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

Magdalena Bennett The University of Texas at Austin

Abstract

Difference-in-Differences (DD) is a commonly-used approach in policy evaluation where, under a parallel trend assumption, we can recover a causal effect by comparing the difference in outcomes between a treatment and a control group, both before and after an intervention was set in place. However, confounders that differentially vary over time often break the identifying assumption, biasing our estimates and rendering our design invalid. In this paper, I identify contexts where matching can help to eliminate or reduce bias, increasing the robustness of estimates under different sensitivity analyses, and show how balancing covariates directly can yield better results than other forms of adjustment or no adjustment at all. I illustrate these results with simulations and a case study of the impact of a new voucher scheme on socioeconomic segregation in Chile.

Keywords Difference-in-Differences, Matching, School vouchers

1. Introduction

Difference-in-Differences (DD) is a commonly-used approach in policy evaluation for identifying the impact of an intervention or treatment. Under a parallel trend assumption (PTA), we can recover a causal effect by comparing the difference in outcomes between a treatment and a control group, both before and after an intervention was set in place. However, time-varying confounders often break the identifying assumption, biasing our estimates and invalidating our study design.

One natural question that stems from the previous limitations is whether we can still recover valid causal estimates for a sub-group of our population that complies with the identification assumption, and, if this is the case, how can we identify such sub-population? In this paper, I propose combining a DD strategy with matching adjustment methods to (i) identify whether groups that potentially follow the PTA exist, and, if so, (ii) estimate a useful causal parameter that can provide insightful information regarding our intervention

The use of matching as a method to recover parallel trends under violations of the PTA is a contentious topic. While some researchers argue that there are clear advantages that stem from combining matching with a DD approach (Basu & Small, 2020; Ham & Miratrix, 2024; Ryan et al., 2015), others argue for a more cautious approach given that matching can also bias estimates depending on the context (Chabé-Ferret, 2017; Daw & Hatfield, 2018a; Zeldow & Hatfield, 2021). In this paper, I identify the different contexts in which matching can help reduce such biases, and show how balancing covariates directly can yield better results for solving some of these issues. I illustrate these findings with simulations and a case study of the impact of a new voucher scheme on socioeconomic segregation in Chile.

Working Paper

Last updated Jan 19, 2025

Key Points

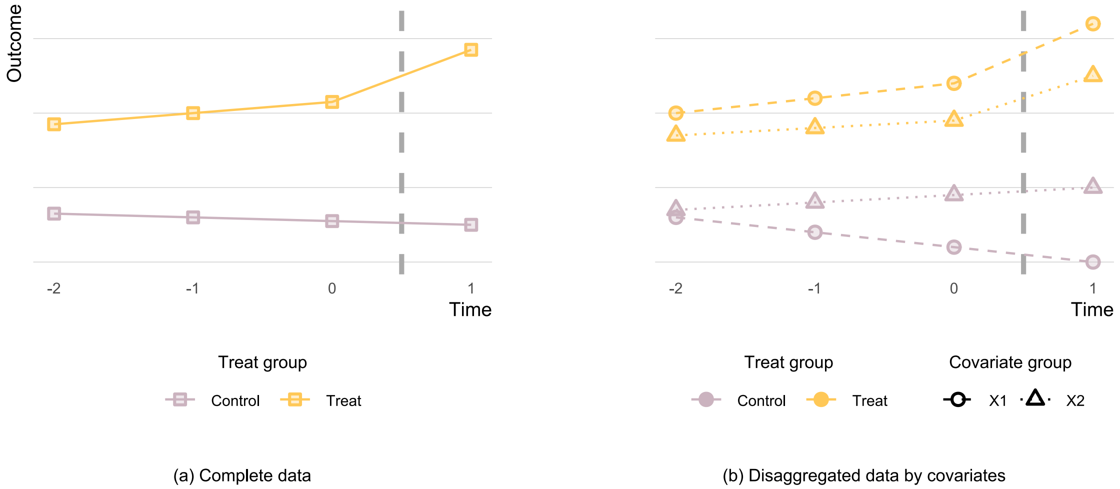
- Difference-in-Differences (DD) is a popular identification strategy in social sciences, which can fail if the main identification assumption (Parallel Trends) does not hold.
- Matching on time-invariant covariates can reduce bias in a DD setting and increase the robustness of additional sensitivity analyses.
- Even under violations of the main identification assumption, bias reduction can help with partial identification.

Correspondence to
Magdalena Bennett
m.bennett@austin.utexas.edu

Data Availability
Associated code will be made available on GitHub.

In particular, the matching procedure I propose for this setting — mixed-integer programming matching — has distinct advantages with respect to other adjustment methods previously addressed in the literature, which have mainly focused on propensity score matching on outcomes. Firstly, by matching on covariates and not pre-intervention outcomes, we are still able to use pre-intervention trends as a robustness check for the main identification assumption, while increasing the robustness of additional sensitivity analyses to violations of the identification assumption (Rambachan & Roth, 2023). Additionally, matching on covariates instead of previous outcomes reduces potential issues associated to regression to the mean, which are more accentuated by matching directly on outcomes before the treatment was set in place. Using a mixed-integer programming (MIP) matching approach we are able to balance covariates directly, also allowing us to identify matched groups that comply with a specific set of balancing constraints in the sample at hand. Finally, using matching as an adjustment procedure lends itself nicely to very transparent sensitivity analysis to hidden bias (Keele et al., 2019).

The main intuition behind the idea of using matching to recover parallel trends is displayed on Figure 1. Figure 1 (a) shows a stylized example of a DD study where the parallel trend assumption is likely to fail. Both groups do not seem to follow similar trajectories even before the intervention was set in place. Figure 1 (b), on the other hand, portrays the same data but disaggregated by covariate profile ($X = X_1$ and $X = X_2$ in this case). If we can identify a sub-population that complies with the parallel trend assumption (covariate profile $X = X_2$), it would be possible to recover a causal estimate for this particular group.



(a) Complete data

(b) Disaggregated data by covariates

Figure 1: Difference-in-differences (DD) example

In order to identify such groups, I propose matching on observed time-invariant characteristics (or characteristics with low temporal variation) to test whether parallel trends hold in the pre-intervention period after matching on contextually-relevant variables. Even though this is not a guarantee that the main identification assumption holds and could be an under-powered robustness check — as exposed by Rambachan & Roth (2023) and others —, it allows us to conduct additional sensitivity analyses that will be more informative given the bias reduction that matching achieves.

I provide a clear case study for the use of DD using matching. By combining a commonly-used identification strategy with MIP matching techniques, I am able to obtain the largest matched sample possible under pre-specified balancing constraints (Zubizarreta, 2012) using time-stable covariates.

This paper contributes to the growing literature of DD in different dimensions. Firstly, it relates to the literature of adjustment in DD settings, both using matching strategies (Basu & Small, 2020; Chabé-Ferret, 2017; Daw &

Hatfield, 2018a; Ham & Miratrix, 2024; Ryan, Burgess, & Dimick, 2015; Zeldow & Hatfield, 2021) as well as other methods of adjustment used in the DD context (Arkhangelsky et al., 2021; Sant'Anna & Zhao, 2020). Unlike previous work in this area, I use a specific method of matching that allows us to get the largest matched sample possible under pre-specified balancing constraints (Bennett et al., 2020; Zubizarreta, 2012) while keeping the unit of analysis intact. This strategy helps to maintain the core idea of the DD strategy for a sub-sample of the original population that is selected by the balancing constraints, without enforcing parallel trends in the pre-intervention period.

Additionally, in line with the current literature, combining MIP matching and DD also allows for some violations in parallel trends (Roth et al., 2023). While the main idea behind selecting a sub-sample through matching is to identify a group for which the main identification assumption holds, sensitivity analyses can be used in the matched sample in the same way that they are used in traditional DD setups. Matched DD designs can be readily used in combination with current sensitivity analyses for violations of the parallel trends (e.g. Rambachan & Roth (2023)) or hidden bias (Keele, Small, Hsu, & Fogarty (2019)).

Finally, this study sets up a framework that allows researchers to assess under whether a DD strategy is reasonable for a particular study. By providing a comparison between the entire sample and matched samples under different balancing constraints, researchers can assess whether the bias reduction through matching is significant in practice or not, and compare the robustness of their conclusions.

The paper is structured in 4 sections, apart from this introduction. Section 2 lays out the main identification strategy for a DD approach combined with matching. In Section 3, I show the main results for simulated scenarios, including results for bias and sensitivity analyses. Section 4 reviews the application used for portraying the previous method, using the Chilean Preferential Voucher system as an example. Finally, Section 5 concludes.

2. Difference-in-Differences using matching

2.1. Difference-in-Differences as an identification strategy

Difference-in-differences (DD) is an identification strategy that relies on two key assumptions: (1) absent the intervention, both treatment and control groups would have the same trend in the post-intervention period (also referred to as the Parallel Trend Assumption), and (2) random shocks do not differentially affect treatment and control groups.

Confounding in a DD setting is related to time-varying differences that are not accounted for in our design. In that sense, time-varying terms that evolve differentially between groups will be the ones that can introduce bias into our analysis if not accounted for correctly. Expanding on Zeldow & Hatfield (2021) notation, I define the structure of potential outcomes in a general additive model as following:

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

where $Y_{it}(z)$ is the potential outcome under a binary treatment $Z = z$ for unit i in period t, with an additive time-dependent effect τ_{it} for the potential outcome under treatment. The parameter α_0 is a general intercept and u_{it} is an idiosyncratic error with mean 0. Time fixed effects are represented by λ_t,[1], assumption I relax on the following section. Finally, X_i is a time-invariant covariate which has both a constant and time-variant effect on the outcome, which can vary by treatment group.

