Difference-in-Differences using Mixed-Integer Programming Matching Approach

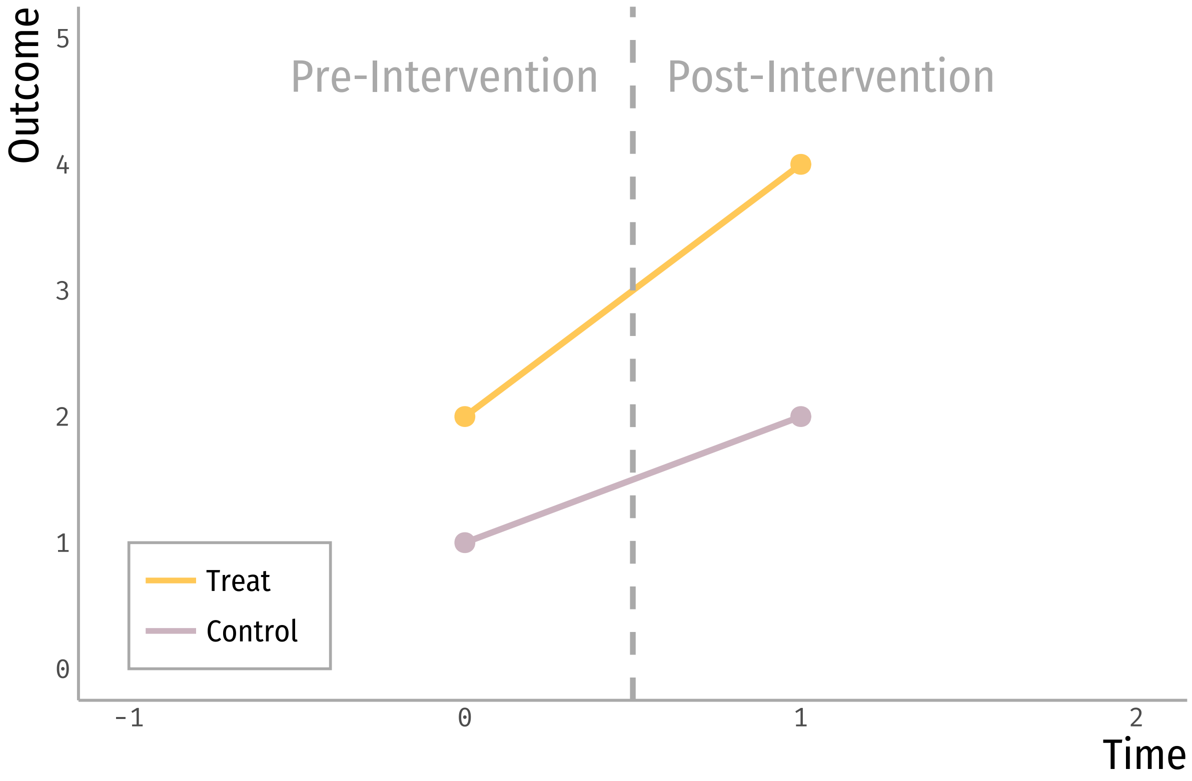
Magdalena Bennett

McCombs School of Business, The University of Texas at Austin

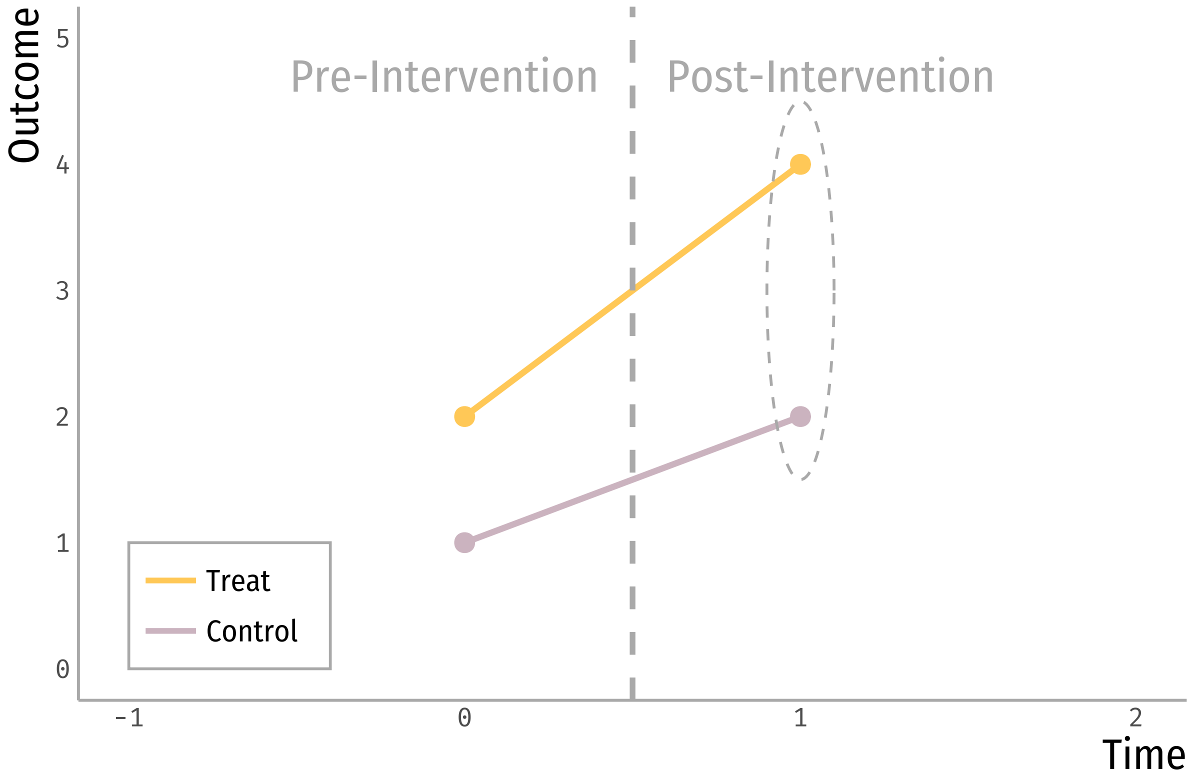
Seminario DIIS - PUC

November 20th, 2024

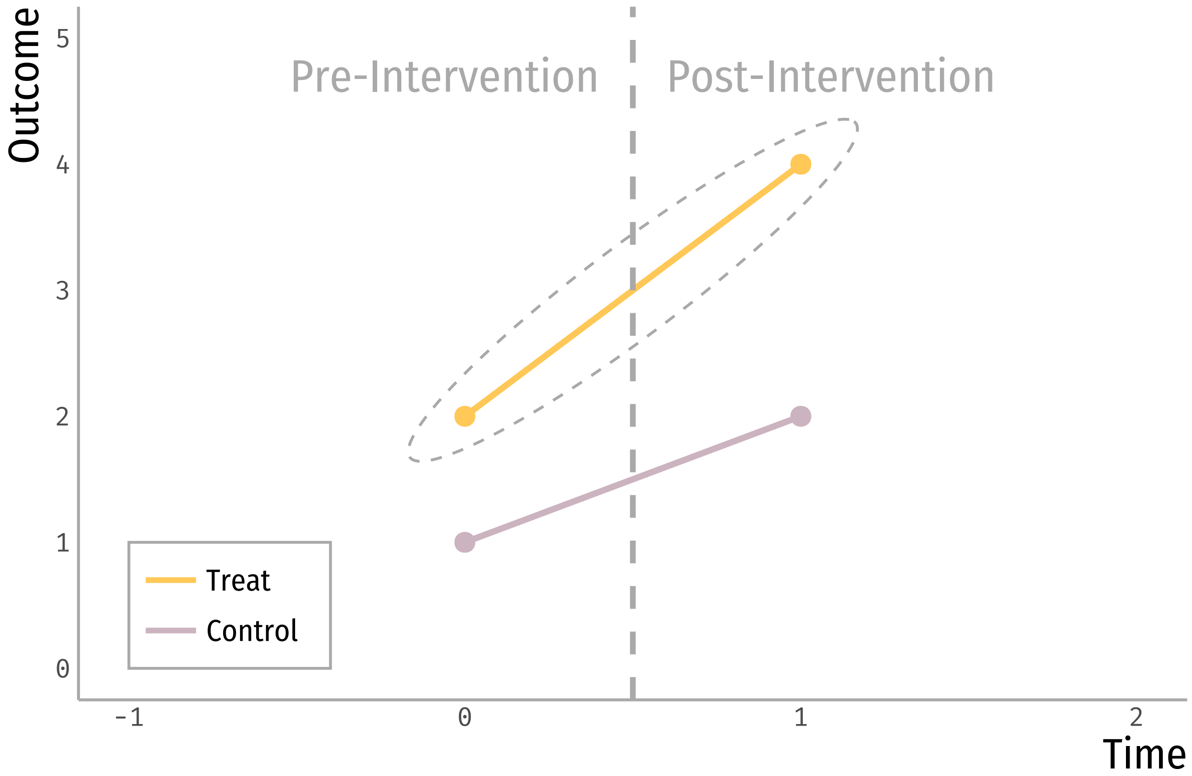
Diff-in-Diff as an identification strategy



