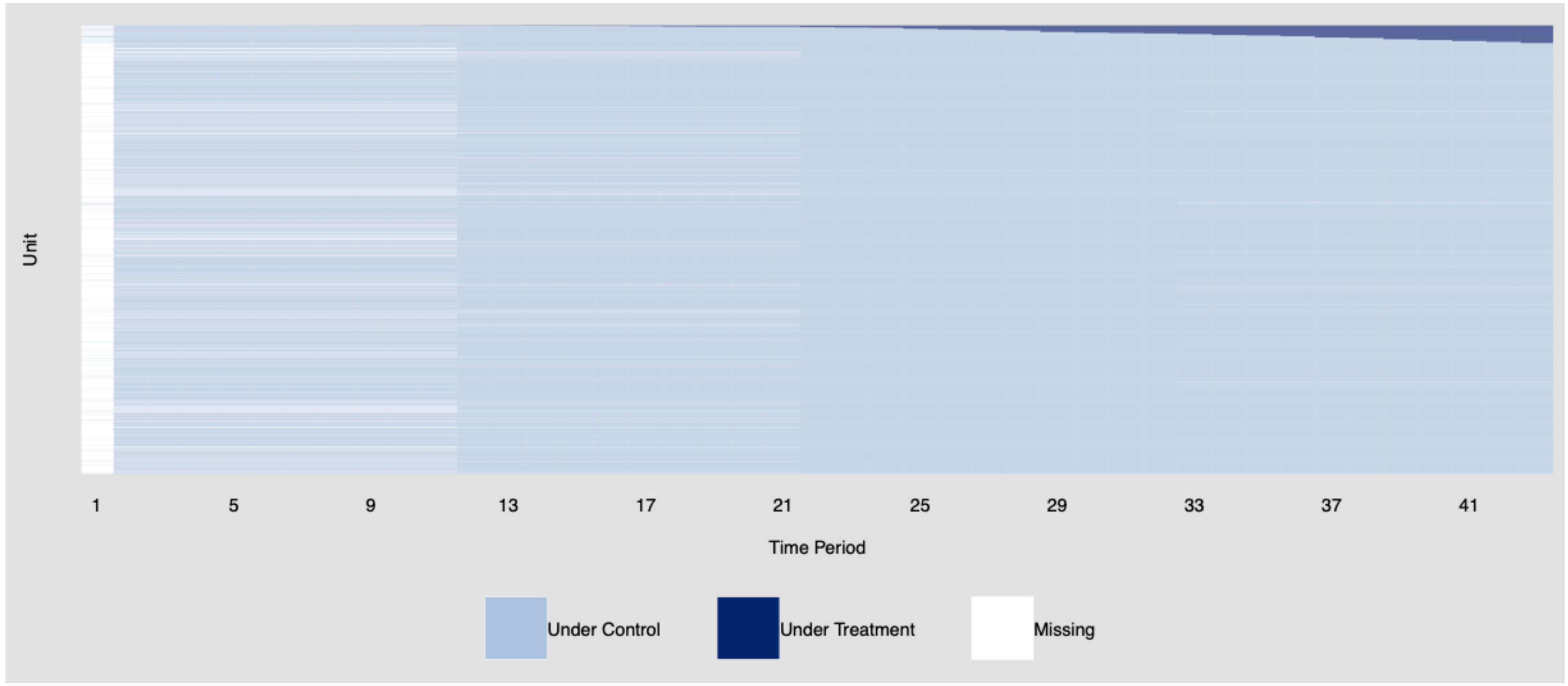
# Various Estimators Still Broadly Agree



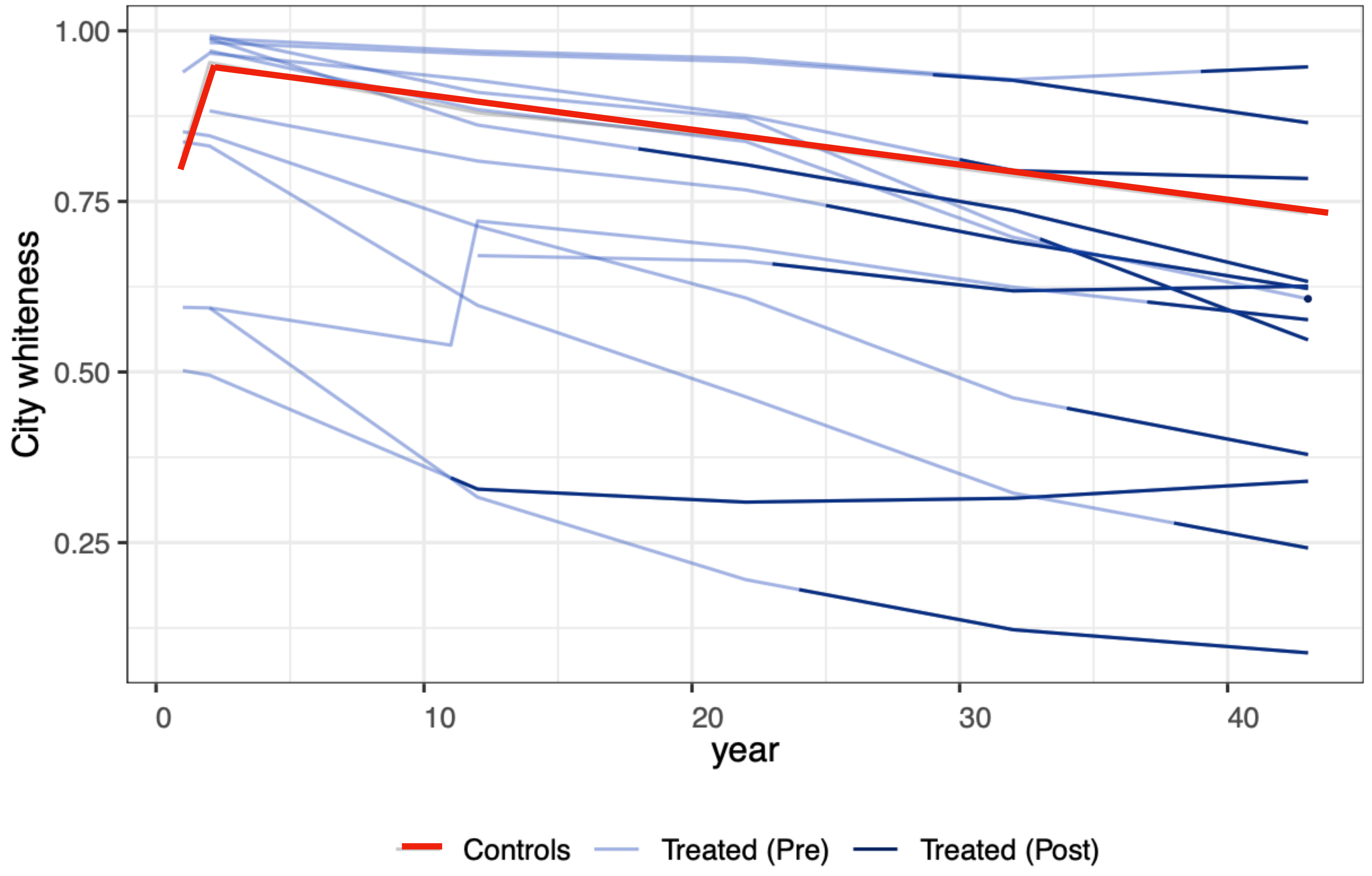
# Various Estimators Still Broadly Agree



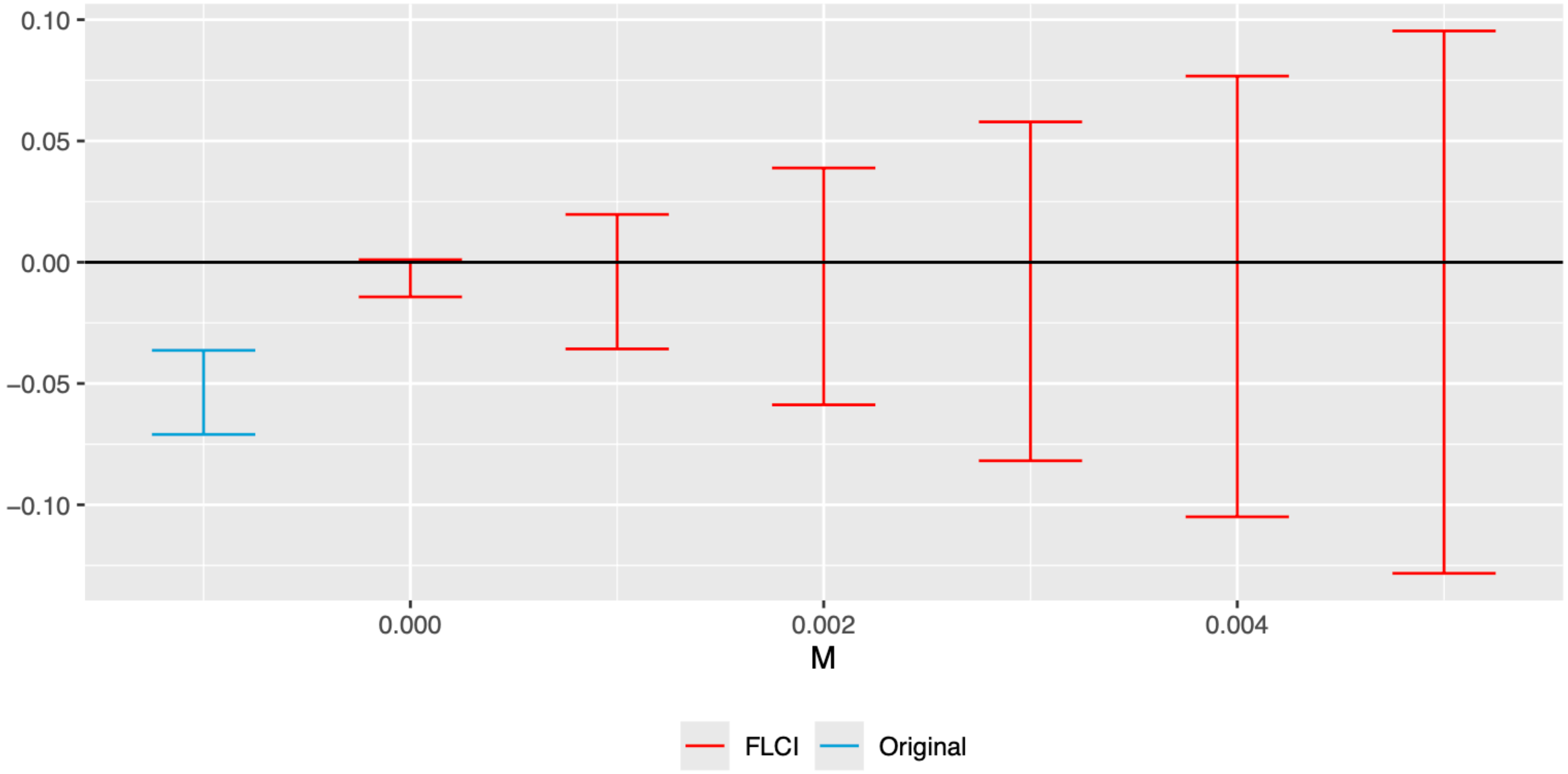

Treatment Status




Outcome Trajectories by Cohort

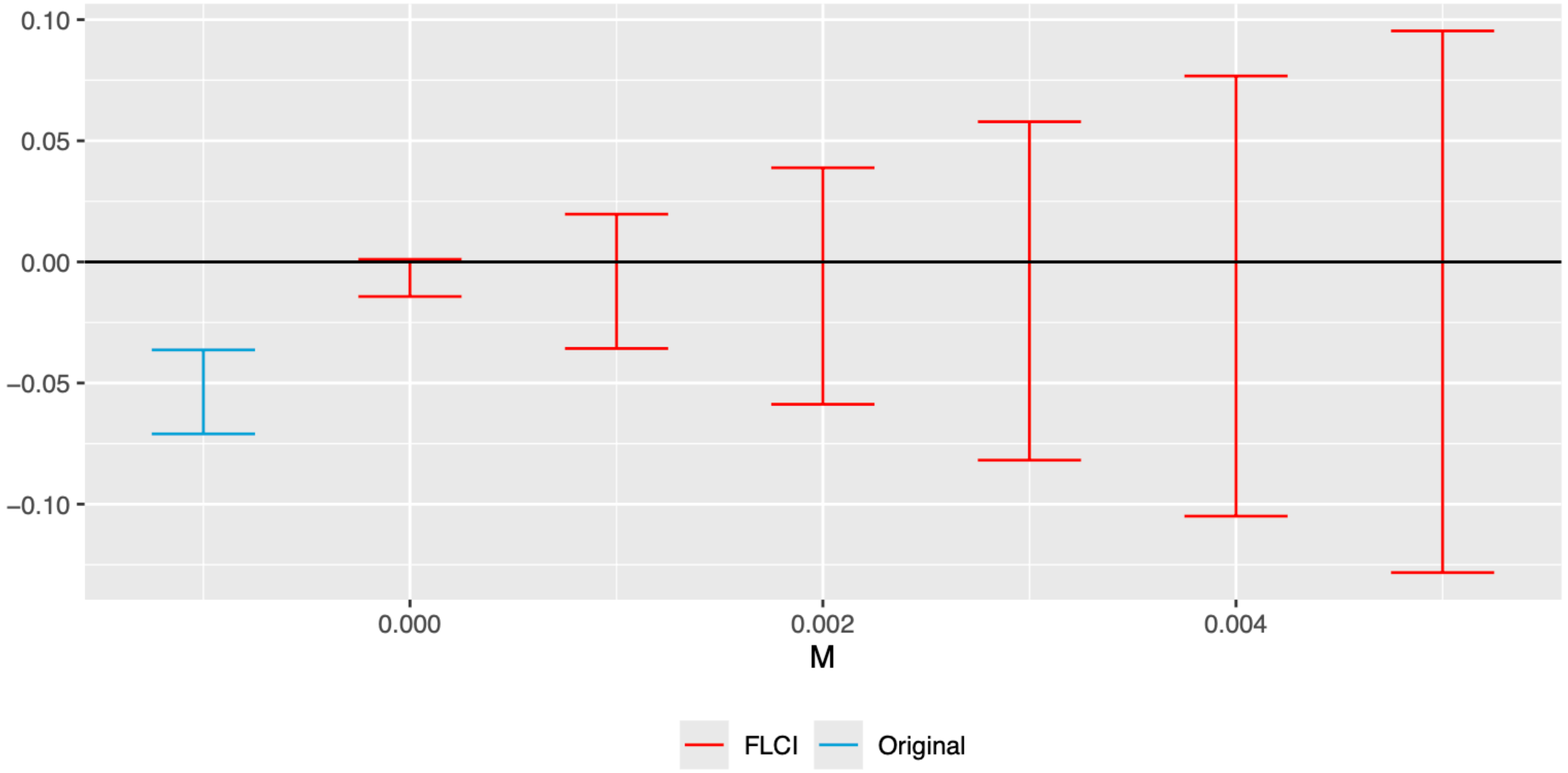


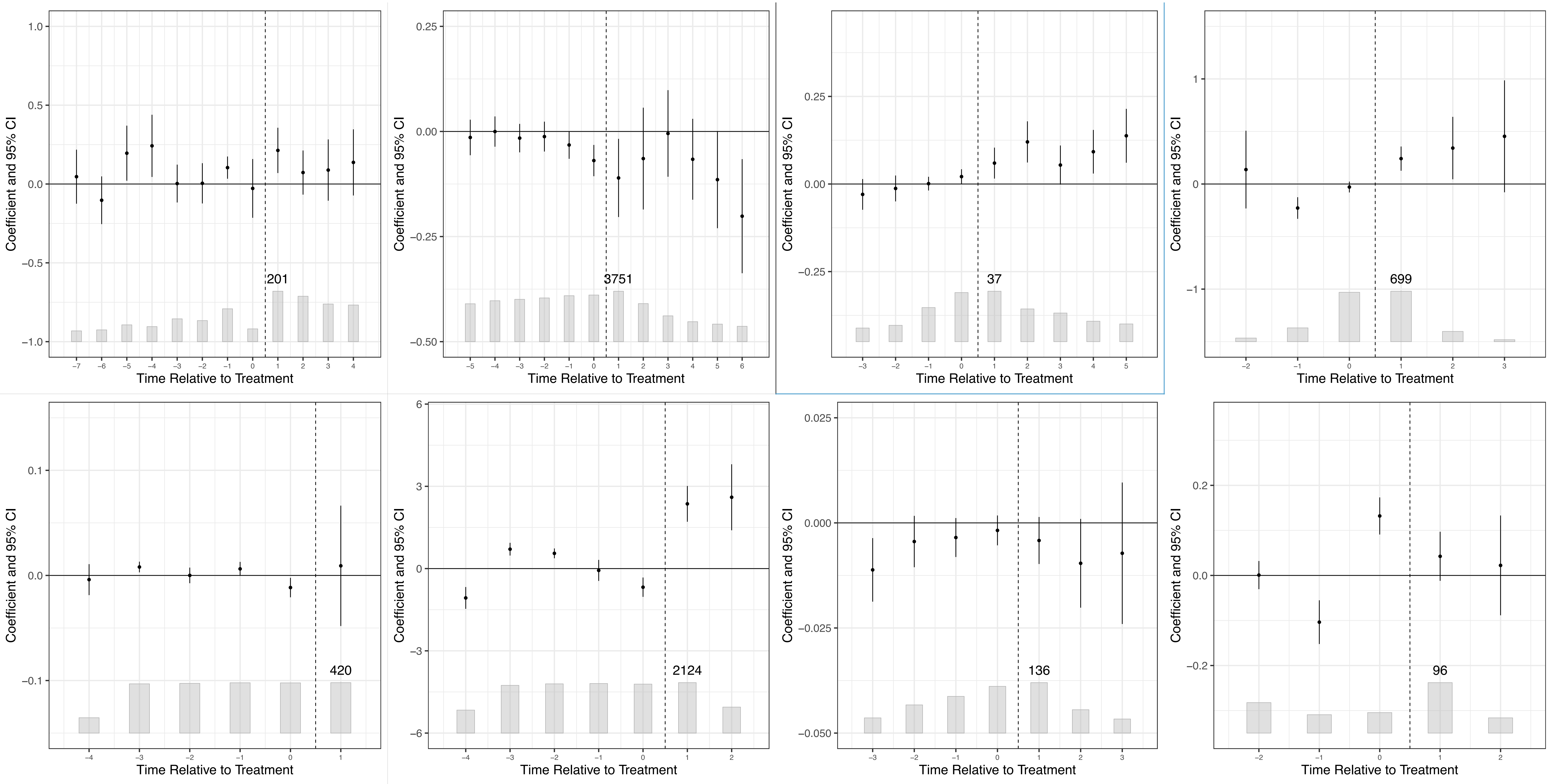
# Sensitivity Analysis w/ Smoothness Restriction

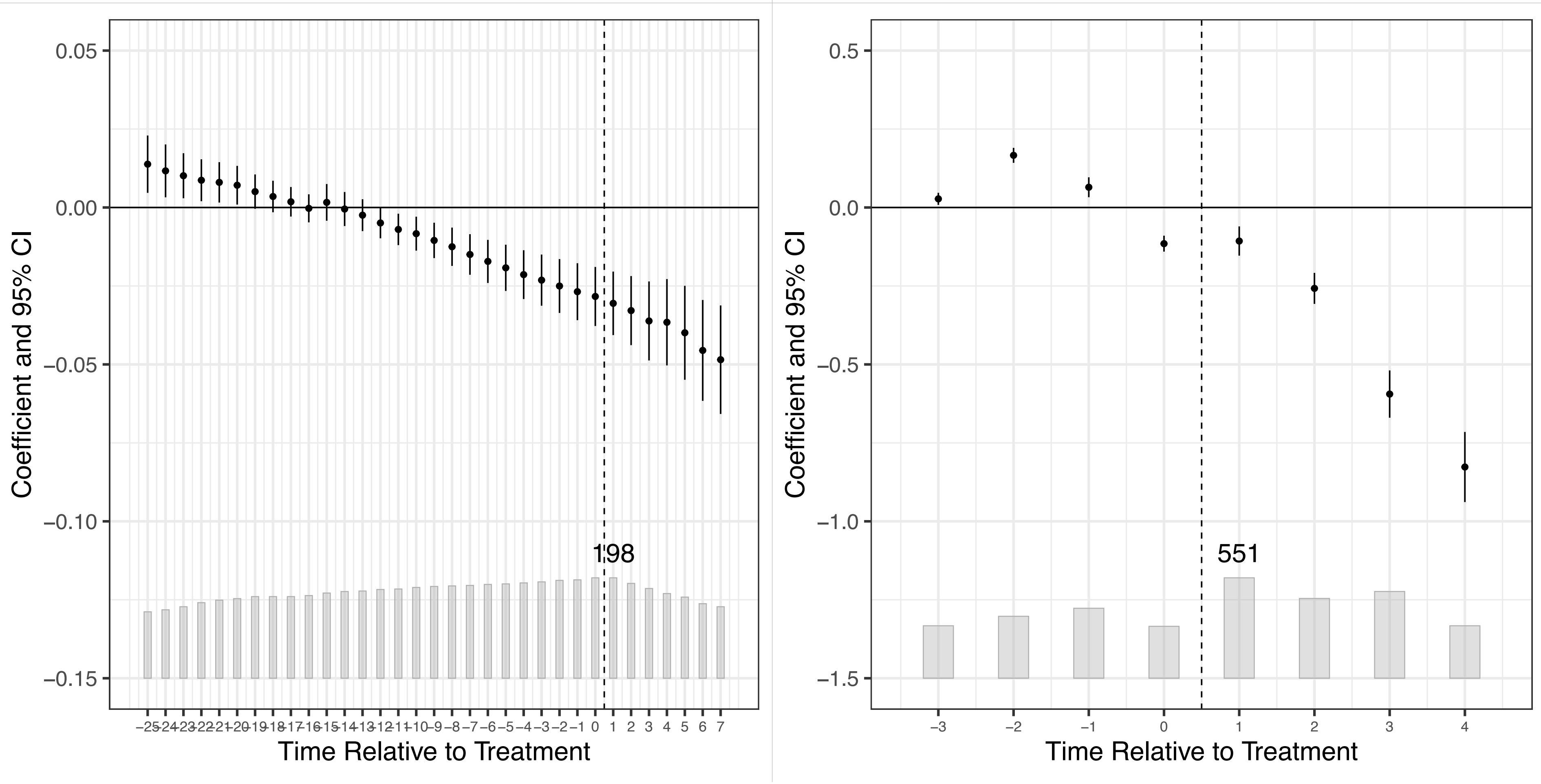


# Sensitivity Analysis w/ Smoothness Restriction

- Sensitivity analysis reveals that the result is not robust to a PT violation with a linear time trend.






# Problematic Pretrends Are Not Rare

# Three Examples

- **Example 1**: Coethnic Mobilization
  - Strong design; HTE matters marginally — estimators (including TWFE) broadly agree
- **Example 2**: Lawsuit against land use restriction

# Three Examples

- **Example 1**: Coethnic Mobilization
  - Strong design; HTE matters marginally — estimators (including TWFE) broadly agree
- **Example 2**: Lawsuit against land use restriction
  - Clear signs of PT violations

# Three Examples

- **Example 1**: Coethnic Mobilization
  - Strong design; HTE matters marginally — estimators (including TWFE) broadly agree
- **Example 2**: Lawsuit against land use restriction
  - Clear signs of PT violations
  - HTE is a second-order issue; agreement does not mean robustness

# Three Examples

- **Example 1**: Coethnic Mobilization
  - Strong design; HTE matters marginally — estimators (including TWFE) broadly agree
- **Example 2**: Lawsuit against land use restriction
  - Clear signs of PT violations
  - HTE is a second-order issue; agreement does not mean robustness
  - Simple plotting (and tests) will help spot the issue

# Example 3: Updating cadastral maps on Tax Revenue

# Example 3: Updating cadastral maps on Tax Revenue

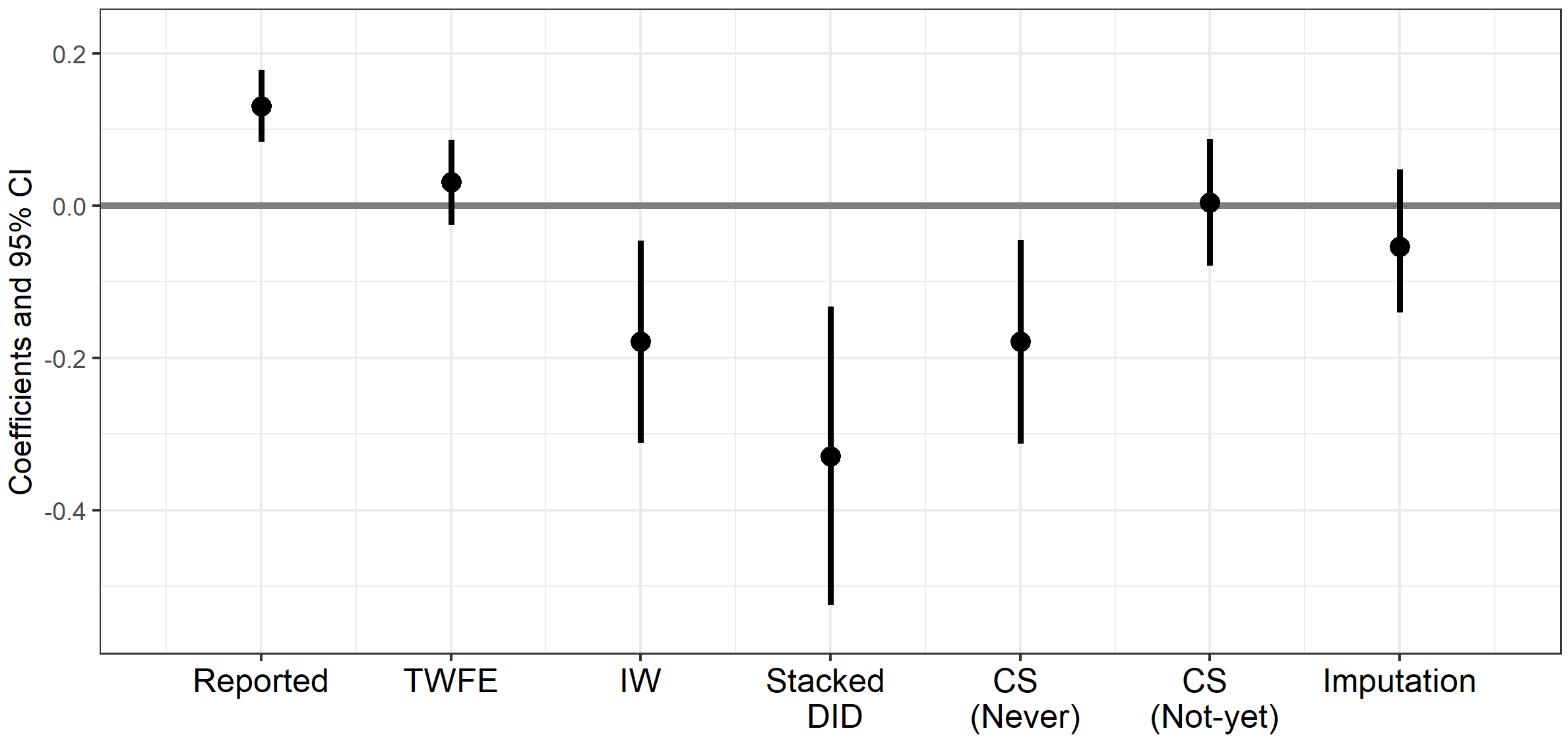
- The authors study the effect of cadastral map updating on property tax revenue in Brazil