[1] I assume that unobserved trends λ_t are the same for both treatment groups, so the main identification assumption holds at least in some cases. If this is not the case, the main identification assumption would be violated and trend analyses for the pre-intervention period in addition to sensitivity analyses would show that a DD approach is not suitable.

The main identification assumption in a DD context, commonly referred to as the Parallel Trend Assumption (PTA), can also be expressed in terms of potential outcomes for a 2x2 scenario as following, where treatment happens at $t = T_0$ and $t < T_0$ and $t' > T_0$:

$$\mathbb{E}[Y_{it'}(0) - Y_{it}(0) \mid Z = 1] = \mathbb{E}[Y_{it'}(0) - Y_{it}(0) \mid Z = 0] \tag{2}$$

Under the PTA (Equation 2), the expected difference in potential outcomes under control for the treatment group before and after the intervention was set in place is the same as the expected difference for the control group.

Plugging in this condition to our potential outcomes framework (Equation 1), we recover the following expression:

$$\mathbb{E}[\gamma_1(X_i, t') + \gamma_2(X_i, t') - \gamma_1(X_i, t) - \gamma_2(X_i, t) \mid Z = 1] = \mathbb{E}[\gamma_1(X_i, t') - \gamma_1(X_i, t) \mid Z = 0] \tag{3}$$

If $\mathbb{E}[\gamma_k(X_i, t) \mid Z = z] = \bar{\gamma}_k(X^z, t)$, then we can replace these values in Equation 3 and under the PTA:

$$\left(\bar{\gamma}_1(X^1, t') - \bar{\gamma}_1(X^0, t')\right) - \left(\bar{\gamma}_1(X^1, t) - \bar{\gamma}_1(X^0, t)\right) - \left(\bar{\gamma}_2(X^1, t') - \bar{\gamma}_2(X^1, t)\right) = 0 \tag{4}$$

Assuming the different biases do not cancel themselves out, from Equation 4, we can see that in order for the parallel trend assumption to hold, then one of the following conditions needs to hold:

1. **No effect or constant effect of X on Y over time**: In this case, if $\gamma_1(X, t) = 0$ or, more generally, $\gamma_1(X, t) = \gamma_1(X)$, then it is trivial to see that the PTA will hold.
2. **Equal distribution of observed covariates between treated and control group**: If $X^0 = X^1 = X$, the differential effect of covariates over time would not be a concern for the PTA.

in addition to the following condition:

3. **No differential time-effect of X on Y by treatment group**: If $\gamma_2(X, t) = \gamma_2(X)$, then the temporal effect of the covariate on the outcome would be the same for both groups, preventing the violations of parallel trends.

Given that we do not observe the functions $\gamma_k(\cdot)$, under potential violations of the parallel trend assumptions we can make sure that (2) holds by matching directly on covariates, and test the robustness of (3) by conducting a sensitivity analysis (Rambachan & Roth, 2023).

2.2. *Relaxation of the parallel trend assumption based on unobservable or time-varying characteristics*

In Equation 1 of Section 2, I assumed that there were no differential trends between the treatment and the control group, where $\lambda_t(Z = 0) = \lambda_t(Z = 1)$. If we relax this assumption, where we allow differential trends between both groups that are not exclusively dependent on the temporal association of time-invariant characteristics (e.g. $\gamma_\cdot$ functions in Equation 1), the previous equation can be expressed as follows:

$$\begin{aligned} Y_{it}(0) = &\quad \alpha_i + \lambda_{tz} + \gamma_0(X_i) + \cdot \gamma_1(X_i, t) + \gamma_2(X_i, t) \cdot Z_i + u_{it} \\ Y_{it}(1) = Y_{it}(0) + \tau_{it} = &\quad \alpha_i + \lambda_{tz} + \gamma_0(X_i) + \cdot \gamma_1(X_i, t) + \gamma_2(X_i, t) \cdot Z_i + \tau_{it} + u_{it} \end{aligned} \tag{5}$$

where λ_{tz} is a group Z specific time trend. In this case, this time trend can depend on unobserved or observed time-varying covariates (i.e. $\lambda_{tz} = g(u_{it}, X_{it})$, where $u_{it} \sim F_u(\mu^u_{zt}, \sigma^u_{zt})$, and $X_{it} \sim F_x(\mu^x_{zt}, \sigma^x_{zt})$).

Under the PTA, now assuming λ_{tz} is a differential trend between both groups, I can re-write Equation 5 as:

$$\begin{aligned}&\overbrace{\Big(\gamma_1^-(X^1,t') - \gamma_1^-(X^0,t')\Big) - \Big(\gamma_1^-(X^1,t) - \gamma_1^-(X^0,t)\Big)}^{Obs.Bias} + \\ &\underbrace{\Big(\gamma_2^-(X^1,t') - \gamma_2^-(X^1,t)\Big)}_{Obs.Diff.Bias} + \underbrace{(\lambda_{t'1} - \lambda_{t'0}) - (\lambda_{t1} - \lambda_{t0})}_{Unobs.Bias} = 0\end{aligned} \tag{6}$$

In Equation 6, I have identified three distinct terms that can introduce bias to the DD setting. The first one, *observed bias (OB)*, refers to the bias introduced by differential distribution of observed time-invariant characteristics X between the treatment and the control group. The second term, *observed differential bias (ODB)*, is related to the heterogeneous association of X on Y by treatment status. Finally, the *unobserved bias (UB)* is introduced by the differential trend of time-varying observable and unobservable characteristics.

In a traditional DD setup, the estimated effect τ^* would be:

$$\tau^* = \hat{\tau} + (OB + ODB + UB) \tag{7}$$

where $\hat{\tau}$ is the unbiased estimated effect for the DD setup for time period t'. In the case we use matching in addition to DD, then the estimated effect τ_M^* would be[2]:

$$\tau_M^* = \hat{\tau} + (ODB + UB) \tag{8}$$

Assuming an informative pre-intervention period, we can assess whether there is absolute bias reduction using placebo tests prior to the intervention to inform us whether following this method would be useful or not. For example, if $ODB + UB > 0$ and OB is such that $-2 \times (ODB + UB) < OB < 0$, then there would not be a decrease in bias by matching.

In Rambachan & Roth (2023), the authors refer to this overall bias as δ_1, where δ_1 is the differential trend between potential outcomes in the post-intervention period. Given that I am assuming a non-staggered DD setting, we can use Rambachan & Roth (2023) notation for sensitivity bounds for relative magnitudes as follows:

$$|d_1| \leq M^- |\delta_{-1}| \tag{9}$$

where $|\delta_{-1}|$ refers to the maximum pre-treatment violation in parallel trends. Then, when conducting a sensitivity analysis for the traditional DD setting, the bounds are as follows:

$$\left[\tau^* - b^{max}, \tau^* - b^{min}\right] = [\hat{\tau} + OB + ODB + UB - b^{max}, \hat{\tau} + OB + ODB + UB - b^{min}] \tag{10}$$

While for the matched DD setup, the sensitivity bounds are defined as:

$$\left[\tau_M^* - b_M^{max}, \tau_M^* - b_M^{min}\right] = [\hat{\tau} + ODB + UB - b_M^{max}, \hat{\tau} + ODB + UB - b_M^{min}] \tag{11}$$

In this case, b^{max} (b_M^{max}) and b^{min} (b_M^{min}) are defined by the maximum violation in the pre-intervention period, in addition to the variability of the estimates (SE), If the overall bias in the pre-intervention period is reduced by matching (as assessed by comparing the original DD and the matched DD estimates for the period prior to the introduction of the treatment), and the reduction is greater than the increase in variability from loss of sample, then the sensitivity bounds in the match DD setting will be tighter than in the traditional DD setting, where $|b^{max}| > |b_M^{max}|$ and $|b^{min}| > |b_M^{min}|$.

[2] For simplicity, I am assuming a constant additive effect τ for all units.

Additionally, if the observed bias in the post-intervention period is different than 0, $OB \neq 0$, and there is an overall bias reduction in the pre-intervention period, then sensitivity bounds for the matched setup will be centered closer to the true treatment effect τ in comparison to the unmatched scenario.

2.3. Matching and Difference-in-Differences

Given that the treatment and control groups used in a DD approach can have significant differences in terms of observed characteristics, matching poses some attractive features that could help reduce these gaps (Ham & Miratrix, 2024; Ryan, Burgess, & Dimick, 2015) and reduce potential bias in DD estimates. Using matching in this setting also has the distinct benefit of making the overlap region explicit between two groups that can be very different in terms of covariates, reducing the risk of relying in a parametric form of adjustment.

However, as many researchers have pointed out, adjustments using matching can also lead to exacerbated bias when using pre-intervention outcomes or even time-varying characteristics as matching covariates (Chabé-Ferret, 2015; 2017; Daw & Hatfield, 2018b; 2018a). One of the main hazards when using matching in combination with a DD approach is that matching on a certain type of covariates can give the false impression that the parallel trend assumption holds, while in fact is just a construction of the matching itself.

To avoid such potential bias and, at the same time, obtain two groups that resemble each other in terms of some key observable covariates, we need to match on time-invariant characteristics (or characteristics that are stable during the pre-intervention period) that we deem relevant for the selection process. Matching on time-invariant characteristics, and not, for instance, on pre-intervention outcomes, has the advantage of avoiding potential regression to the mean issues and overall bias (Chabé-Ferret, 2017; Daw & Hatfield, 2018b).