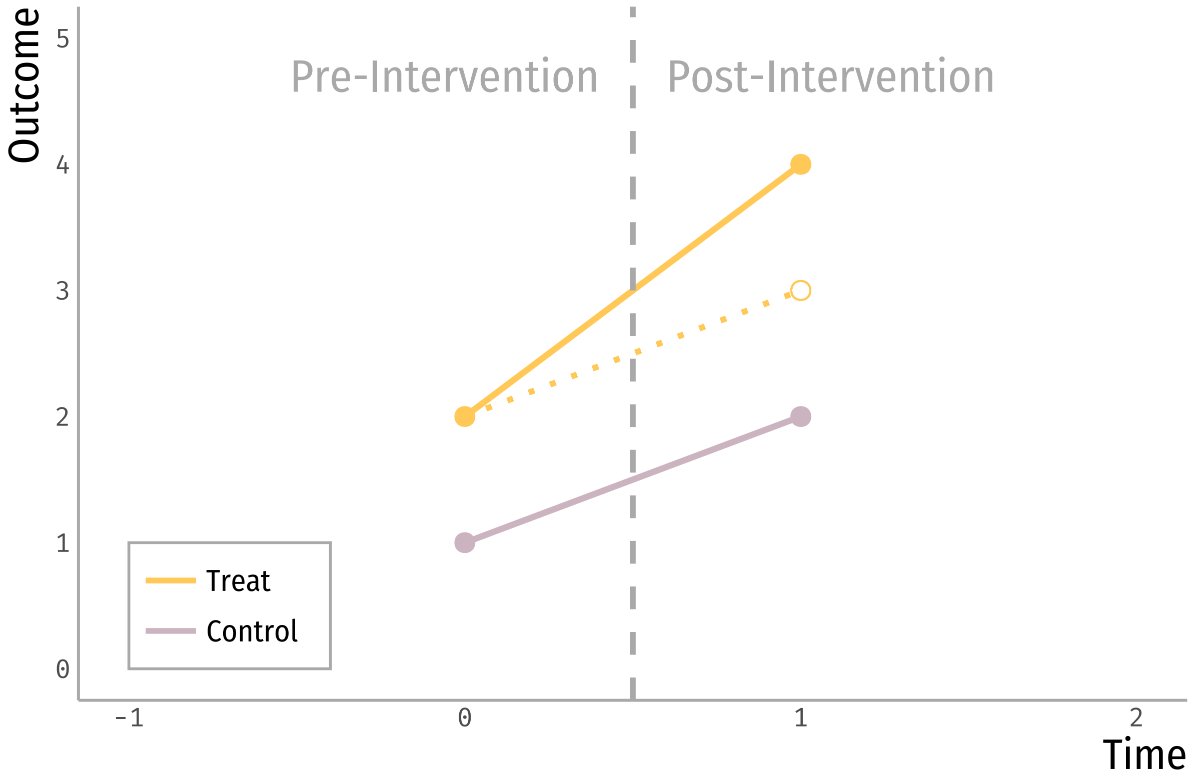
Cannot compare treated vs control



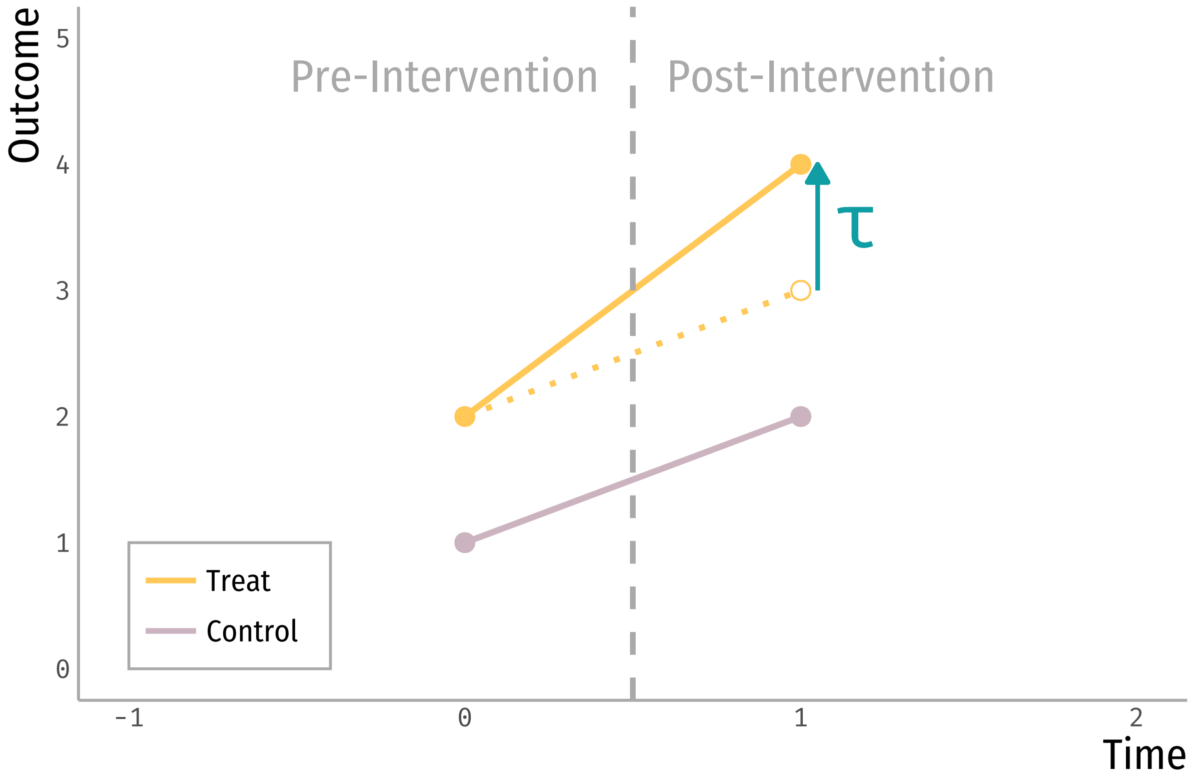
Cannot compare before and after



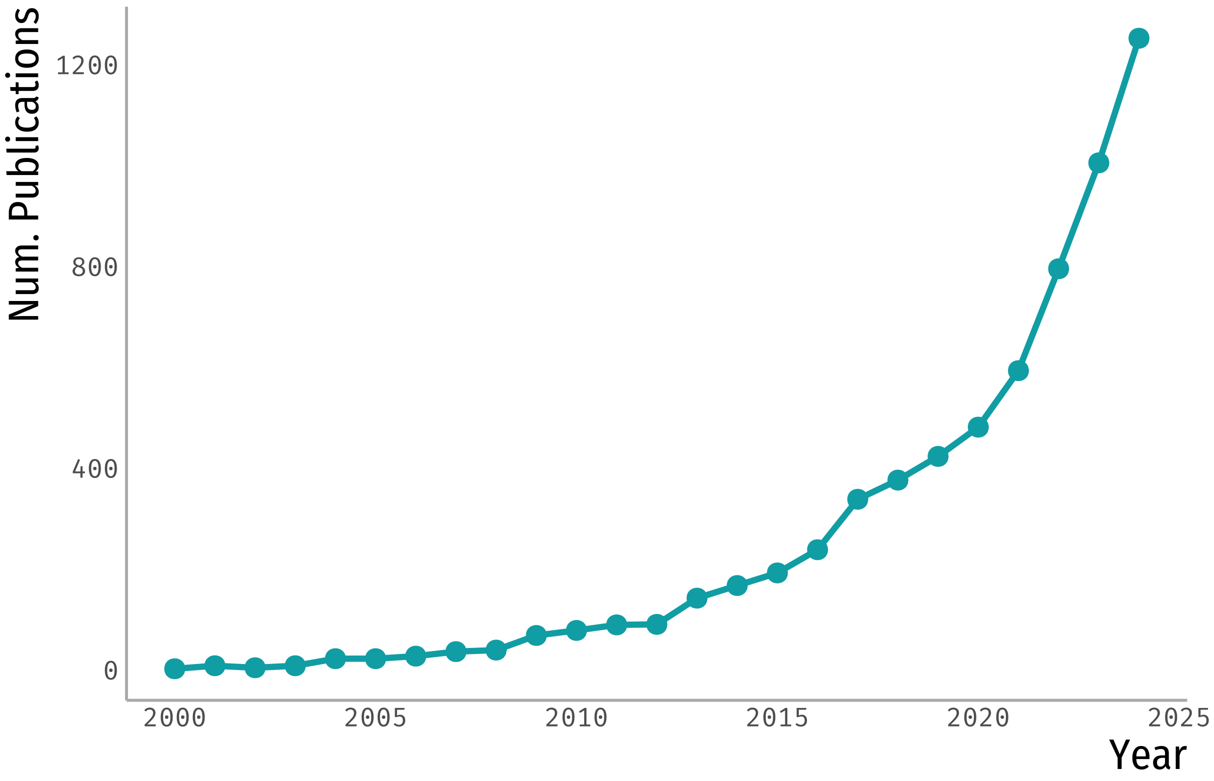
Parallel trend assumption (PTA)



Estimate Average Treatment Effect on the Treated (ATT)

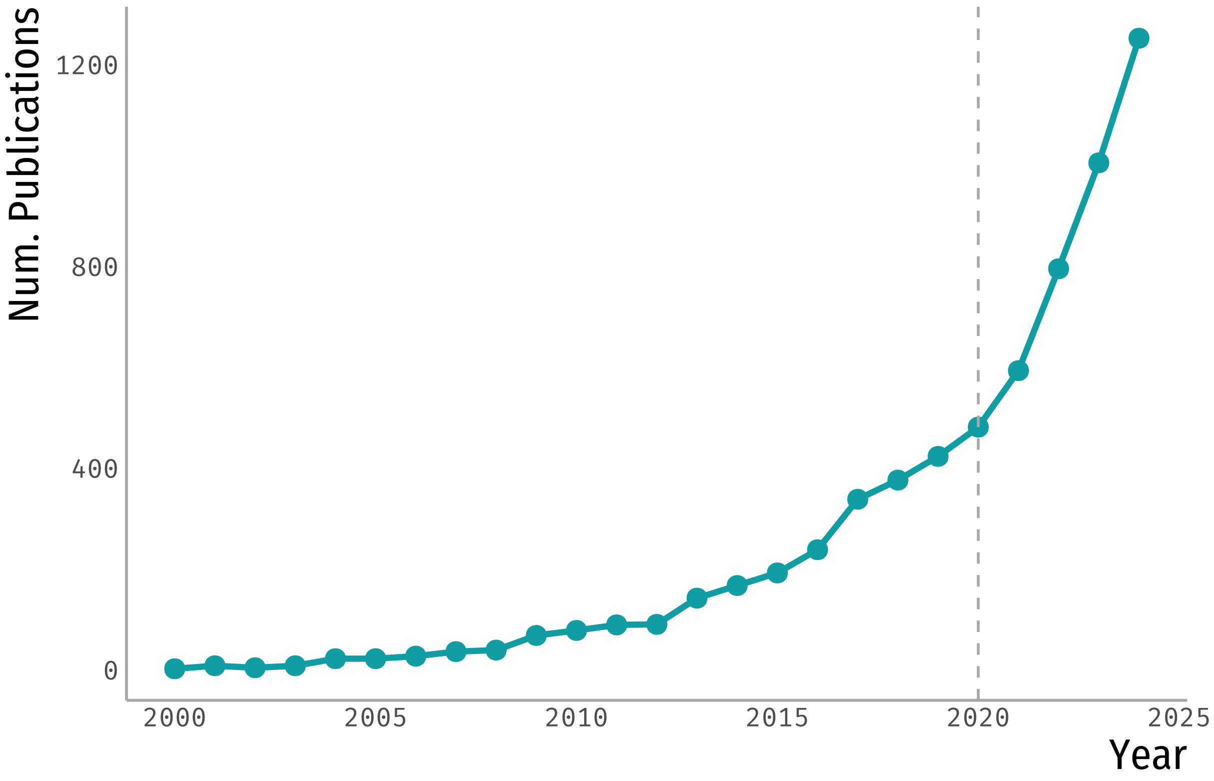


Diff-in-Diff is very popular in Economics



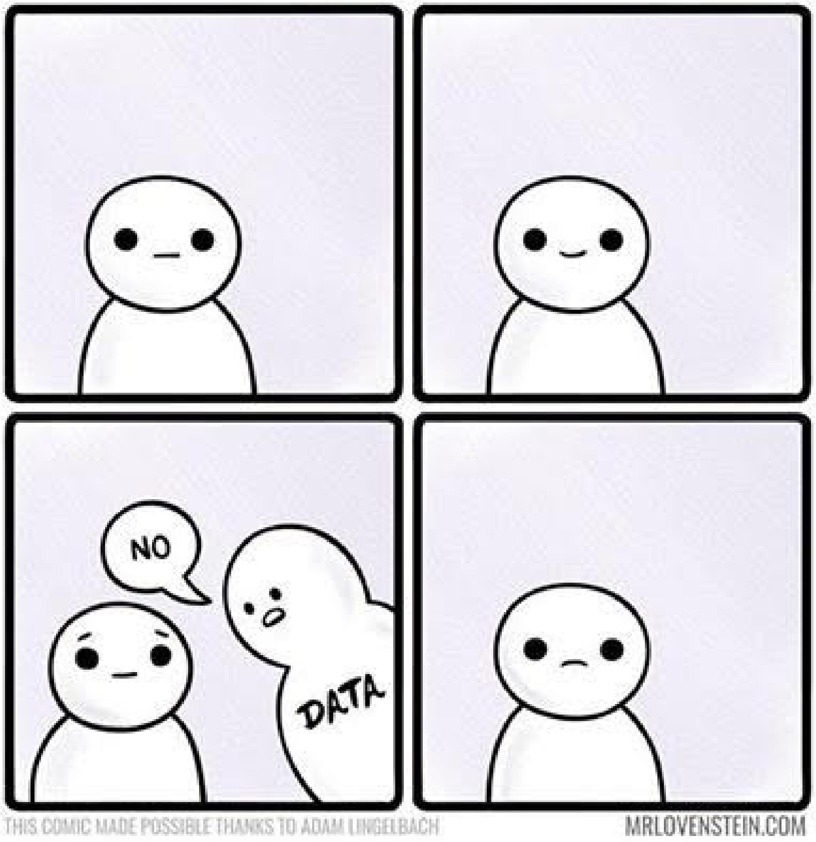
Source: Web of Science (11/18/2024)

Diff-in-Diff is very popular in Economics

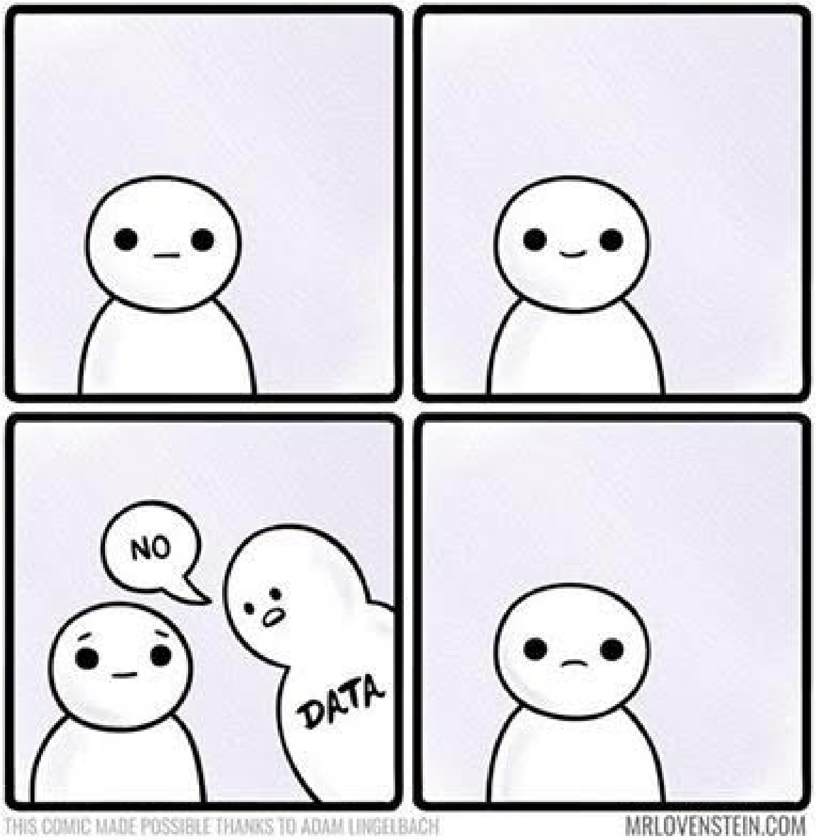


Source: Web of Science (11/18/2024)

What about parallel trends?

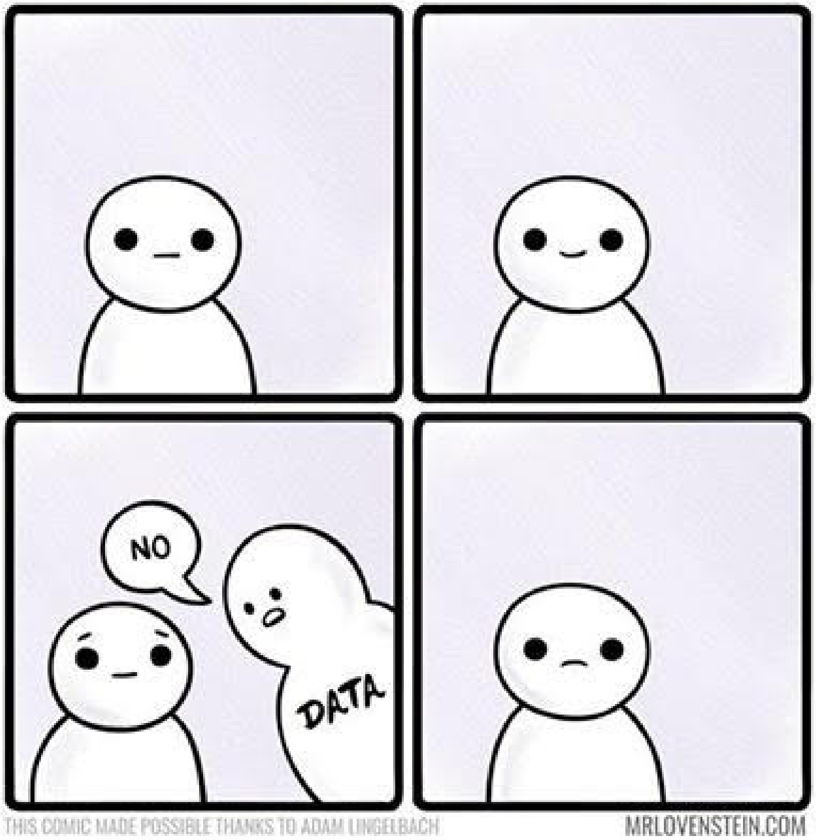


What about parallel trends?



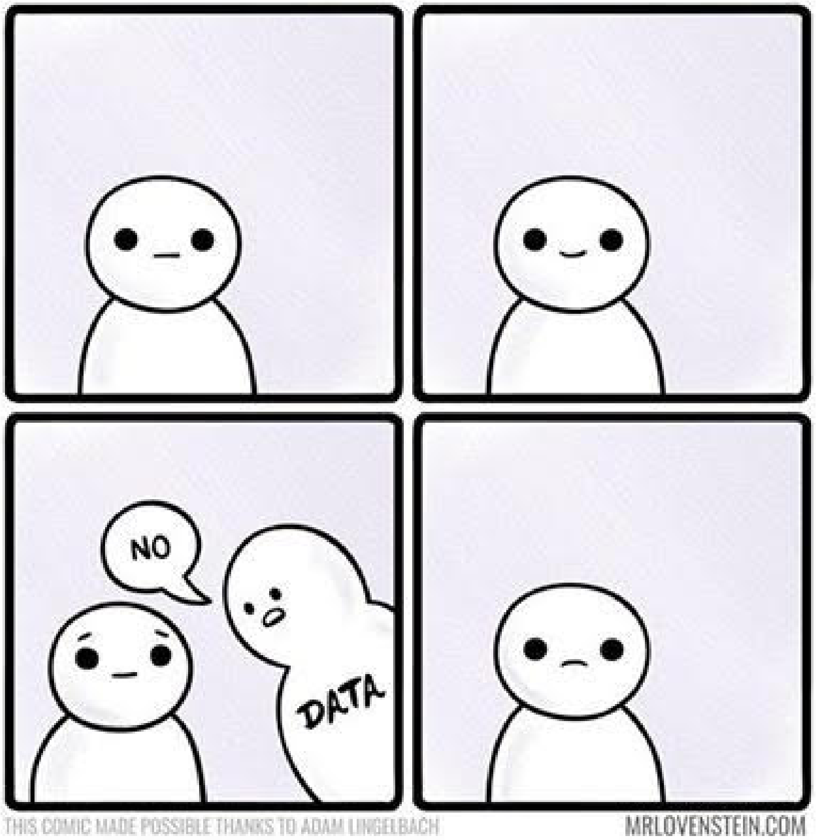
- Main identification assumption **fails**

What about parallel trends?



- Main identification assumption fails
- Find sub-groups that potentially follow PTA
  - E.g. similar units in treatment and control
  - Similar to synthetic control intuition.

What about parallel trends?