# Example 3: Updating cadastral maps on Tax Revenue

- The authors study the effect of cadastral map updating on property tax revenue in Brazil
- Disagreement among estimators in the full sample

# Example 3: Updating cadastral maps on Tax Revenue

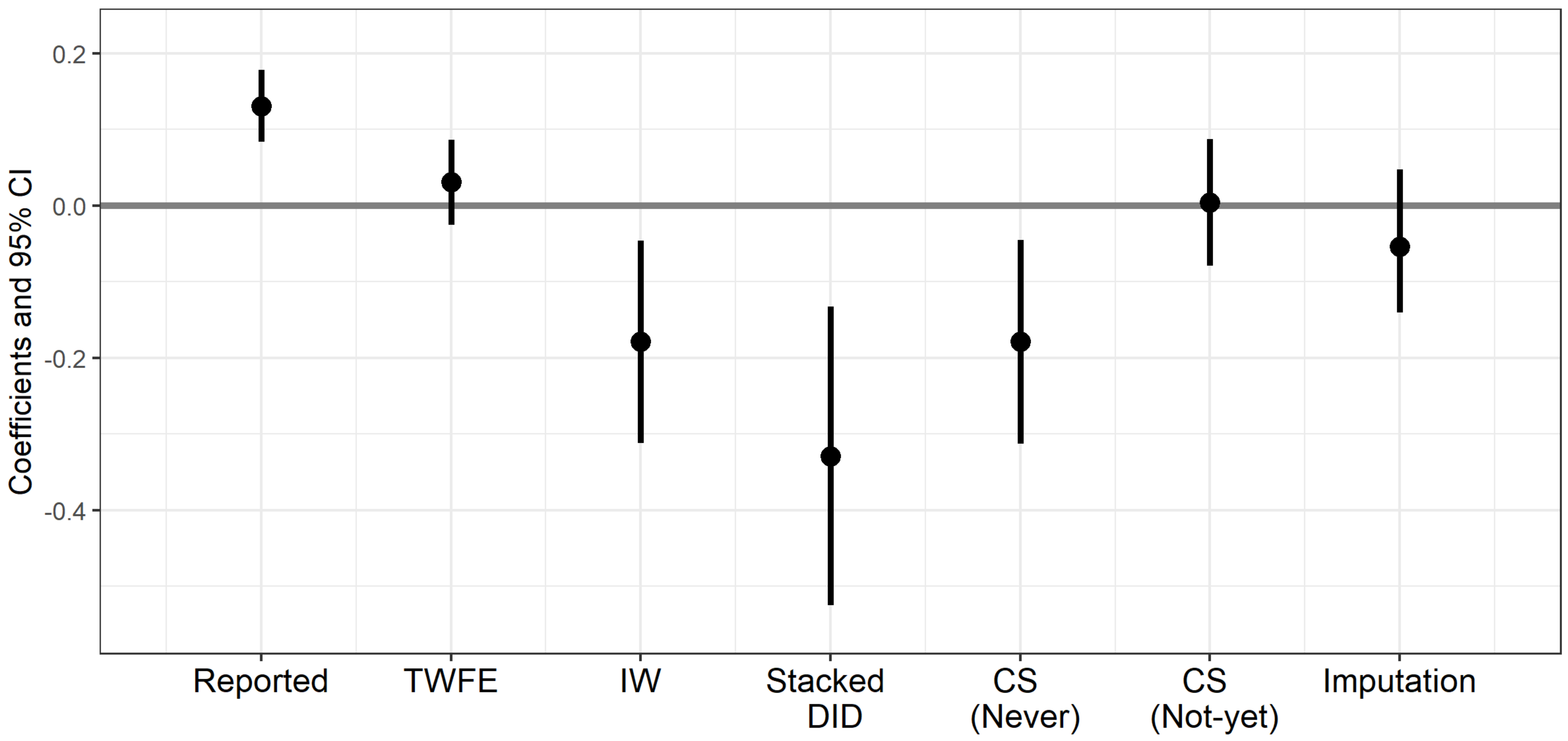
- The authors study the effect of cadastral map updating on property tax revenue in Brazil
- Disagreement among estimators in the full sample




Replicated based on the Full Sample

# Example 3: Updating cadastral maps on Tax Revenue

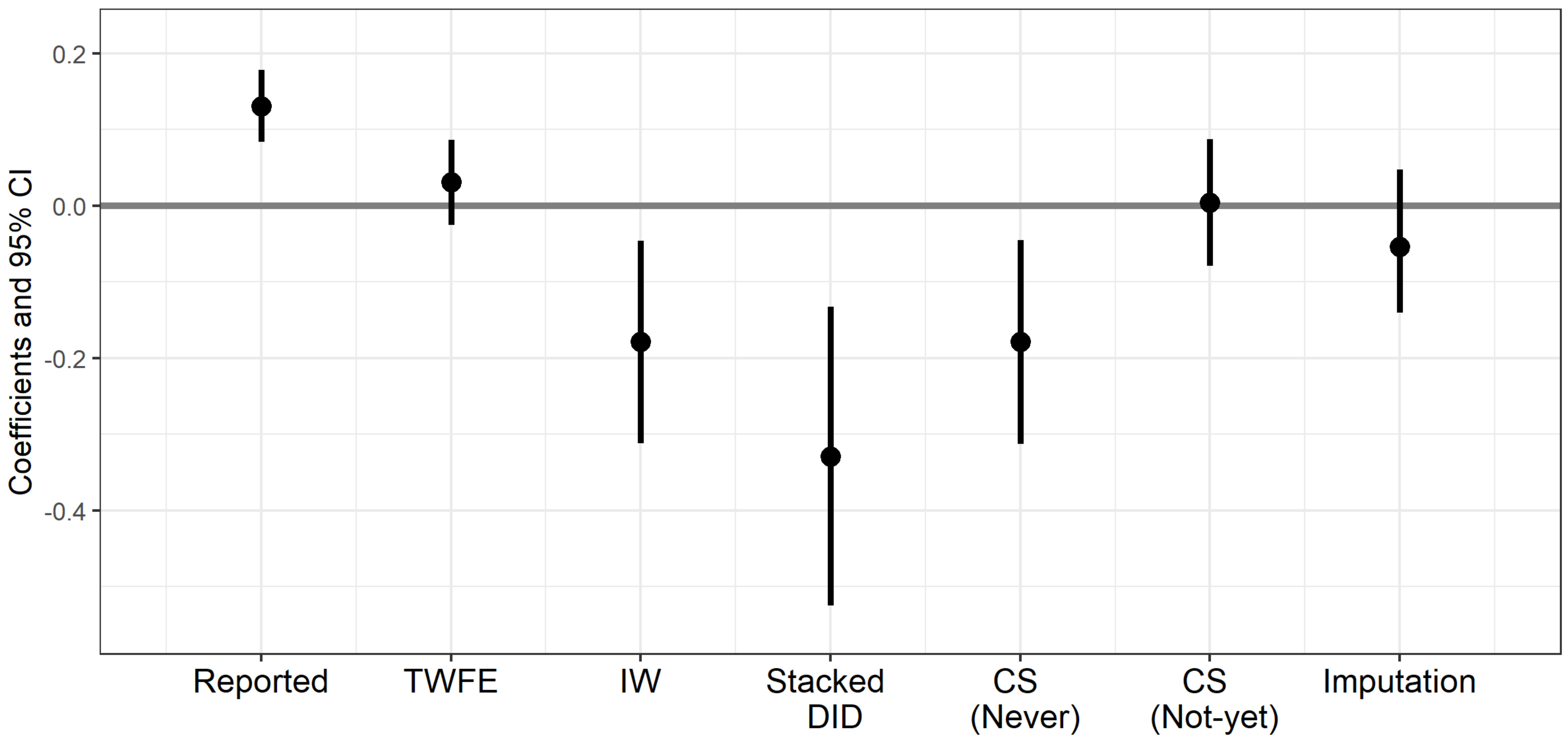
- The authors study the effect of cadastral map updating on property tax revenue in Brazil
- Disagreement among estimators in the full sample
- Event study plot based on a subsample suggests a positive effect




Replicated based on the Full Sample

# Example 3: Updating cadastral maps on Tax Revenue

- The authors study the effect of cadastral map updating on property tax revenue in Brazil
- Disagreement among estimators in the full sample
- Event study plot based on a subsample suggests a positive effect




Replicated based on the Full Sample

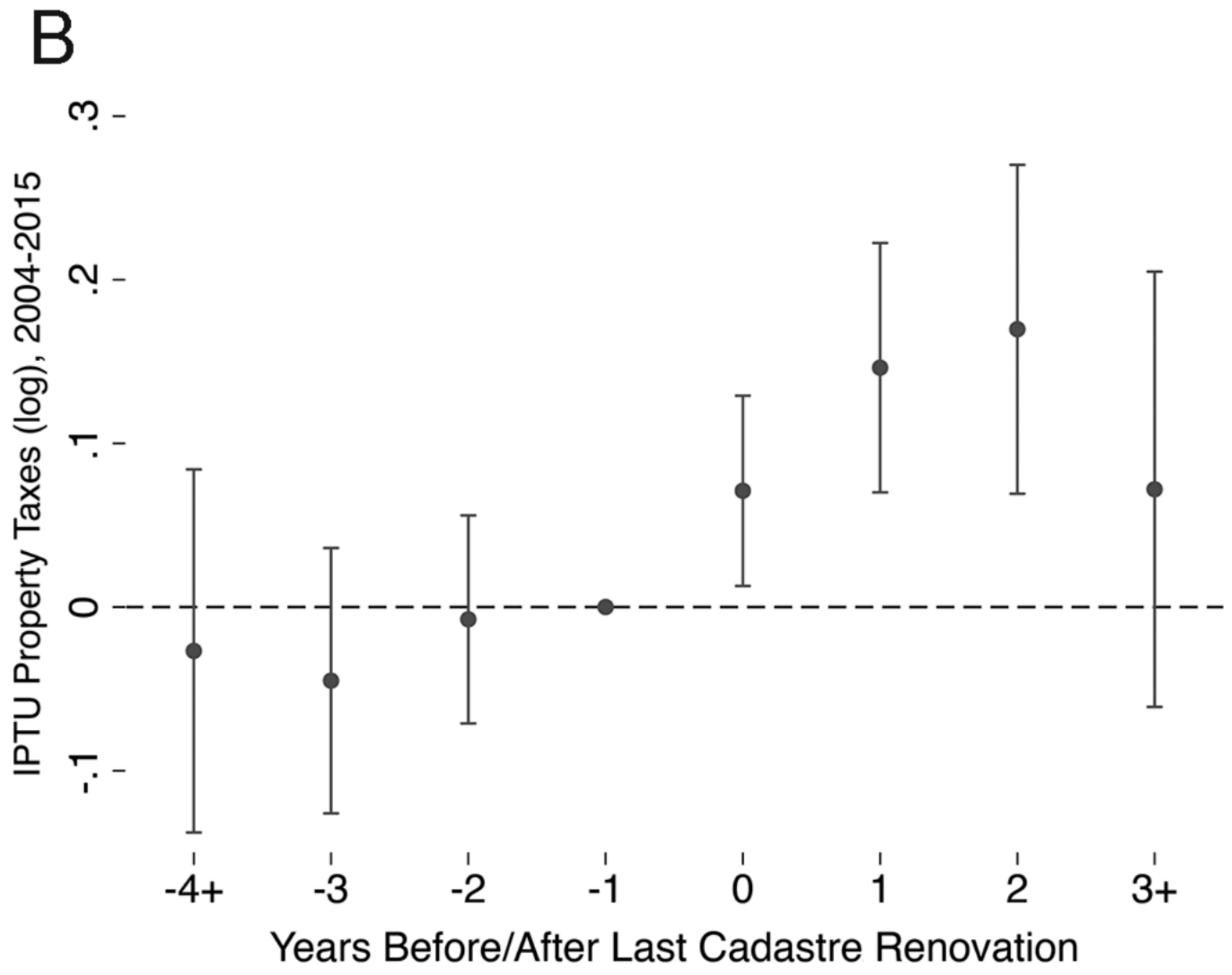
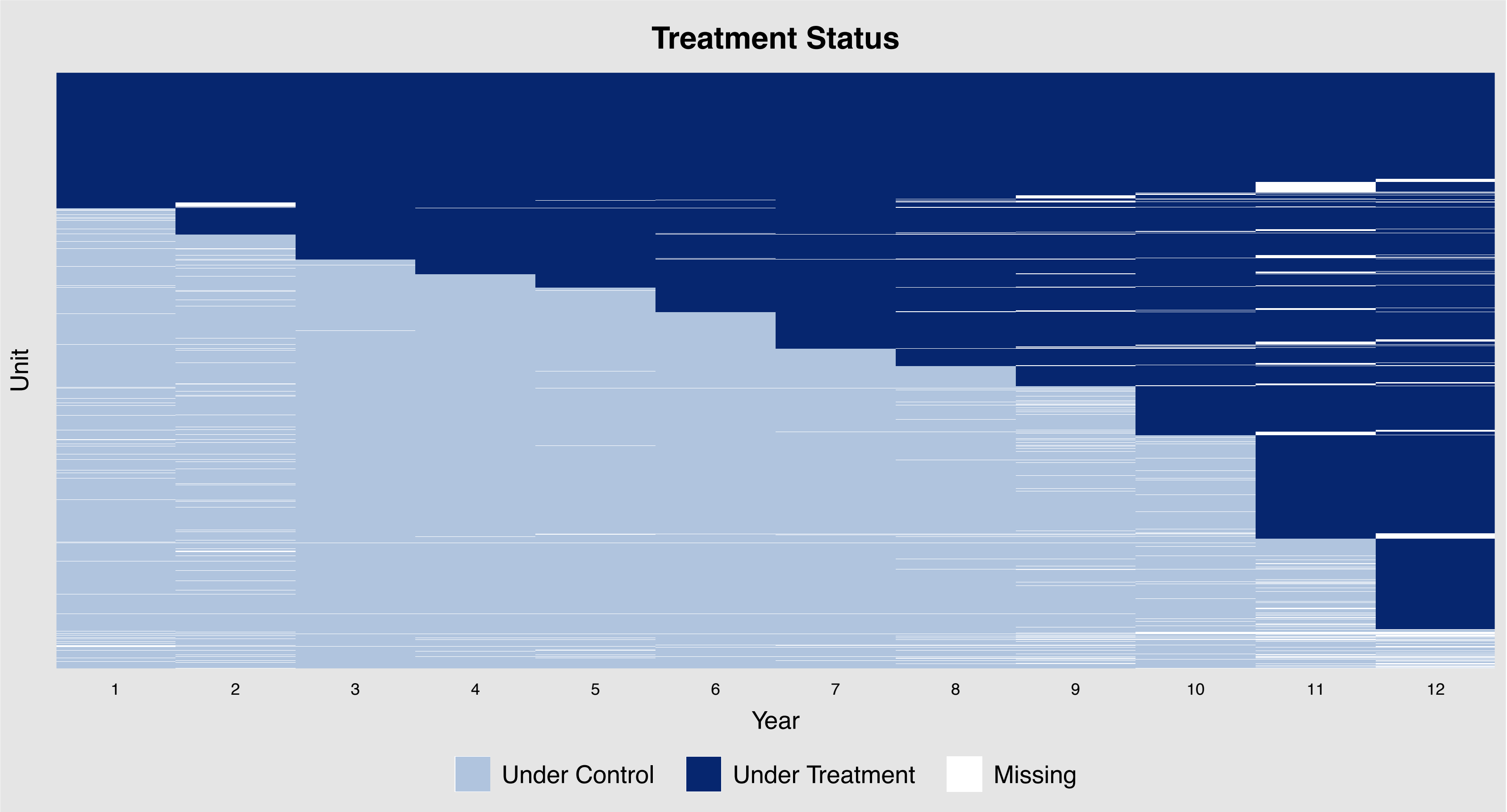



Figure 1B (based on a Subsample)

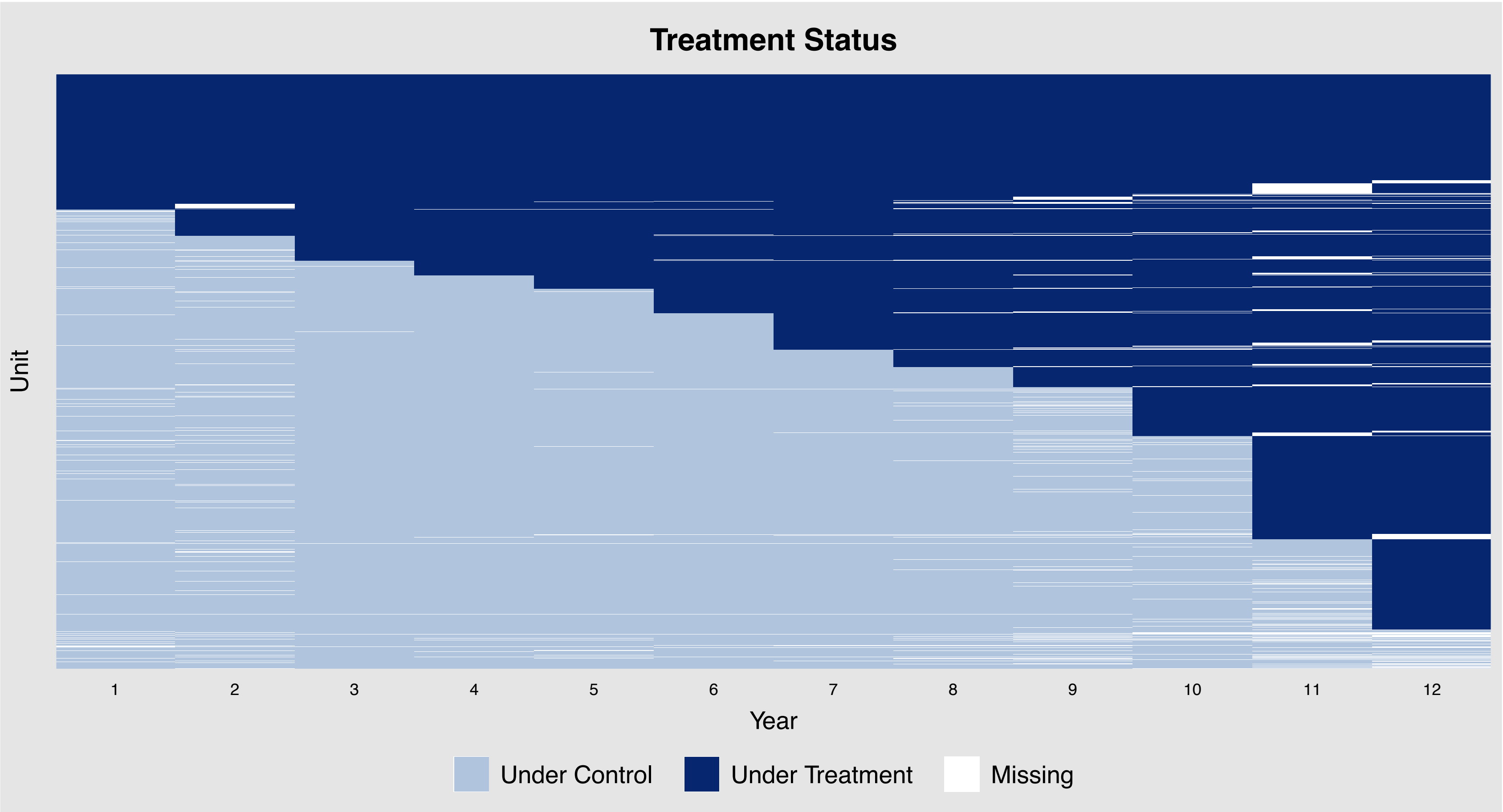
# Look Deeper...

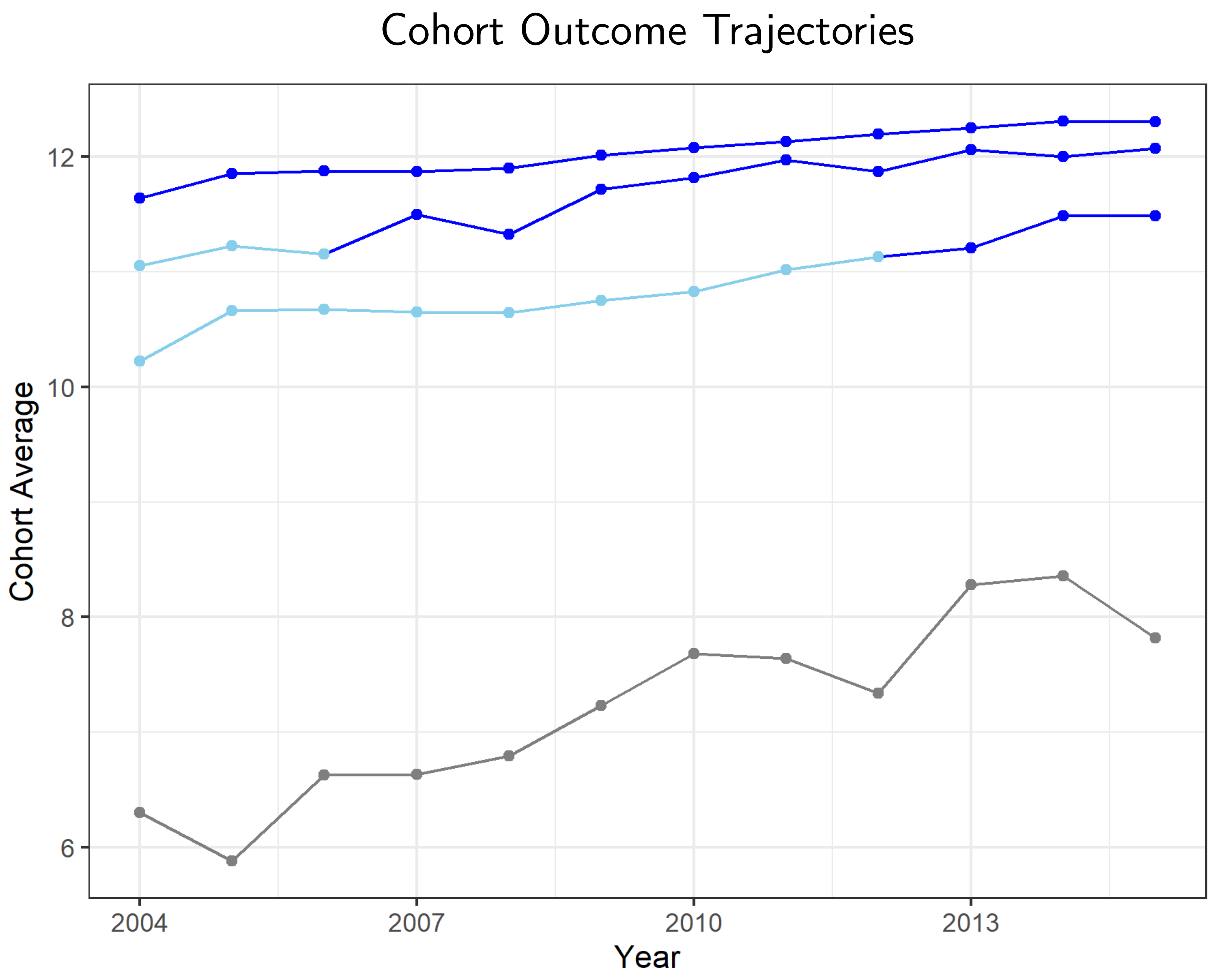
# Look Deeper...