Matching for a set of covariates $\mathbf{X}_i$, I estimate the average treatment effect on the treated on the matched set (SATT) of K pairs as:

$$\hat{\tau}_{SATT} = \frac{1}{K}\sum_{k=1}^{K}\Big(Y_{k(1)T} - Y_{k(1)0} - \big(Y_{k(0)T} - Y_{k(0)0}\big)\Big) \tag{12}$$

Where $Y_{k(j)t}$ is the outcome for unit $j = \{0, 1\}$ (control or treated) in matched pair k for time period t, where $t = 0$ is the pre-intervention period and $t = T$ is some period after the intervention was put in place. It is important to note that the matched $SATT$ might not be the same as the complete sample ATT, because we might be in the presence of heterogeneous effects. However, using this strategy, we are able to clearly characterize our matched sample and compare it to the entire population of the study.

3. Simulations

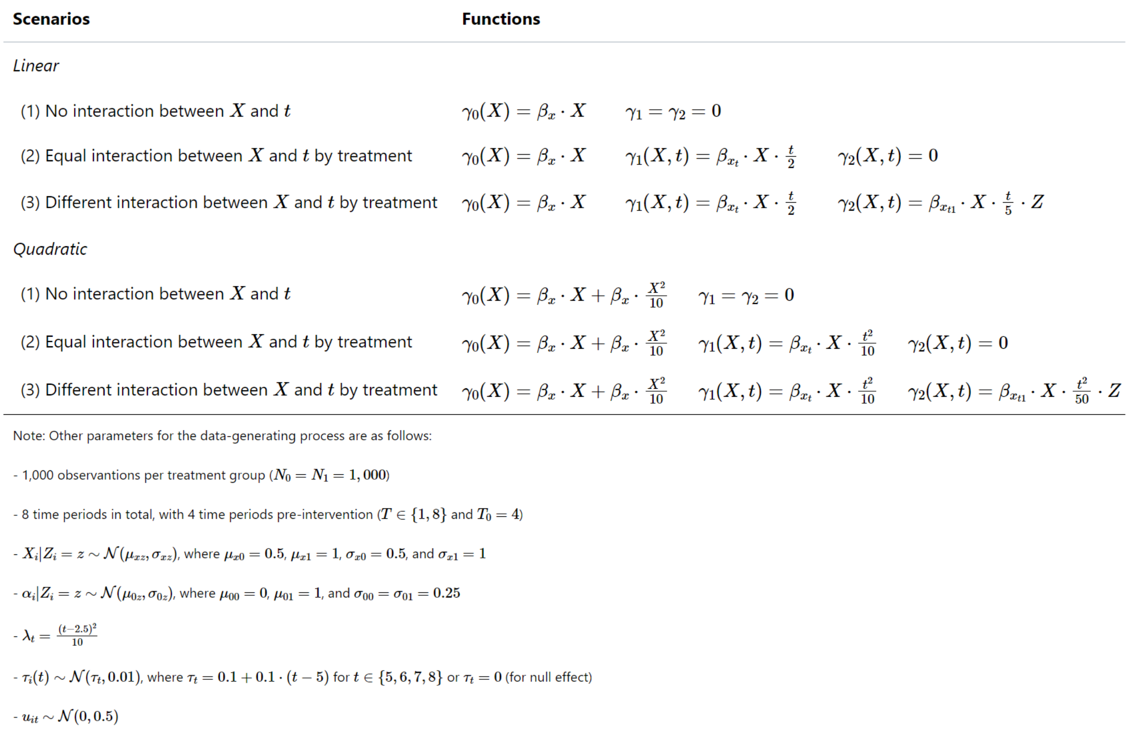
For simulations, expanding on Daw & Hatfield (2018b) setup, I assume the following data generating process previously described:

$$Y_{it} = \alpha_i + \lambda_t + \gamma_0(X_i) + \cdot\,\gamma_1(X_i, t) + \gamma_2(X_i, t)\cdot Z_i + \tau_i(t)\cdot Z_i \cdot \mathrm{I}(\mathrm{t} > \mathrm{T}_0) + u_{it} \tag{13}$$

Where Y_{it} is the outcome for unit i in time period t, Z_i is the treatment status for the individual, and X_i is a time-invariant covariate. Functions γ_0, γ_1, and γ_2 capture the association between Y and X, allowing for time interactions, and λ_t represents a common time trend. Finally, $\tau_i(t)$ represents and additive treatment effect, which is included only after the intervention is set on $t > T_0$.

Using the general structure in Equation 13, I focus on the following scenarios:

Table 1: Data generating process for different simulation scenarios



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	$\gamma_0(X) = \beta_x \cdot X + \beta_x \cdot \frac{X^2}{10}$	$\gamma_1(X,t) = \beta_{x_t} \cdot X \cdot \frac{t^2}{10}$	$\gamma_2(X,t) = \beta_{x_{t1}} \cdot X \cdot \frac{t^2}{50} \cdot Z$

Note: Other parameters for the data-generating process are as follows:

- 1,000 observantions per treatment group ($N_0 = N_1 = 1,000$)

- 8 time periods in total, with 4 time periods pre-intervention ($T \in \{1, 8\}$ and $T_0 = 4$)

- $X_i | Z_i = z \sim \mathcal{N}(\mu_{xz}, \sigma_{xz})$, where $\mu_{x0} = 0.5$, $\mu_{x1} = 1$, $\sigma_{x0} = 0.5$, and $\sigma_{x1} = 1$

- $\alpha_i | Z_i = z \sim \mathcal{N}(\mu_{0z}, \sigma_{0z})$, where $\mu_{00} = 0$, $\mu_{01} = 1$, and $\sigma_{00} = \sigma_{01} = 0.25$

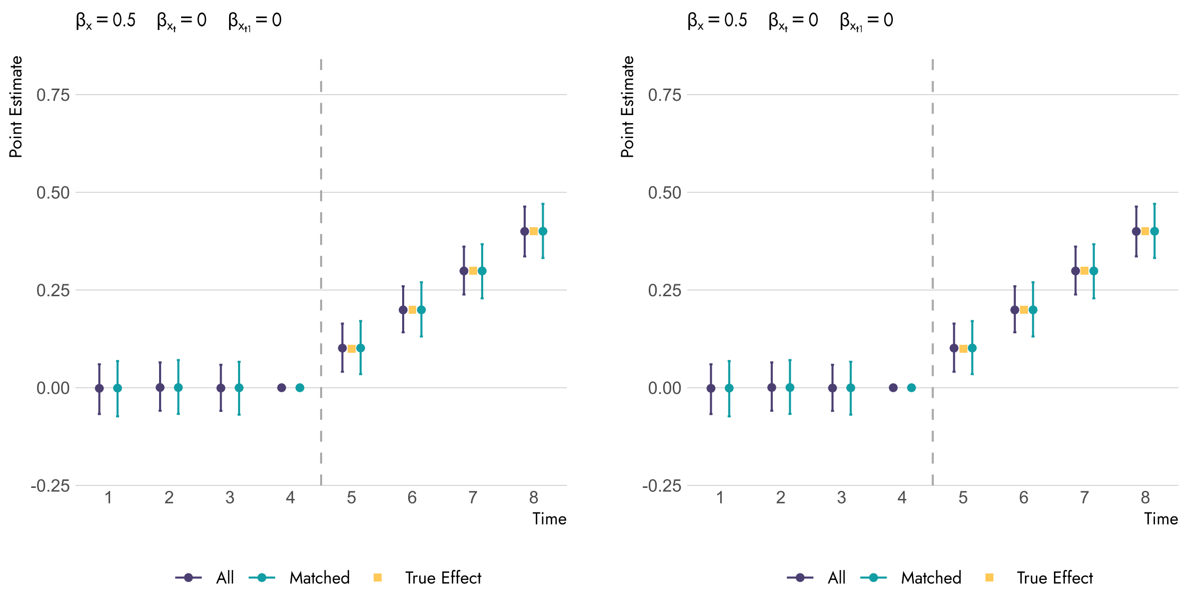
- $\lambda_t = \frac{(t-2.5)^2}{10}$

- $\tau_i(t) \sim \mathcal{N}(\tau_t, 0.01)$, where $\tau_t = 0.1 + 0.1 \cdot (t-5)$ for $t \in \{5, 6, 7, 8\}$ or $\tau_t = 0$ (for null effect)

- $u_{it} \sim \mathcal{N}(0, 0.5)$

For all scenarios and values of β_x, β_{x_t}, and $\beta_{x_{t1}}$, I ran 1,000 simulations and estimate the treatment effect over time using an event study approach for the entire sample and the matched sample. For matching, I use a balance restriction of a 0.01SD of the mean and use a cardinality matching approach (Zubizarreta et al., 2014) to obtain the largest matched sample possible under this balancing constraint.

For the first scenario, where there is no interaction between the covariates and the time variable, there is no parallel trend assumption violation, so the DD strategy should recover unbiased estimates of the true treatment effect. Figure 2 shows the event study plot for Scenario 1, for some values of β_x, β_{x_t}, and $\beta_{x_{t1}}$, both under linear and quadratic associations with the outcome. When there is no interaction between the time period and observed covariates, using the complete sample as well as the match sample recover unbiased estimates of the true increasing treatment effect. In this case, the matching strategy is slightly less efficient because of the loss of sample size, but still produces an unbiased and consistent estimate.