- Main identification assumption fails
- Find sub-groups that potentially follow PTA
  - E.g. similar units in treatment and control
  - Similar to synthetic control intuition.
- Can matching help?
  - It's complicated (Ham & Miratrix, 2022; Zeldow & Hatfield, 2021; Basu & Small, 2020; Lindner & McConnell, 2018; Daw & Hatfield, 2018 (x2); Ryan, 2018; Ryan et al., 2018)

This paper

- Identify contexts when matching can recover causal estimates under certain violations of the parallel trend assumption.
  - Overall bias reduction and increase in robustness for sensitivity analysis.
- Use mixed-integer programming matching (MIP) to balance covariates directly.

Simulations:
Different DGP scenarios

Application:
School segregation & vouchers

Let's get started

DD Setup

- Let $Y_{it}(z)$ be the potential outcome for unit i in period t under treatment z.
- Intervention implemented in $T_0 \rightarrow$ No units are treated in $t \leq T_0$
- Difference-in-Differences (DD) focuses on ATT for $t > T_0$:

$$ATT(t) = E[Y_{it}(1) - Y_{it}(0)|Z = 1]$$

Expected difference in potential outcomes
if the treatment hadn't happened
for the treatment group

DD Setup

- Let $Y_{it}(z)$ be the potential outcome for unit i in period t under treatment z.
- Intervention implemented in $T_0 \to$ No units are treated in $t \leq T_0$
- Difference-in-Differences (DD) focuses on ATT for $t > T_0$:

$$ATT(t) = E[Y_{it}(1) - Y_{it}(0)|Z = 1]$$

- Assumptions for DD:
  - Parallel-trend assumption (PTA)
  - Common shocks

$$E[Y_{i1}(0) - Y_{i0}(0)|Z = 1] = E[Y_{i1}(0) - Y_{i0}(0)|Z = 0]$$

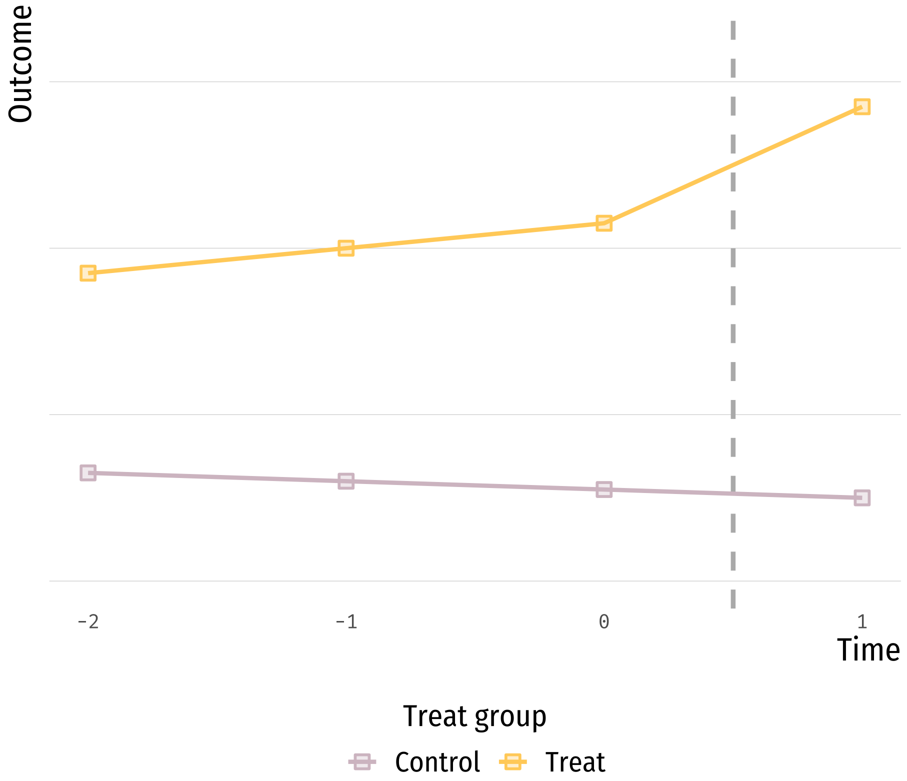
DD Setup (cont.)

- Under these assumptions:

$$\hat{\tau}^{DD} = \overbrace{E[Y_{i1}|Z=1] - E[Y_{i1}|Z=0]}^{\Delta_{post}} - \underbrace{(E[Y_{i0}|Z=1] - E[Y_{i0}|Z=0])}_{\Delta_{pre}}$$

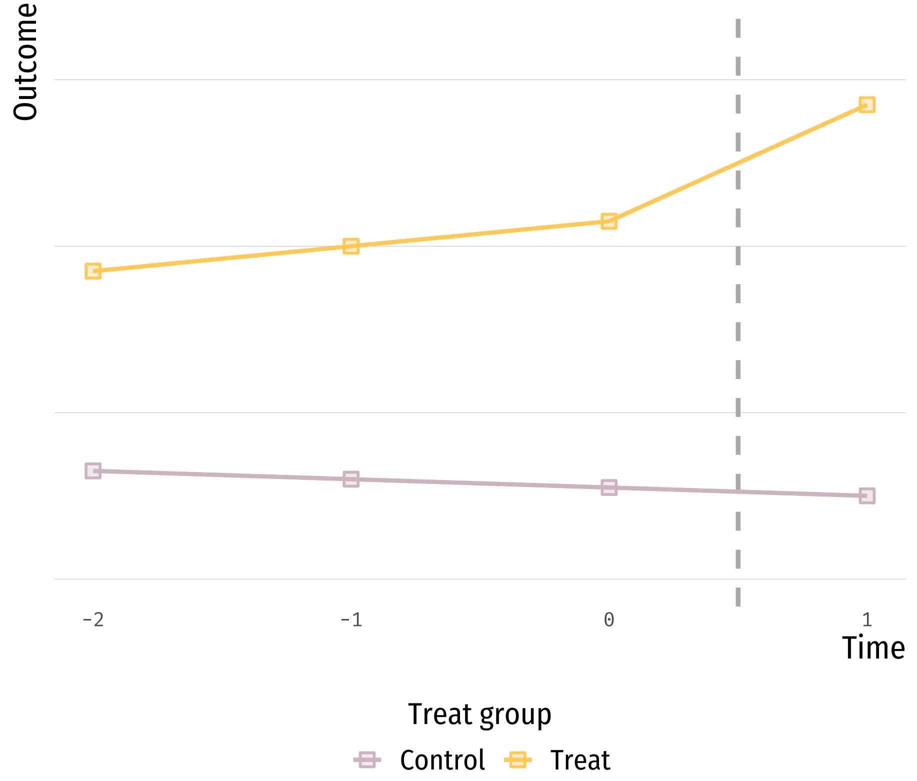
  - Where $t = 0$ and $t = 1$ are the pre- and post-intervention periods, respectively.
  - $Y_{it} = Y_{it}(1) \cdot Z_i + (1 - Z_i) \cdot Y_{it}(0)$ is the observed outcome.

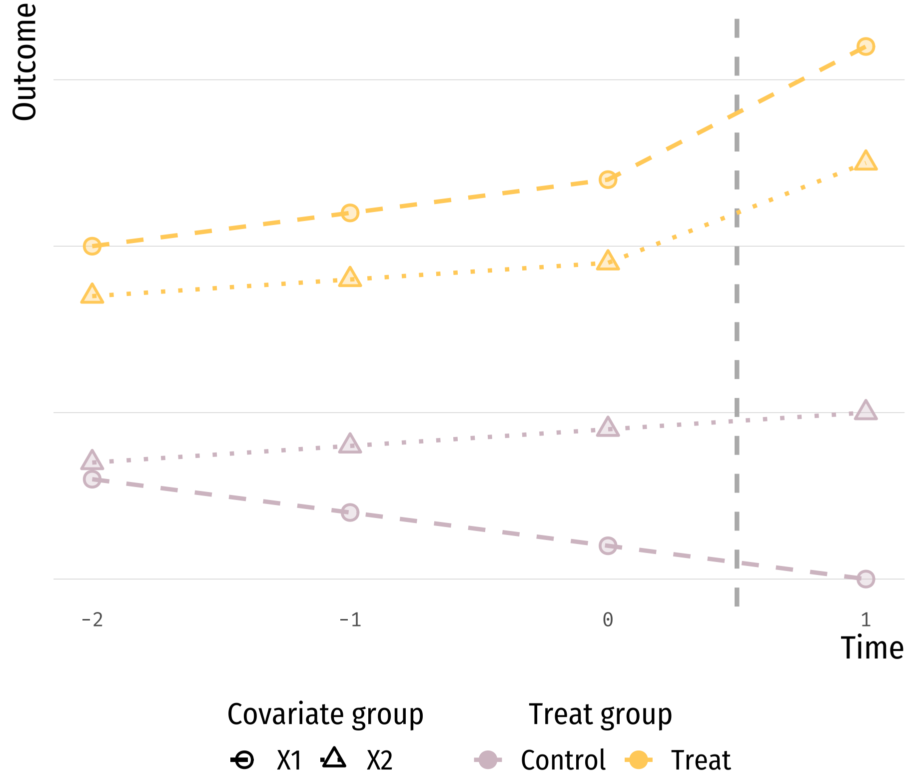
But what if the PTA doesn't hold?



But what if the PTA doesn't hold?

We can potentially remove [part of] the bias by matching on $X^s{}_{it}=X_i$





General form of potential outcomes

We can write a general form of the potential outcomes $Y(0)$ and $Y(1)$ as follows:

$$Y_{it}(0) = \alpha_i + \lambda_t + \gamma_0(X_i) + \gamma_1(X_i, t) + \gamma_2(X_i, t) \cdot Z_i + u_{it}$$

$$Y_{it}(1) = Y_{it}(0) + \tau_{it} = \alpha_i + \lambda_t + \gamma_0(X_i) + \gamma_1(X_i, t) + \gamma_2(X_i, t) \cdot Z_i + \tau_{it} + u_{it}$$

Covariate distribution between groups can be **different**

$$X_i | Z \sim F_x(z)$$

General form of potential outcomes

We can write a general form of the potential outcomes $Y(0)$ and $Y(1)$ as follows:

$$Y_{it}(0) = \alpha_{\mathbf{i}} + \lambda_{\mathbf{t}} + \gamma_0(X_i) + \gamma_1(X_i, t) + \gamma_2(X_i, t) \cdot Z_i + u_{it}$$

- α_i and λ_t are individual and time FE, respectively.
  - If $\lambda_t | Z \sim F_\lambda(z, t)$, then PTA fails.

General form of potential outcomes

We can write a general form of the potential outcomes $Y(0)$ and $Y(1)$ as follows:

$$Y_{it}(0) = \alpha_i + \lambda_t + \boldsymbol{\gamma_0(X_i)} + \gamma_1(X_i, t) + \gamma_2(X_i, t) \cdot Z_i + u_{it}$$

- α_i and λ_t are individual and time FE, respectively.
    - If $\lambda_t | Z \sim F_\lambda(z, t)$, then PTA fails.

- $\gamma_0(X_i)$ is a time-invariant function that associates X and Y

General form of potential outcomes

We can write a general form of the potential outcomes $Y(0)$ and $Y(1)$ as follows:

$$Y_{it}(0) = \alpha_i + \lambda_t + \gamma_0(X_i) + \gamma_1\mathbf{(X_i, t)} + \gamma_2(X_i, t) \cdot Z_i + u_{it}$$

- α_i and λ_t are individual and time FE, respectively.
    - If $\lambda_t | Z \sim F_\lambda(z, t)$, then PTA fails.
- $\gamma_0(X_i)$ is a time-invariant function that associates X and Y.
- $\gamma_1(X_i, t)$ is a time-dependent function.

General form of potential outcomes

We can write a general form of the potential outcomes $Y(0)$ and $Y(1)$ as follows:

$$Y_{it}(0) = \alpha_i + \lambda_t + \gamma_0(X_i) + \gamma_1(X_i, t) + \boldsymbol{\gamma_2(X_i, t) \cdot Z_i} + u_{it}$$

- α_i and λ_t are individual and time FE, respectively.
  - If $\lambda_t | Z \sim F_\lambda(z, t)$, then PTA fails.
- $\gamma_0(X_i)$ is a time-invariant function that associates X and Y.
- $\gamma_1(X_i, t)$ is a time-dependent function.

- $\gamma_2(X_i, t) \cdot Z_i$ is a *differential* time-dependent function **only** for the treatment group.

If the PTA holds...

Then, for a 2x2 DD, where $t < T_0$ (pre) and $t' > T_0$ (post):

$$\mathbb{E}[\gamma_1(X_i, t') + \gamma_2(X_i, t') - \gamma_1(X_i, t) - \gamma_2(X_i, t)|Z = 1] = \mathbb{E}[\gamma_1(X_i, t') - \gamma_1(X_i, t)|Z = 0]$$

One of the two conditions need to hold:

1) No effect or constant effect of X on Y over time: $\mathbb{E}[\gamma_1(X, t)] = \mathbb{E}[\gamma_1(X)]$

2) Equal distribution of observed covariates between groups: $X_i|Z = 1 \stackrel{d}{=} X_i|Z = 0$

in addition to:

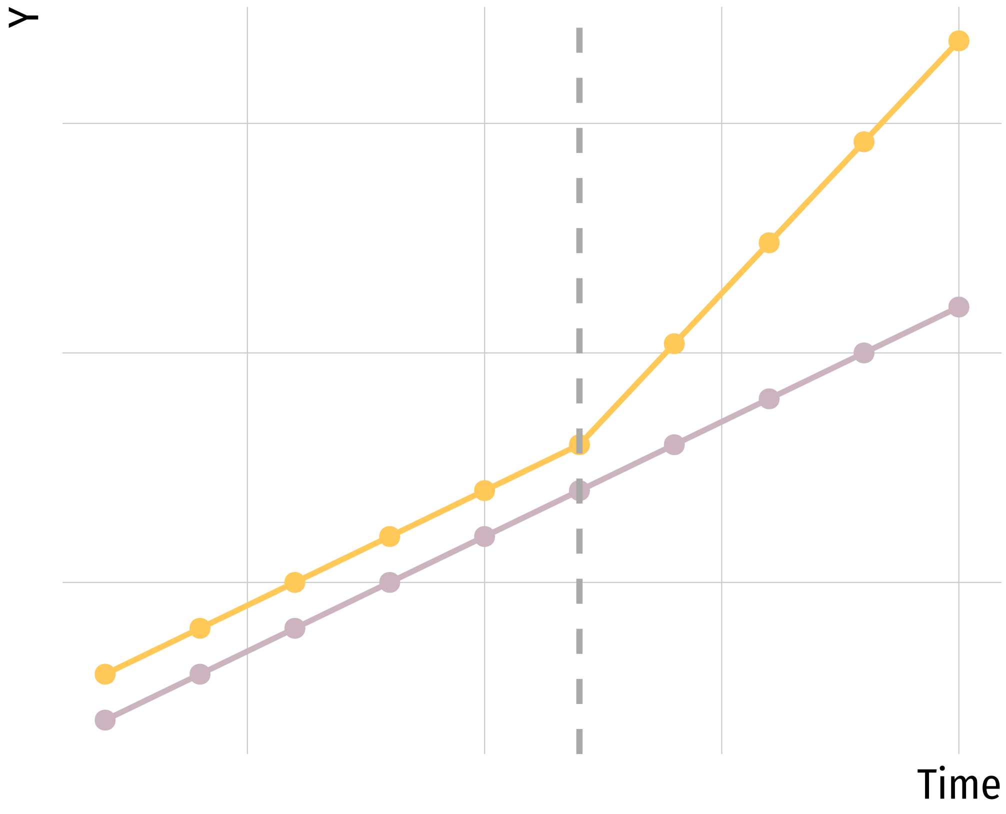
3) No differential time effect of X on Y by treatment group: $\mathbb{E}[\gamma_2(X, t)] = 0$