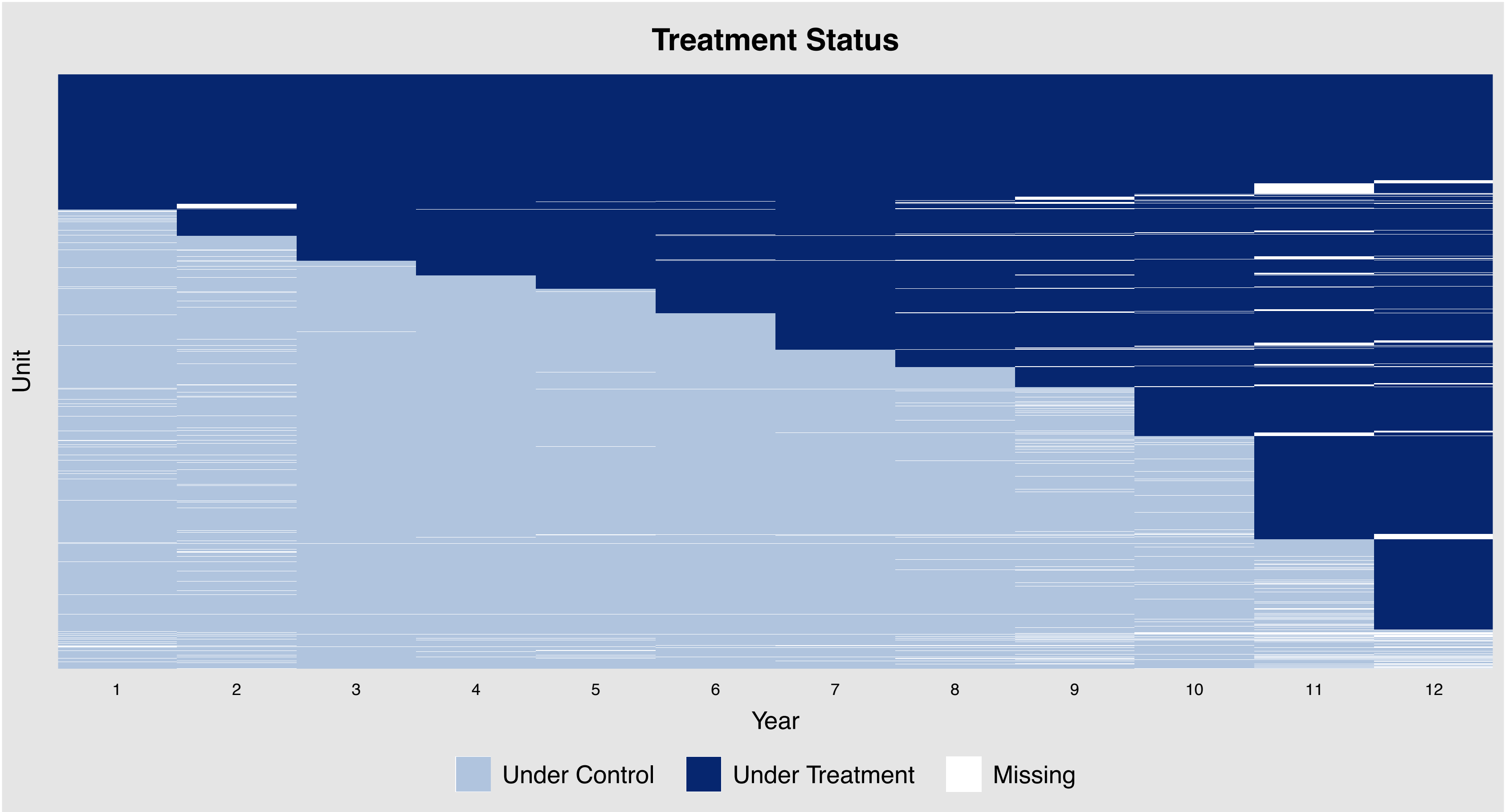
Treatment Status
Unit
1
2
3
4
5
6
7
8
9
10
11
12
Year
Under Control
Under Treatment
Missing

# Look Deeper...

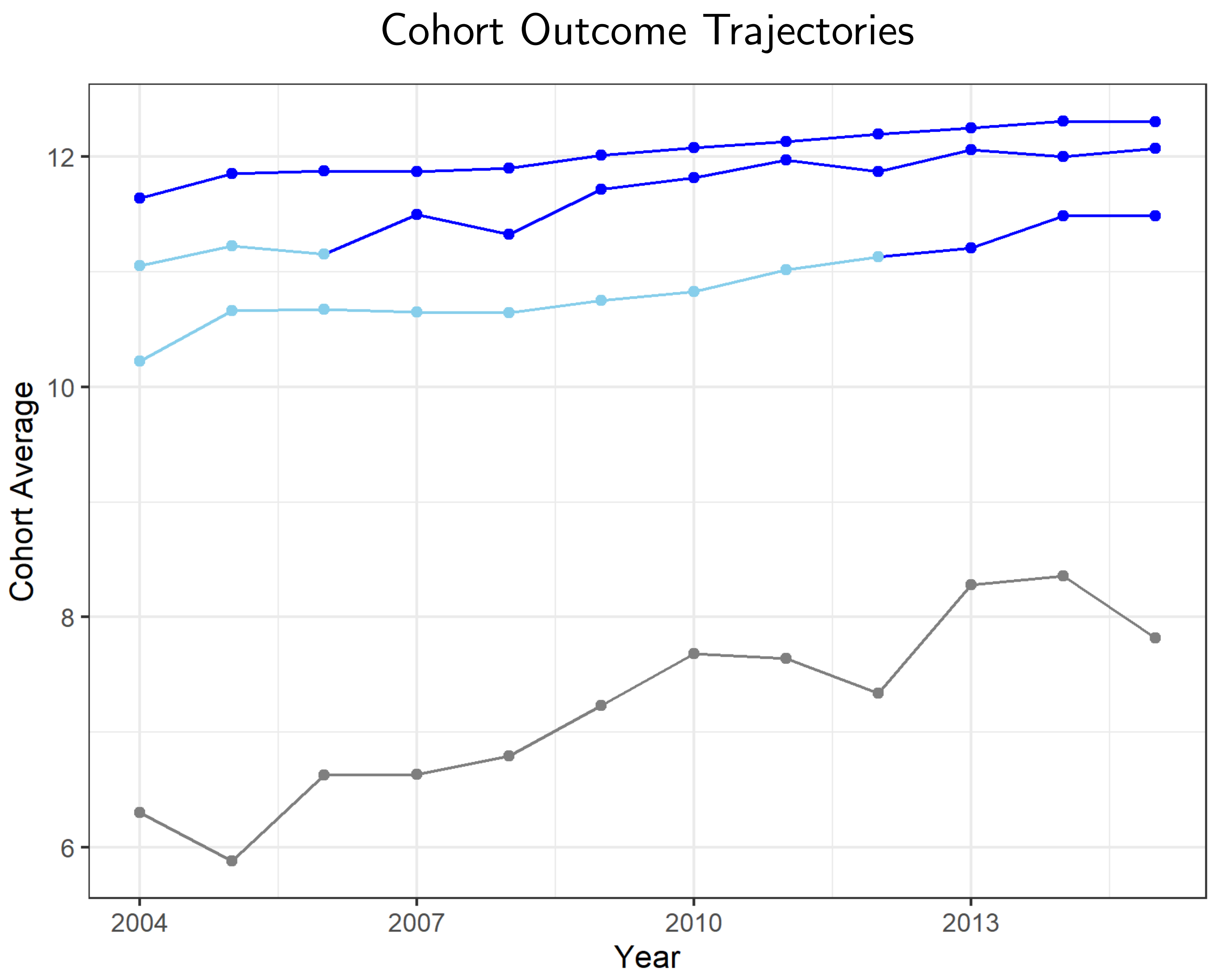

Treatment Status
Unit
1
2
3
4
5
6
7
8
9
10
11
12
Year
Under Control
Under Treatment
Missing


Cohort Outcome Trajectories
Cohort Average
12
10
8
6
2004
2007
2010
2013
Year

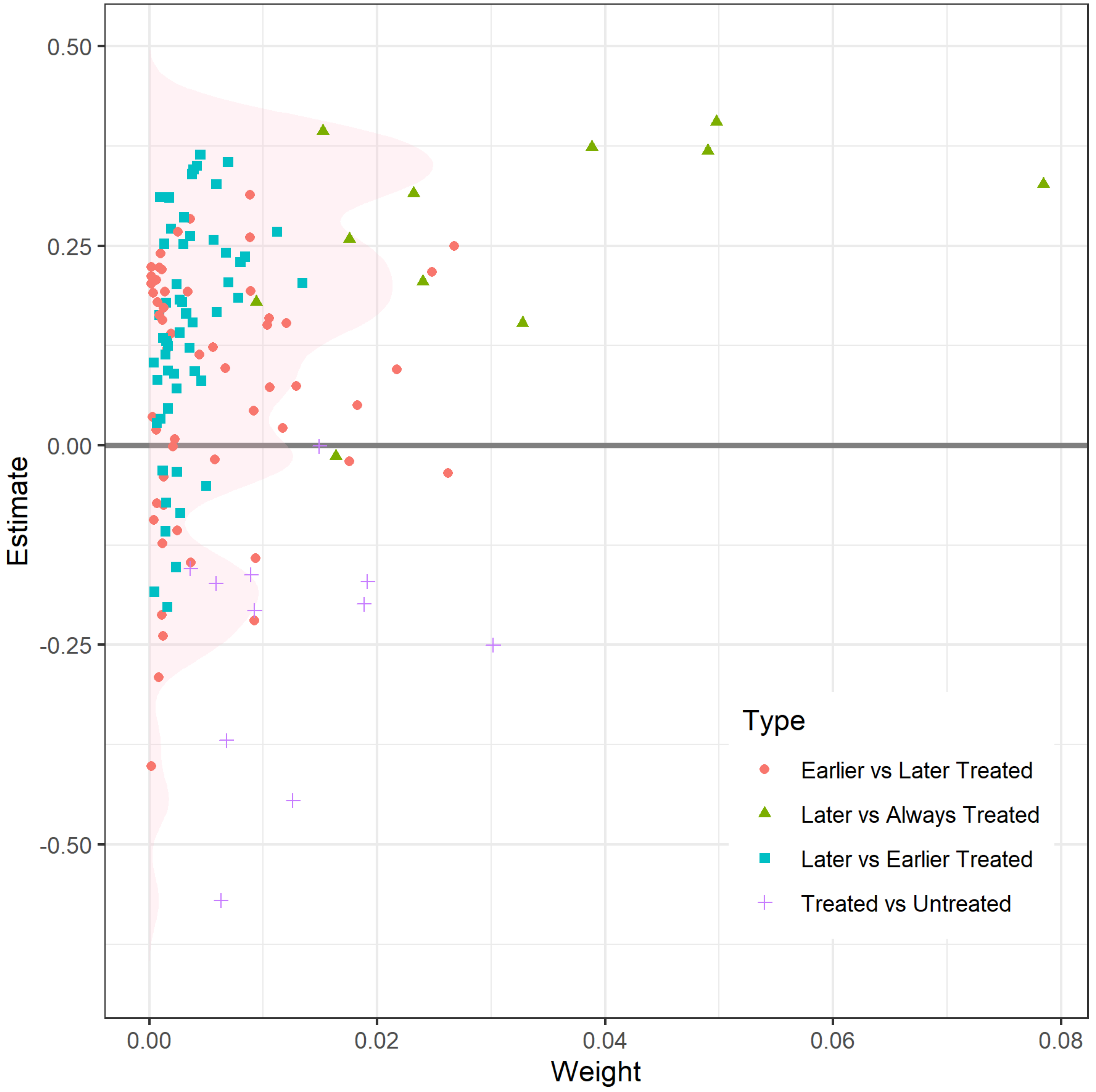
# Look Deeper...



Treatment Status
Unit
1
2
3
4
5
6
7
8
9
10
11
12
Year
Under Control
Under Treatment
Missing

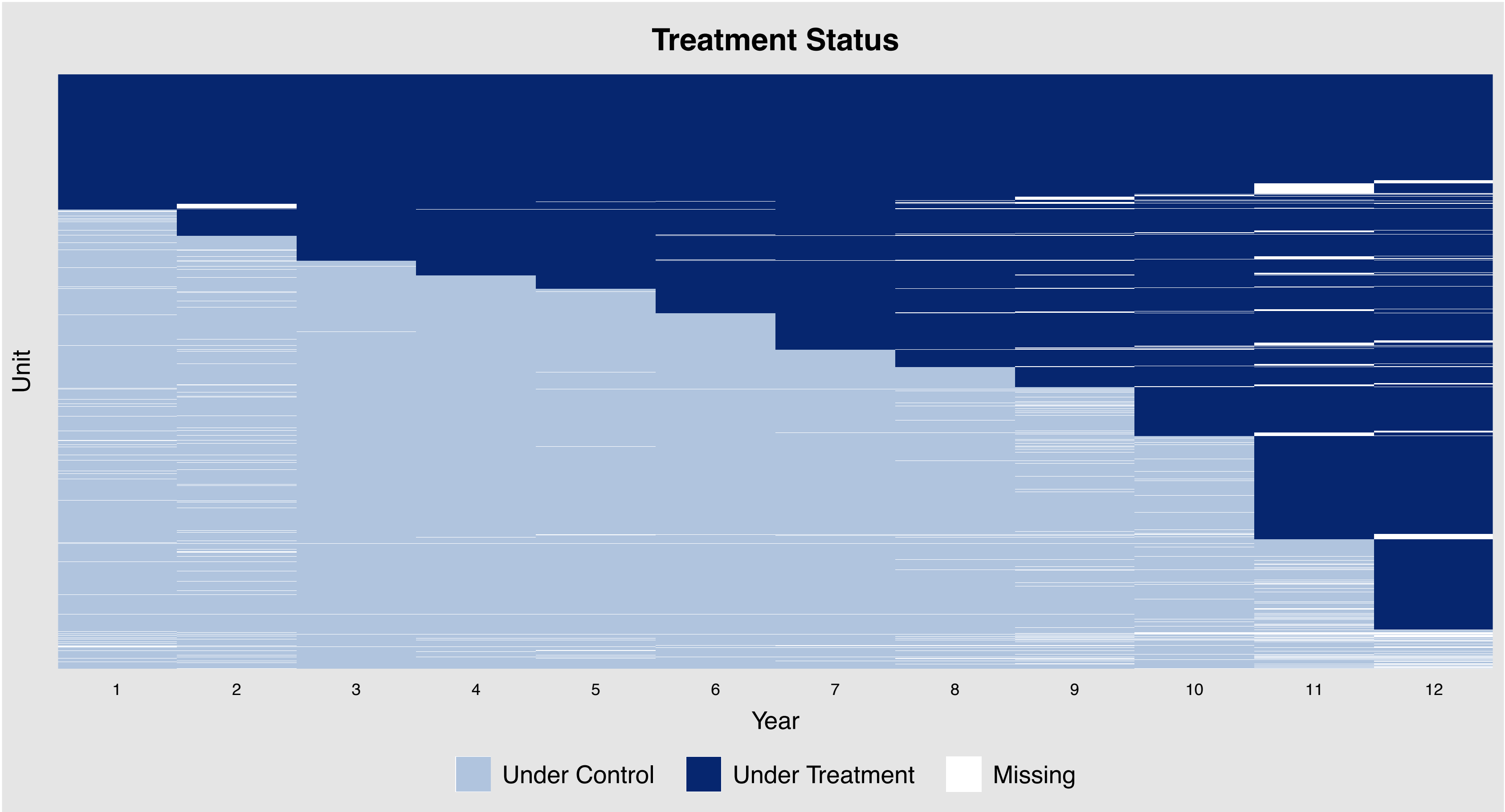



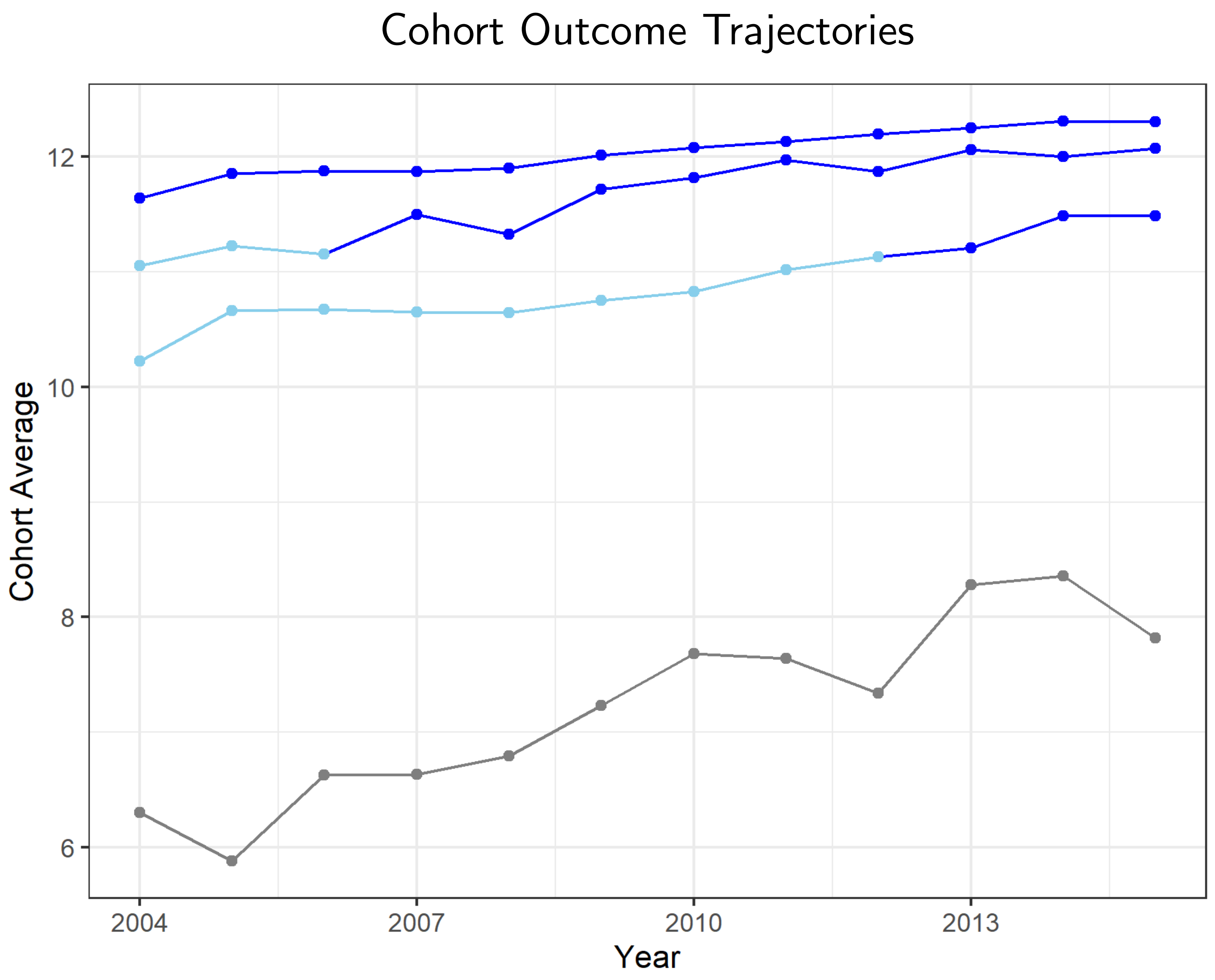
Cohort Outcome Trajectories
Cohort Average
12
10
8
6
2004
2007
2010
2013
Year

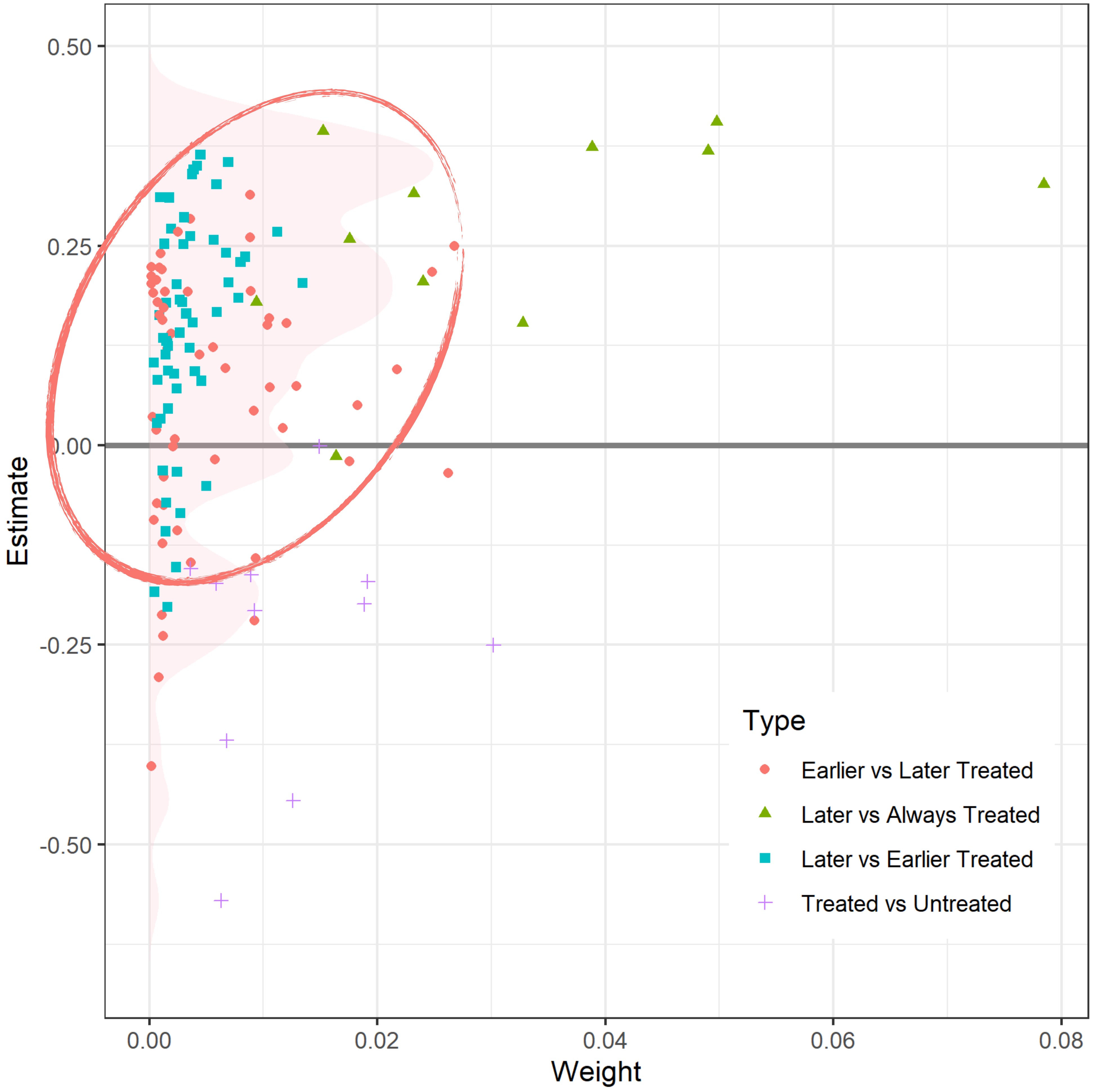



0.50
0.25
0.00
-0.25
-0.50
Estimate
0.00
0.02
0.04
0.06
0.08
Weight
Type
Earlier vs Later Treated
Later vs Always Treated
Later vs Earlier Treated
Treated vs Untreated

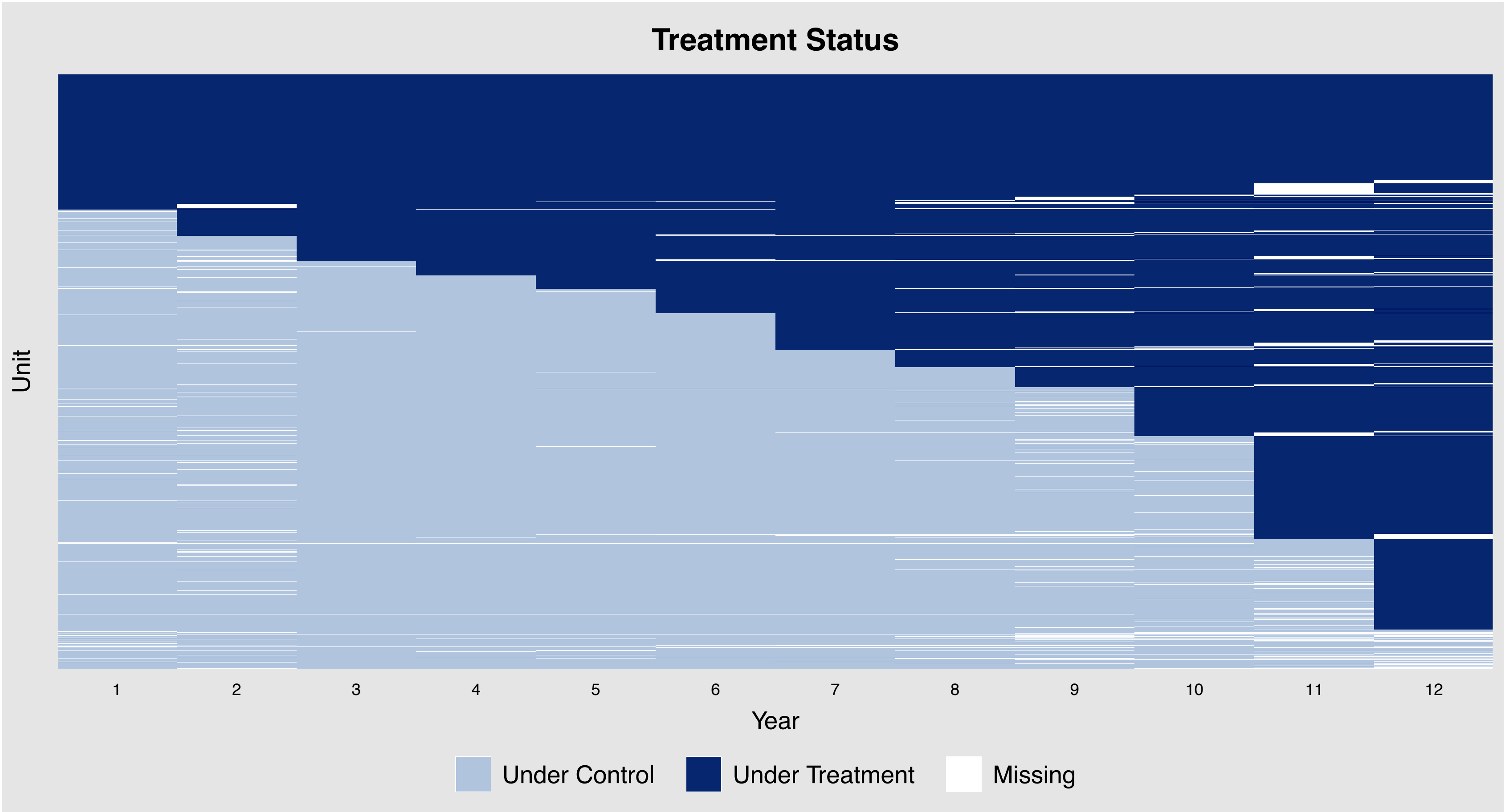
# Look Deeper...