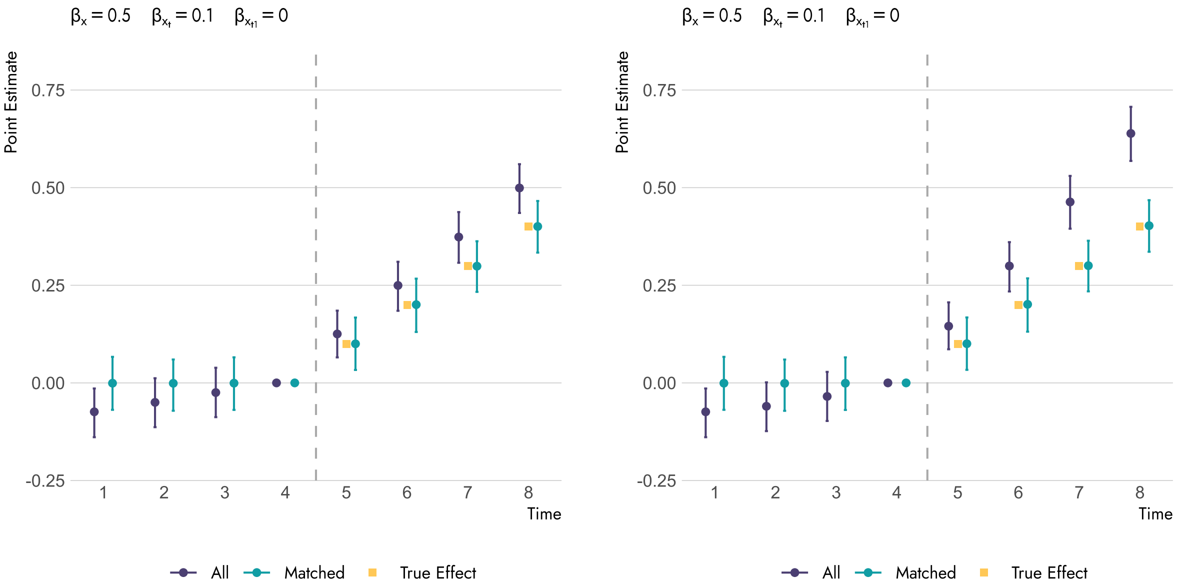


(a) Linear

(b) Quadratic

Figure 2: Event study estimates by time period (wrt T=4) for no interaction between X and t

In Scenario 2, there is an interaction between covariates and time period, but the functional form is the same for both groups. However, the covariate distribution can differ between both groups, introducing potential bias to a DD estimation. While a traditional DD approach would produce a biased estimate, which would depend on the functional form of the interaction, matching prior to estimation in a DD setting returns unbiased estimates of the true treatment effects (Figure 3). These results hold for different functional forms and values of βs, as I am not relying on parametric assumptions for adjusting for covariates.

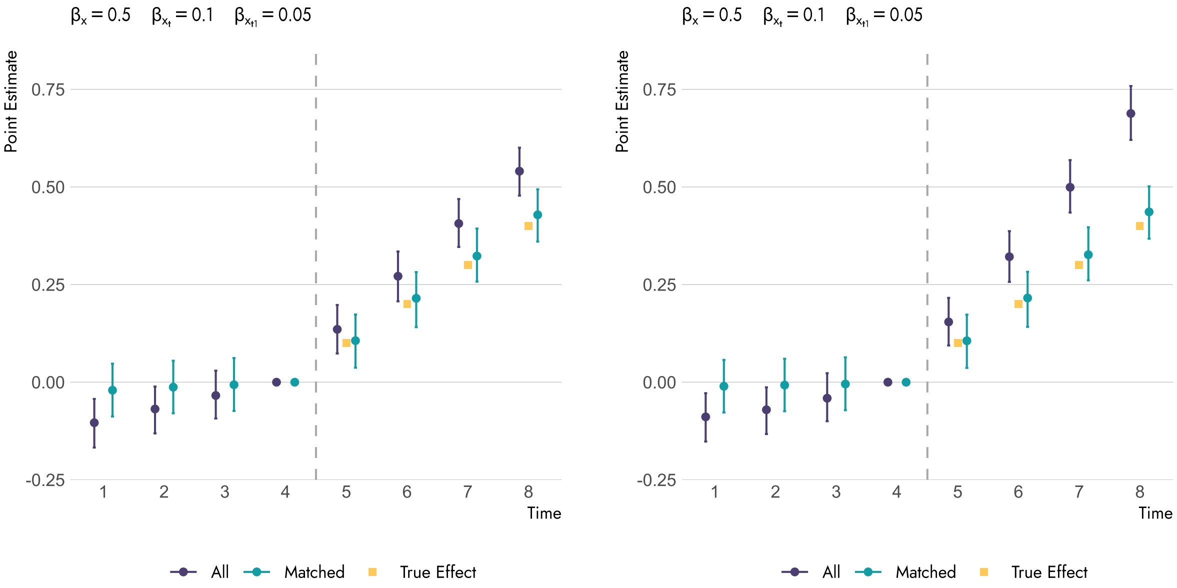


(a) Linear

(b) Quadratic

Figure 3: Event study estimates by time period (wrt T=4) for equal interaction between X and t

In the case where there is a differential interaction between covariate X and time period t (Scenario 3), both methods produce biased estimates of the true treatment effects. However, matching reduces the estimates bias significantly compared to the unmatched version, but whether this reduction is useful or not from a practical standpoint depends on the magnitude of the heterogeneity.

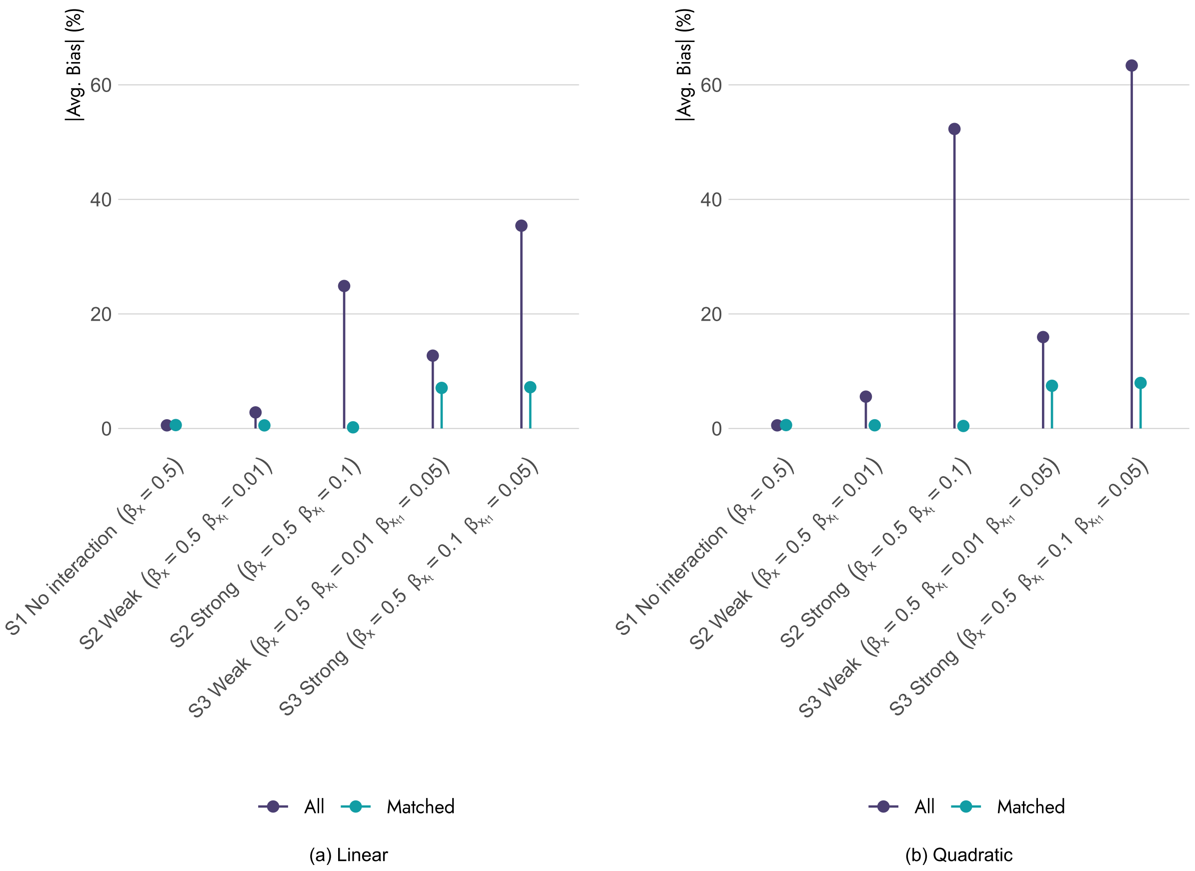


(a) Linear

(b) Quadratic

Figure 4: Event study estimates by time period (wrt T=4) for differential interaction between X and t

Figure 5 shows the average bias for the different scenarios presented in Table 1, under different values of β parameters. As previously described, matching on time-stable covariates is a useful strategy to complement a DD design and recover unbiased estimates when the violations of parallel trends is due to differences in covariate distribution between both groups. Even under other types of violations of parallel trends, like differential evolution over time between groups, matching could help reduce the magnitude of the bias, which also improves the usefulness of posterior sensitivity analyses.