Cond. 2 can hold through **matching**

Cond. 3 can be tested with **sensitivity analysis**

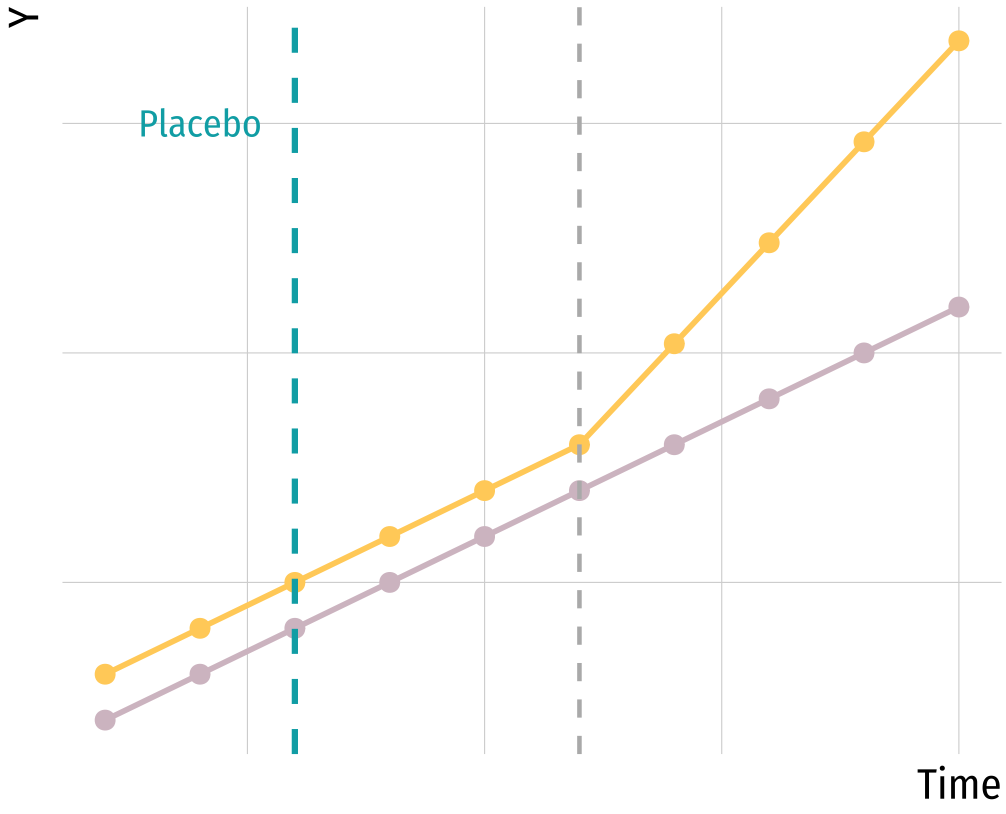
Sensitivity analysis for Diff-in-Diff

- Use of pre-trends to test plausibility of the PTA:



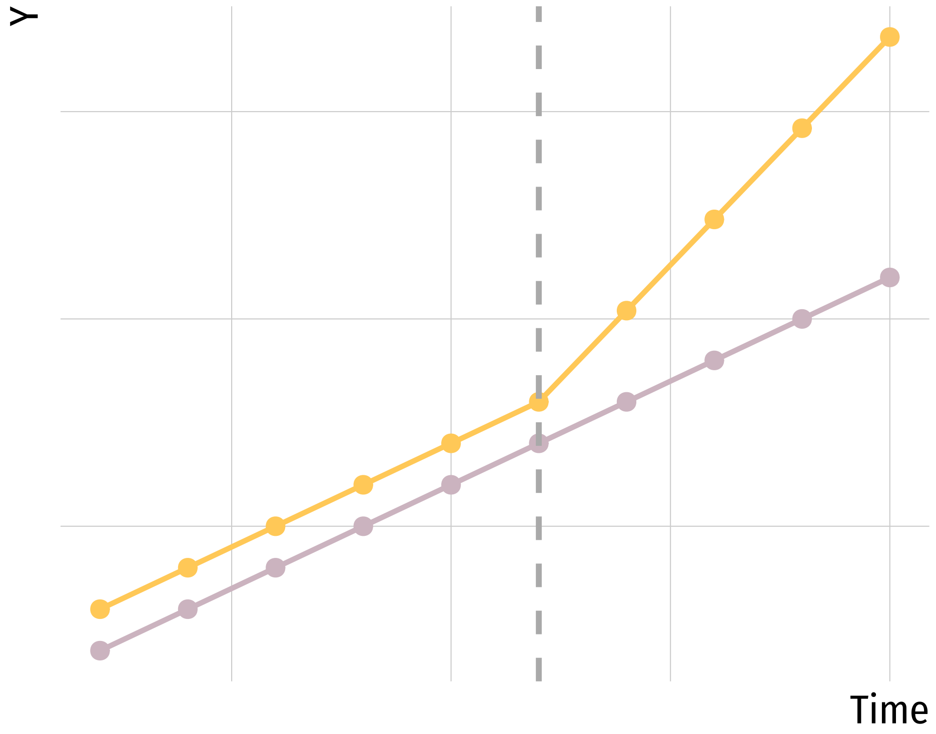
Sensitivity analysis for Diff-in-Diff

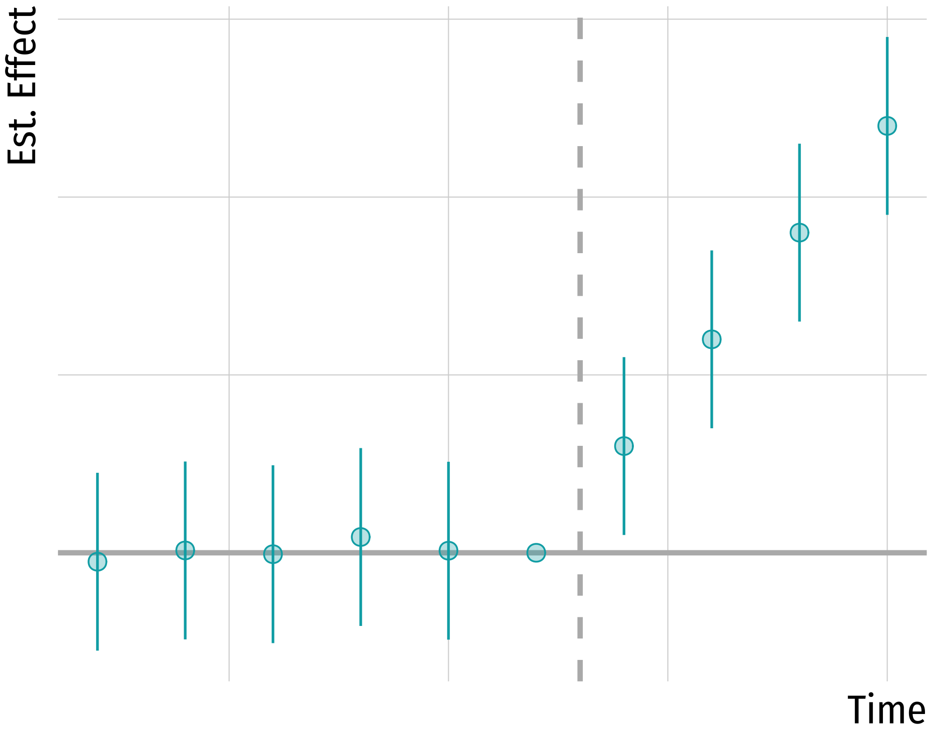
- Using a diff-in-diff strategy, we shouldn't find an effect



Sensitivity analysis for Diff-in-Diff

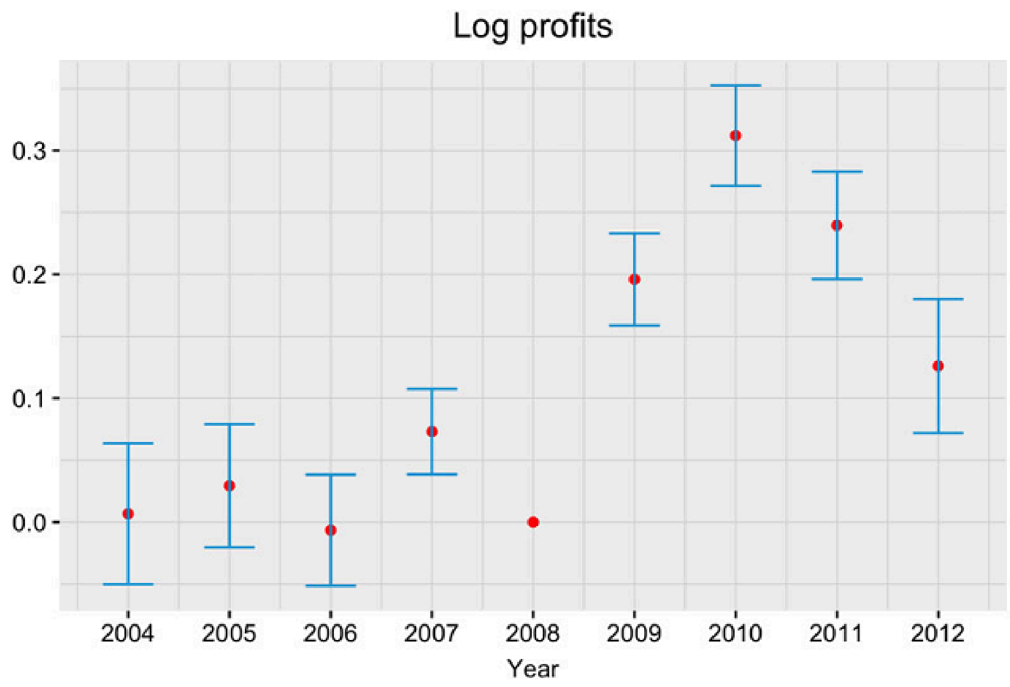
- In an event study → null effects prior to the intervention:

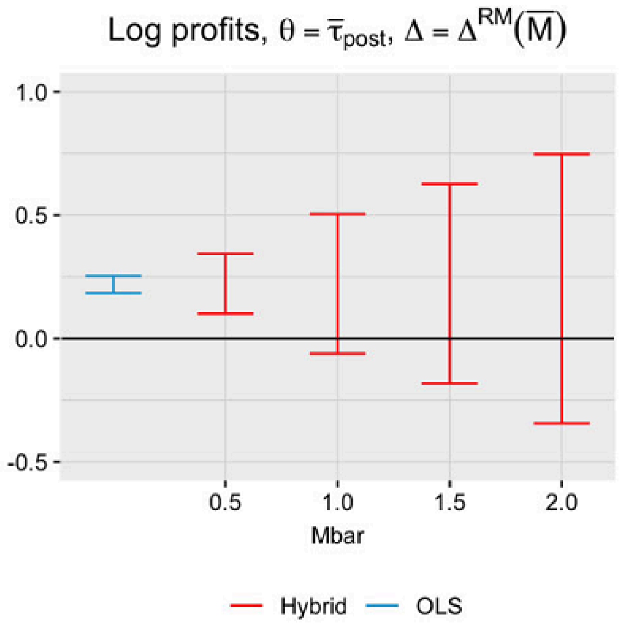




Honest approach to test pretrends

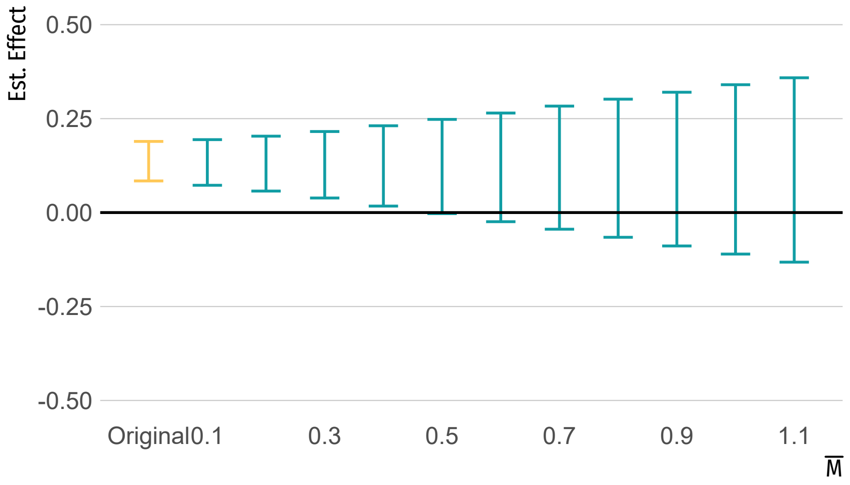
- One main issue with the previous test $\rightarrow$ **Underpowered**
- Rambachan & Roth (2023) propose **sensitivity bounds** to allow pre-trends violations:
  - E.g. Violations in the post-intervention period can be *at most* M times the max violation in the pre-intervention period.



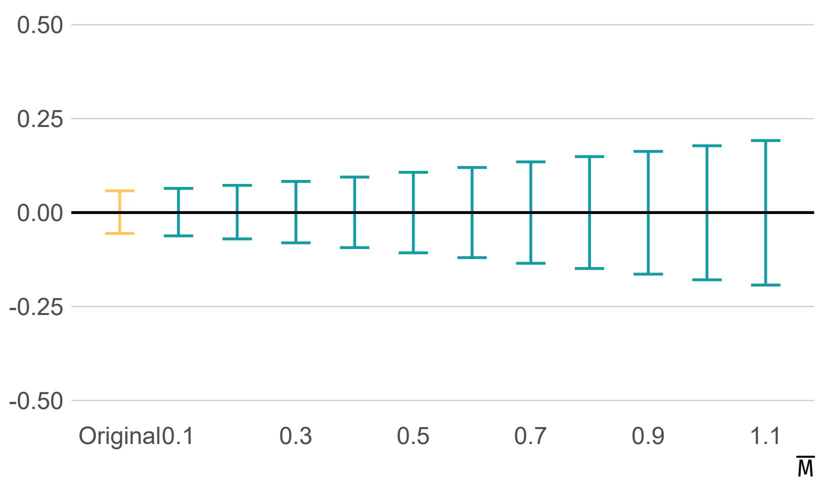


Honest approach to test pretrends

- One drawback of the previous method is that it can **overstate** (or understate) the robustness of findings if the point estimate is biased.
  - Honest CIs depend on the **magnitude of the point estimate** as well as the **pre-trend violations**.
- Matching can **reduce the overall bias** of the point estimate



(a) Biased estimate



(b) Unbiased estimate

How do we match?

- Match on covariates or outcomes? Levels or trends?
- Propensity score matching? Optimal matching? etc.