Treatment Status
Unit
1
2
3
4
5
6
7
8
9
10
11
12
Year
Under Control
Under Treatment
Missing

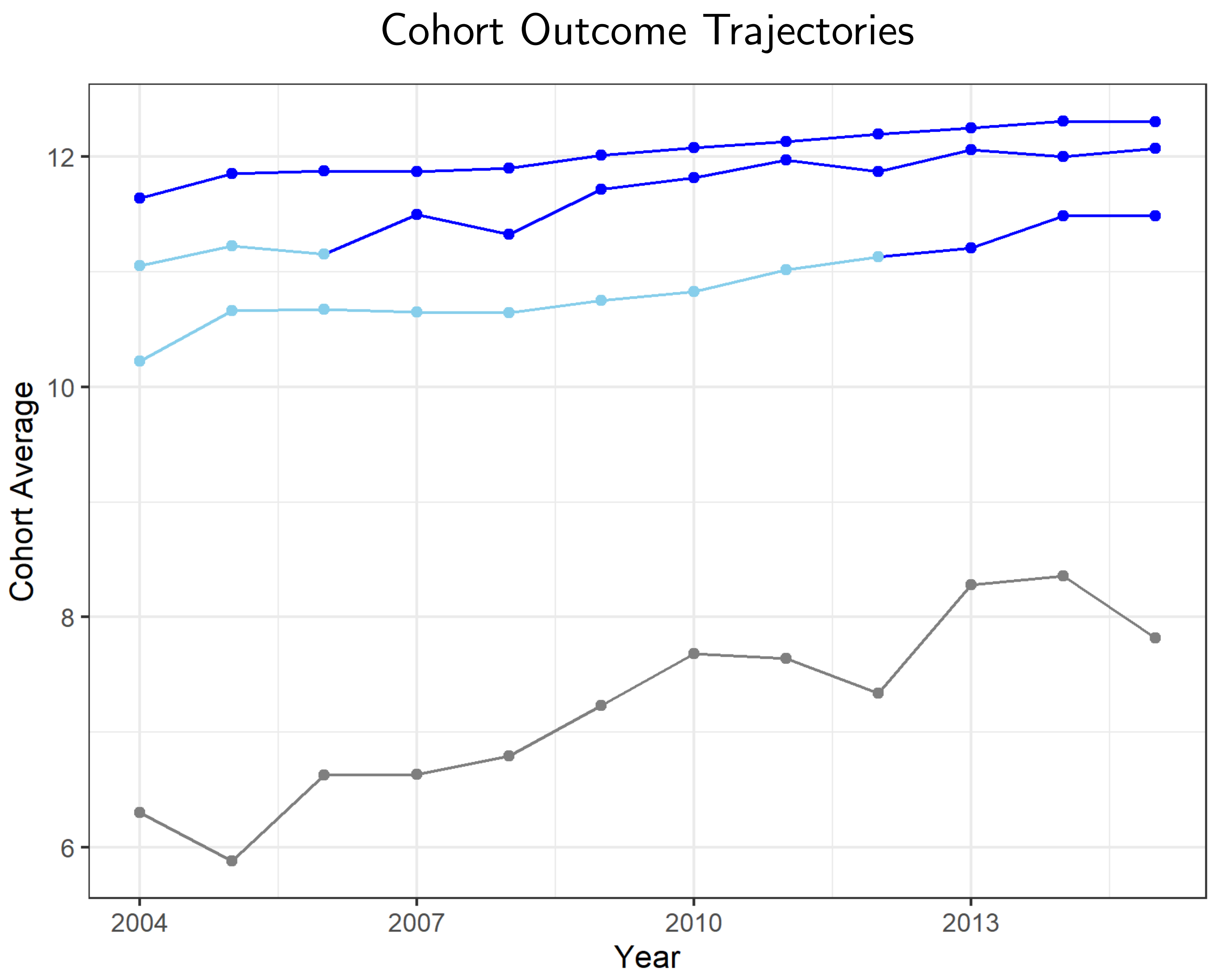

Cohort Outcome Trajectories
Cohort Average
12
10
8
6
2004
2007
2010
2013
Year


Estimate
0.50
0.25
0.00
-0.25
-0.50
0.00
0.02
0.04
0.06
0.08
Weight
Type
Earlier vs Later Treated
Later vs Always Treated
Later vs Earlier Treated
Treated vs Untreated

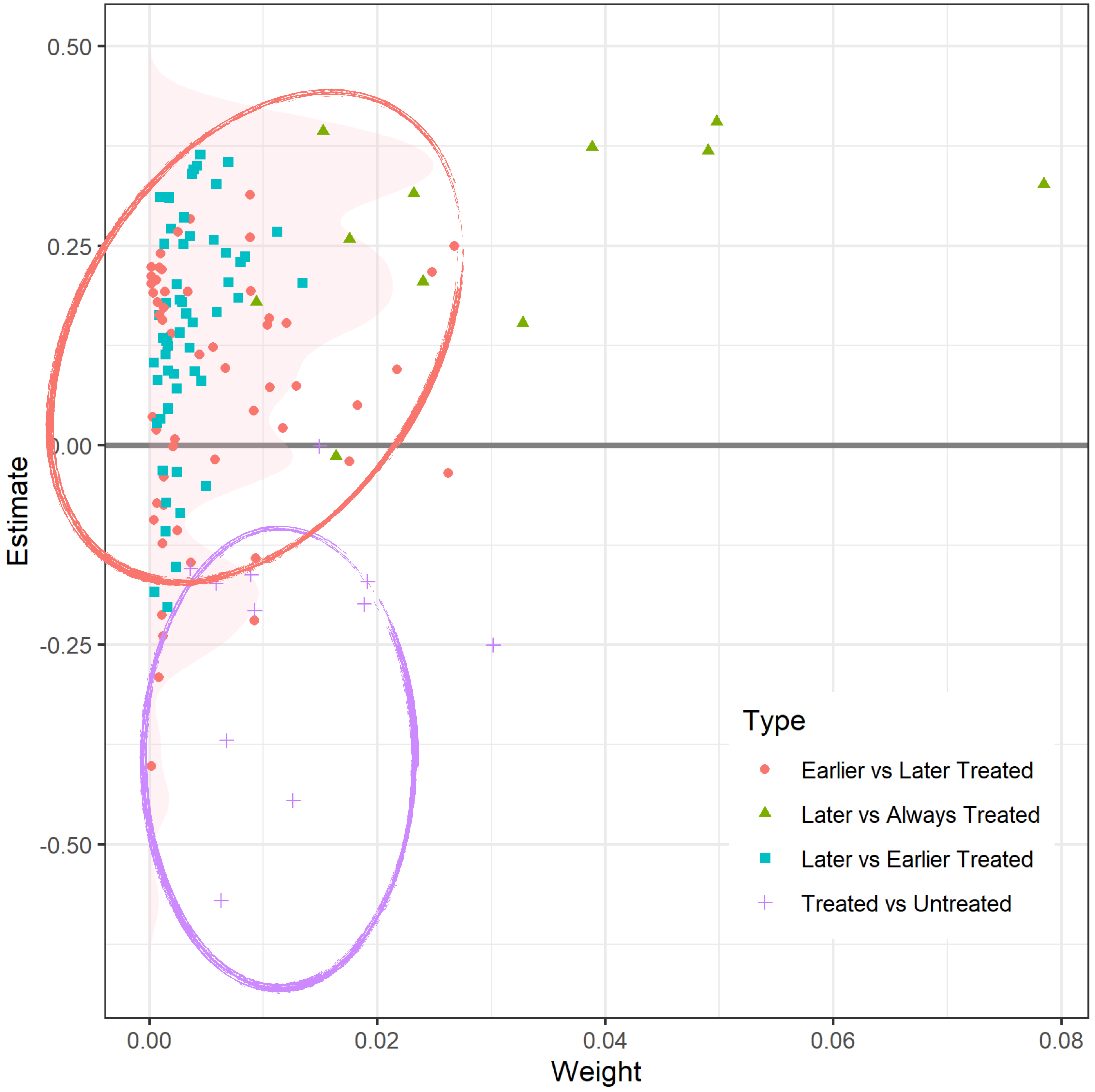
# Look Deeper...



Treatment Status
Unit
1
2
3
4
5
6
7
8
9
10
11
12
Year
Under Control
Under Treatment
Missing

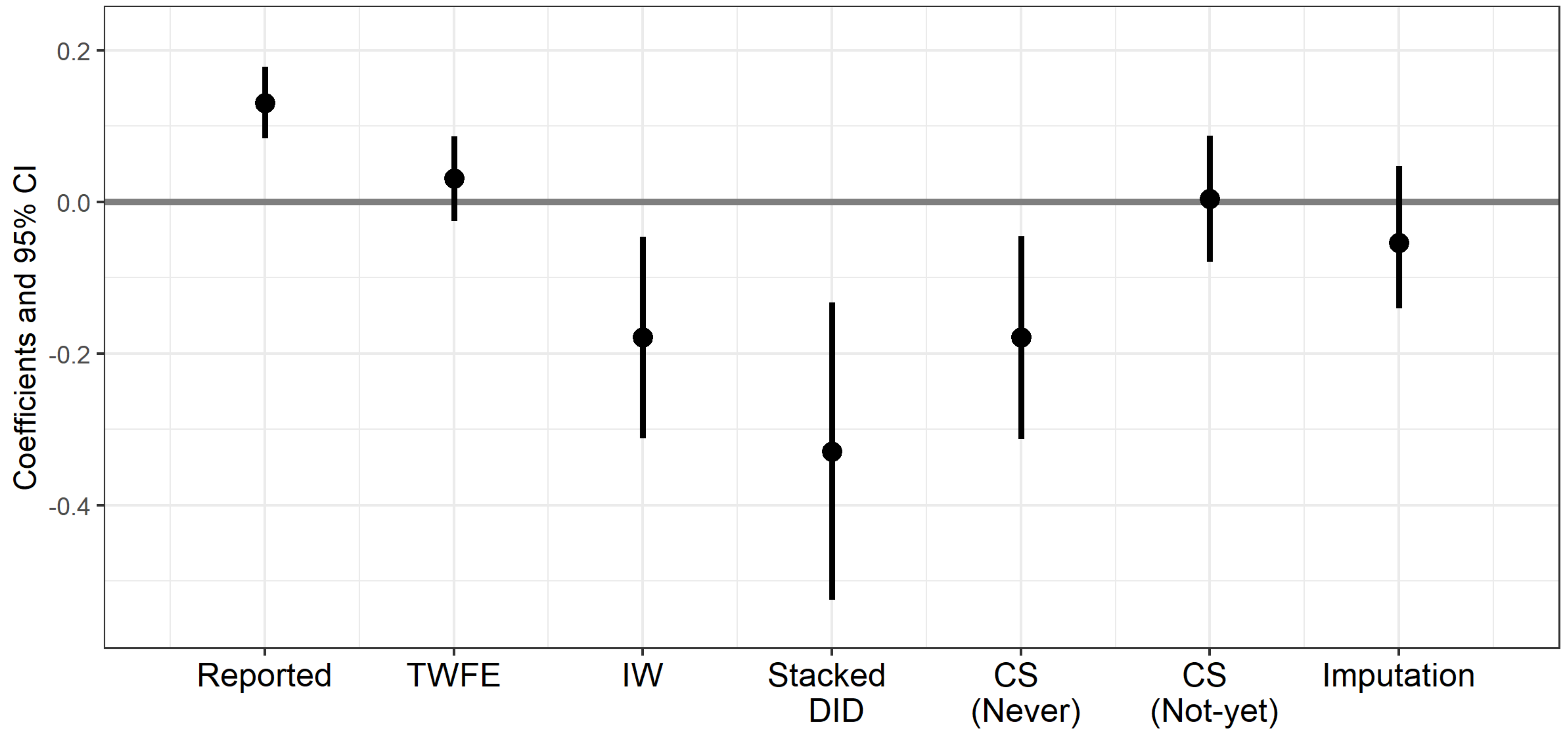
Cohort Outcome Trajectories
Cohort Average
12
10
8
6
2004
2007
2010
2013
Year

Estimate
0.50
0.25
0.00
-0.25
-0.50
0.00
0.02
0.04
0.06
0.08
Weight
Type
Earlier vs Later Treated
Later vs Always Treated
Later vs Earlier Treated
Treated vs Untreated