(a) Linear

(b) Quadratic

Figure 5: Average bias by different data generating processes and scenarios

3.1. *Sensitivty Analysis*

One enticing characteristic of a DD analysis is that it allows us to conduct robustness tests to the main identification assumption by using the trend between the treatment and control group prior to the intervention ("pre-trends"). However, these tests tend to be under-powered and could provide a false sense of robustness. To overcome some of these limitations, I implement the approach proposed by Rambachan & Roth (2023) in the `HonestDiD` R package to include a sensitivity parameter for the test of parallel pre-trends.

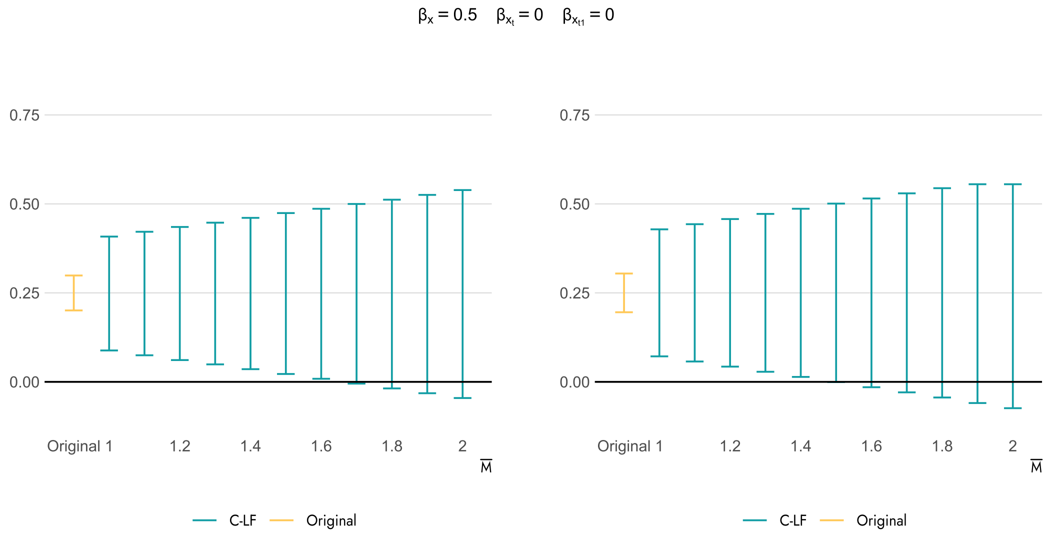
For assessing the robustness of both methods (DD with the entire sample and with the matched sample) under different DGPs, I focus on two different measures: (1) The value M^* of the relative magnitude (M) of pre-trends violations that change the qualitative conclusions of the study[3], and (2) the width of the sensitivity bounds.

[3] In this case, I assume that the qualitative conclusions change when the estimate is no longer statistically significant at a 5% confidence level.

The first measure tells us how large of a violation in the pre-parallel trends is allowed, while still maintaining the main conclusions of the study. The width of the bounds, on the other hands, also provide information about the robustness of the study, regardless of the point estimate.

One important issue that arises with this type of sensitivity analysis (and sensitivity analyses in general) is the fact that results are still skewed based on the magnitude of the bias for the point-estimate. For example, in the case of Rambachan & Roth (2023) `HonestDiD` relative magnitudes approach, the breakout value[4] M^* for M will be determined partly by the magnitude of the pre-trend violations but also by the magnitude of the point estimate. In this case, if the point-estimate is upwards-biased, results for the sensitivity analysis might lead us to incorrect conclusions.

If we observe the sensitivity bounds for the scenario where there is no bias in the point estimate (see Figure 2 (b)), then using the entire sample produces a slightly more robust finding (Figure 6). This is because in the case of no DD confounding, matching reduces the sample size, increasing the width of confidence intervals.



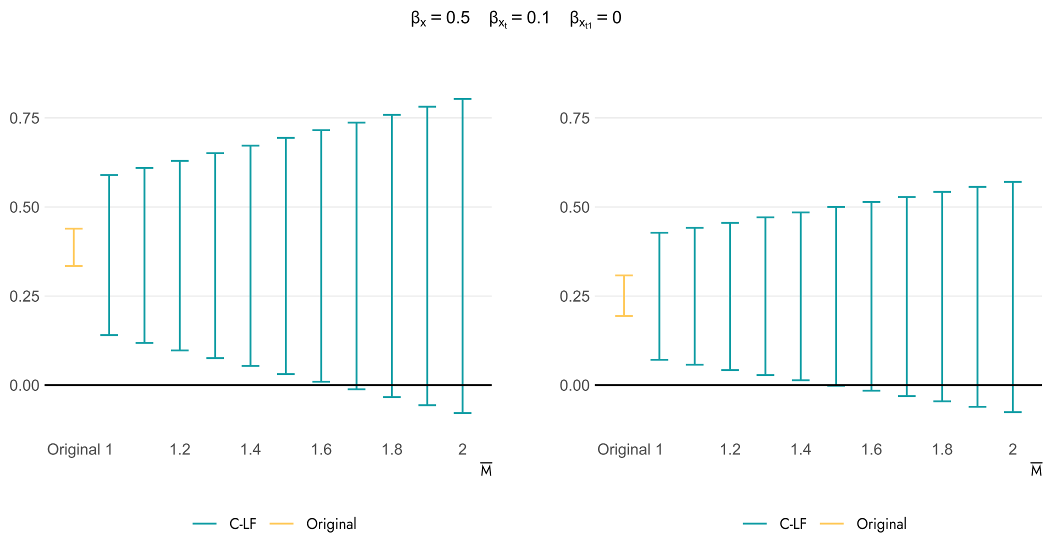
(a) Entire Sample

(b) Matched Sample

Figure 6: Relative Magnitude Sensitivity Bounds on relative magnitudes for Scenario 1 (quadratic)

Figure 7 shows the confidence intervals for a sensitivity analysis under scenario 2 (quadratic) of the previous simulations. This case also shows that the entire sample produces a more robust finding, seeing that the breakout value for M is higher compared to the matched sample. This result, however, is due to the upwards bias of the point estimate (Figure 3 (b)). The width of the CIs in Figure 7 (a) compared to Figure 7 (b) indicate that the first estimate is less robust due to trend violations in the pre-intervention period, but the increase in CI width does not overcome the impact the upwards bias has on this analysis.

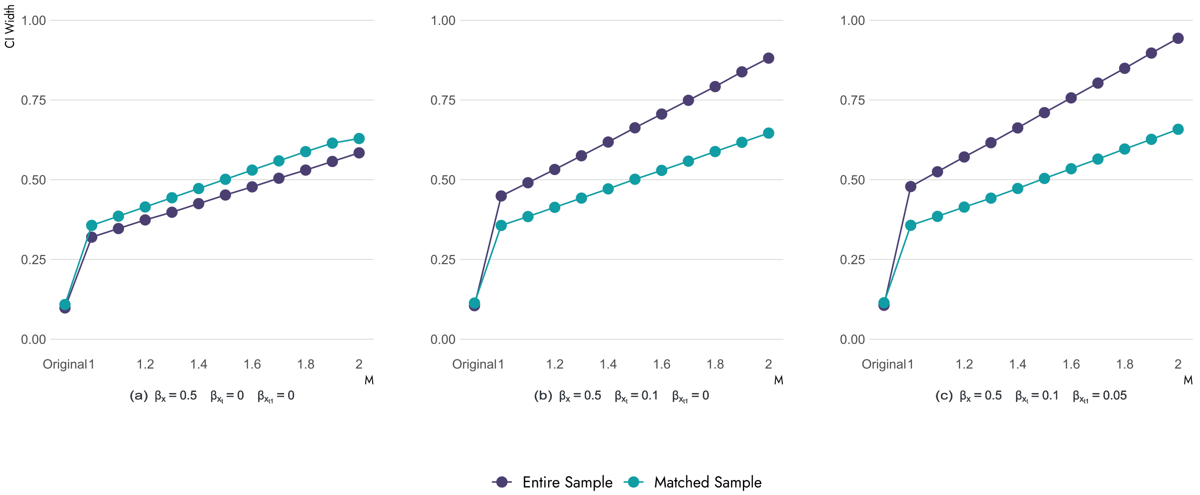
[4] The breakout value refers to the minimum value of M for which the confidence intervals include 0.



(a) Entire Sample

(b) Matched Sample

Figure 7: Relative Magnitude Sensitivity Bounds on relative magnitudes for Scenario 2 (quadratic)



(a) $\beta_x = 0.5 \quad \beta_{x_t} = 0 \quad \beta_{x_{t1}} = 0$

(b) $\beta_x = 0.5 \quad \beta_{x_t} = 0.1 \quad \beta_{x_{t1}} = 0$

(c) $\beta_x = 0.5 \quad \beta_{x_t} = 0.1 \quad \beta_{x_{t1}} = 0.05$

Figure 8: Confidence interval width for bounds on relative magnitudes (quadratic)

This is particularly salient in the case were there is no actual effect. Figure 9 shows the event study for scenario 2 (interaction between X and t) when there is *no treatment effect*. Under this magnitude of bias, if using the entire sample, we would assume that violations of the parallel trend assumption up to 60% of the magnitude of the violations in the pre-intervention period are allowed to still identify a positive treatment effect.