This paper:

- Match on time-invariant covariates that could make groups behave differently.
  - Use distribution of covariates to match on a template.
- Use of Mixed-Integer Programming (MIP) Matching (Zubizarreta, 2015; Bennett, Zubizarreta, & Vielma, 2020):
  - Balance covariates directly
  - Yield largest matched sample under balancing constraints (cardinality matching)
  - Works fast with large samples

Simulations

Different scenarios

For linear and quadratic functions:

S1: No interaction between X and t

S2: Equal interaction between X and t

S3: Differential interaction between X and t

Additional tests:

S1b-S3b: Including time-varying covariates

- For all scenarios, differential distribution of covariates X between groups

Data Generating Processes

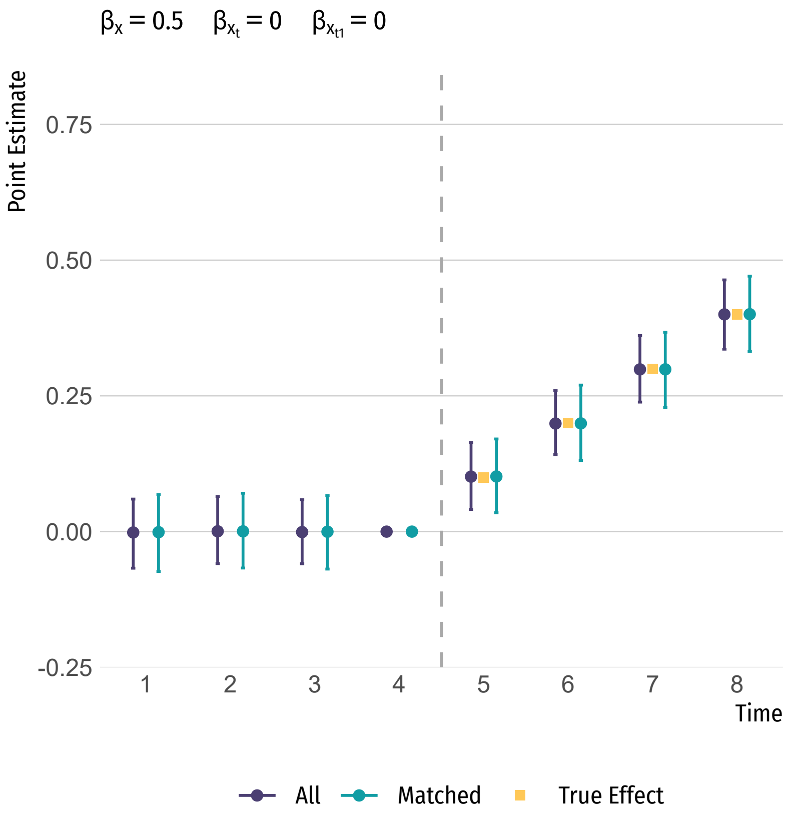
Scenarios	Functions		
Linear			
(1) No interaction between X and t	$\gamma_0(X) = \beta_x \cdot X$	$\gamma_1 = \gamma_2 = 0$	
(2) Equal interaction between X and t by treatment	$\gamma_0(X) = \beta_x \cdot X$	$\gamma_1(X,t) = \beta_{x_t} \cdot X \cdot \frac{t}{2}$	$\gamma_2(X,t) = 0$
(3) Different interaction between X and t by treatment	$\gamma_0(X) = \beta_x \cdot X$	$\gamma_1(X,t) = \beta_{x_t} \cdot X \cdot \frac{t}{2}$	$\gamma_2(X,t) = \beta_{x_{t1}} \cdot X \cdot \frac{t}{5} \cdot Z$
Quadratic			
(1) No interaction between X and t	$\gamma_0(X) = \beta_x \cdot X + \beta_x \cdot \frac{X^2}{10}$	$\gamma_1 = \gamma_2 = 0$	
(2) Equal interaction between X and t by treatment	$\gamma_0(X) = \beta_x \cdot X + \beta_x \cdot \frac{X^2}{10}$	$\gamma_1(X,t) = \beta_{x_t} \cdot X \cdot \frac{t^2}{10}$	$\gamma_2(X,t) = 0$
(3) Different interaction between X and t by treatment	$\gamma_0(X) = \beta_x \cdot X + \beta_x \cdot \frac{X^2}{10}$	$\gamma_1(X,t) = \beta_{x_t} \cdot X \cdot \frac{t^2}{10}$	$\gamma_2(X,t) = \beta_{x_{t1}} \cdot X \cdot \frac{t^2}{50} \cdot Z$

Parameters:

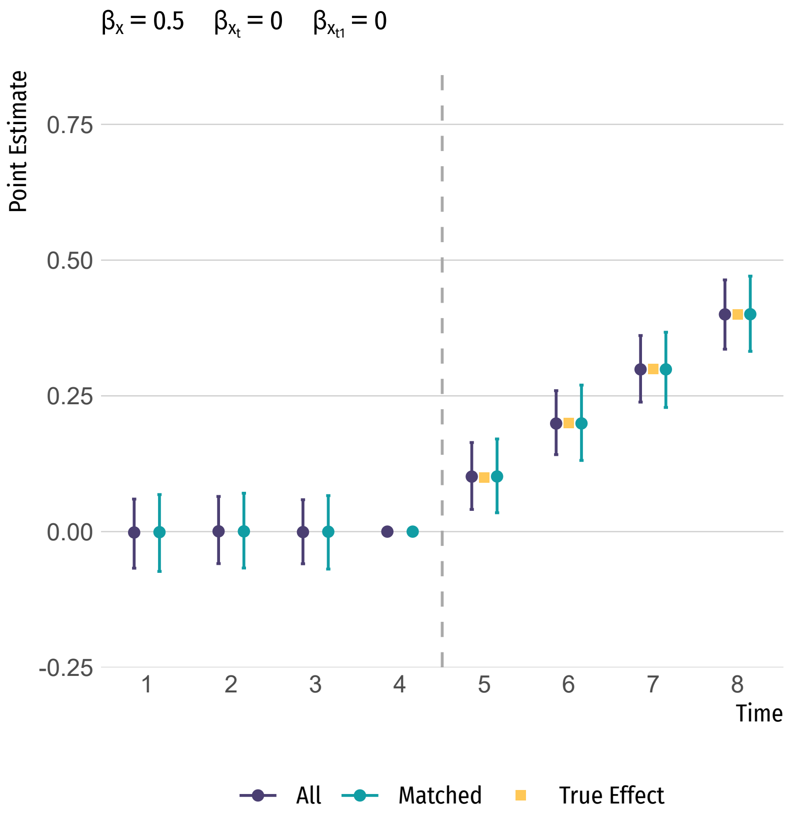
Parameter	Value
Number of obs (N)	1,000
`Pr(Z=1)`	0.5
Time periods (T)	8
Last pre-intervention period (T_0)	4
Matching PS	Nearest neighbor (using calipers)
MIP Matching tolerance	.01 SD
Number of simulations	1,000

- Estimate compared to sample ATT (*can be different for matching*)