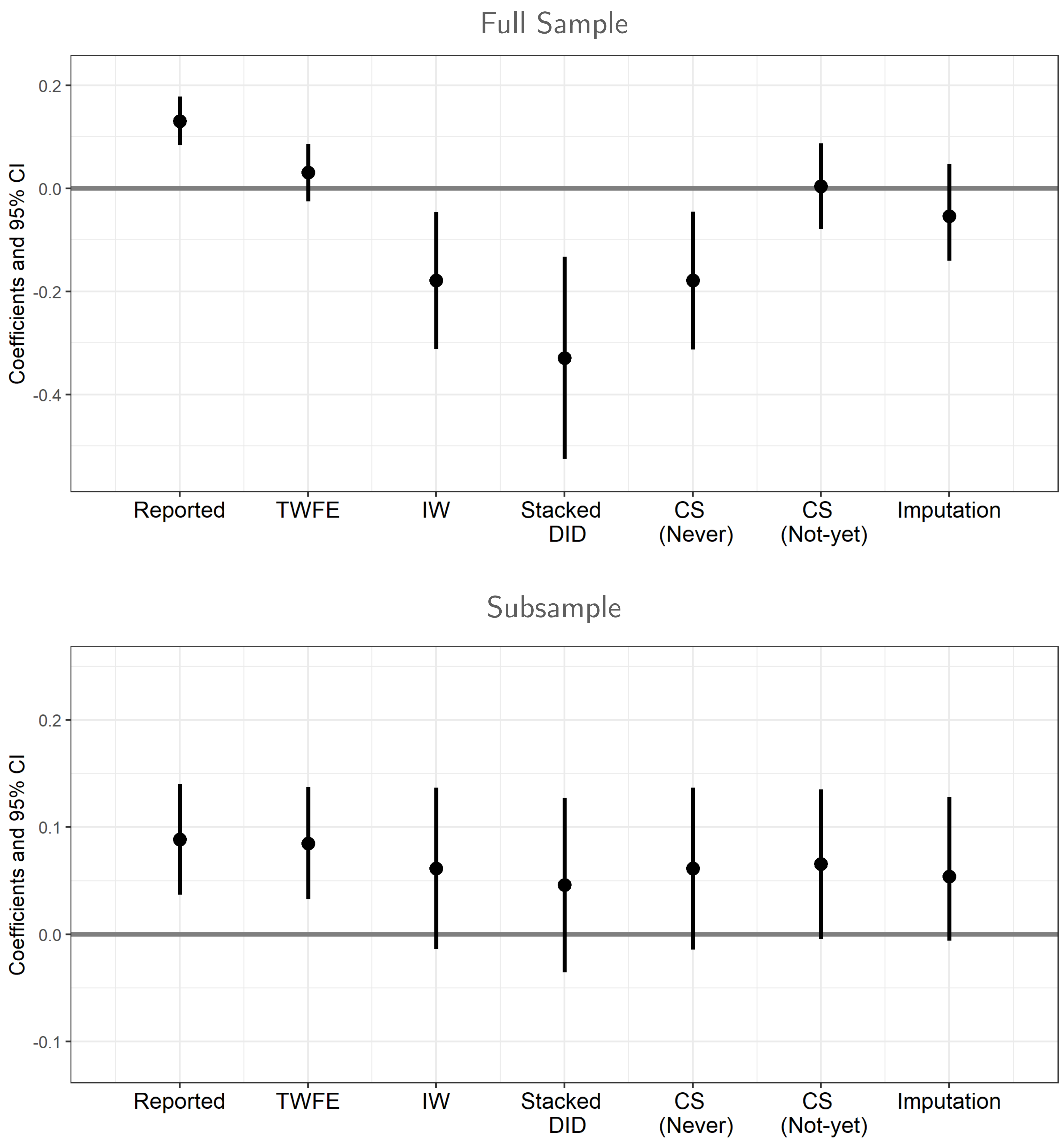
# Trimmed Subsample

Full Sample

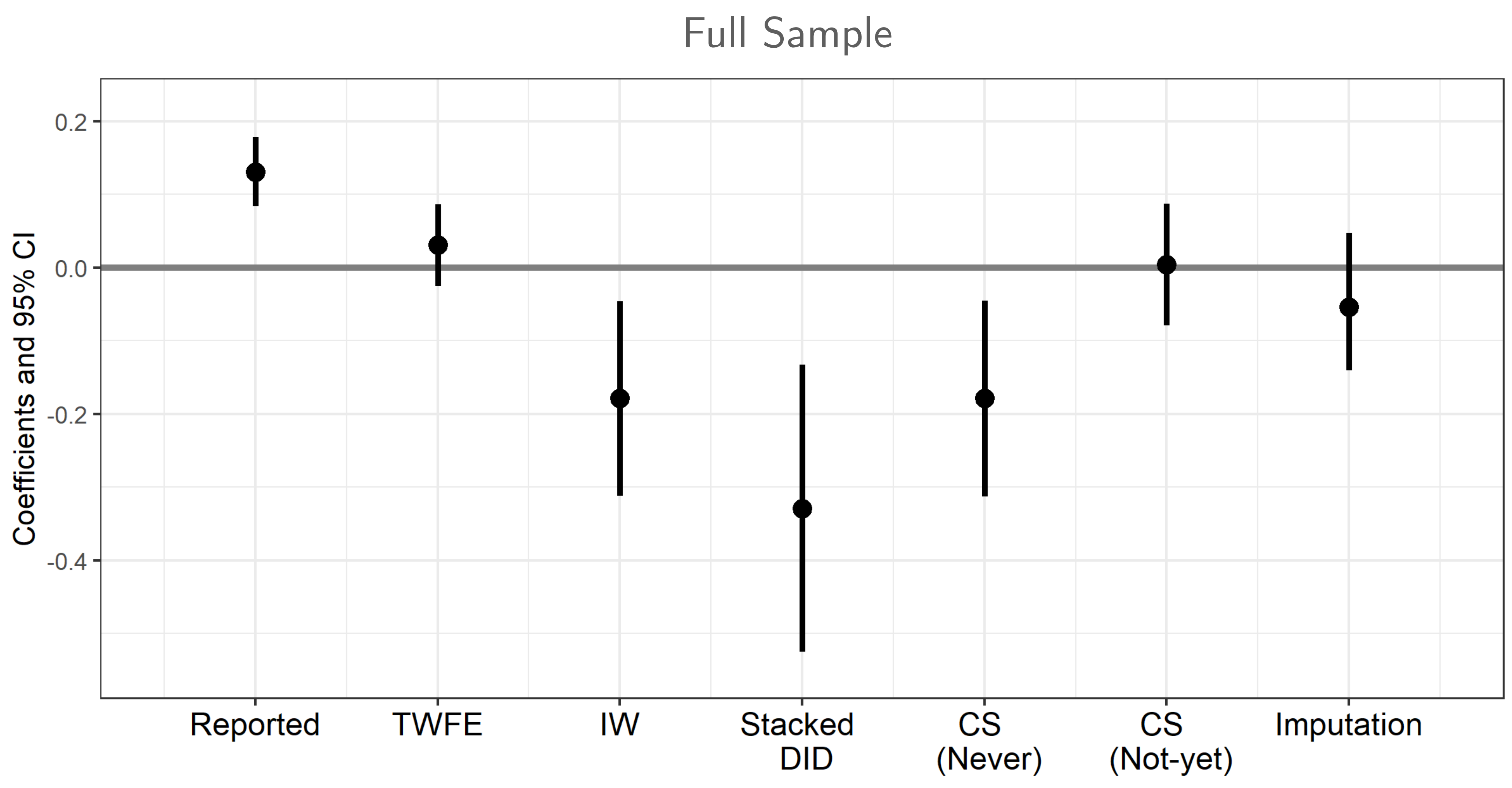


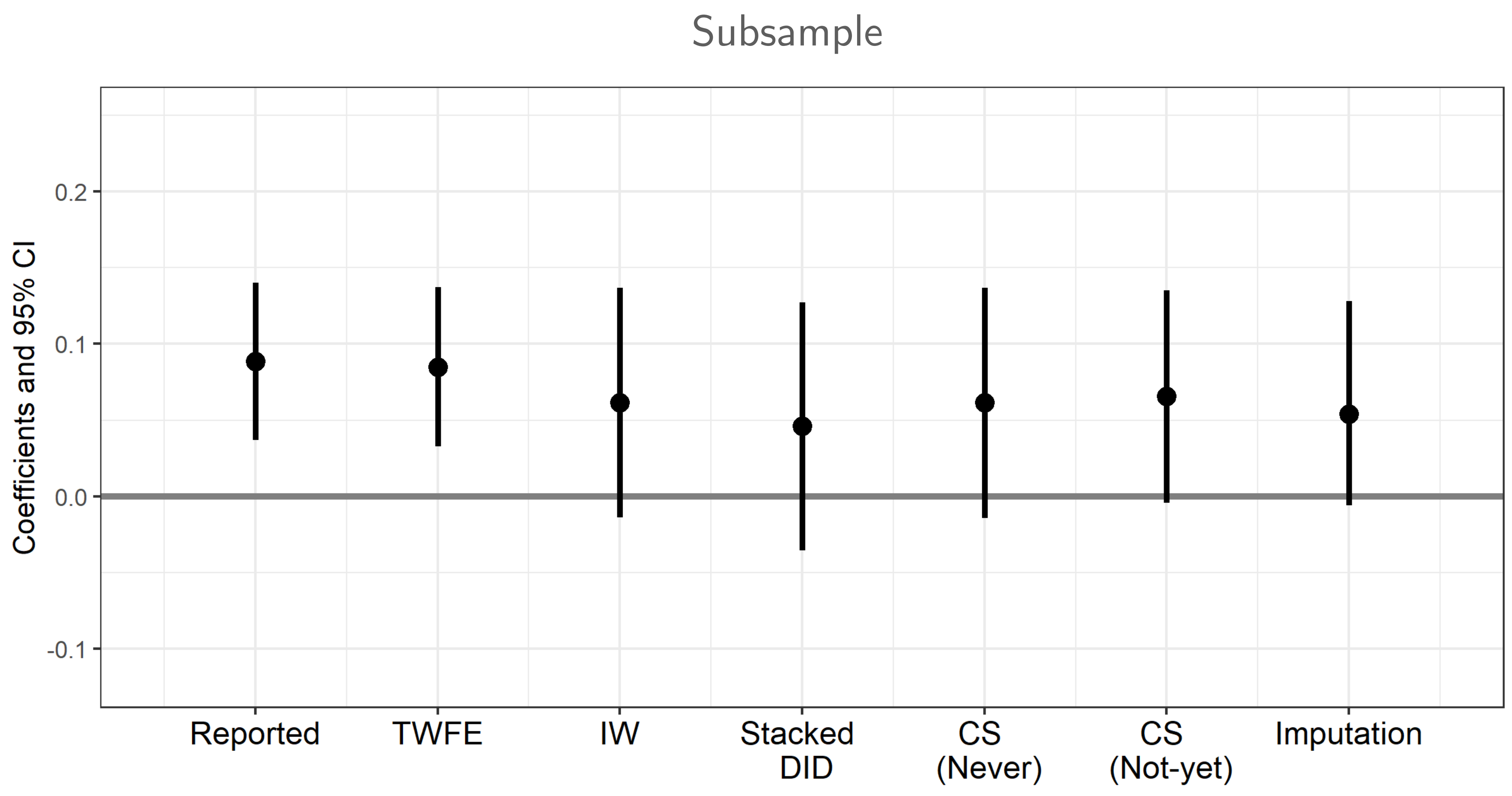

Subsample

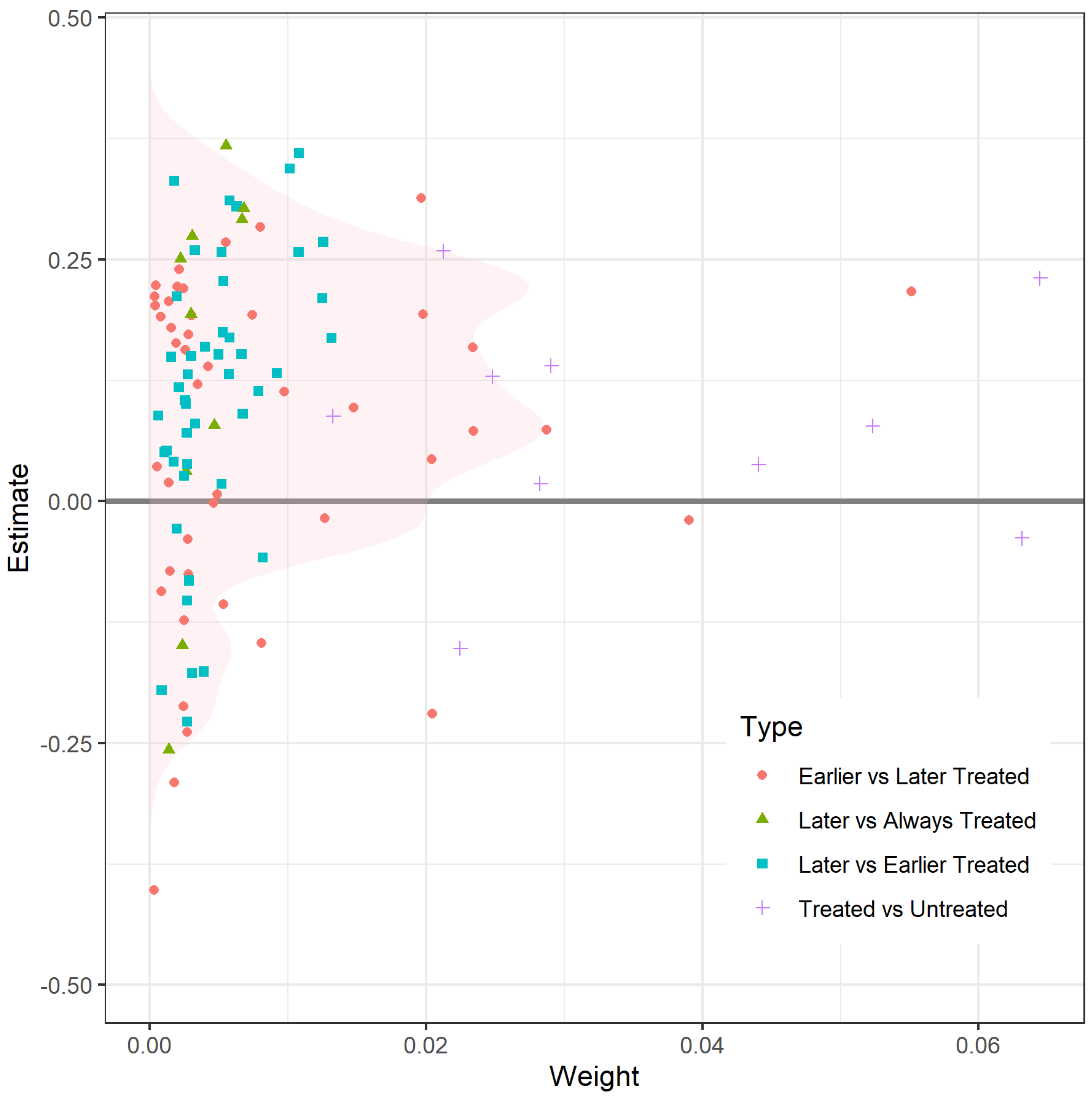
# Trimmed Subsample



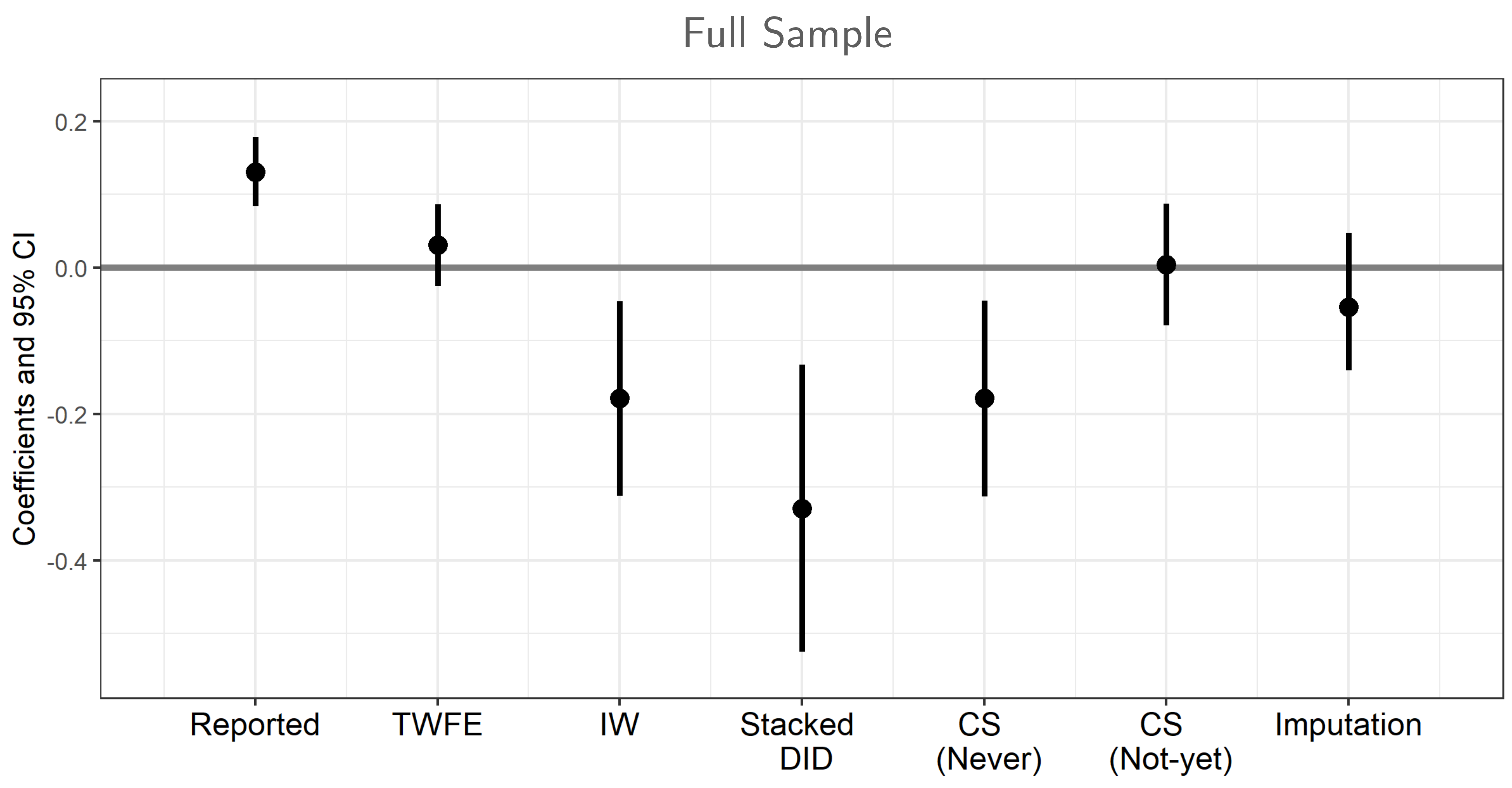
# Trimmed Subsample

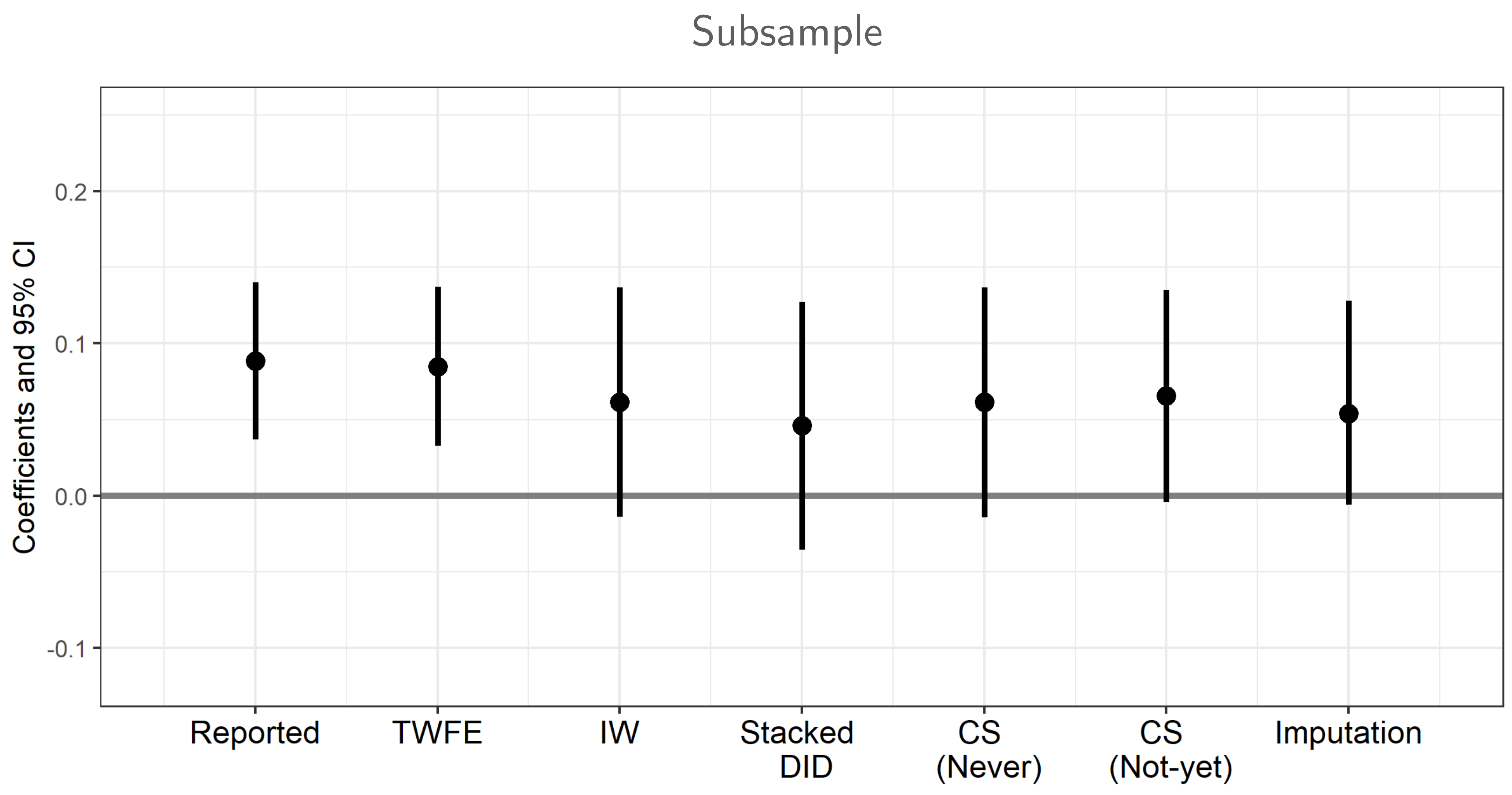


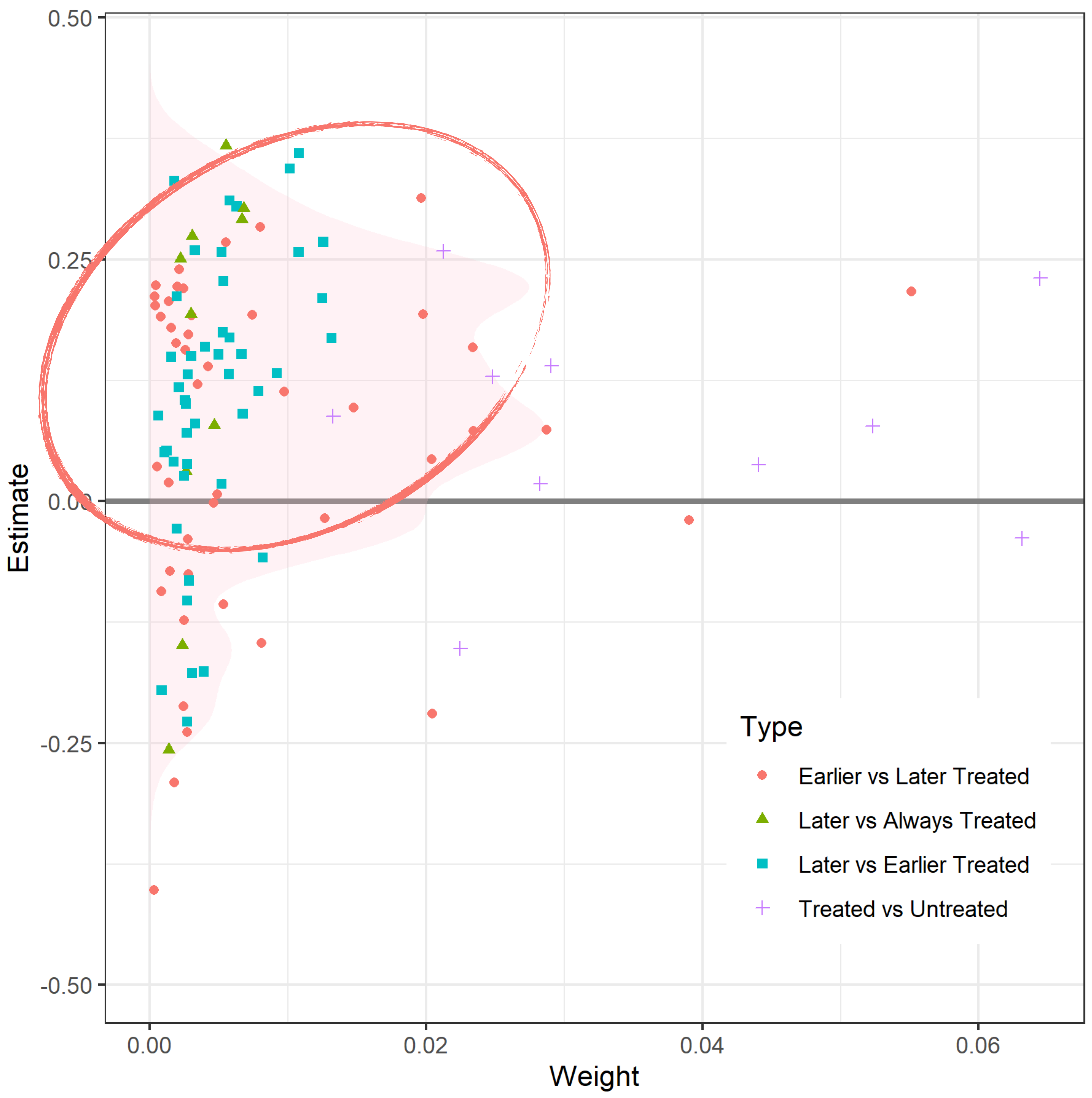



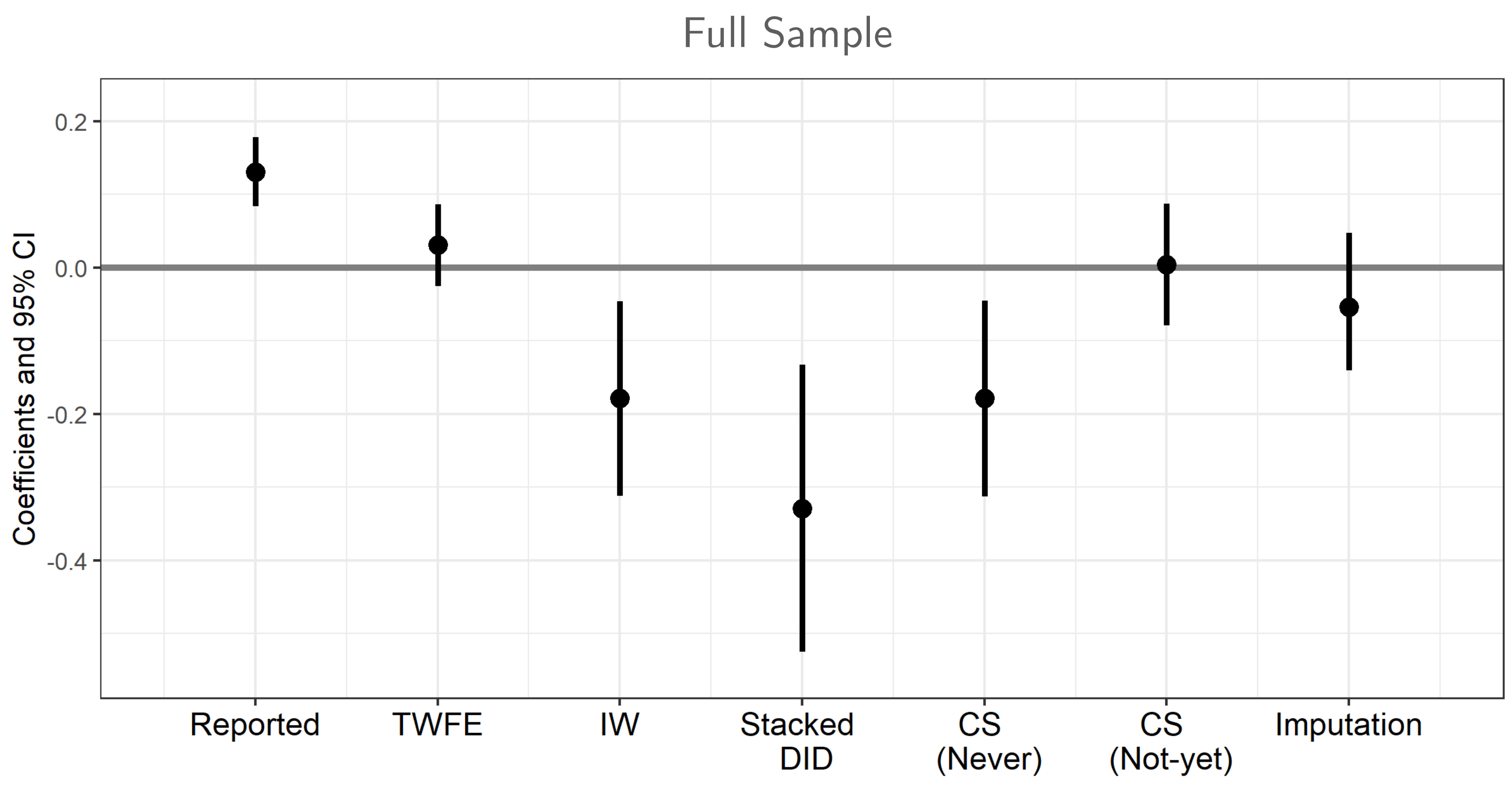
# Trimmed Subsample

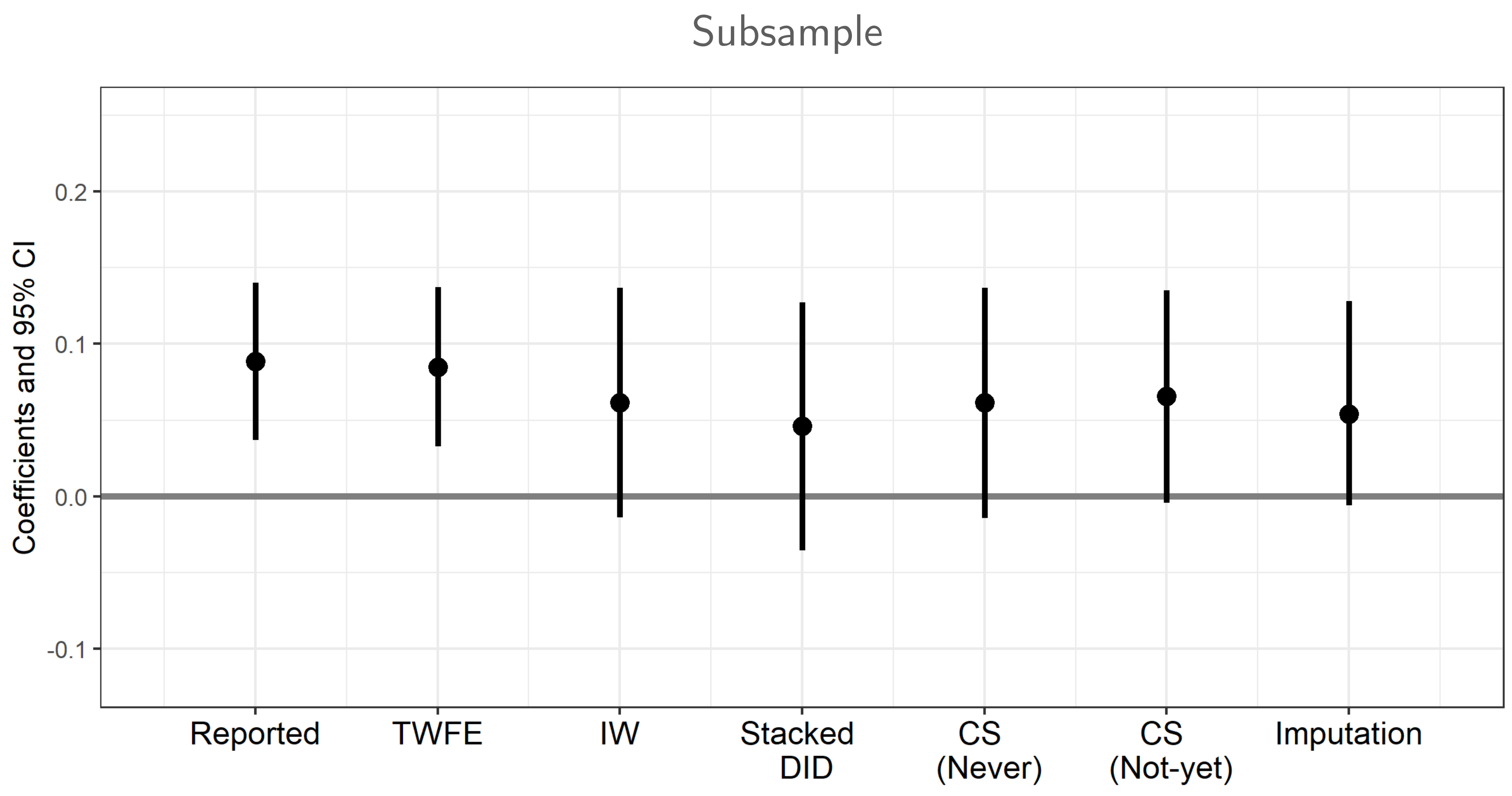


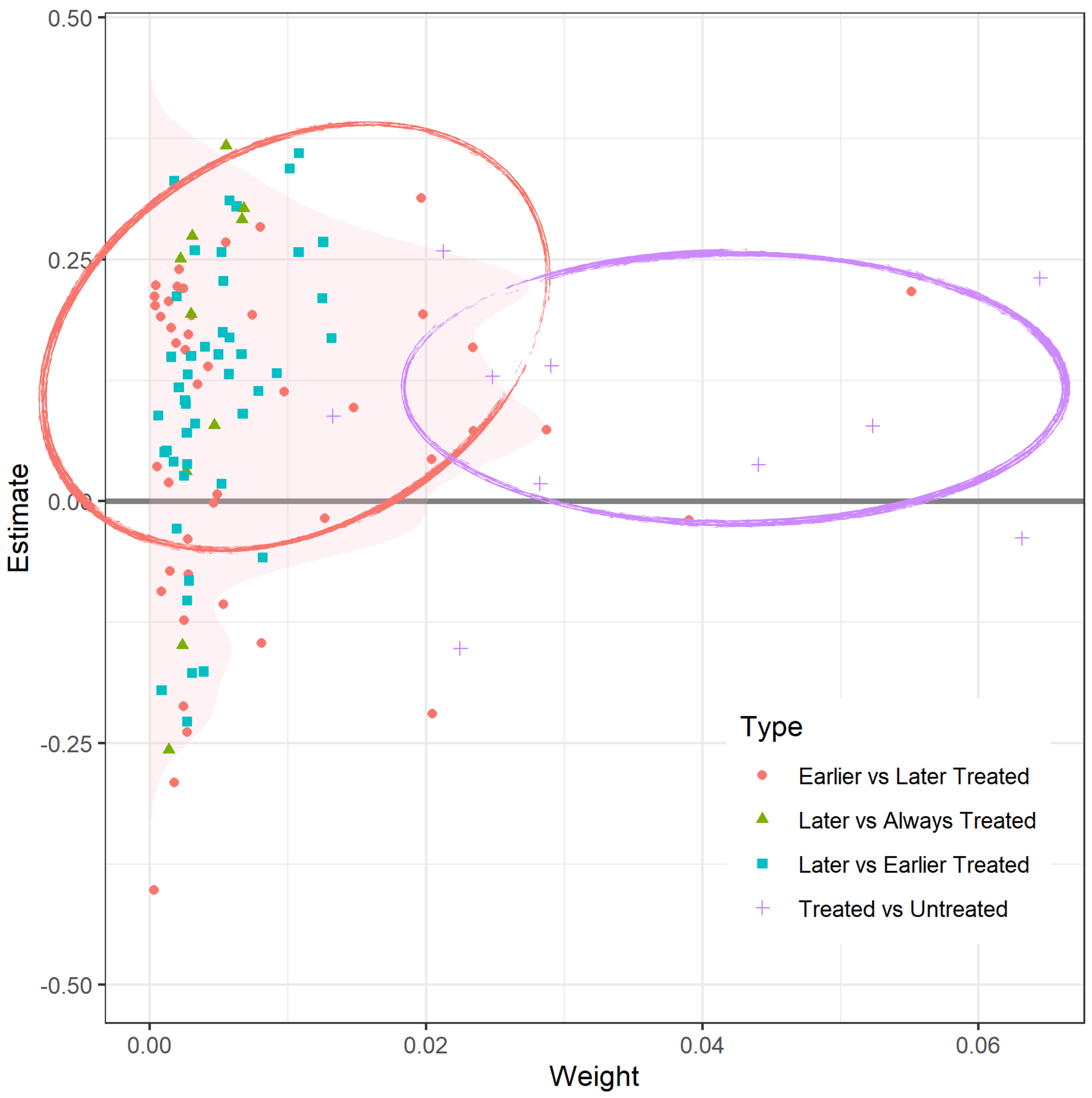



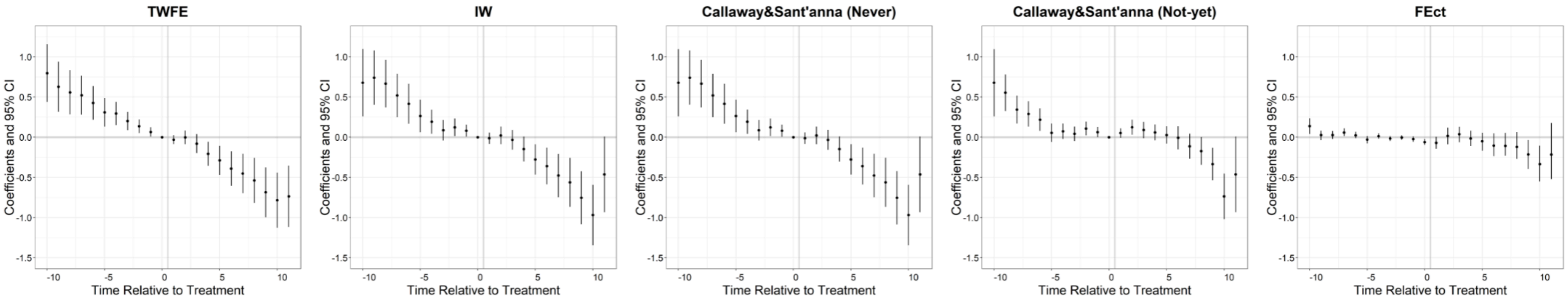
# Trimmed Subsample



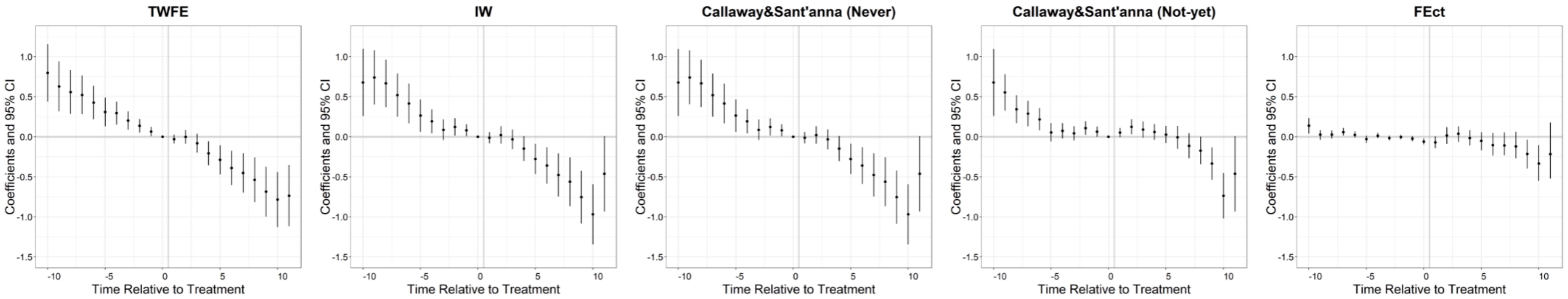



# Event Study Plots

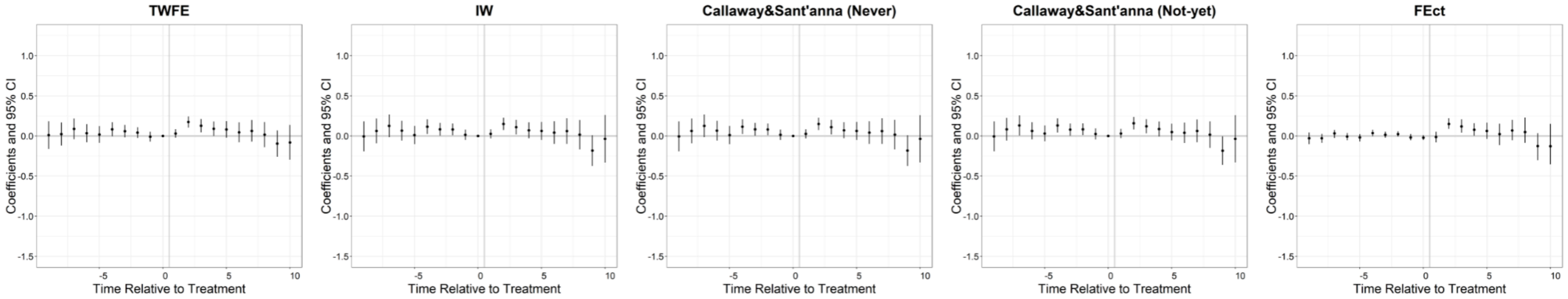



Full Sample

# Event Study Plots




Full Sample




Trimmed Sample

# Three Examples

- **Example 1**: Coethnic Mobilization
  - Strong design; HTE matters marginally — estimators (including TWFE) broadly agree
- **Example 2**: Lawsuit against land use restriction
  - Clear signs of PT violations
  - HTE is a second-order issue; agreement does not mean robustness
  - Simple plotting (and tests) will help spot the issue
- **Example 3**: Updating cadastral maps on tax revenue

# Three Examples

- **Example 1**: Coethnic Mobilization
  - Strong design; HTE matters marginally — estimators (including TWFE) broadly agree
- **Example 2**: Lawsuit against land use restriction
  - Clear signs of PT violations
  - HTE is a second-order issue; agreement does not mean robustness
  - Simple plotting (and tests) will help spot the issue
- **Example 3**: Updating cadastral maps on tax revenue
  - When estimators disagree, it may be a sign of PT violations

# Three Examples

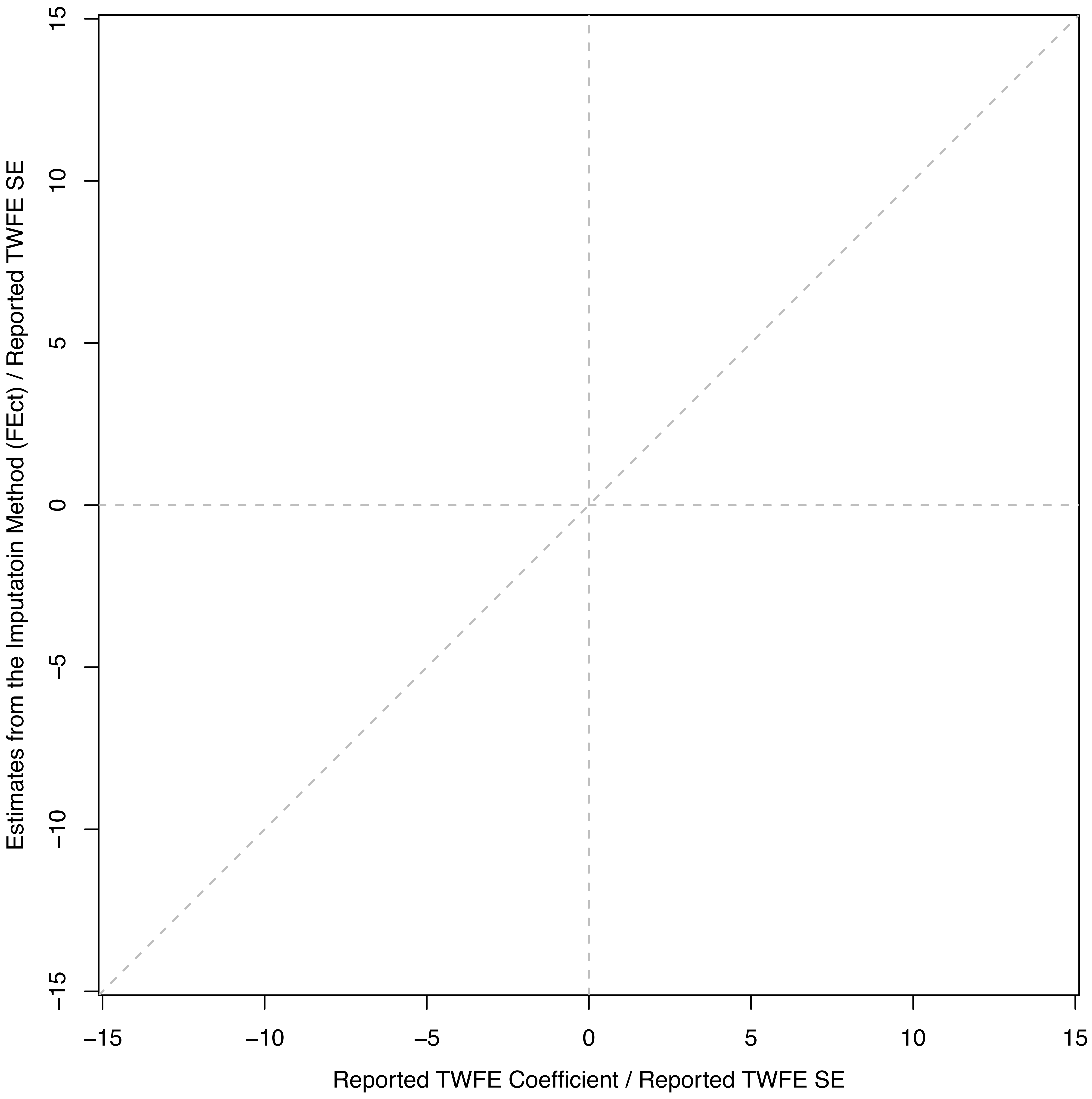
- **Example 1**: Coethnic Mobilization
  - Strong design; HTE matters marginally — estimators (including TWFE) broadly agree
- **Example 2**: Lawsuit against land use restriction
  - Clear signs of PT violations
  - HTE is a second-order issue; agreement does not mean robustness
  - Simple plotting (and tests) will help spot the issue
- **Example 3**: Updating cadastral maps on tax revenue
  - When estimators disagree, it may be a sign of PT violations
  - Design phrase, e.g. trimming, help improve inference (Imbens & Rubin 2015)