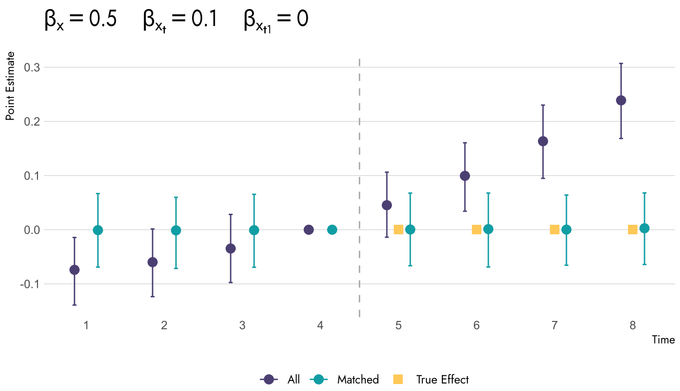
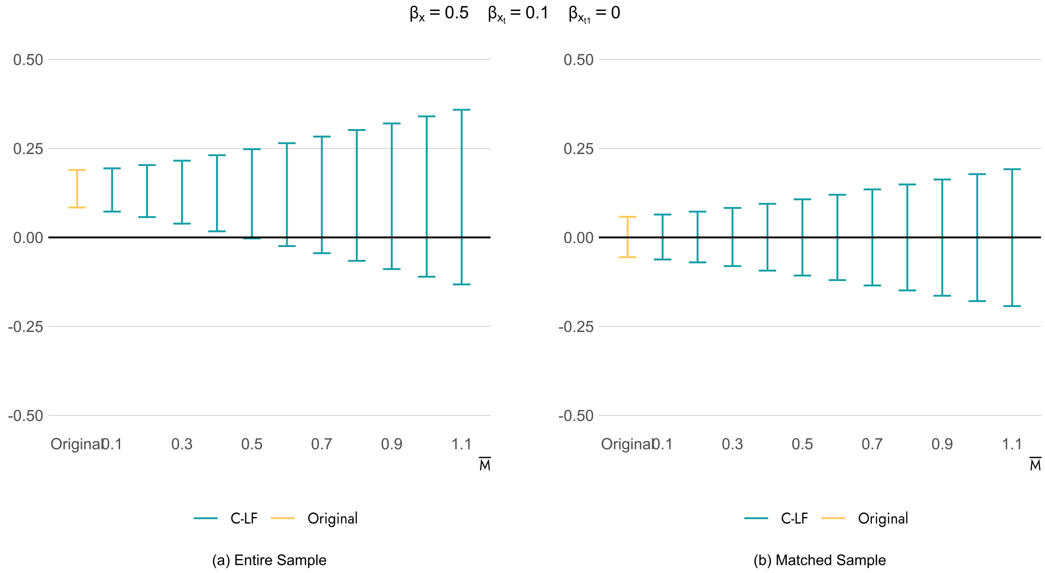


Figure 9: Event study estimates by time period (wrt T=4) for interaction between X and t (Quadratic)



(a) Entire Sample

(b) Matched Sample

Figure 10: Relative Magnitude Sensitivity Bounds on relative magnitudes for Scenario 2 (quadratic) - No effect

In the case that our estimates suffer from downward bias, even under proper sensitivity analyses, we would be more likely to conclude that there is no significant effect, even if there is, increasing our Type II error in this case.

In terms of hypothesis testing, the following figures show the proportion of rejections for the null hypothesis $H_0 : \hat{\tau}_{ATT} = 0$ in the case were there is *no treatment effect* ($\tau_t = 0 \ \forall \ t \geq T_0$). Figure 11 shows the rejection rate for 200 simulations for each value of β_{x_t}, from 0 to 0.3 (x axis)[5]. As we can see, while the matched DD provides a low (or even null) rate of rejection for the null hypothesis, even under modest bias ($\beta_{x_t} = 0.02$), we would reject the null around 20% of the time under no true treatment effect.

[5] In this case, the same parameters as shown in Table 1 - *Quadratic*: Scenario 2 are used, fixing $\beta_x = 0.5$ and $\beta_{x_{t1}} = 0$.

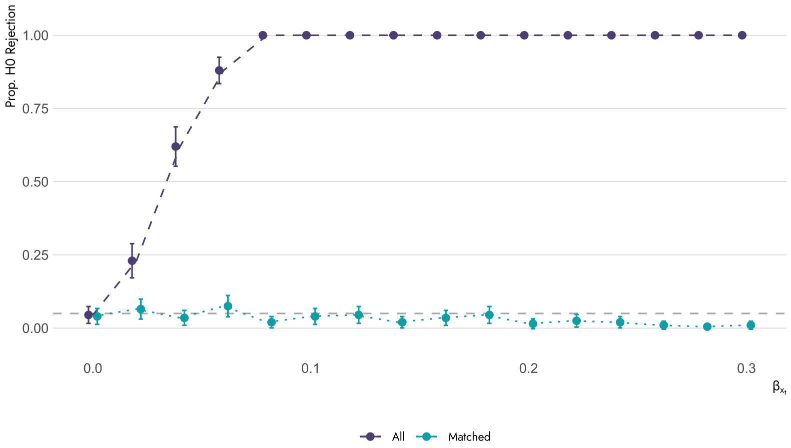


Figure 11: Rejection rate of null hypothesis for different values of β_{x_t}

Figure 12 shows similar plots for different values of relative magnitude bias M for sensitivity analysis purposes. In this case, even when conducting a sensitivity analysis to check the robustness of our findings, we would reject the null hypothesis 25% of the times under moderate bias ($\beta_{x_t} = 0.12$) when considering violations up to 60% of the pre-intervention period trend.

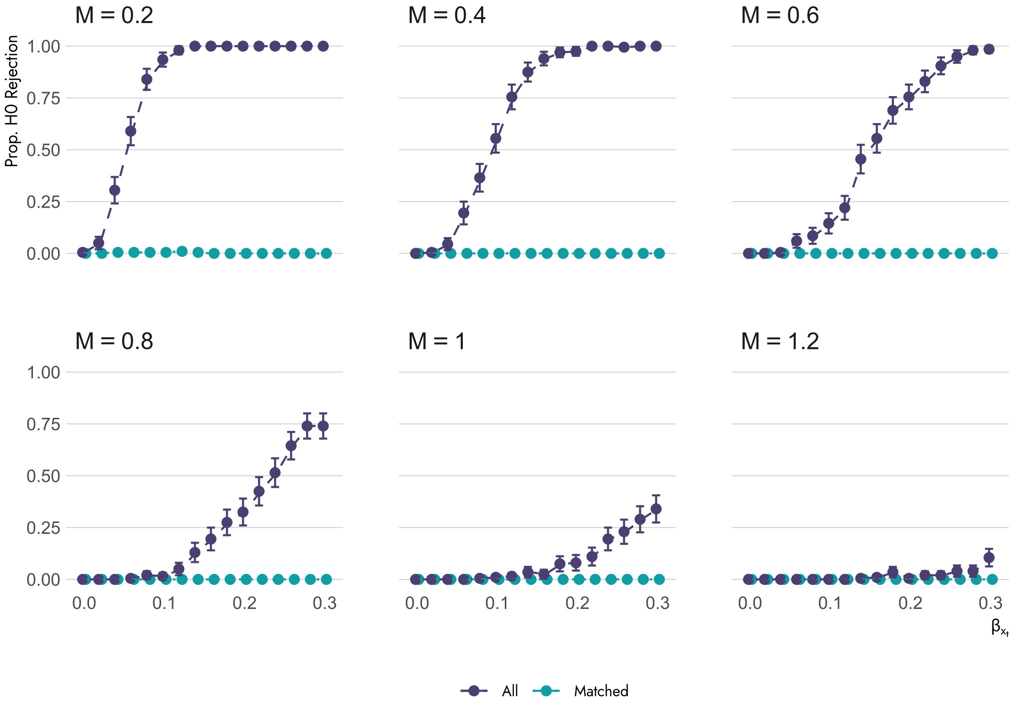
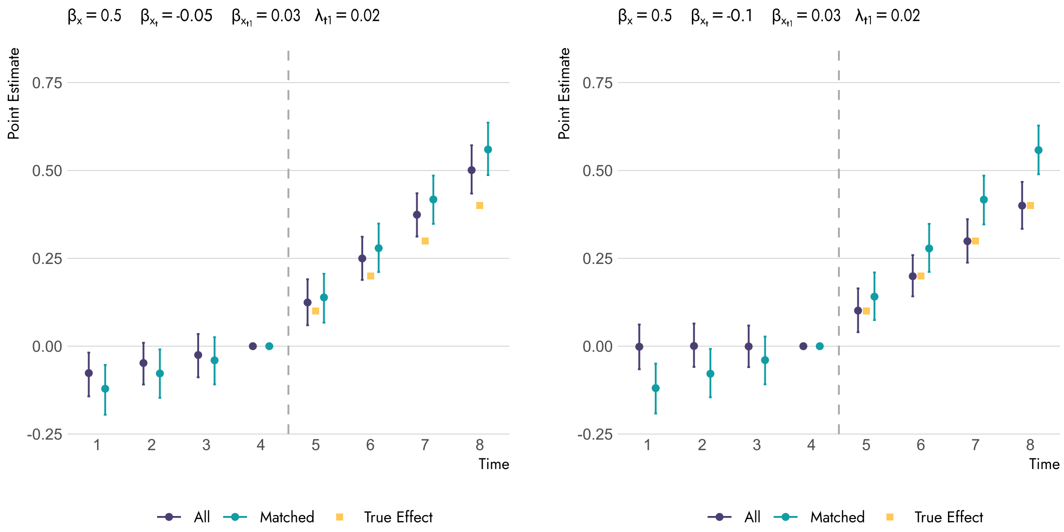


Figure 12: Rejection rate of null hypothesis for different values of β_{x_t} and Relative Magnitude Bias M

These results show how sensitivity analysis can be a powerful tool for assessing the robustness of our findings, but its usefulness can be skewed by the presence of bias in our point estimates. In this case, matching can help reduce bias and still allow us to construct informative bounds for DD findings under violations of the parallel trend assumption, as long as there is bias reduction in the pre-intervention period as well.