S1 - No interaction between X and t

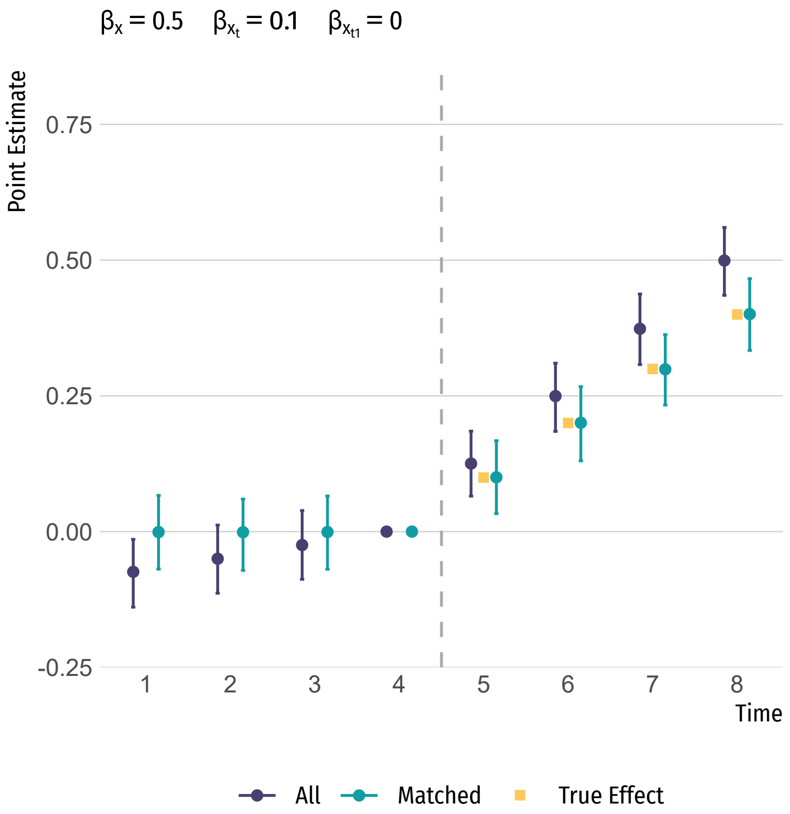


(a) Linear

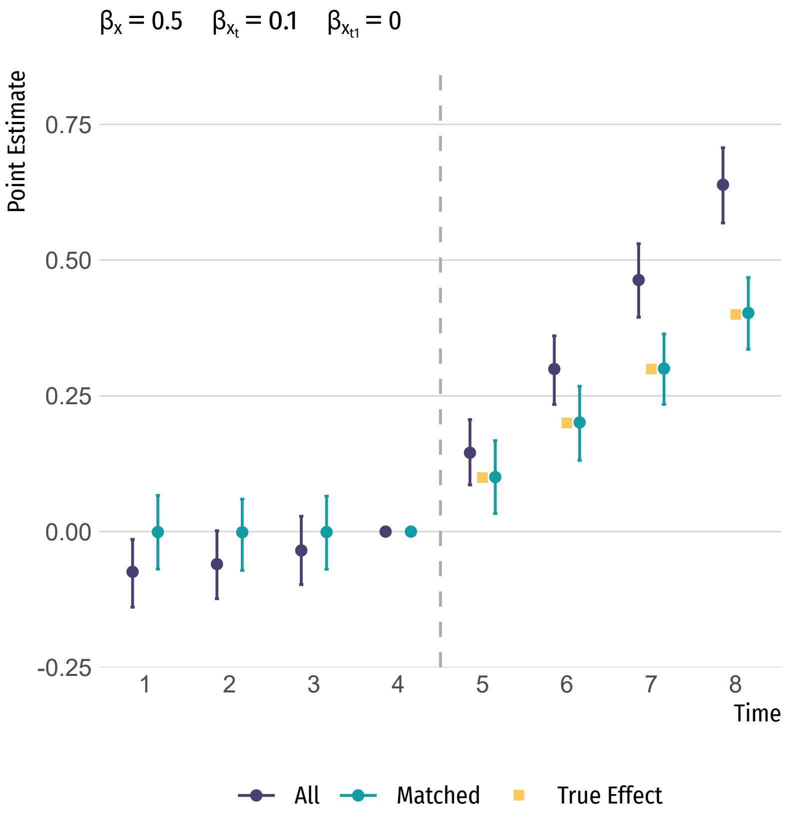


(b) Quadratic

S2 - Equal interaction between X and t by treatment

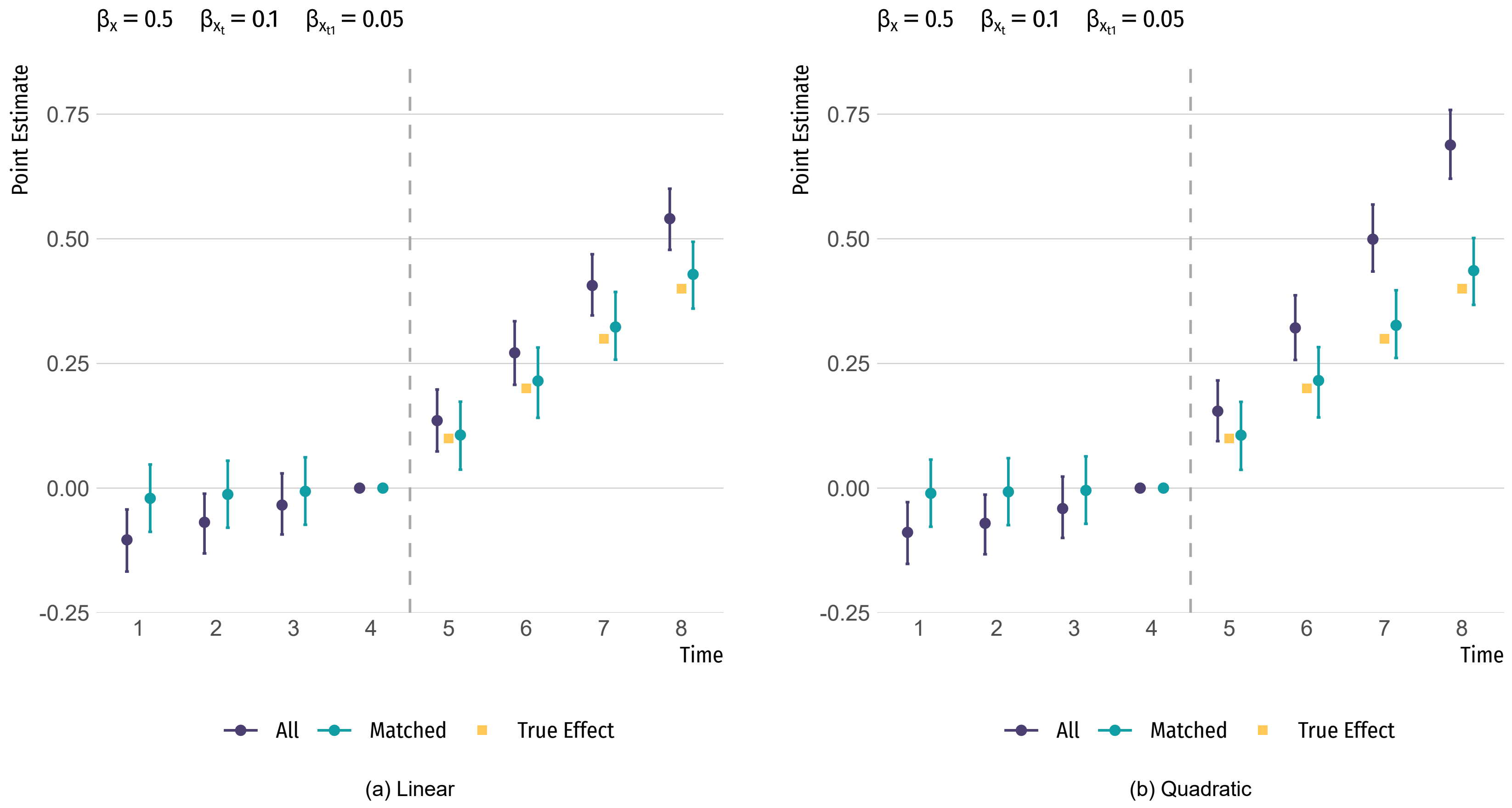


(a) Linear

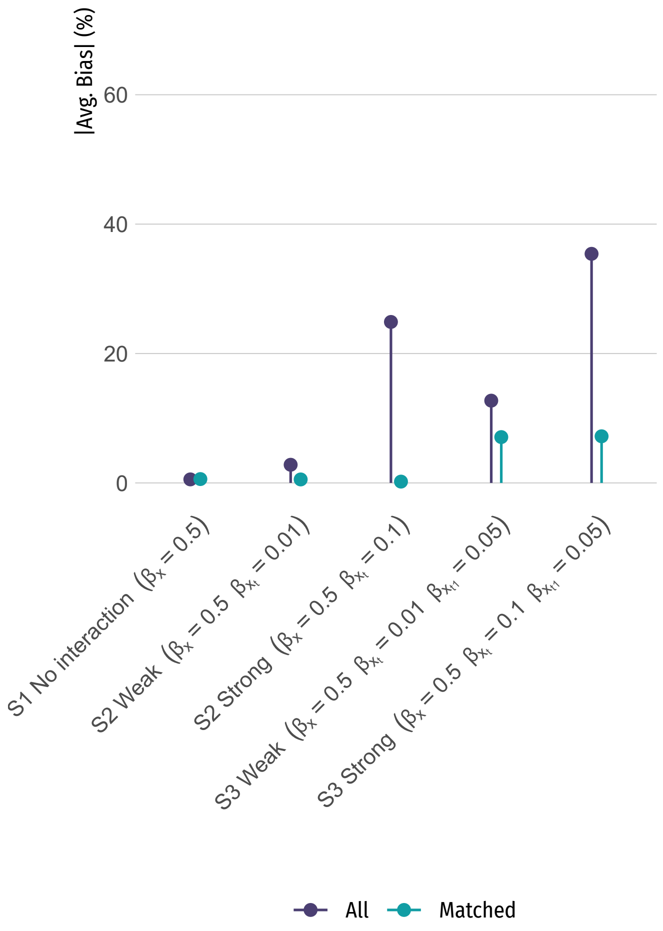


(b) Quadratic

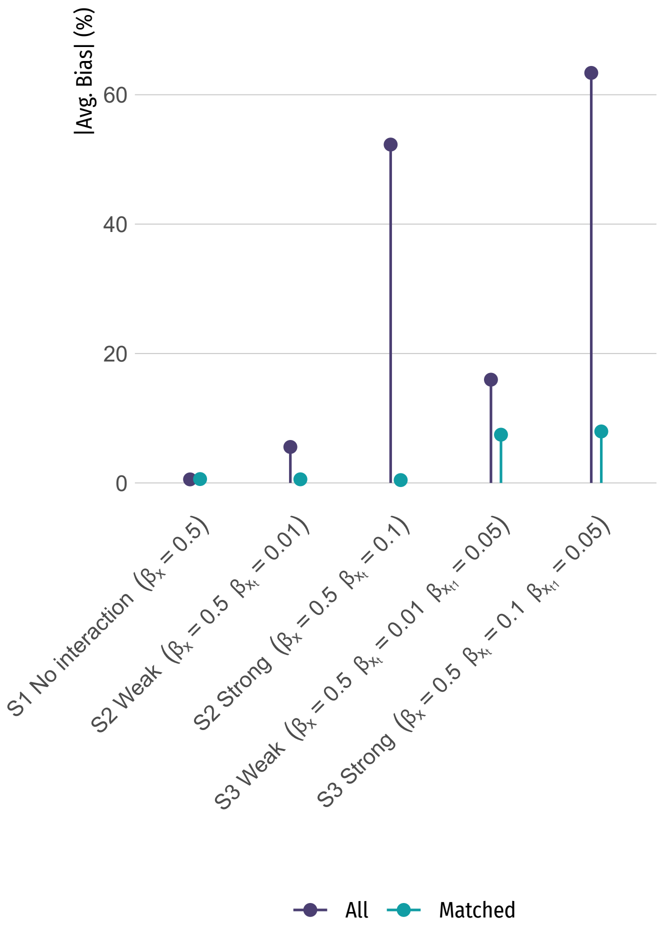
S3 - Differential interaction between X and t by treatment



Matching as an adjustment method for reducing/eliminating bias



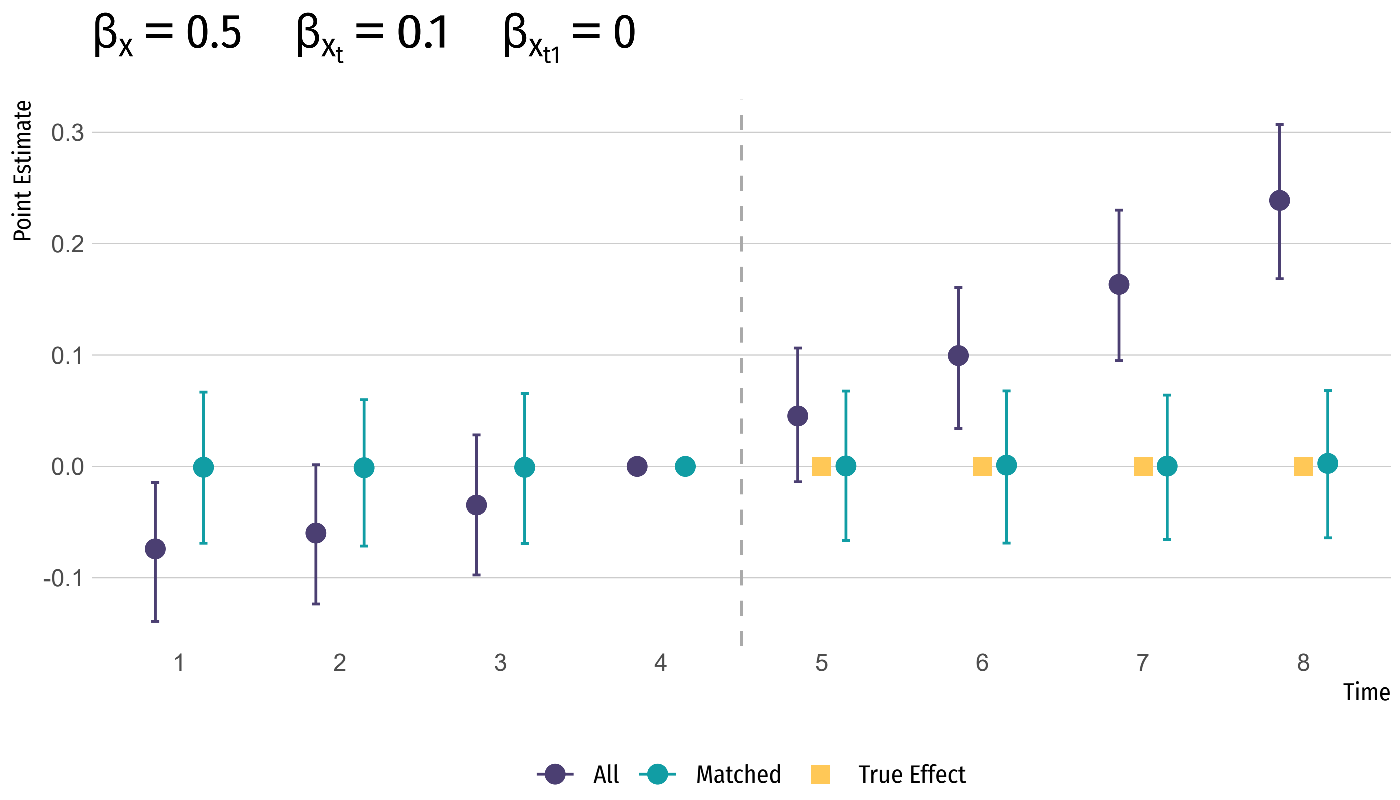
(a) Linear



(b) Quadratic

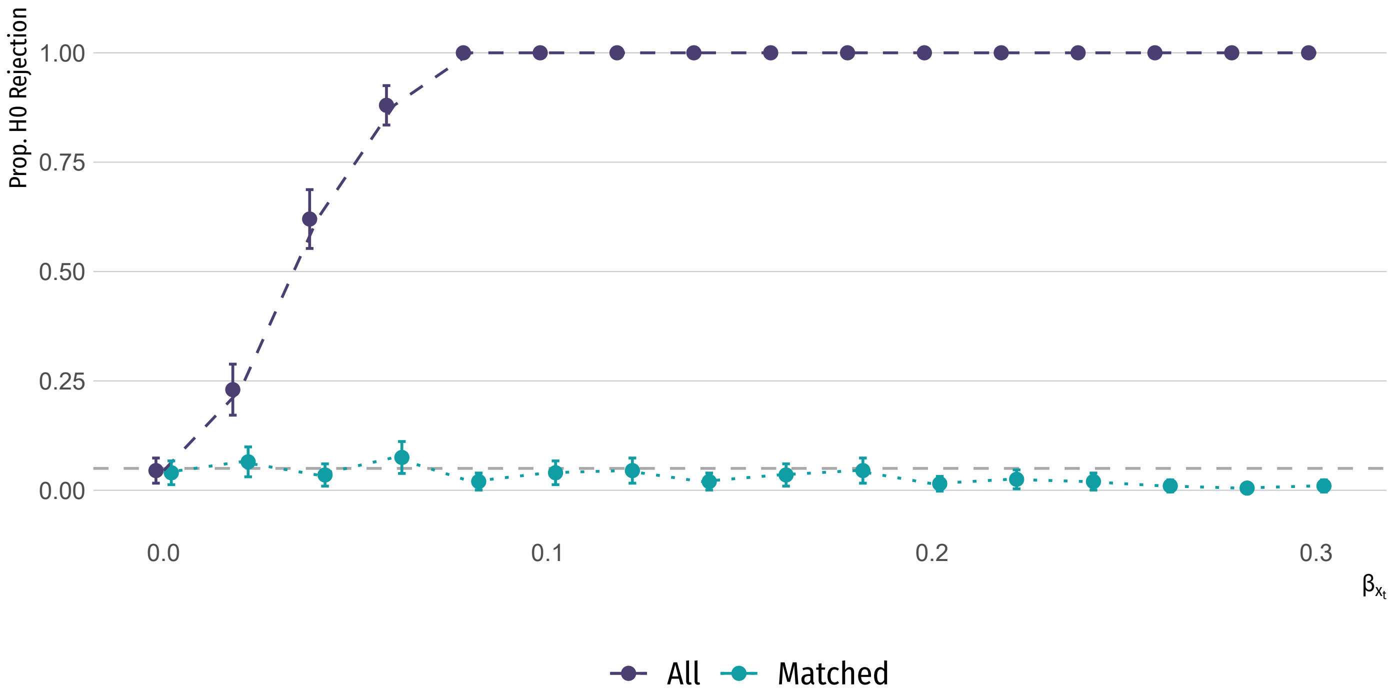
Why is this bias reduction important?

- Example of S2 (Quadratic) with no true effect:



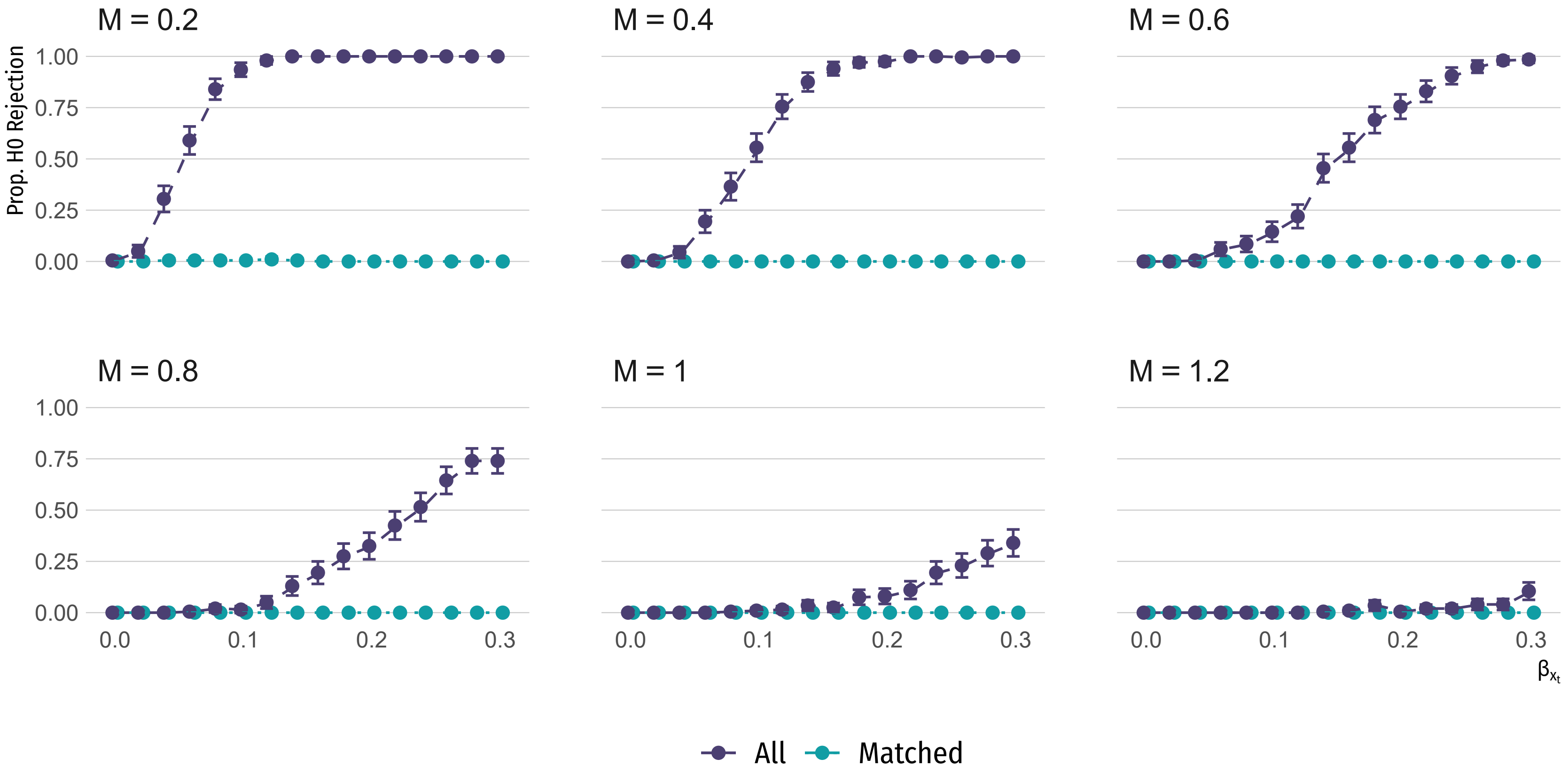
Why is this bias reduction important?

- Even under modest bias, we would incorrectly reject the null 20% of the time.



Why is this bias reduction important?

- Sensitivity analysis results are skewed by the magnitude of the bias.



Application

Preferential Voucher Scheme in Chile

- Universal flat voucher scheme $\xrightarrow{\textbf{2008}}$ Universal + preferential voucher scheme
- Preferential voucher scheme:
    - Targeted to bottom 40% of vulnerable students
    - Additional 50% of voucher per student
    - Additional money for concentration of SEP students.