# Overal Assessment

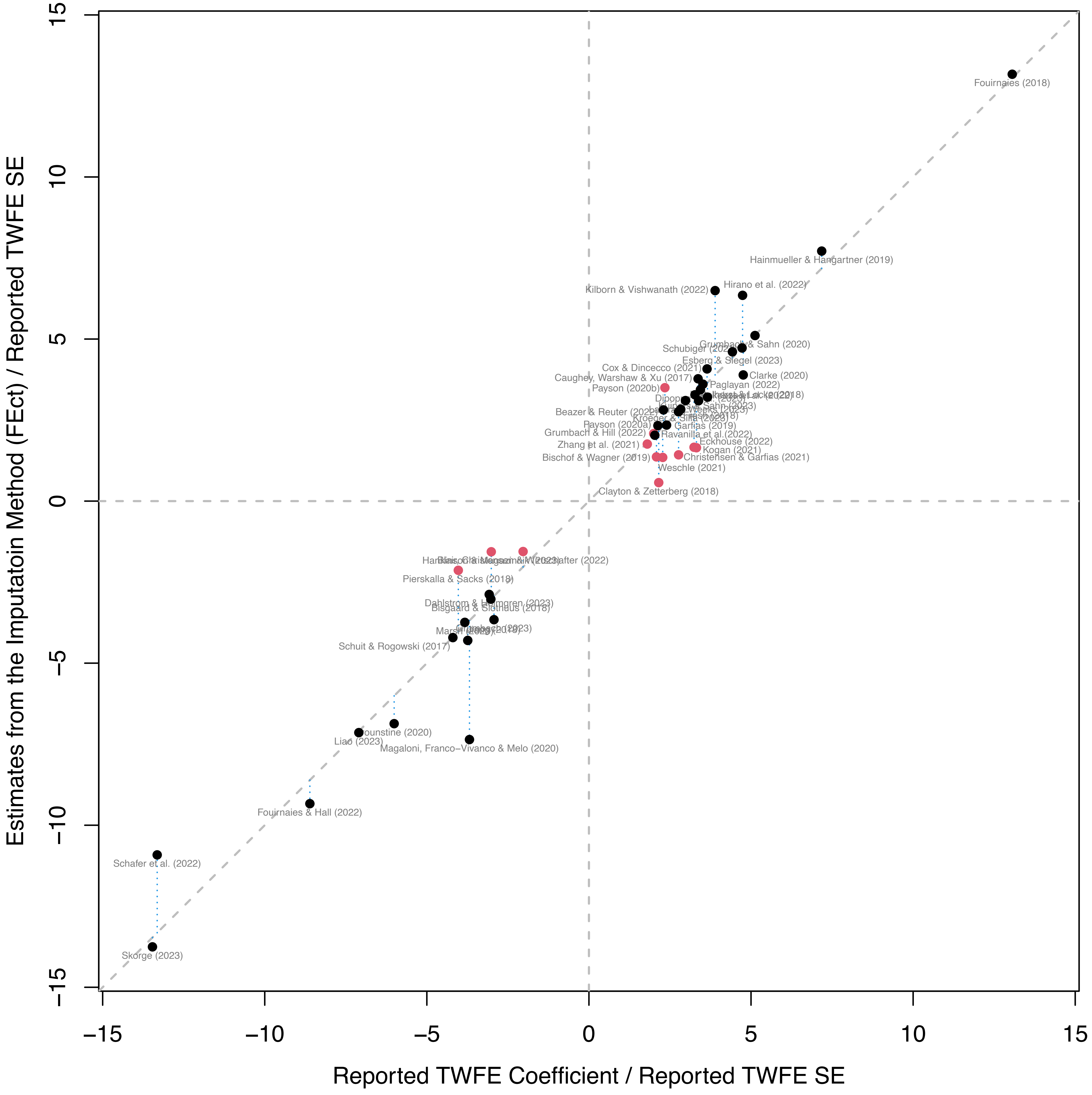
How much does HTE matter?

Why does “robust DID” require so much power?

# Do HTE-Robust Estimators Overturn Original Findings?

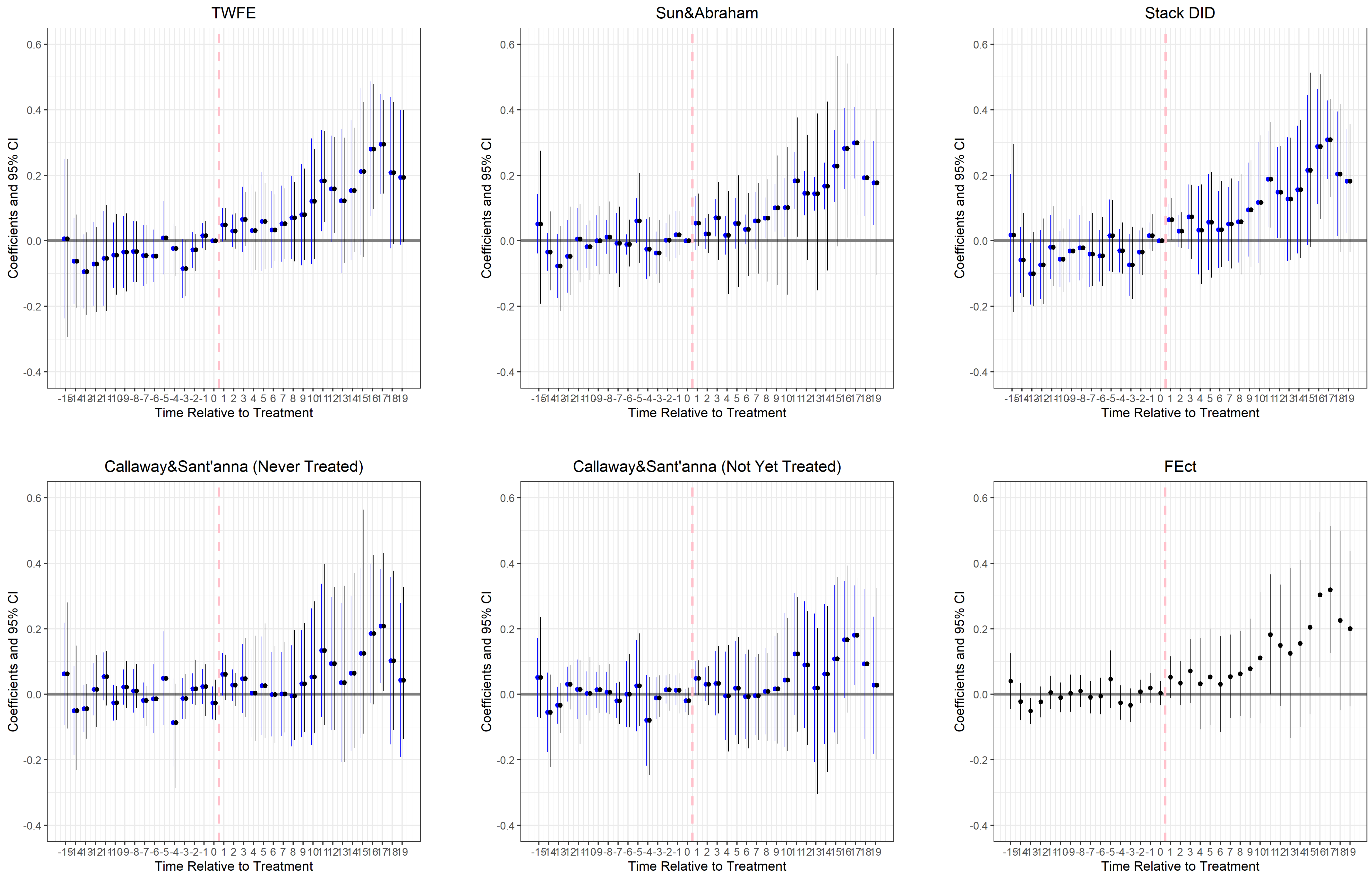


# Estimates from TWFE and Imputation Method Broadly Aligned



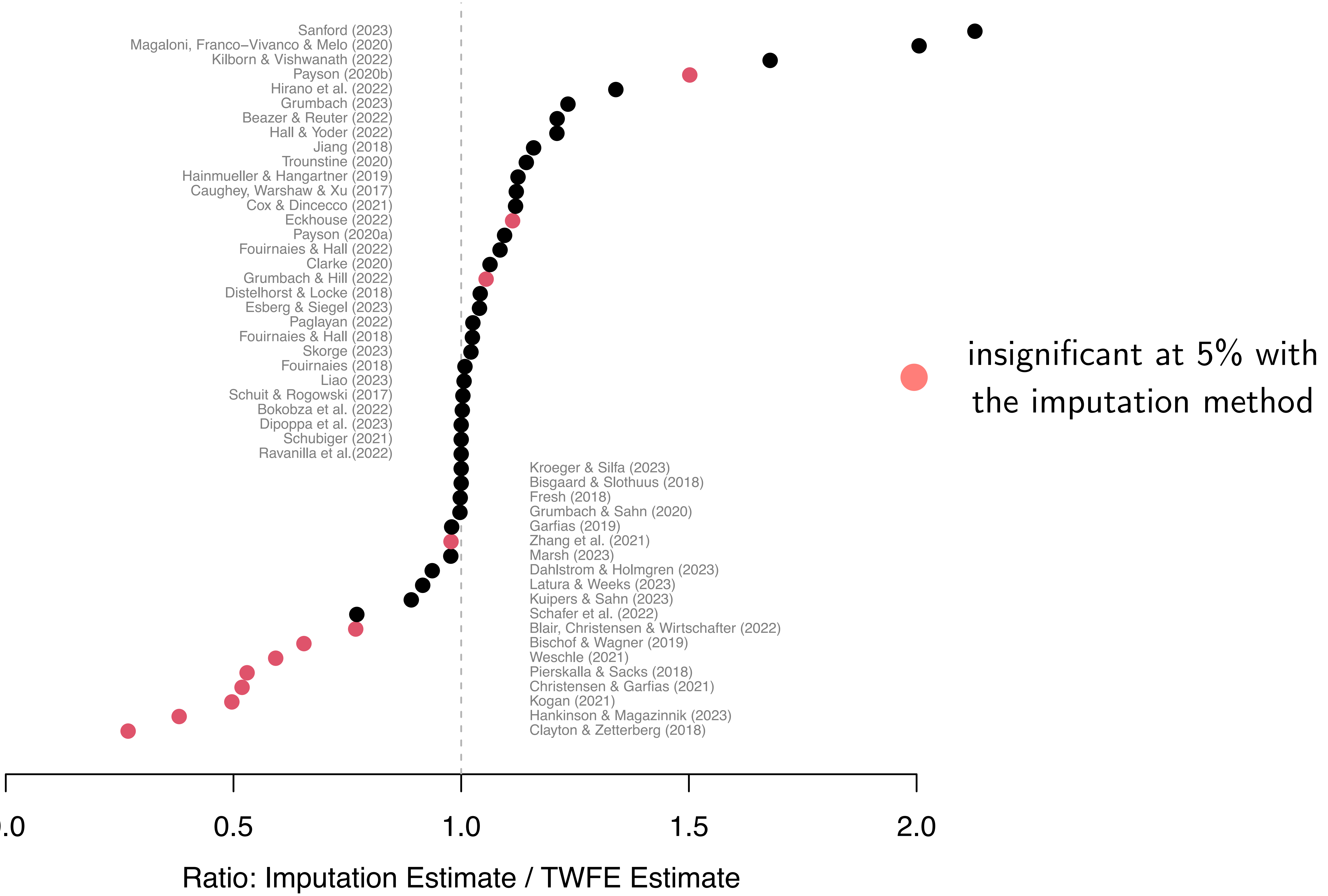

insignificant at 5% with the imputation method

# When PT Seems Plausible, Estimators Tend to Agree



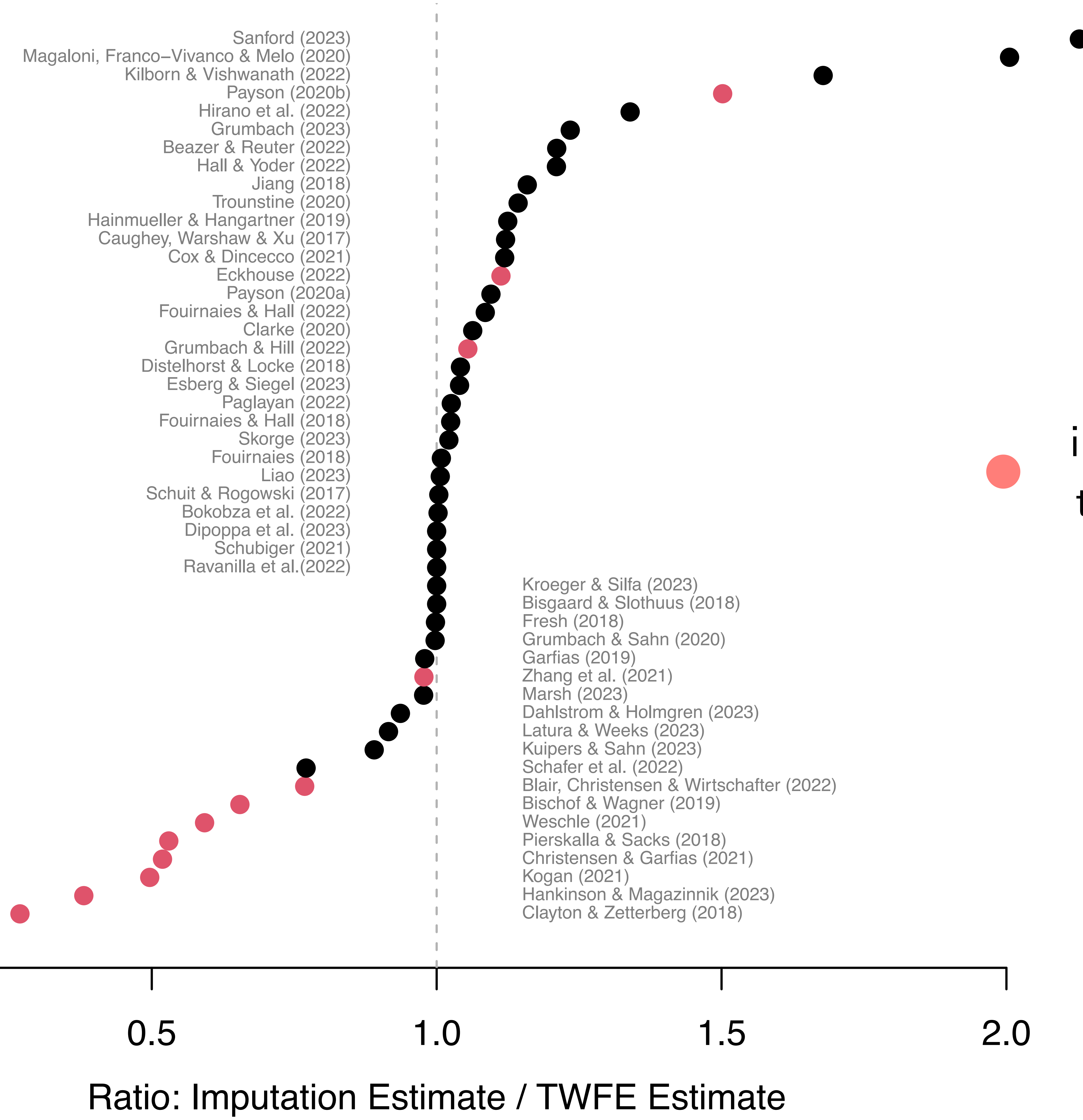

Bischof and Wagner (2019)

# However, Variability Cannot Be Overlooked



# However, Variability Cannot Be Overlooked

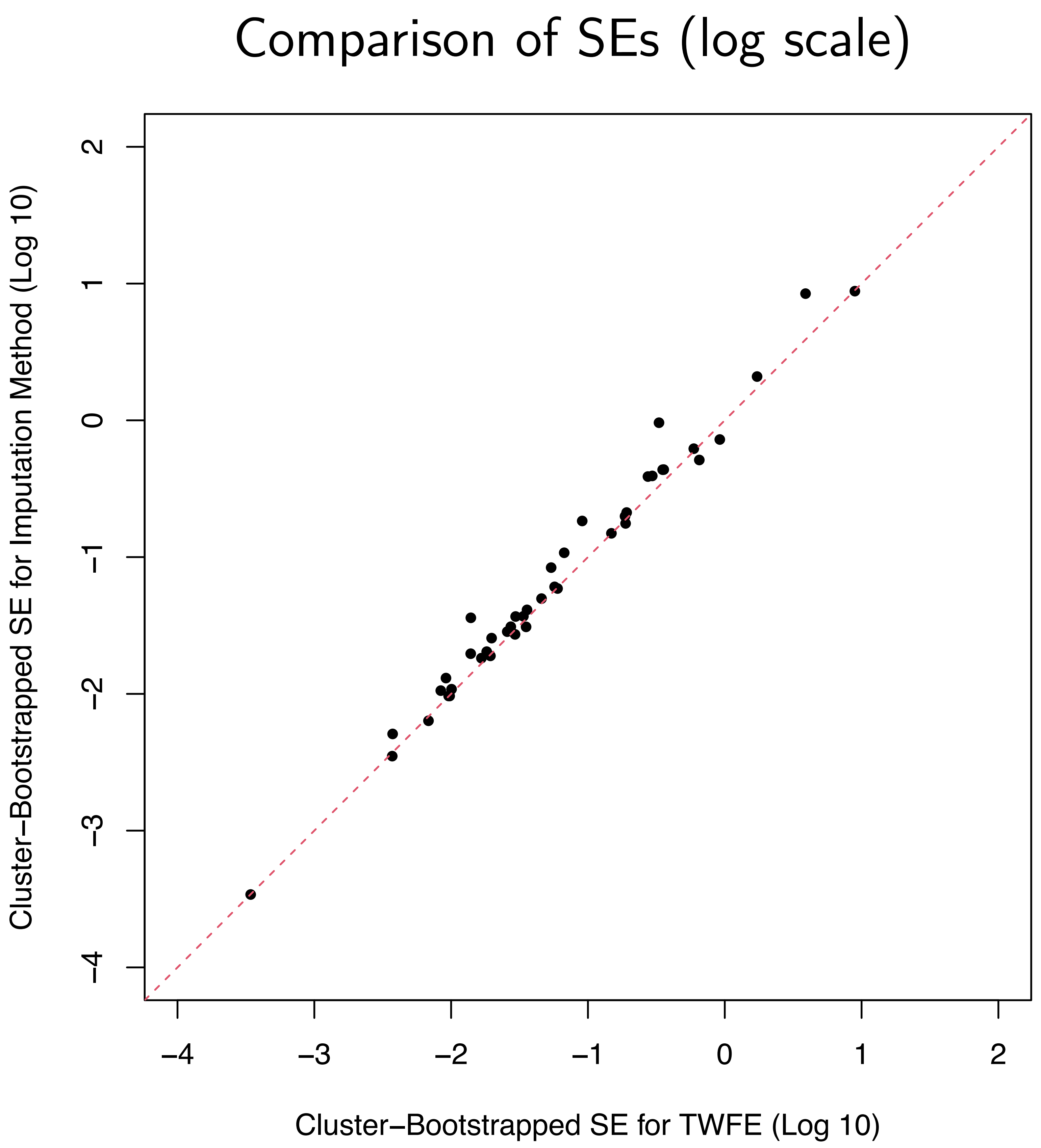
Mean(Ratio) = 1.02
Median(Ratio) = 1.00



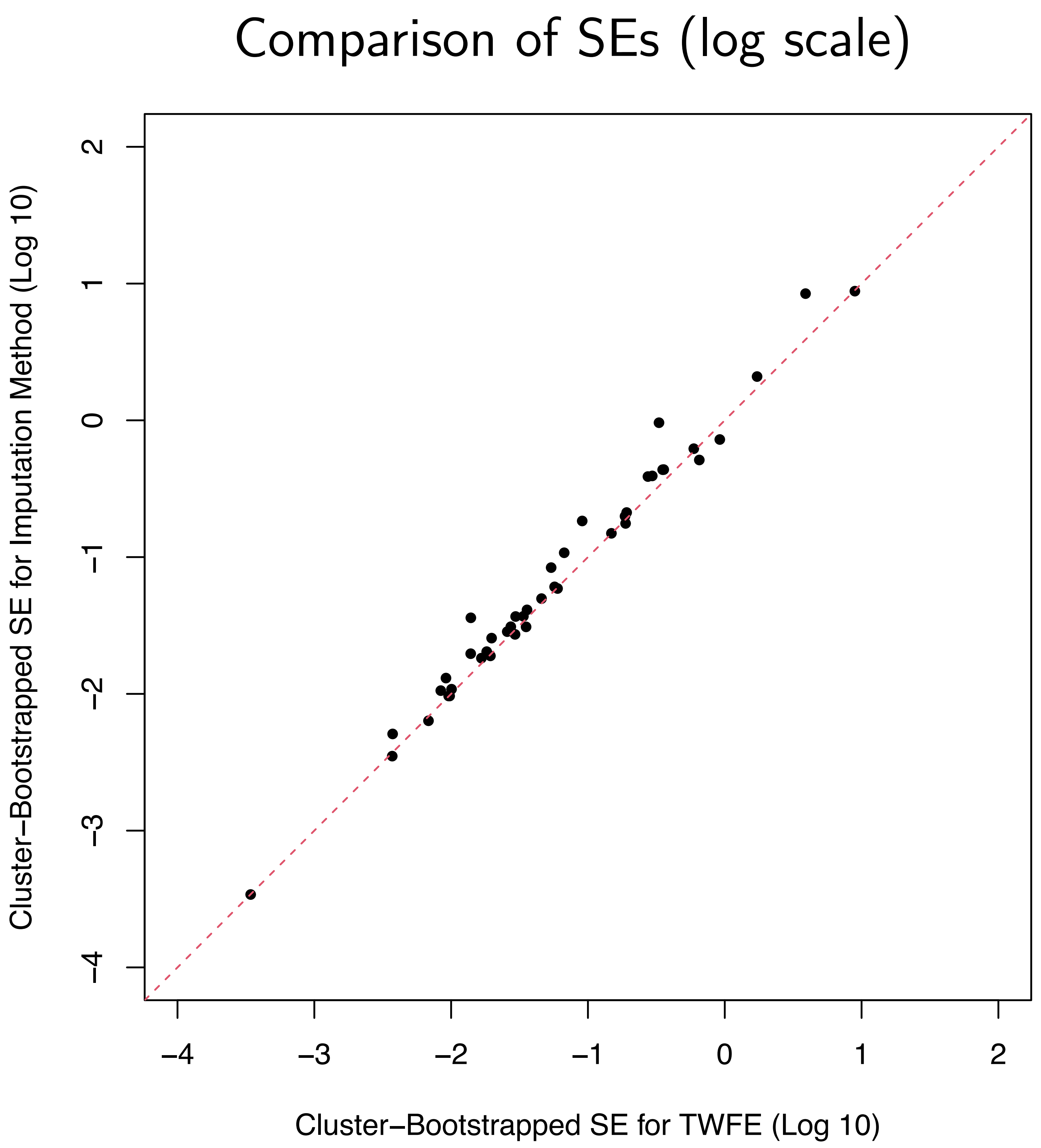

insignificant at 5% with the imputation method

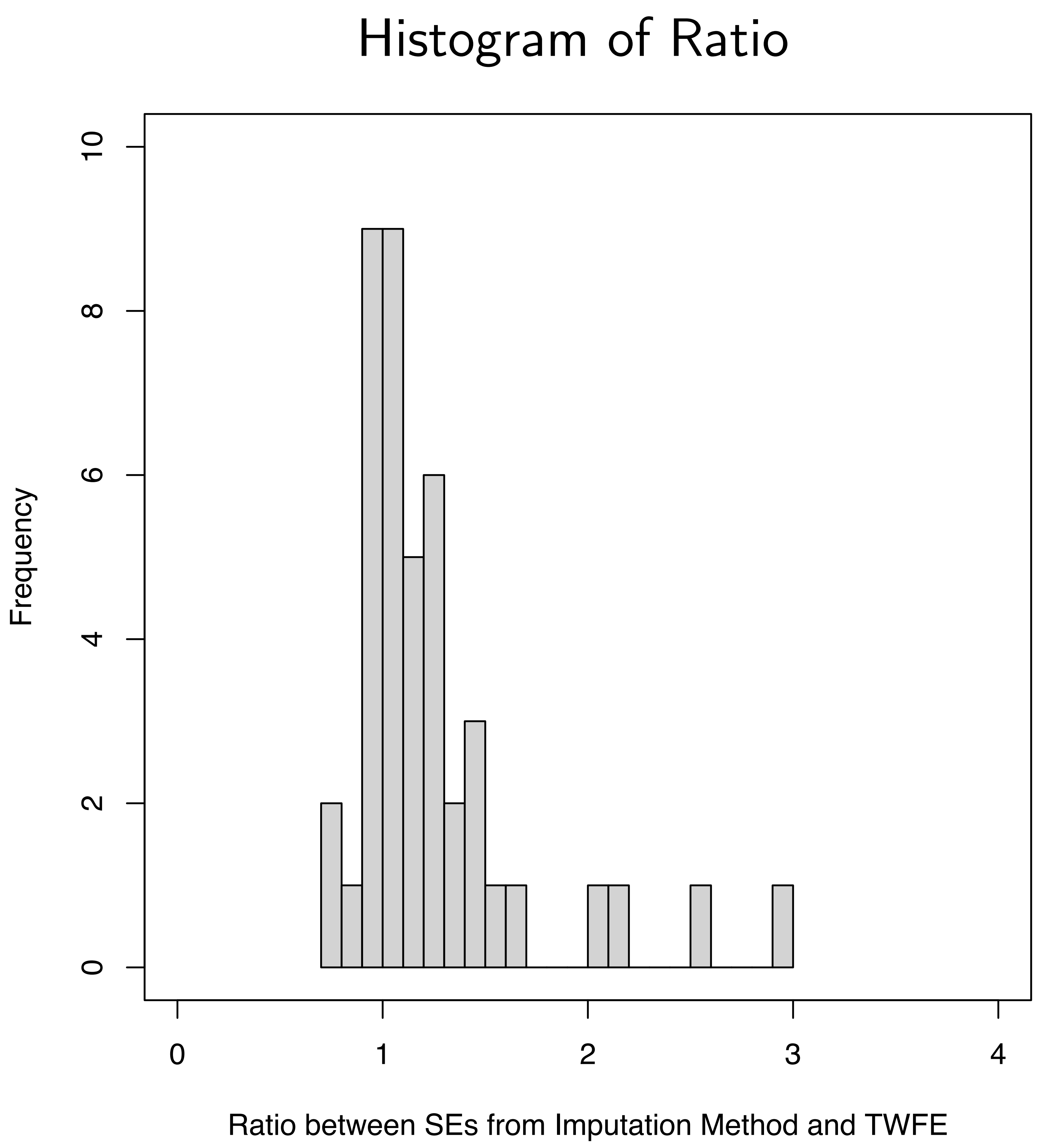
# Cost of Efficiency: TWFE versus Imputation SEs (Both Bootstrapped)