3.2. When is matching in a DD not useful?

As it was previously mentioned, matching can be an overall good strategy for bias reduction, as long as other unobserved biases do not counteract the effect of the observed bias. For example, Figure 13 shows a scenario[6] where there is no overall reduction in the pre-intervention period, and matched estimates have, in fact, higher bias than the unmatched setting. Figure 13 (a) shows the setting where both DD and matched DD are biased, but the bias is larger after matching, and Figure 13 (b) shows the particular scenario where different biases actually cancel each other. Then, analyzing the bias in the pre-intervention period can be useful for knowing how to proceed with the analysis.



(a) Larger bias in matched DD vs Smaller bias in unmatched DD

(a) Larger bias in matched DD vs No bias in unmatched DD

Figure 13: Event study estimates by time period (wrt T=4) for differential interaction between X and t and bias cancellation

3.3. MIP Matching vs Propensity Score matching

Another important decision when deciding to combine matching a DD strategy is what *kind* of matching to use and how to match different units. I argue that in this setup, MIP matching has some desirable advantages over Propensity Score (PS) matching, one of the most popular matching methods used in social sciences, such as being able to balance covariates directly while obtaining the largest matched sample possible under a specified set of balancing constraints.

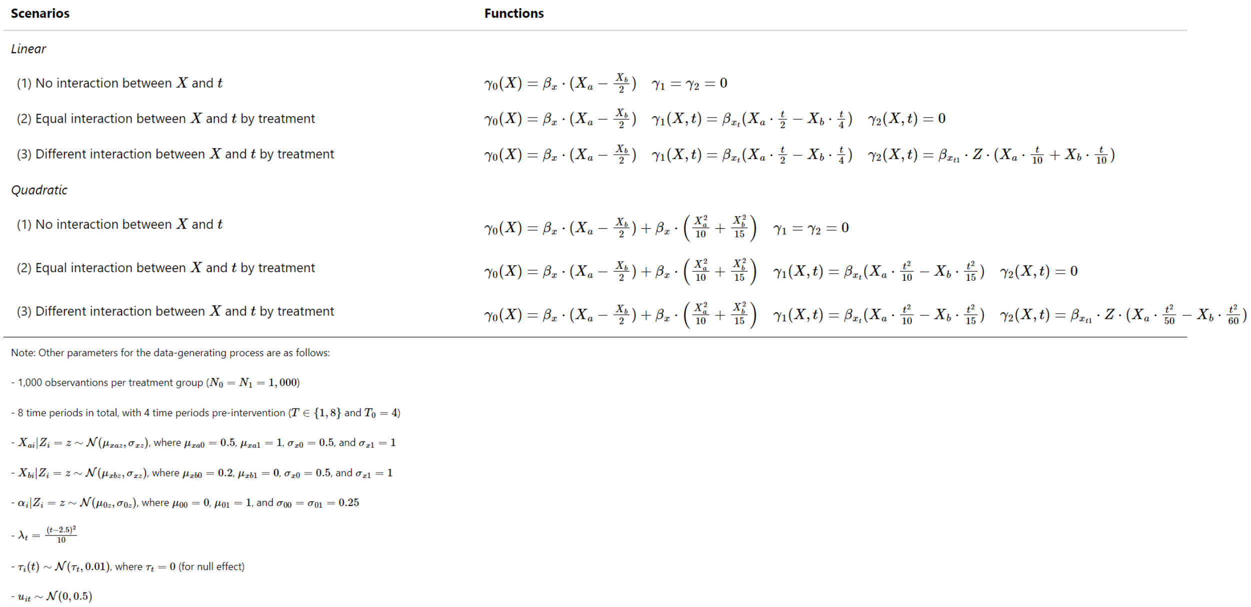
Conceptually, PS matching matches units using the estimated probability of belonging to the treatment group (i.e. $Z = 1$) given a set of observed covariates X. Matching on one dimension (distance) instead on multiple characteristics has the advantage of reducing the dimensionality of the problem while, *in expectation*, balancing covariates X (Rosenbaum & Rubin, 1983). However, PS matching is not able to balance covariates directly through distance matching, and other restrictions need to be imposed — like a caliper for the distance metric or a specific caliper for each covariate — in order to achieve specific balancing constraints.

[6] DGP in this case is $Y = \alpha + \beta_x \cdot x + 0.1 \cdot t \cdot Z + \beta_{xt} \cdot x \cdot t + \beta_{xt_1} \cdot x \cdot t \cdot Z + \mathrm{I}(t > T_0) \cdot \tau + u$.

MIP matching, on the other hand, is able to set balancing restrictions at the covariate level (e.g. restricted mean balance, fine balance, etc.) while still identifying the largest possible matched set to improve the efficiency of our estimates (Zubizarreta, Paredes, & Rosenbaum, 2014).

To show the difference between MIP matching and similar PS matching methods, I conduct similar simulation scenarios as before, but now including two covariates for matching, X_a and X_b (see Table 2). In this case, I will use three different matching procedures: (1) MIP matching using a 0.01 SD mean balance restriction, (2) PS matching using a 0.01 SD caliper on the distance metric, and (3) PS matching using a 0.01 SD caliper on each covariate, X_a and X_b.

Table 2: Data generating process for different simulation scenarios for comparing MIP and PS matching



Scenarios	**Functions**
Linear	
(1) No interaction between X and t	$\gamma_0(X) = \beta_x \cdot (X_a - \frac{X_b}{2}) \quad \gamma_1 = \gamma_2 = 0$
(2) Equal interaction between X and t by treatment	$\gamma_0(X) = \beta_x \cdot (X_a - \frac{X_b}{2}) \quad \gamma_1(X,t) = \beta_{x_t}(X_a \cdot \frac{t}{2} - X_b \cdot \frac{t}{4}) \quad \gamma_2(X,t) = 0$
(3) Different interaction between X and t by treatment	$\gamma_0(X) = \beta_x \cdot (X_a - \frac{X_b}{2}) \quad \gamma_1(X,t) = \beta_{x_t}(X_a \cdot \frac{t}{2} - X_b \cdot \frac{t}{4}) \quad \gamma_2(X,t) = \beta_{x_{t1}} \cdot Z \cdot (X_a \cdot \frac{t}{10} + X_b \cdot \frac{t}{10})$
Quadratic	
(1) No interaction between X and t	$\gamma_0(X) = \beta_x \cdot (X_a - \frac{X_b}{2}) + \beta_x \cdot \left(\frac{X_a^2}{10} + \frac{X_b^2}{15}\right) \quad \gamma_1 = \gamma_2 = 0$
(2) Equal interaction between X and t by treatment	$\gamma_0(X) = \beta_x \cdot (X_a - \frac{X_b}{2}) + \beta_x \cdot \left(\frac{X_a^2}{10} + \frac{X_b^2}{15}\right) \quad \gamma_1(X,t) = \beta_{x_t}(X_a \cdot \frac{t^2}{10} - X_b \cdot \frac{t^2}{15}) \quad \gamma_2(X,t) = 0$
(3) Different interaction between X and t by treatment	$\gamma_0(X) = \beta_x \cdot (X_a - \frac{X_b}{2}) + \beta_x \cdot \left(\frac{X_a^2}{10} + \frac{X_b^2}{15}\right) \quad \gamma_1(X,t) = \beta_{x_t}(X_a \cdot \frac{t^2}{10} - X_b \cdot \frac{t^2}{15}) \quad \gamma_2(X,t) = \beta_{x_{t1}} \cdot Z \cdot (X_a \cdot \frac{t^2}{50} - X_b \cdot \frac{t^2}{60})$

Note: Other parameters for the data-generating process are as follows:

- 1,000 observantions per treatment group ($N_0 = N_1 = 1,000$)
- 8 time periods in total, with 4 time periods pre-intervention ($T \in \{1, 8\}$ and $T_0 = 4$)
- $X_{ai}|Z_i = z \sim \mathcal{N}(\mu_{xaz}, \sigma_{xz})$, where $\mu_{xa0} = 0.5$, $\mu_{xa1} = 1$, $\sigma_{x0} = 0.5$, and $\sigma_{x1} = 1$
- $X_{bi}|Z_i = z \sim \mathcal{N}(\mu_{xbz}, \sigma_{xz})$, where $\mu_{xb0} = 0.2$, $\mu_{xb1} = 0$, $\sigma_{x0} = 0.5$, and $\sigma_{x1} = 1$
- $\alpha_i|Z_i = z \sim \mathcal{N}(\mu_{0z}, \sigma_{0z})$, where $\mu_{00} = 0$, $\mu_{01} = 1$, and $\sigma_{00} = \sigma_{01} = 0.25$
- $\lambda_t = \frac{(t-2.5)^2}{10}$
- $\tau_i(t) \sim \mathcal{N}(\tau_t, 0.01)$, where $\tau_t = 0$ (for null effect)
- $u_{it} \sim \mathcal{N}(0, 0.5)$

Table 3 shows the means of X_a and X_b for the treatment and the control group *after matching* for all three procedures, including also the difference between groups and the average matched units per group for 1,000 simulations. We can see that MIP matching obtains the largest sample while maintaining the balance restriction. Using a caliper for each covariate, unlike MIP matching, yields a very small matched sample.