Students:
- Verify SEP status
- Attend a SEP school

Schools:
- Opt-into the policy
- No selection, no fees
- Resources ~ performance

Impact of the SEP policy

- Mixed evidence of impact on test scores for lower-income students (Aguirre, 2022; Feigenberg et al., 2019; Neilson, 2016; Mizala & Torche, 2013)

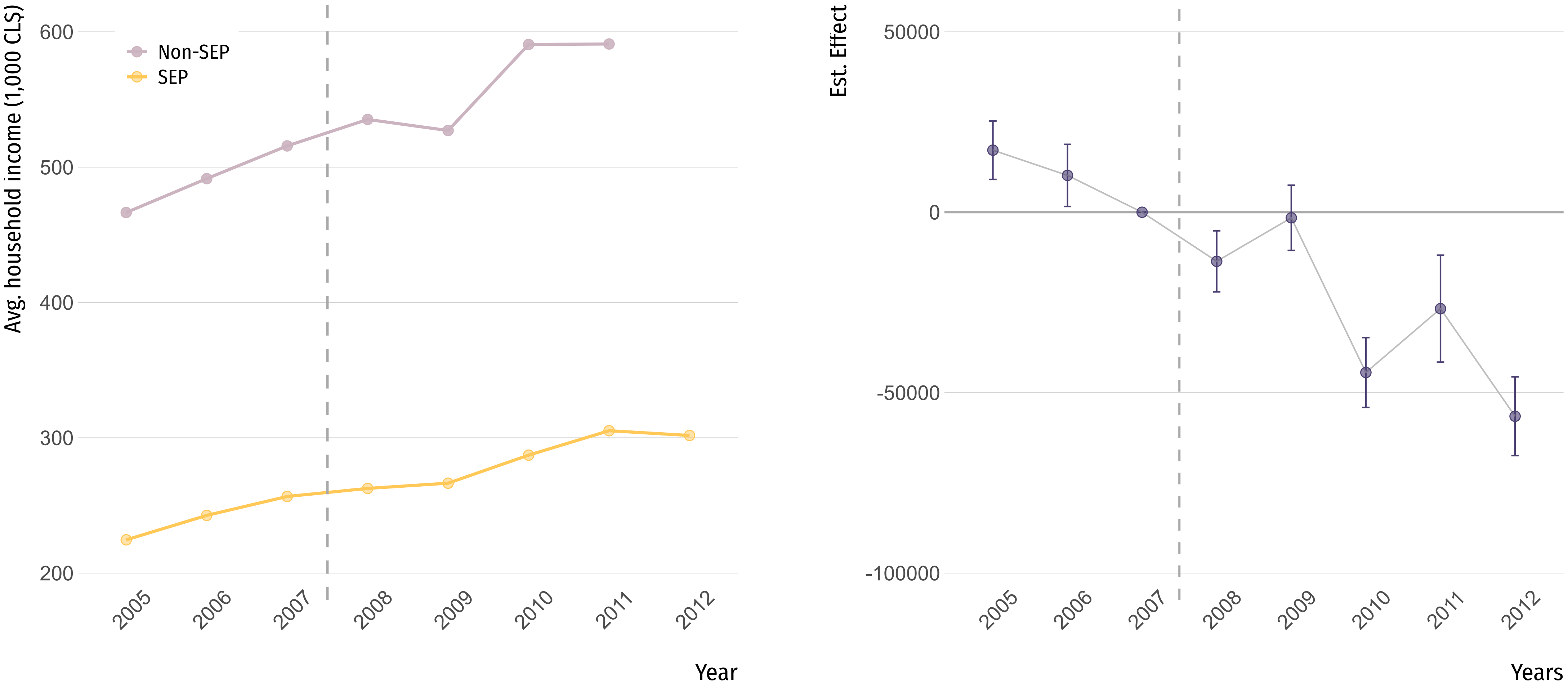
Impact of the SEP policy

- Mixed evidence of impact on test scores for lower-income students (Aguirre, 2022; Feigenberg et al., 2019; Neilson, 2016; Mizala & Torche, 2013)
- Design could have increased socioeconomic segregation (E.g. Incentives for concentration of SEP students)

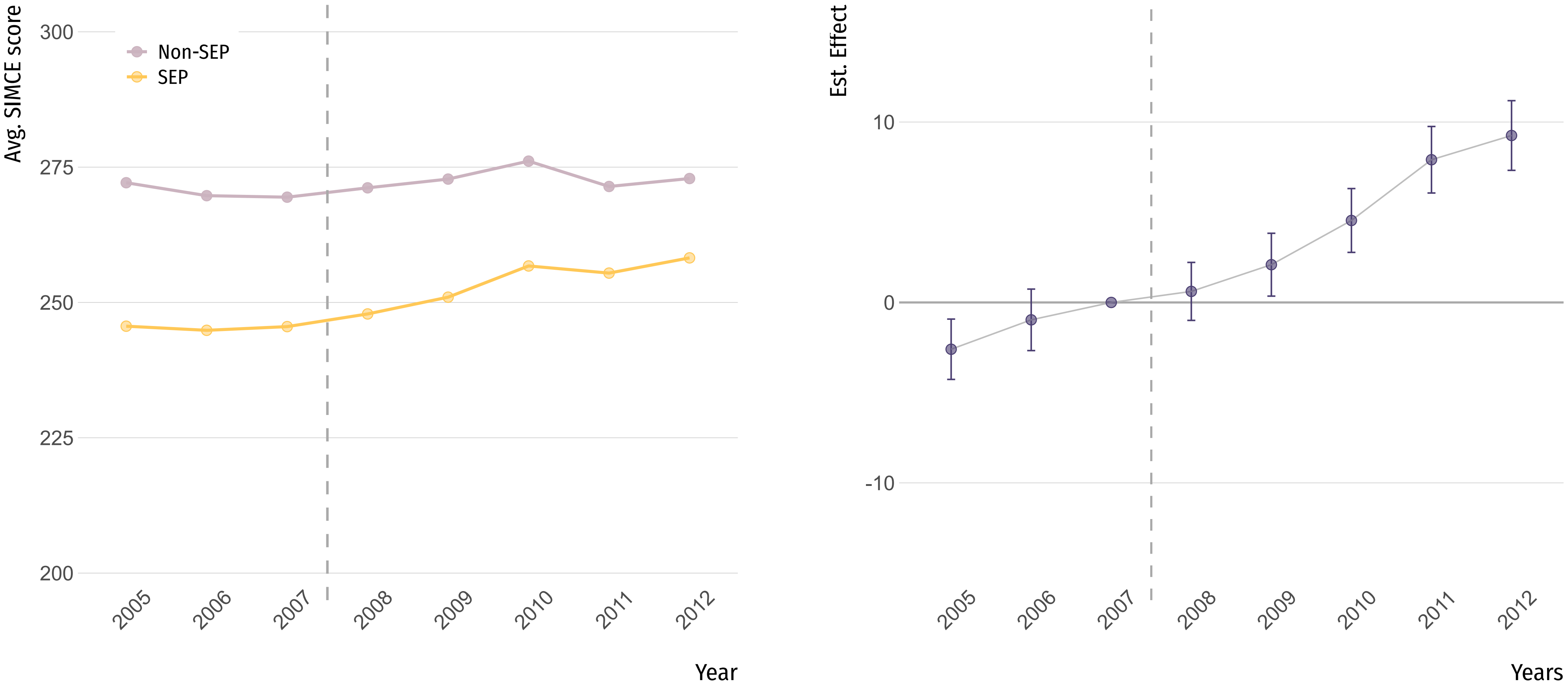
Impact of the SEP policy

- Mixed evidence of impact on test scores for lower-income students (Aguirre, 2022; Feigenberg et al., 2019; Neilson, 2016; Mizala & Torche, 2013)
- Design could have increased socioeconomic segregation (E.g. Incentives for concentration of SEP students)
- Key decision variables for schools: Performance, current SEP students, competition, add-on fees.
- Diff-in-diff (w.r.t. 2007) for SEP and non-SEP schools:
    - Only for private-subsidized schools
    - Matching using 2007 variables (similar results when using 2005-2007).
    - Outcome: Average students' household income and SIMCE score

Before matching: Household income



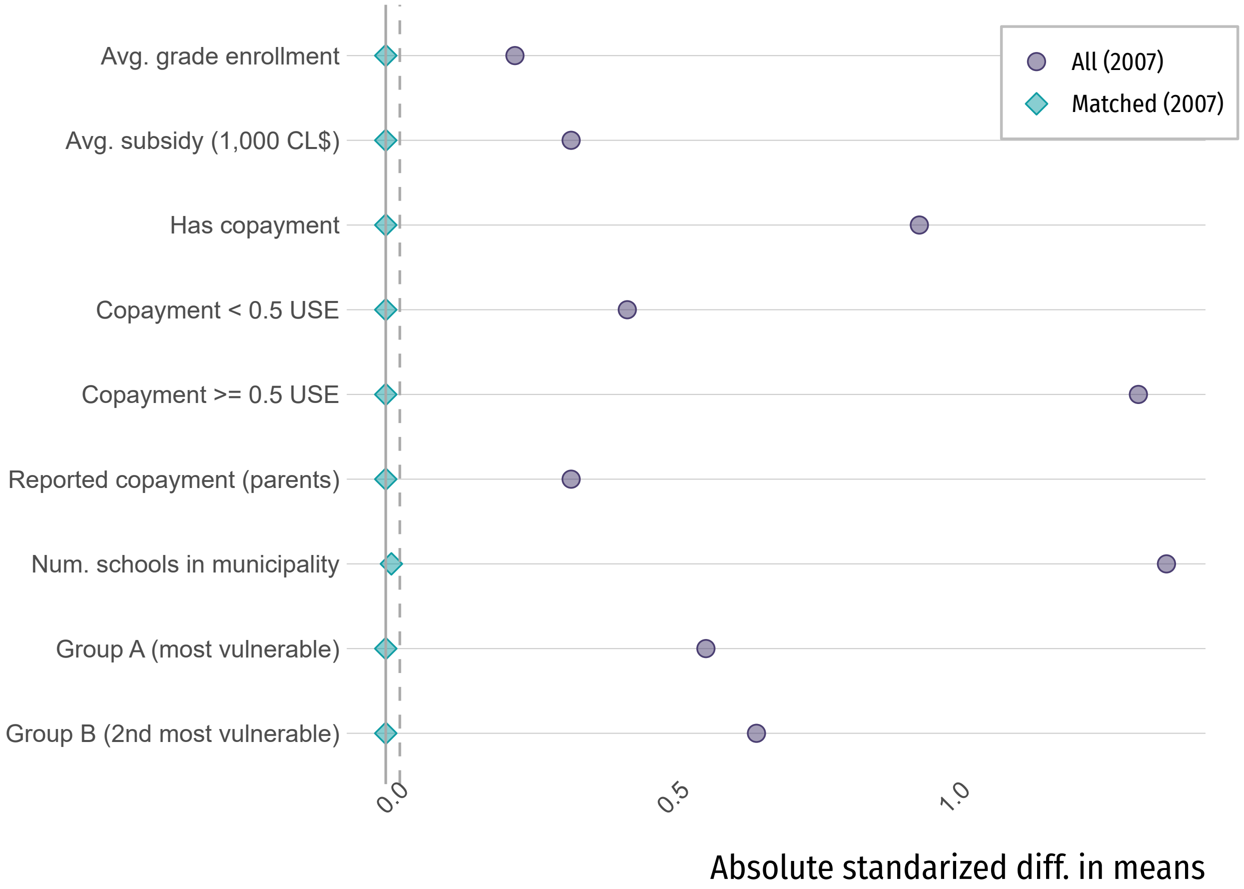
Before matching: Average SIMCE



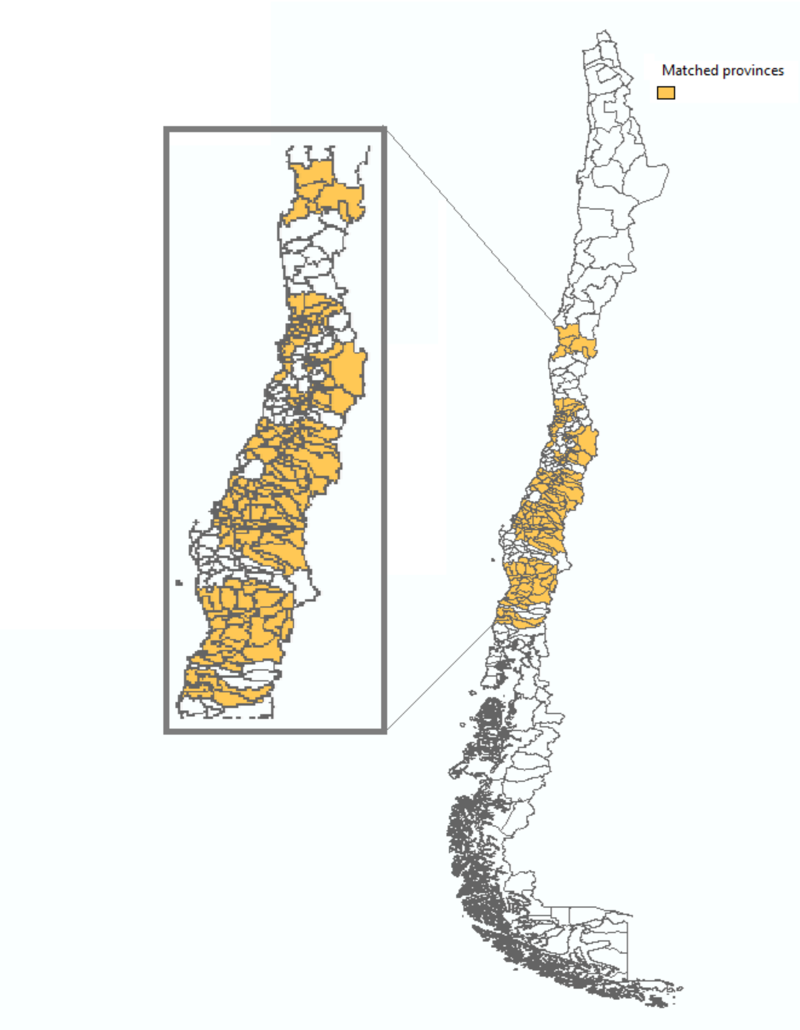
Matching + DD

- **Prior to matching**: No parallel pre-trend
- **Different types of schools**:
  - Schools that charge high co-payment fees.
  - Schools with low number of SEP student enrolled.
- **MIP Matching** using constant or "sticky" covariates:
  - Mean balance (0.025 SD): Enrollment, average yearly subsidy, number of voucher schools in county, charges add-on fees
  - Exact balance: Geographic province