# Cost of Efficiency: TWFE versus Imputation SEs (Both Bootstrapped)

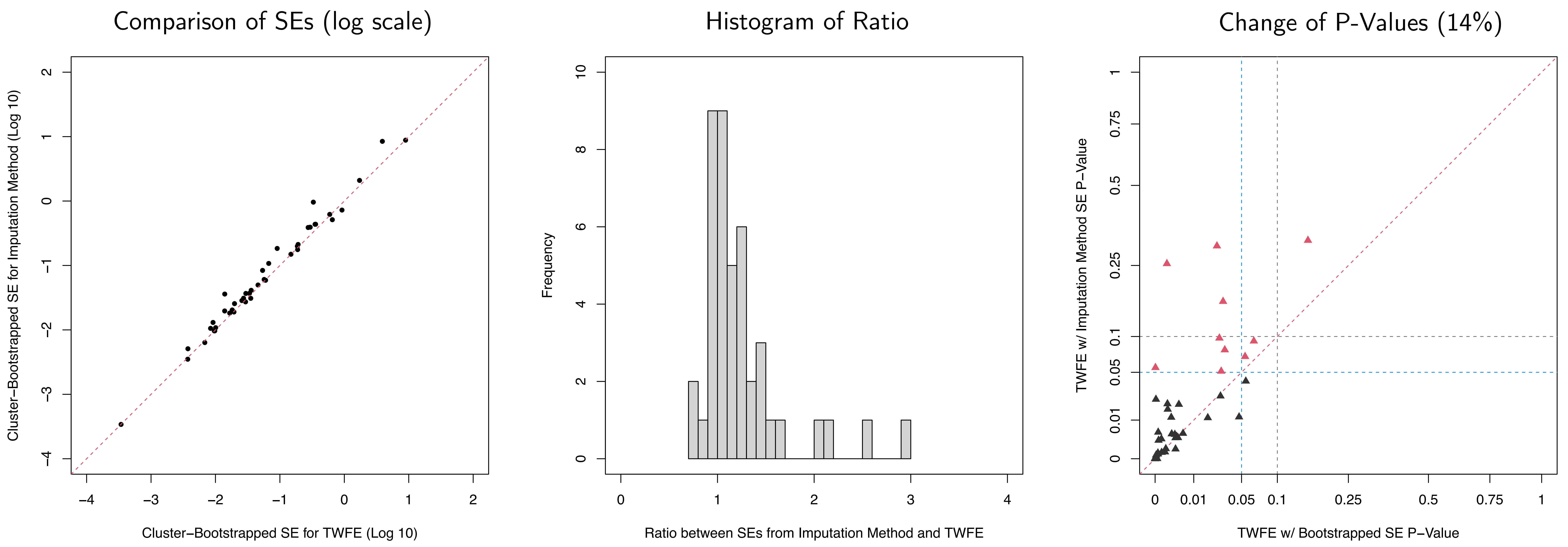


# Cost of Efficiency: TWFE versus Imputation SEs (Both Bootstrapped)

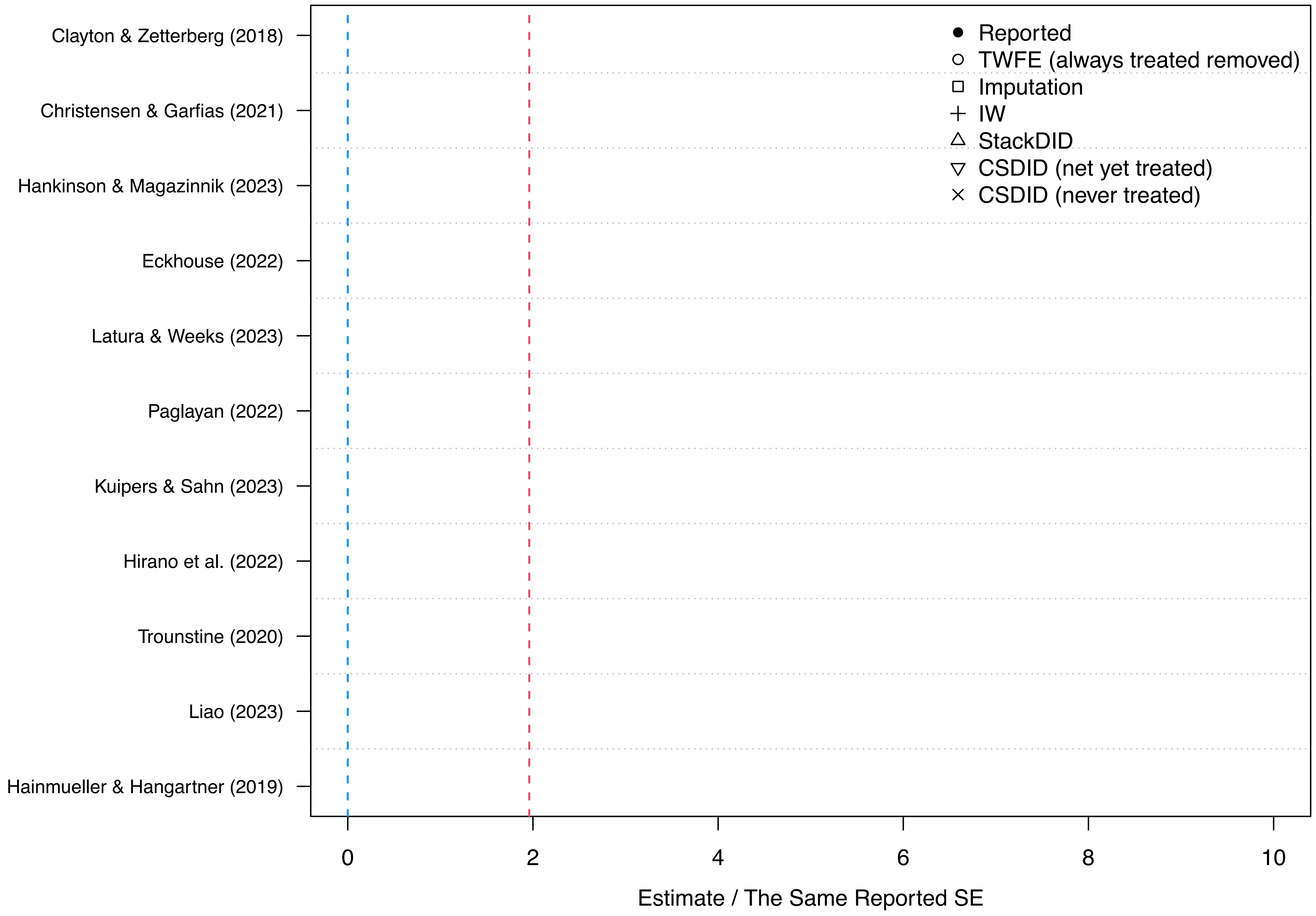


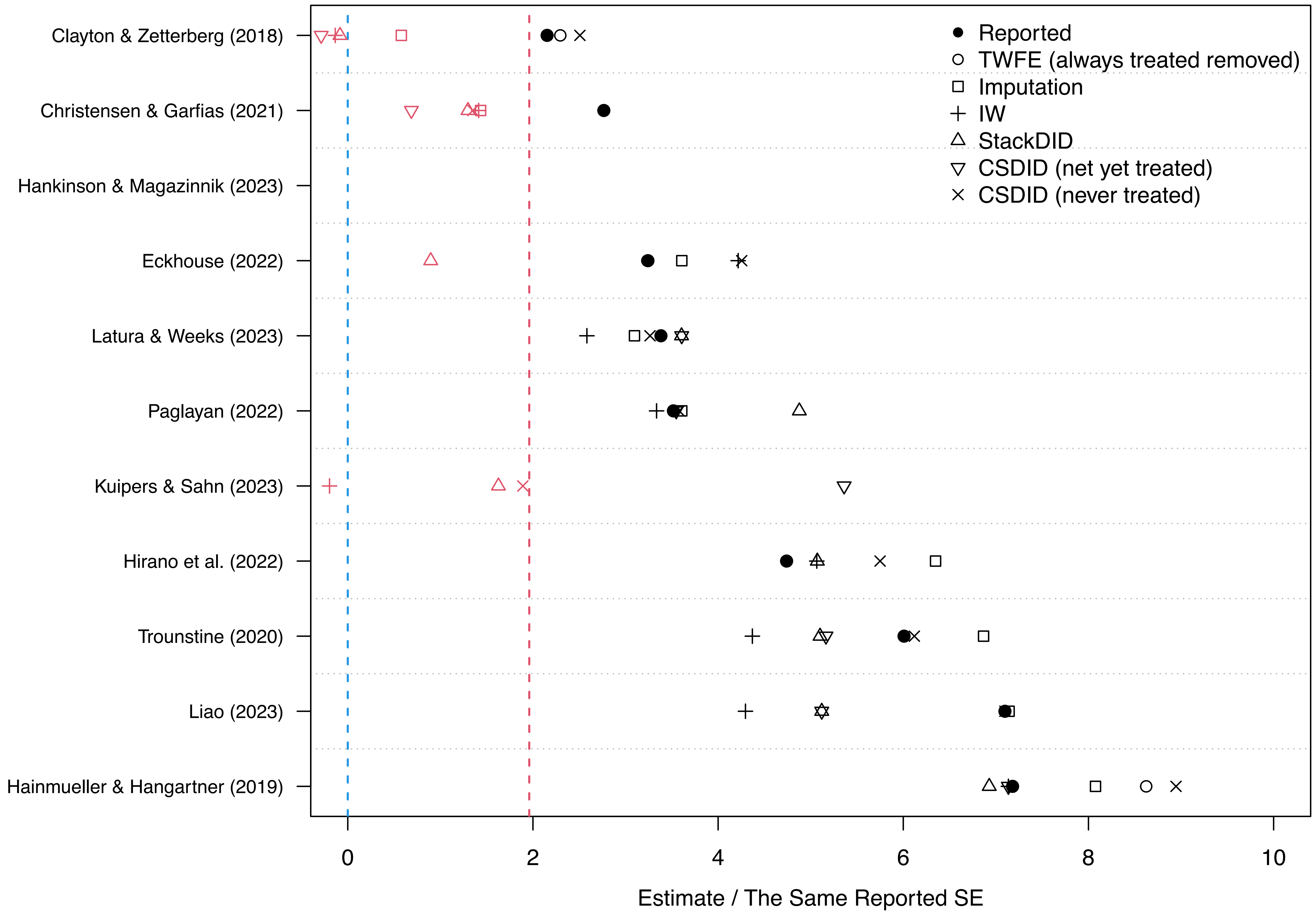
# Cost of Efficiency: TWFE versus Imputation SEs (Both Bootstrapped)



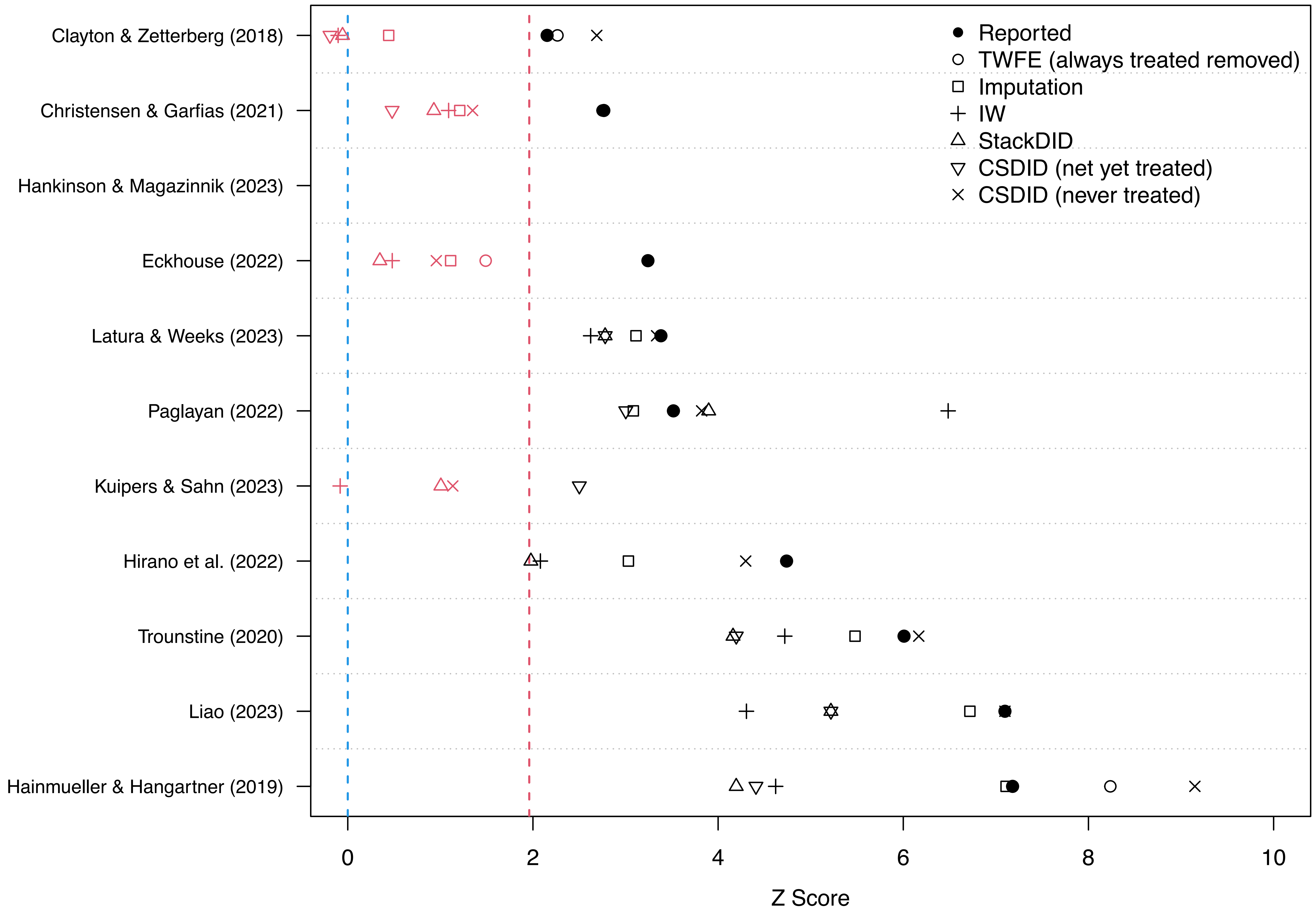
# The Staggered Cases — Coefficients



# The Staggered Cases — Coefficients

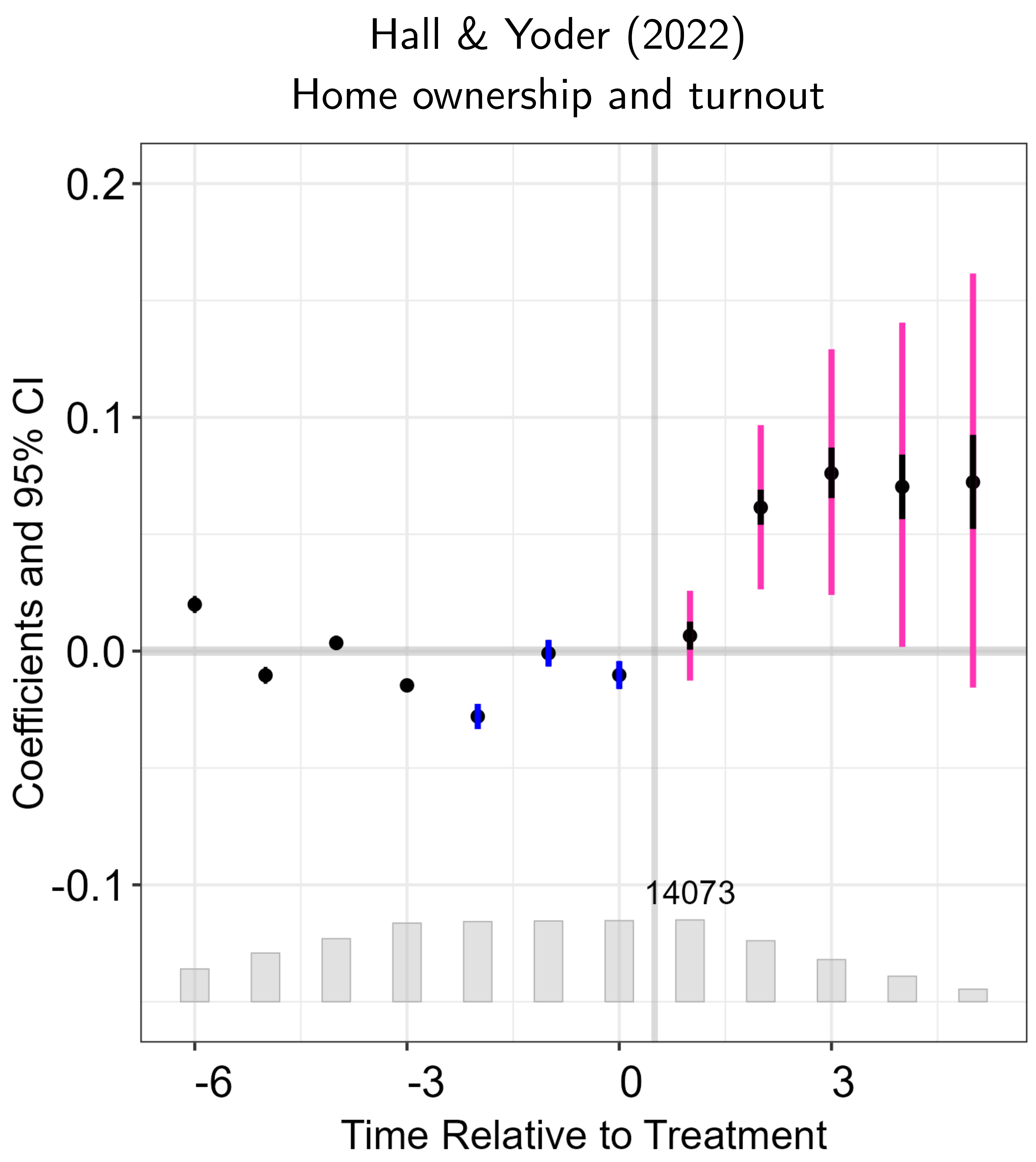


# The Staggered Cases — Z Scores

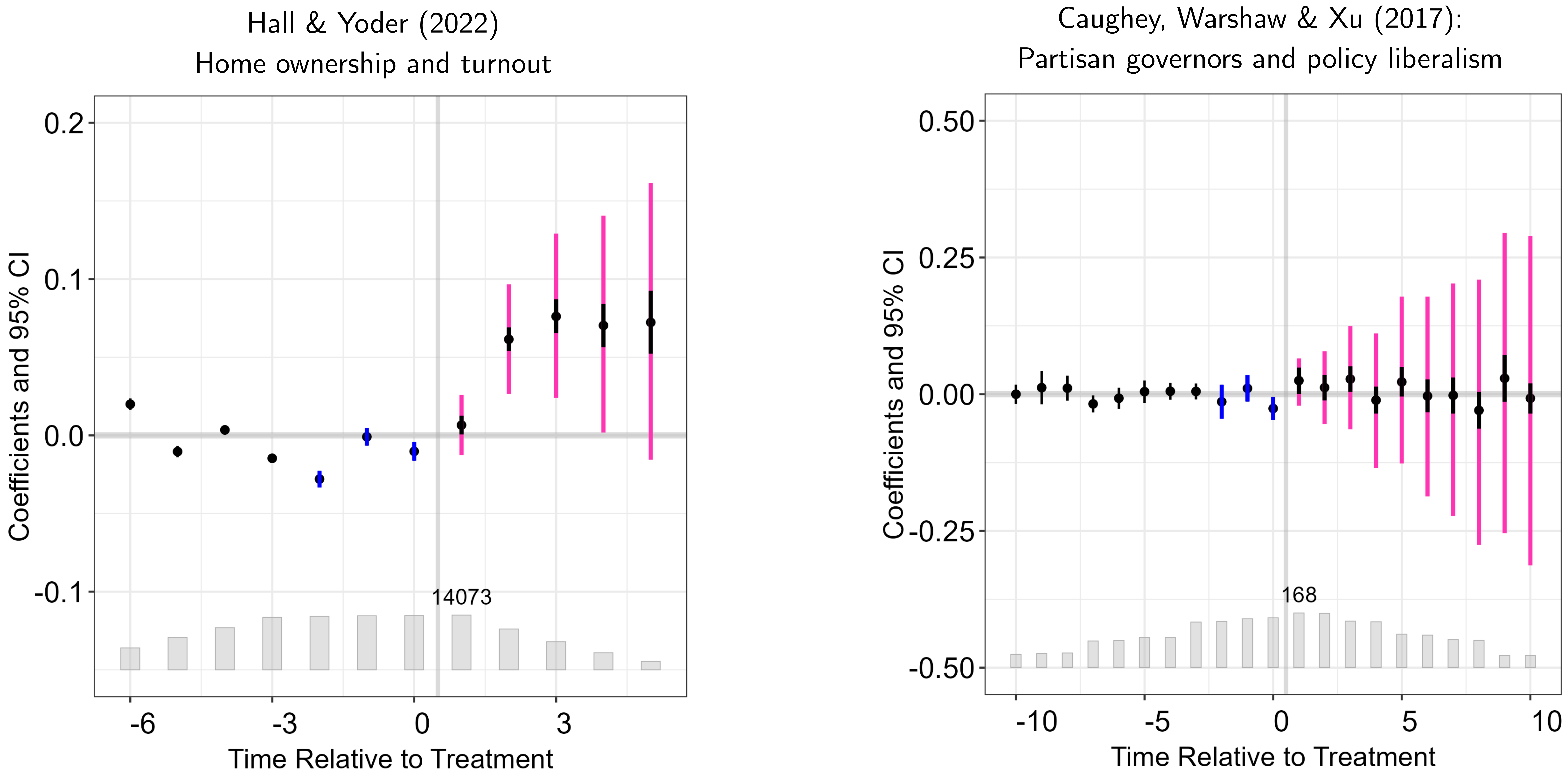


# Sensitivity Analysis with Relaxed PT $(M = 0.5)$

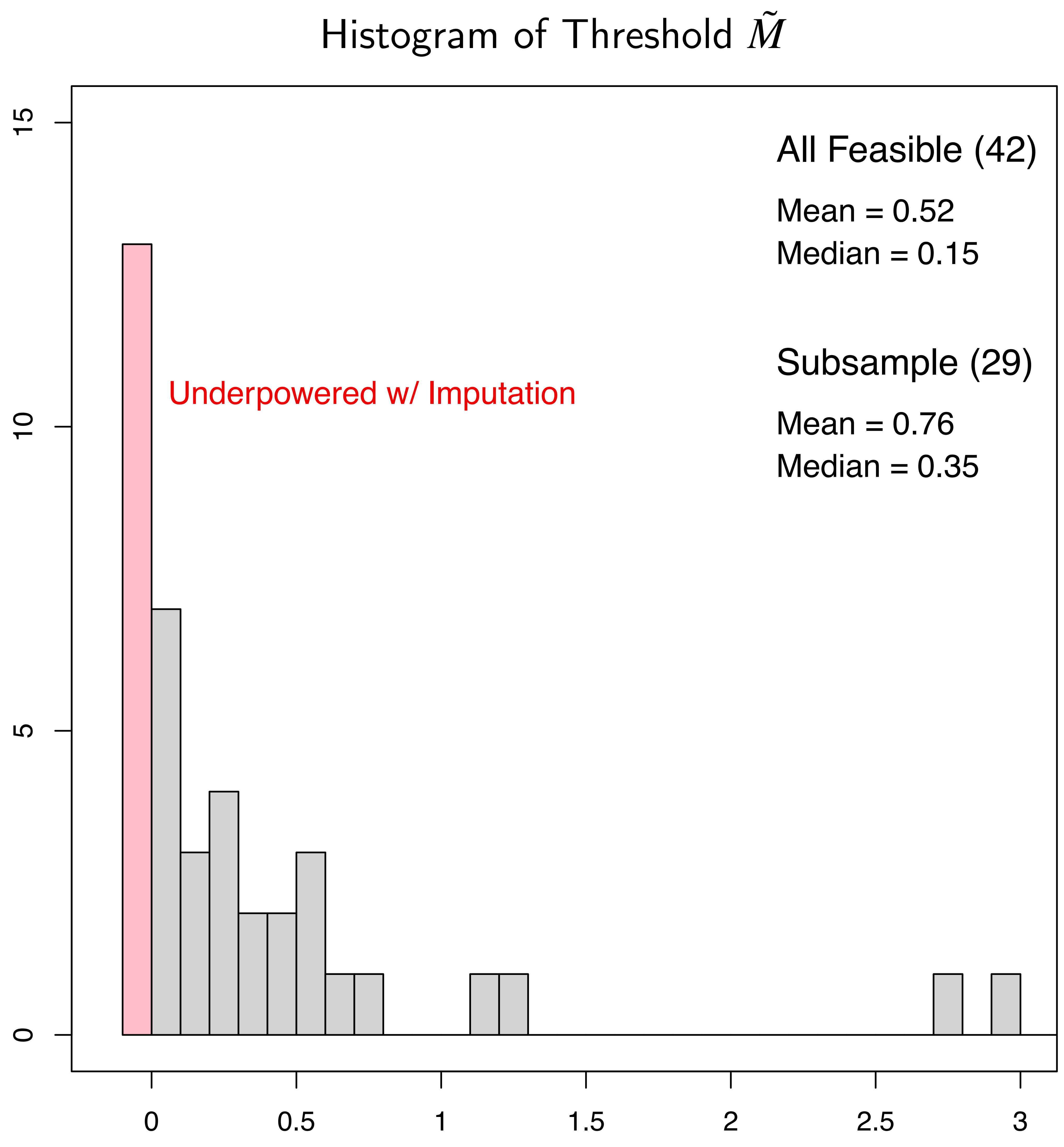
# Sensitivity Analysis with Relaxed PT (M = 0.5)



# Sensitivity Analysis with Relaxed PT (M = 0.5)



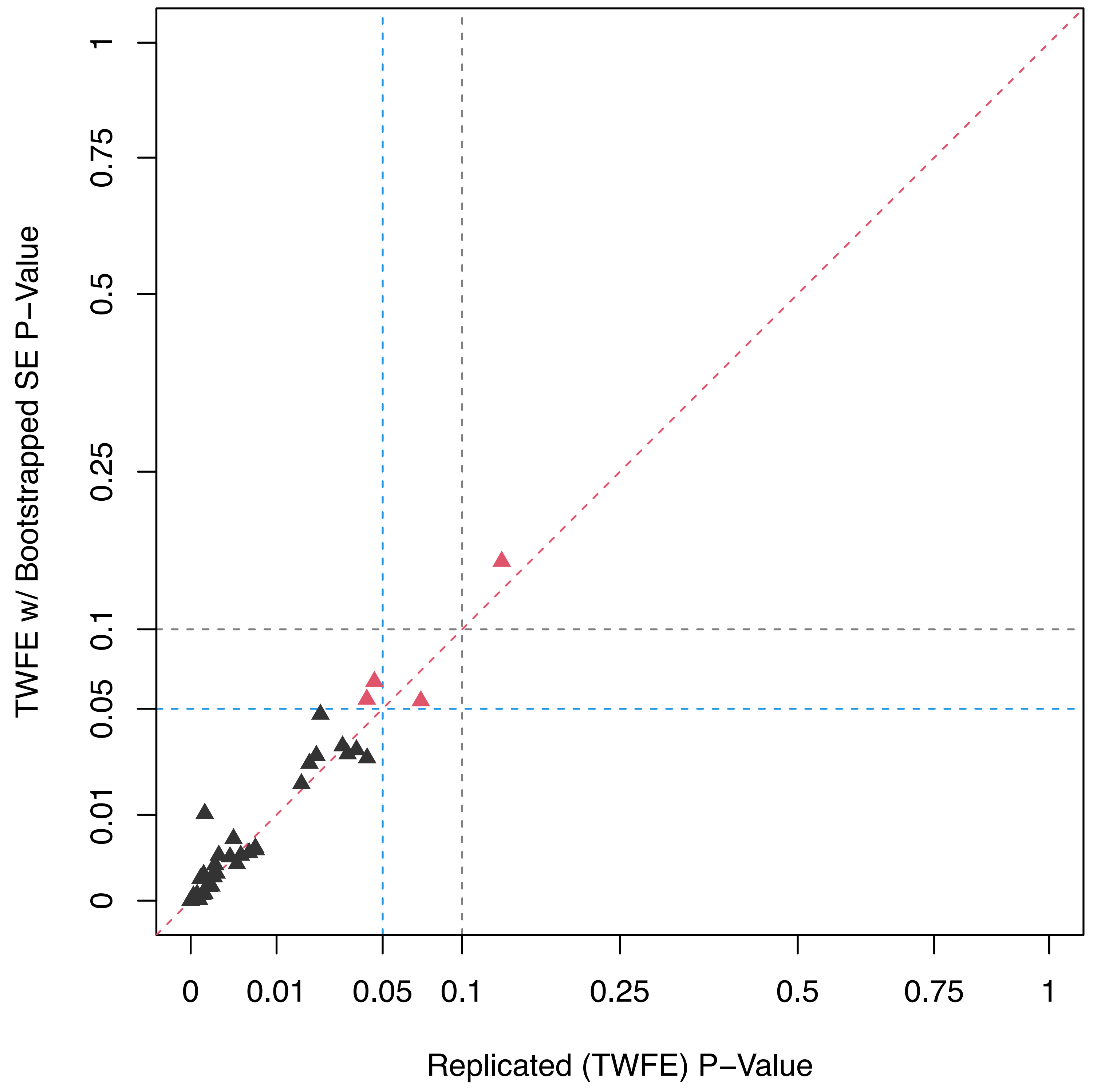
# Sensitivity Analysis with Relaxed PT



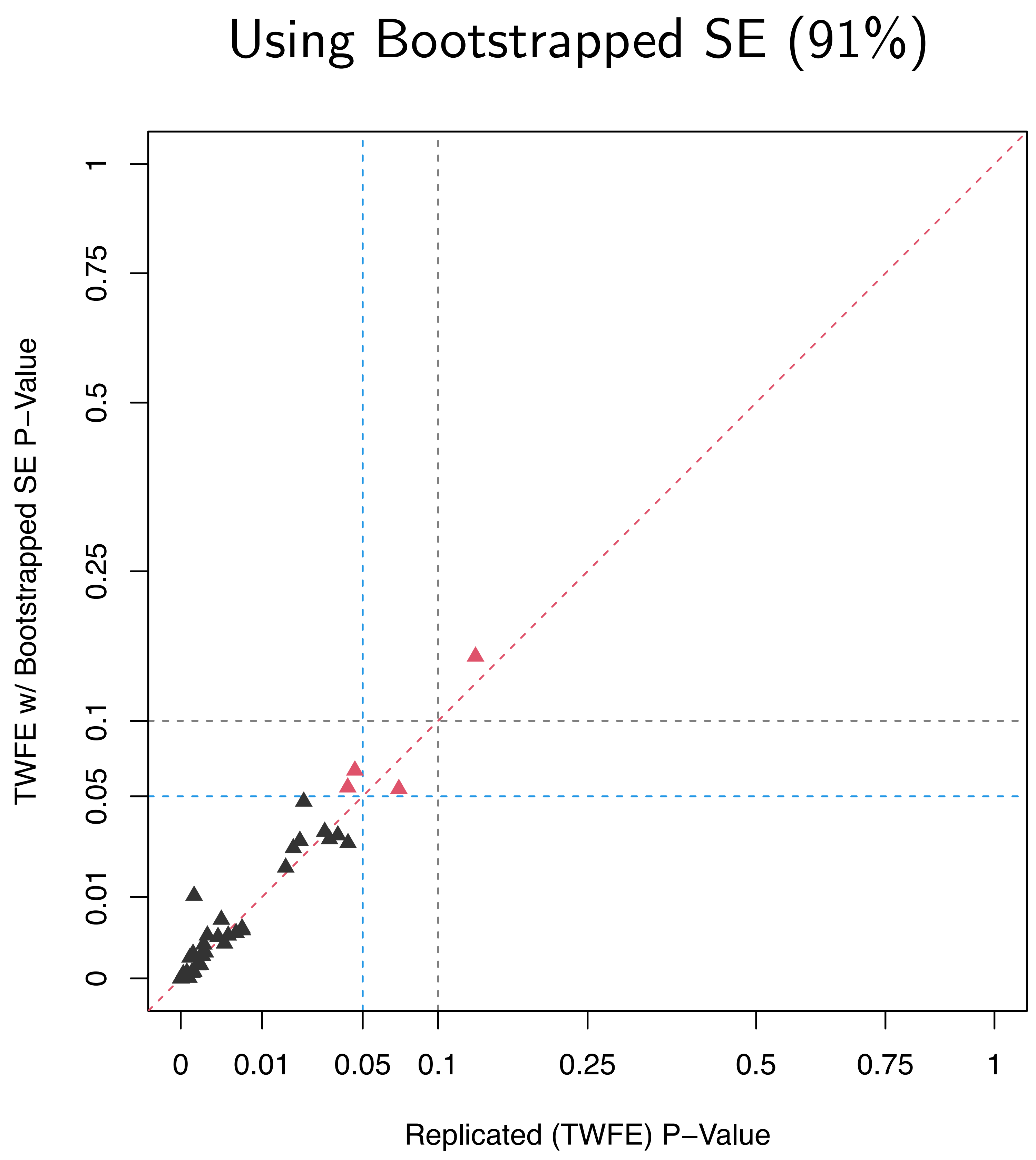
# “Robust” DID Is Power Hungry

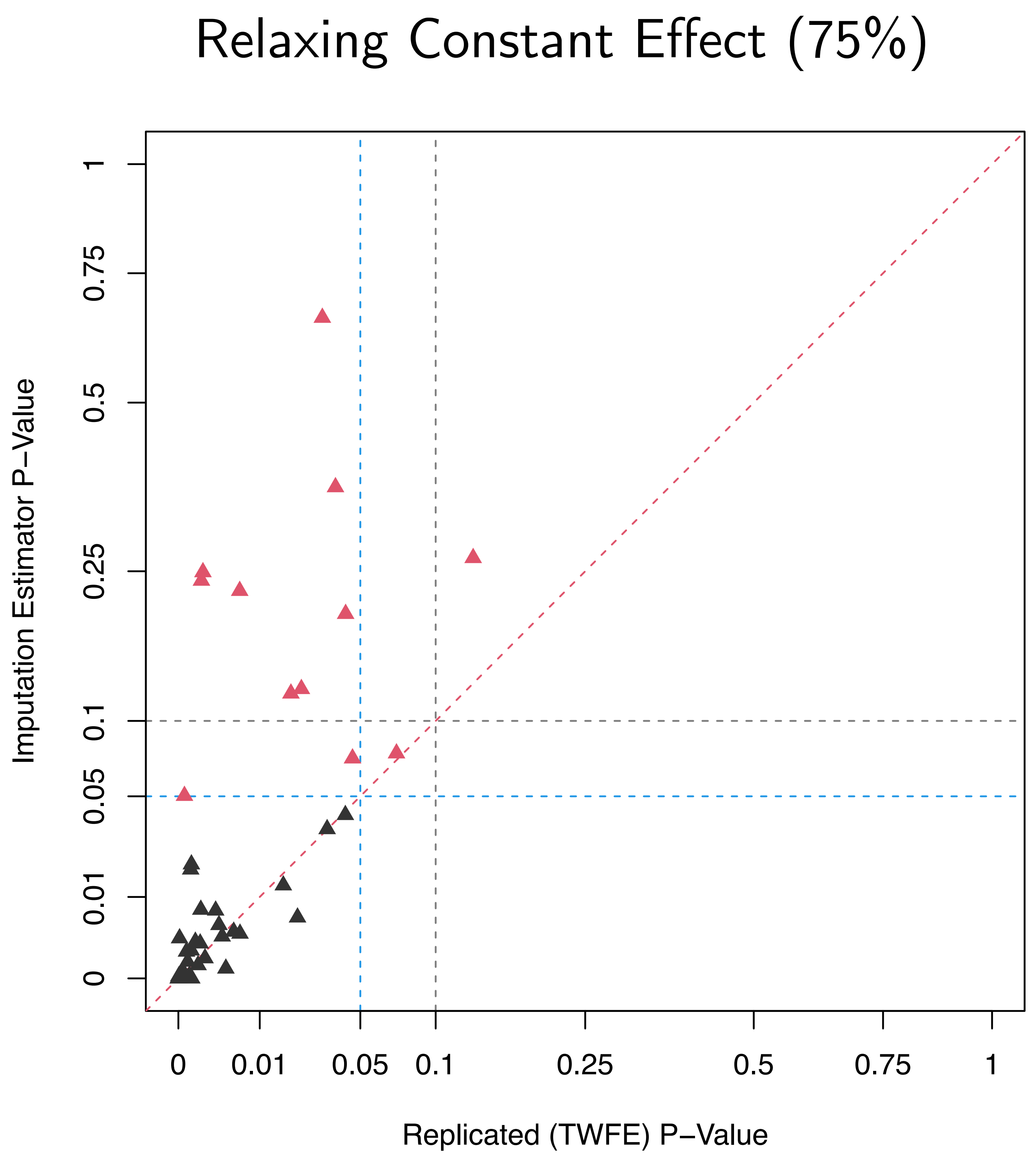
# "Robust" DID Is Power Hungry

Using Bootstrapped SE (91%)

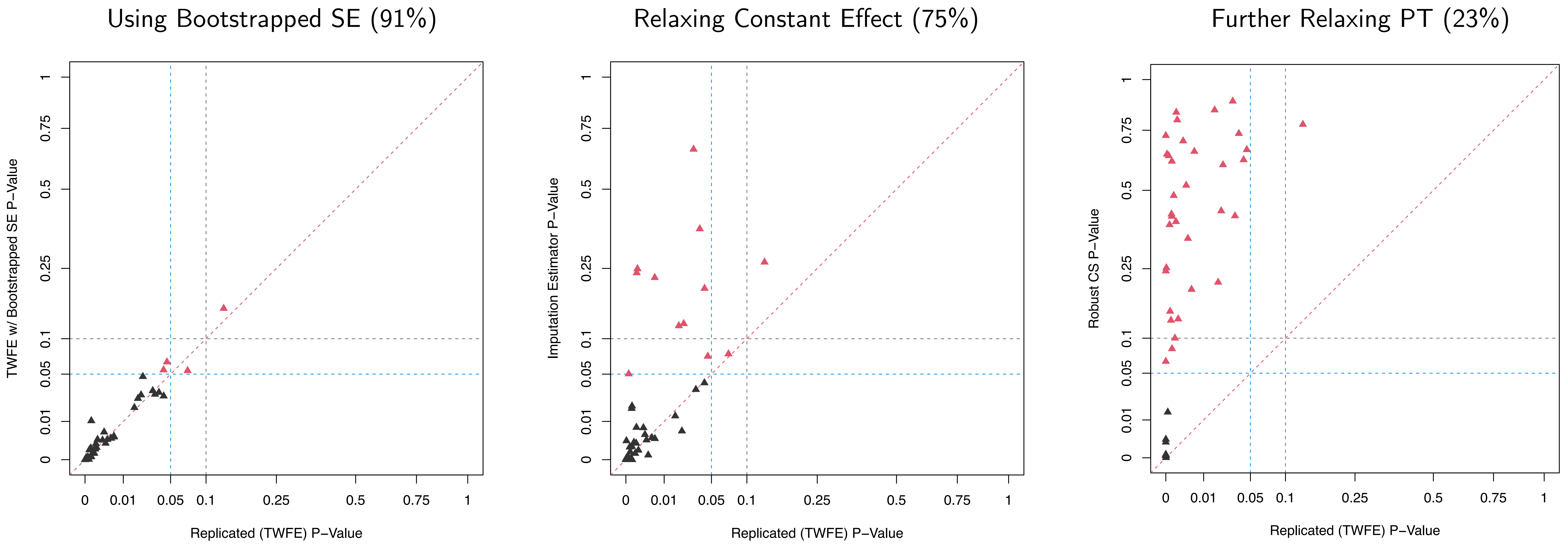


# "Robust" DID Is Power Hungry



# “Robust” DID Is Power Hungry



# Big Picture

# Big Picture

- Clear signs of PT violations still common (~25%)
  - In >50% cases, we cannot tell b/c too few pre-periods or low power

# Big Picture

- Clear signs of PT violations still common (~25%)
  - In >50% cases, we cannot tell b/c too few pre-periods or low power
- HTE matters (but it's complicated)
  - Few sign-flipping
  - Estimators tend to agree when PT seems plausible
  - Large variability in some cases, likely driven by sparse data & PT violations

# Big Picture

- Clear signs of PT violations still common (~25%)
  - In >50% cases, we cannot tell b/c too few pre-periods or low power
- HTE matters (but it's complicated)
  - Few sign-flipping
  - Estimators tend to agree when PT seems plausible
  - Large variability in some cases, likely driven by sparse data & PT violations

Relaxing constant effects or the PT leads to insufficient power

# Big Picture

- Clear signs of PT violations still common (~25%)
  - In >50% cases, we cannot tell b/c too few pre-periods or low power
- HTE matters (but it's complicated)
  - Few sign-flipping
  - Estimators tend to agree when PT seems plausible
  - Large variability in some cases, likely driven by sparse data & PT violations

Relaxing constant effects or the PT leads to insufficient power

"paradox" of committing to TWFE

w/ enough power, can afford HTE-robust estimators

w/o enough power, cannot validate TWFE assumptions

# Big Picture

- Clear signs of PT violations still common (~25%)
  - In >50% cases, we cannot tell b/c too few pre-periods or low power
- HTE matters (but it's complicated)
  - Few sign-flipping
  - Estimators tend to agree when PT seems plausible
  - Large variability in some cases, likely driven by sparse data & PT violations
- Other Issues
  - Missing data (unlikely Missing-At-Random)
  - Carryover effects are common

Relaxing constant effects or the PT leads to insufficient power

"paradox" of committing to TWFE

w/ enough power, can afford HTE-robust estimators

w/o enough power, cannot validate TWFE assumptions

# Recommendations

# Recommendations

| | Do's | Don'ts |
|---|---|---|
| **Design trumps analysis** | Start empirical analysis with a research design; proceed if "feedback" from past outcomes to treatment assignment is not a major concern | Start empirical analysis by blindly running regressions using existing data |
| **Discussion of designs** | Clearly specify designs and their corresponding identification assumptions | Equate designs with outcome models |
| **Plot raw data** | Plot raw data to better understand the research setting, missingness, sources of variations in the treatment and outcome variables, and univariate/bivariate distributions | Run regressions without looking at the data |
| **Estimators** | Choose HTE-robust estimators and always plot the estimated dynamic treatment effects | Choose models solely based on your beliefs; report regression coefficients only; no results visualization or diagnostics |
| **Diagnostics** | Conduct both visual and statistical tests to gauge the validity the identification and modeling assumptions | |
| **Level of clustering** | Cluster SEs at the level of treatment assignment or higher to account for potential spatial spillover | Cluster SEs at a level lower than treatment assignment |
| **Bootstrapping** | Use cluster-bootstrap procedures when the number of clusters is small (e.g., <50) | Use asymptotic SEs when the number of clusters is small |
| **Explore HTE** | Explore HTE along theoretically important pretreatment covariates with flexible estimation strategies and visualize your findings (future work) | Explore HTE through rigid regression models with interactions without visual aid |

# Recommendations

| | Do's | Don'ts |
|---|---|---|
| **Design trumps analysis** | Start empirical analysis with a research design; proceed if "feedback" from past outcomes to treatment assignment is not a major concern | Start empirical analysis by blindly running regressions using existing data |
| **Discussion of designs** | Clearly specify designs and their corresponding identification assumptions | Equate designs with outcome models |
| **Plot raw data** | Plot raw data to better understand the research setting, missingness, sources of variations in the treatment and outcome variables, and univariate/bivariate distributions | Run regressions without looking at the data |
| **Estimators** | Choose HTE-robust estimators and always plot the estimated dynamic treatment effects | Choose models solely based on your beliefs; report regression coefficients only; no results visualization or diagnostics |
| **Diagnostics** | Conduct both visual and statistical tests to gauge the validity the identification and modeling assumptions | |
| **Level of clustering** | Cluster SEs at the level of treatment assignment or higher to account for potential spatial spillover | Cluster SEs at a level lower than treatment assignment |
| **Bootstrapping** | Use cluster-bootstrap procedures when the number of clusters is small (e.g., <50) | Use asymptotic SEs when the number of clusters is small |
| **Explore HTE** | Explore HTE along theoretically important pretreatment covariates with flexible estimation strategies and visualize your findings (future work) | Explore HTE through rigid regression models with interactions without visual aid |

# Recommendations

- Come up with a plausible research design ... estimators $\neq$ designs; “shocking” element; justify $\Delta_{s,t} Y_{i,t}(0) \perp\!\!\!\perp D_{i,t}, \forall s, t$

| | Do's | Don'ts |
|---|---|---|
| Design ... | ... outcomes to treatment assignment is not a major concern | ... data |
| Discussion of designs | Clearly specify designs and their corresponding identification assumptions | Equate designs with outcome models |
| Plot raw data | Plot raw data to better understand the research setting, missingness, sources of variations in the treatment and outcome variables, and univariate/bivariate distributions | Run regressions without looking at the data |
| Estimators | Choose HTE-robust estimators and always plot the estimated dynamic treatment effects | Choose models solely based on your beliefs; report regression coefficients only; no results visualization or diagnostics |
| Diagnostics | Conduct both visual and statistical tests to gauge the validity the identification and modeling assumptions | |
| Level of clustering | Cluster SEs at the level of treatment assignment or higher to account for potential spatial spillover | Cluster SEs at a level lower than treatment assignment |
| Bootstrapping | Use cluster-bootstrap procedures when the number of clusters is small (e.g., <50) | Use asymptotic SEs when the number of clusters is small |
| Explore HTE | Explore HTE along theoretically important pretreatment covariates with flexible estimation strategies and visualize your findings (future work) | Explore HTE through rigid regression models with interactions without visual aid |

# Recommendations

- Come up with a plausible research design ... estimators $\neq$ designs; "shocking" element; justify $\Delta_{s,t} Y_{i,t}(0) \perp\!\!\!\perp D_{i,t}, \forall s, t$
- Understand your data better ...before typing "reghdfe" in Stata

| | Do's | Don'ts |
|---|---|---|
| Design [illegible] analysis | [illegible] outcomes to treatment assignment is not a major concern | [illegible] data |
| Discussion of designs | Clearly specify designs and their corresponding identification assumptions | Equate designs with outcome models |
| Plot raw data | Plot raw data to better understand the research setting, missingness, sources of variations in the treatment and outcome variables, and univariate/bivariate distributions | Run regressions without looking at the data |
| Estimators | Choose HTE-robust estimators and always plot the estimated dynamic treatment effects | Choose models solely based on your beliefs; report regression coefficients only; no results visualization or diagnostics |
| Diagnostics | Conduct both visual and statistical tests to gauge the validity the identification and modeling assumptions | |
| Level of clustering | Cluster SEs at the level of treatment assignment or higher to account for potential spatial spillover | Cluster SEs at a level lower than treatment assignment |
| Bootstrapping | Use cluster-bootstrap procedures when the number of clusters is small (e.g., <50) | Use asymptotic SEs when the number of clusters is small |
| Explore HTE | Explore HTE along theoretically important pretreatment covariates with flexible estimation strategies and visualize your findings (future work) | Explore HTE through rigid regression models with interactions without visual aid |

# Recommendations

- Come up with a plausible research design ... estimators $\neq$ designs; “shocking” element; justify $\Delta_{s,t}Y_{i,t}(0) \perp\!\!\!\perp D_{i,t}, \forall s, t$
- Understand your data better ...before typing “reghdfe” in Stata
- Trimming your data (to “compare like with like”) is not forbidden ...as long as Y is not being used

# Recommendations

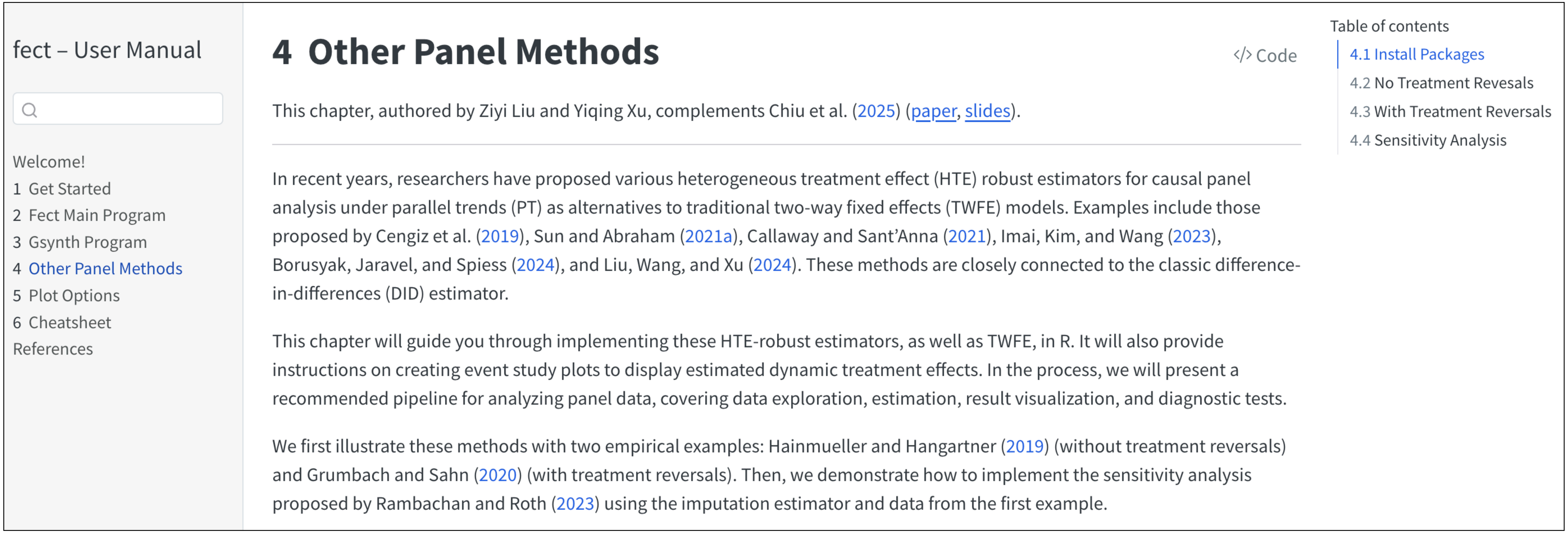
- Come up with a plausible research design ... estimators $\neq$ designs; "shocking" element; justify $\Delta_{s,t} Y_{i,t}(0) \perp\!\!\!\perp D_{i,t}, \forall s, t$
- Understand your data better ...before typing "reghdfe" in Stata
- Trimming your data (to "compare like with like") is not forbidden ...as long as Y is not being used
- Using HTE-robust estimators is safer ...and the choice of estimators shouldn't matter much

# Recommendations

- Come up with a plausible research design ... estimators ≠ designs; "shocking" element; justify $\Delta_{s,t}Y_{i,t}(0) \perp\!\!\!\perp D_{i,t}, \forall s, t$
- Understand your data better ...before typing "reghdfe" in Stata
- Trimming your data (to "compare like with like") is not forbidden ...as long as Y is not being used
- Using HTE-robust estimators is safer ...and the choice of estimators shouldn't matter much
- Validate your assumptions ...knowing that power is a major concern

# Tools

- panelView (R & Stata), fect (R & Stata)
- Tutorial: https://yiqingxu.org/packages/fect/05-panel.html



fect – User Manual

Welcome!
1 Get Started
2 Fect Main Program
3 Gsynth Program
4 Other Panel Methods
5 Plot Options
6 Cheatsheet
References

# 4 Other Panel Methods

</> Code

This chapter, authored by Ziyi Liu and Yiqing Xu, complements Chiu et al. (2025) (paper, slides).

In recent years, researchers have proposed various heterogeneous treatment effect (HTE) robust estimators for causal panel analysis under parallel trends (PT) as alternatives to traditional two-way fixed effects (TWFE) models. Examples include those proposed by Cengiz et al. (2019), Sun and Abraham (2021a), Callaway and Sant'Anna (2021), Imai, Kim, and Wang (2023), Borusyak, Jaravel, and Spiess (2024), and Liu, Wang, and Xu (2024). These methods are closely connected to the classic difference-in-differences (DID) estimator.

This chapter will guide you through implementing these HTE-robust estimators, as well as TWFE, in R. It will also provide instructions on creating event study plots to display estimated dynamic treatment effects. In the process, we will present a recommended pipeline for analyzing panel data, covering data exploration, estimation, result visualization, and diagnostic tests.

We first illustrate these methods with two empirical examples: Hainmueller and Hangartner (2019) (without treatment reversals) and Grumbach and Sahn (2020) (with treatment reversals). Then, we demonstrate how to implement the sensitivity analysis proposed by Rambachan and Roth (2023) using the imputation estimator and data from the first example.

Table of contents

Thank you!