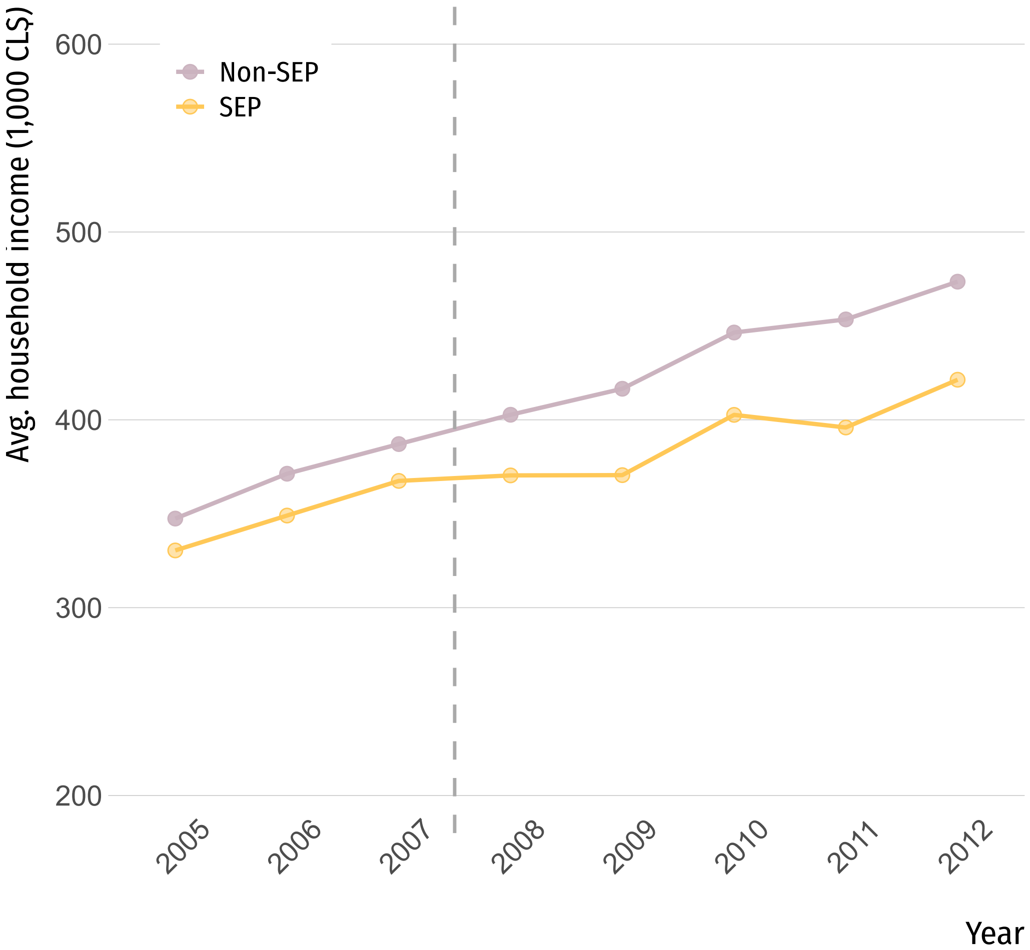
Groups are balanced in specific characteristics

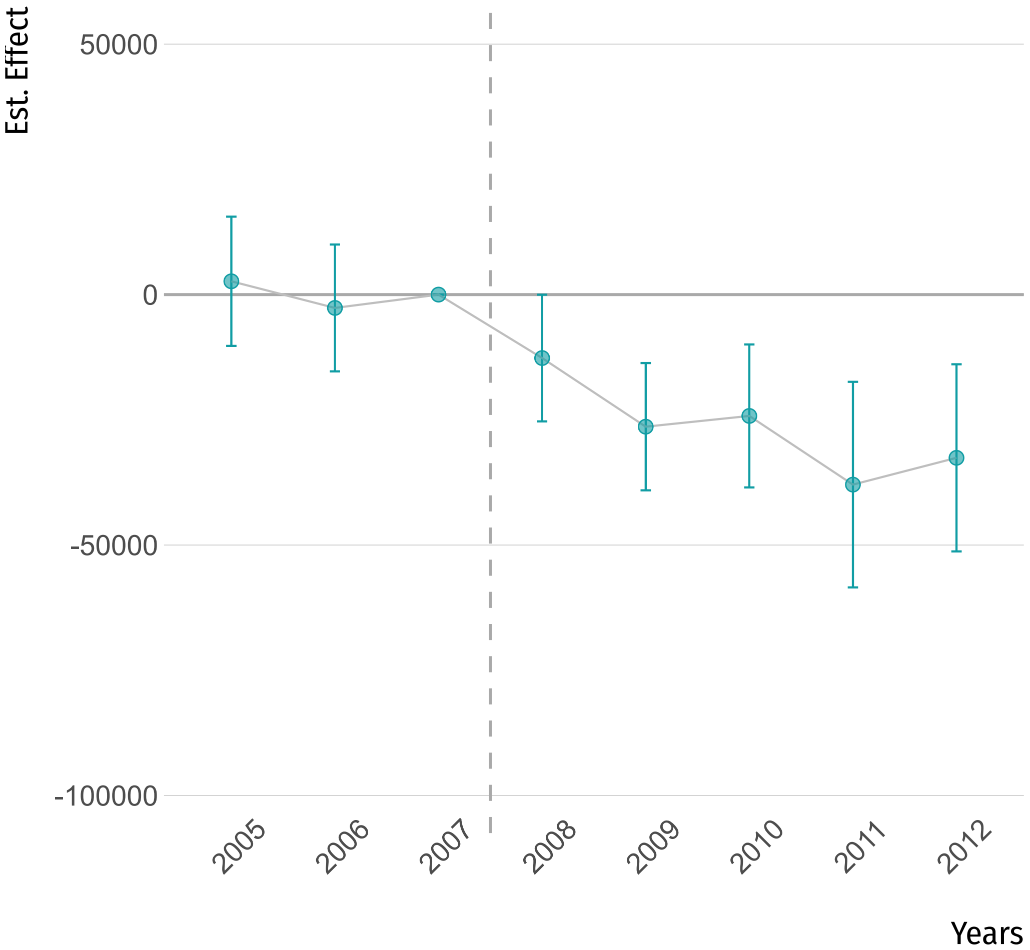


Matching in 16 out of 53 provinces

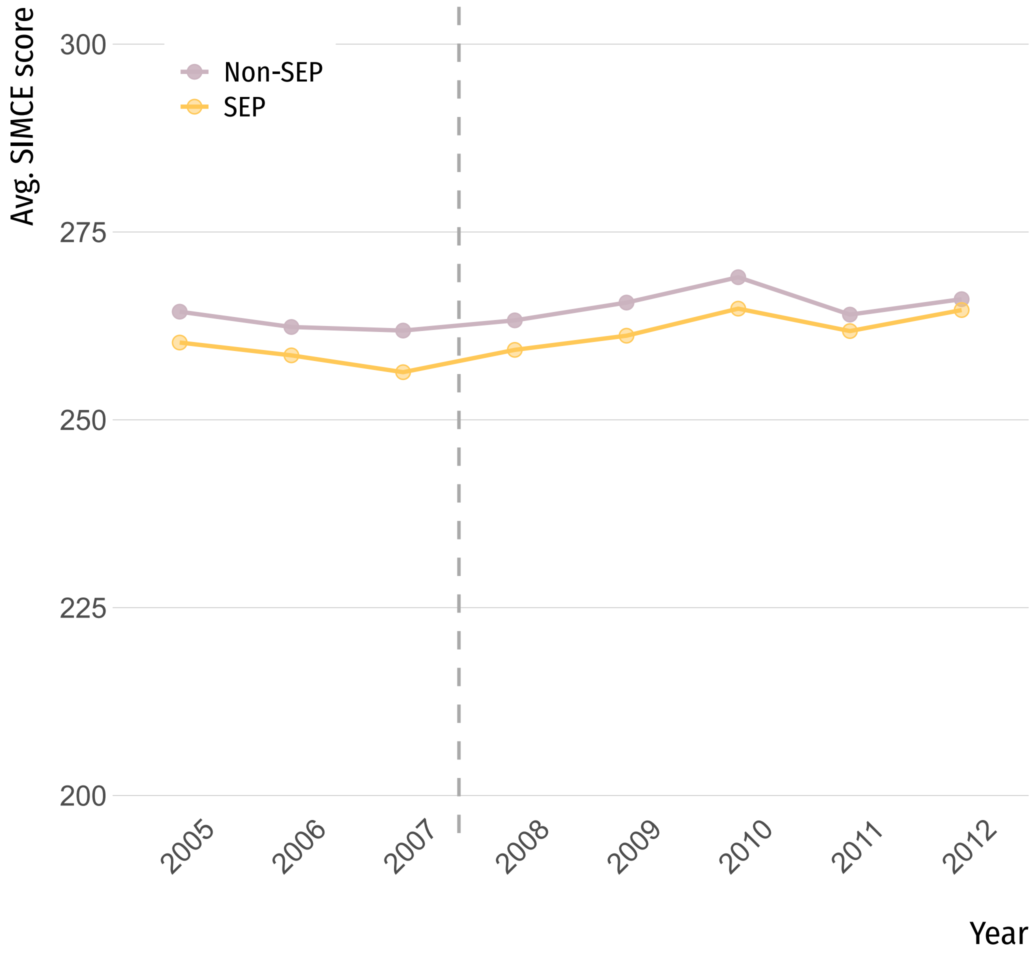


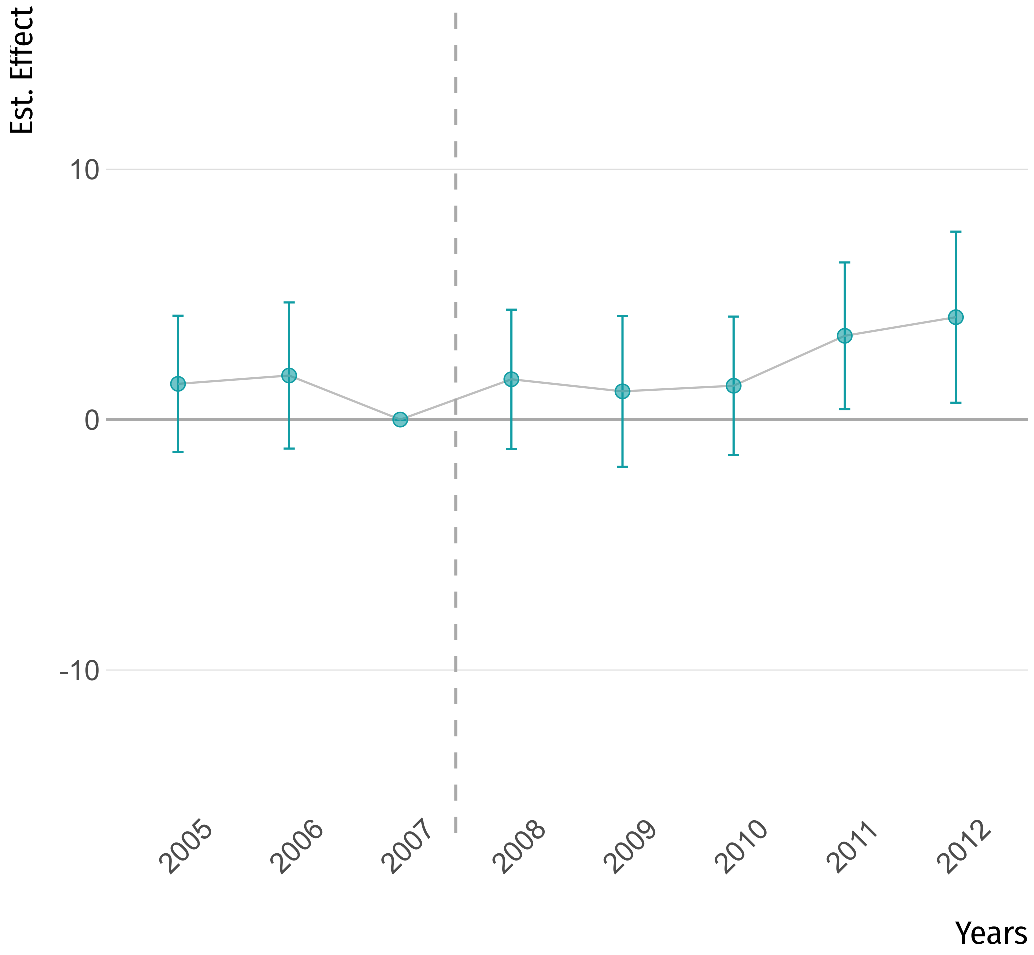
After matching: Household income





After matching: Average SIMCE



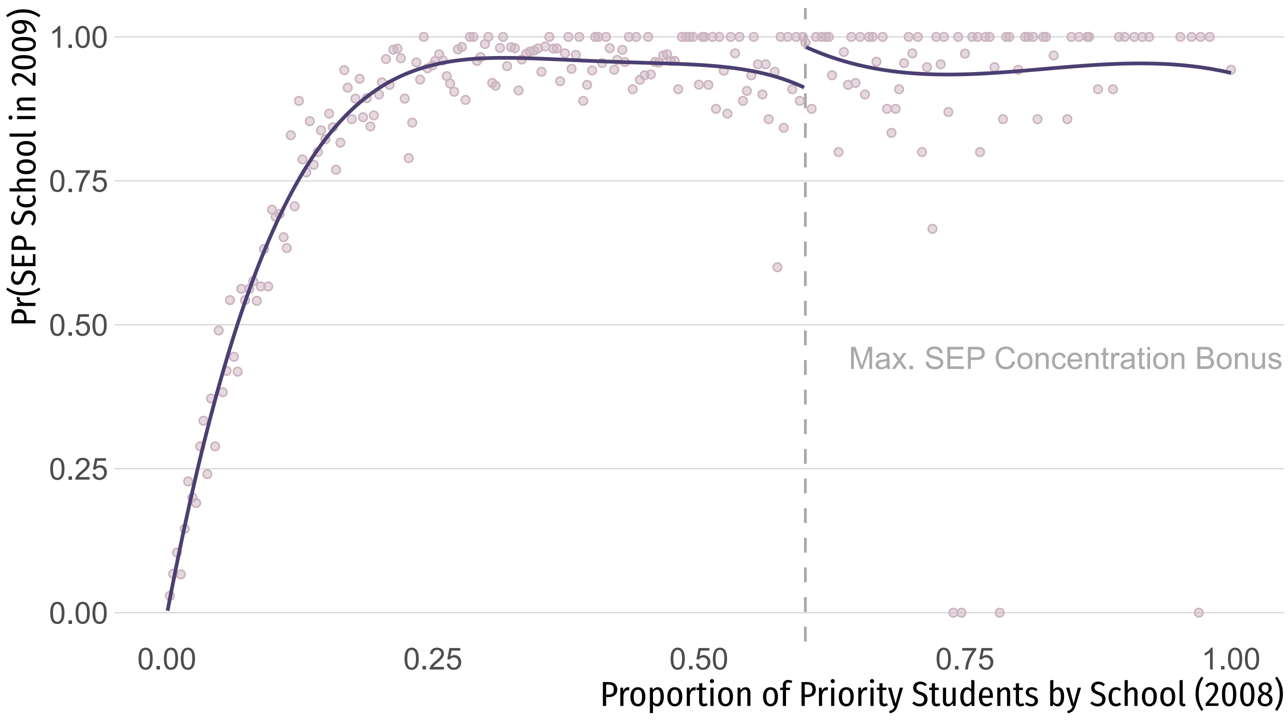


Results

- Matched schools:
  - More vulnerable and lower test scores than the population mean.
- 9pp increase in the income gap between SEP and non-SEP schools in matched DD:
  - SEP schools attracted even more vulnerable students.
  - Non-SEP schools increased their average family income.
- No evidence of increase in SIMCE score:
  - Could be a longer-term outcome.
- Findings in segregation are moderately robust to hidden bias (Keele et al., 2019):
  - $\Gamma_c = 1.76 \rightarrow$ Unobserved confounder would have to change the probability of assignment from 50% vs 50% to 32.7% vs 67.3%.
  - Allows up to 70% of the maximum deviation in the pre-intervention period (*M = 0.7*) vs 50% without matching (Rambachan & Roth, 2023)

Potential reasons?

- Increase in probability of becoming SEP in 2009 jumps discontinuously at 60% of SEP student concentration in 2008 (4.7 pp; SE = 0.024)



Let's wrap it up

Conclusions and Next Steps

- Matching can be an important tool to address violations in PTA.
- Bias reduction is very important for sensitivity analysis.
- Serial correlation also plays an important role: Don't match on random noise.
- Next steps: Partial identification using time-varying covariates

Difference-in-Differences using Mixed-Integer Programming Matching Approach

Magdalena Bennett

McCombs School of Business, The University of Texas at Austin

Seminario DIIS - PUC

November 20th, 2024

SEP adoption over time