Table 3: Covariate balance between MIP matching, PS matching using caliper for distance, and PS matching using caliper for each variable

(a) MIP matching

Covariates	Control	Treat	Diff
xa	0.6635	0.6711	0.0076
xb	0.1371	0.1297	−0.0074
N matched (avg)	797	797	·

(b) PS matching (distance caliper)

Covariates	Control	Treat	Diff
xa	0.5818	0.5832	0.0014
xb	0.1684	0.1678	−0.0006
N matched (avg)	562	562	·

(c) PS matching (covariates caliper)

Covariates	Control	Treat	Diff
xa	0.6074	0.6074	0.0000
xb	0.1668	0.1668	0.0000
N matched (avg)	28	28	·

Finally, Table 4 shows the average bias for the ATT in the post-intervention period, including the 95% confidence intervals.[7] Similarly to prior simulations, when there are no differential trends across time for both groups (Table 4 (a)), all methods are unbiased. When there is a differential time trend between treatment and control group due to observed covariates (Table 4 (b)), MIP matching estimates have the lowest bias and are the most precise out of the other two matching methods tested. Finally, under time-varying trends that are different by group, all estimates are biased, though MIP and PS matching reduce the amount of bias significantly. It is important to note that in all cases, confidence intervals for PS matching using a caliper on covariates produce very wide intervals that would not be informative.

[7] In these cases, for Scenario 1: $\beta_x = 0.5$, $\beta_{x_t} = 0$ and $\beta_{x_{t1}} = 0$. For Scenario 2: $\beta_x = 0.5$, $\beta_{x_t} = 0.1$ and $\beta_{x_{t1}} = 0$, and for Scenario 3: $\beta_x = 0.5$, $\beta_{x_t} = 0.1$ and $\beta_{x_{t1}} = 0.05$

Table 4: Average bias for no treatment effect between MIP matching, PS matching using caliper for distance, and PS matching using caliper for each variable

(a) Scenario 1 - Quadratic

Method	Bias	95% CI
All	0.0006	[−0.0619, 0.0632]
MIP Match	0.0006	[−0.0691, 0.0702]
PS Match (dist)	0.0009	[−0.081, 0.0827]
PS Match (each var)	0.0035	[−0.3674, 0.3744]

(b) Scenario 2 - Quadratic

Method	Bias	95% CI
All	0.1002	[0.0338, 0.1665]
MIP Match	0.0004	[−0.0708, 0.0717]
PS Match (dist)	0.0011	[−0.0833, 0.0855]
PS Match (each var)	0.0028	[−0.3645, 0.3701]

(c) Scenario 3 - Quadratic

Method	Bias	95% CI
All	0.126	[0.0622, 0.1898]
MIP Match	0.0144	[−0.0518, 0.0806]
PS Match (dist)	0.0109	[−0.0725, 0.0943]
PS Match (each var)	0.0096	[−0.3612, 0.3804]

4. Application: Preferential School Vouchers in Chile

4.1. Context

School vouchers are usually seen as a contentious policy in the educational world. While its advocates argue that vouchers increase the choice set of schools for parents and improve quality through competition, its detractors highlight some of its negative consequences, such as "cream skimming" and enhanced segregation, attracting higher-ability and higher-income students to specific schools (Epple et al., 2002; Epple & Romano, 2008; Hsieh & Urquiola, 2006; Urquiola, 2016). Targeted voucher schemes, or preferential vouchers, intend to correct some of these pernicious effects by allocating more resources to vulnerable students. By recognizing the differential cost of educating different types of students, targeted vouchers should ameliorate some of the unintended consequences that are usually reported in the voucher literature, particularly in terms of segregation. However, the design and implementation of these subsidies policies play an important role in the end result, and depending on the incentives schemes that are provided, could actually promote increasing segregation between schools.

In this application, I tackle the specific question about potential unintended consequences of educational voucher policies, assessing whether socioeconomic and income segregation increased due to the implementation of a new school subsidy. I focus on the introduction of the Chilean preferential voucher scheme in 2008, and by using a DD approach combined with matching, I measure to what extent socioeconomic diversity changed between schools which opted into the program versus those which did not.

The issue of school segregation is particularly important in the Chilean context. Chile is a highly stratified and segregated country, due mainly to its great income inequality. In fact, Chile was ranked the 21st most unequal nation (out of 147) according to its the Gini coefficient in 2015 World Bank (2015)[8]. Such income inequality is not only replicated but also enhanced by its educational system, as educational opportunities are closely related to the student's socioeconomic status (Arteaga et al., 2016; OECD, 2014).

Most research regarding the introduction of the preferential voucher scheme in Chile relates to the impact of the policy on academic performance (Correa et al., 2014; Feigenberg et al., 2019; Mizala & Torche, 2013; Navarro-Palau, 2017) or school choice quality (Aguirre, 2022; Neilson, 2021). The objective of this application is to contribute to the previous literature by providing evidence of the effect of preferential vouchers on a different dimension, assessing the impact of a potential externality on schools' composition and potential reduction in socioeconomic diversity within institutions.

4.1.a. *Chilean Educational System:* The Chilean educational system is composed of three types of schools: (i) public, (ii) private-subsidized (voucher), and (ii) private-unsubsidized schools (non-voucher). Public schools are funded by the Government and were operated by municipalities,[9] receiving resources per student and also base contributions. Private-subsidized schools, or voucher schools, are privately owned and managed institutions, run either by for-profit organizations, NGOs, or other non-profit associations. They receive public funding through vouchers, so their resources are linked to students' enrollment and attendance. These schools are also allowed to charge add-ons (copayment) to parents, which usually varied between US$5 and US$140 a month in 2008. Finally, private-unsubsidized schools are privately owned, managed, and funded.

In the 1990s, public schools were the ones that concentrated the largest amount of students. However, there has been a constant decrease in public enrollment since the mid-90s, and since 2007, voucher schools have the lead in enrollment. Figure 14 shows the evolution of enrollment by type of school between 1990 and 2012.

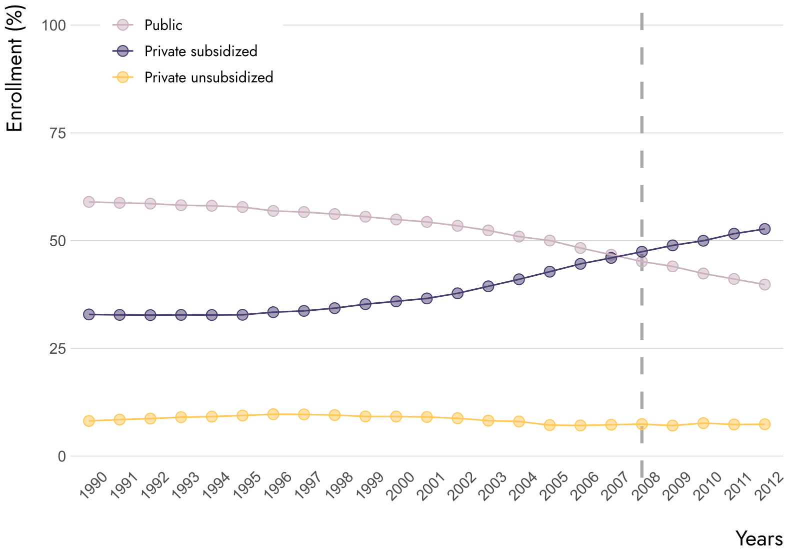


Figure 14: Enrollment from 1990 to 2011 by school depedence

4.1.b. *Preferential Schooling Subsidy:* The preferential schooling subsidy (SEP) was introduced in 2008, and its objective was to allocate more resources to vulnerable students. The amount of the increase of the voucher was

[8] Ranking of countries with Gini coefficients data for the 2006-2015 period

[9] In 2018, the Inclusion Law (``Ley de Inclusion'') started a shift in public schools' management from municipalities to local educational services, directly dependent from the central government.

significant compared to the previously universal flat voucher: Preferential vouchers represented approximately a 50% increase with respect of the previous amount per child, increasing from US$70 to US$105 in 2008, on average (Romaguera & Gallegos, 2010). In addition to the extra resources given for each vulnerable student, schools received a concentration bonus associated with the percentage of these type of students that they enrolled, which ranged between US$2.5 per student for schools with a concentration between 15% and 30% of vulnerable students, to US$6.5 per student for schools with over 60% of vulnerable students. Thus, this concentration bonus increased the complementary voucher per student by an additional 20% for schools with higher shares of vulnerable students.

As it was previously mentioned, the SEP policy was focused specifically on vulnerable students or "priority" students . Priority students were identified using a means test called *Ficha de Protección Social* (Social Protection Questionnaire), which was a form all households had to complete in order to apply for Government welfare. The questionnaire included different socioeconomic variables, such as income, household composition, educational attainment of the head of the household, assets, and housing conditions, among others, giving households a score which was then used to identify vulnerable students (the lower 40% of the distribution).[10]

Schools had the option to opt-in to this program in order to receive the extra resources for the vulnerable students that were enrolled in their institution, meaning that a key feature of this policy is that it was voluntary. However, in order to receive these funds, schools also had to comply with certain Government requirements: (i) accountability, (ii) no discrimination, and (iii) educational quality.

SEP schools (i.e. schools which opted into the program) had to present a Plan for Educational Improvement (PEI), which limited the use of SEP funds to activities or resources that directly benefited vulnerable students. This clause was put in place to prevent diversion of SEP funds to other activities, especially in for-profit schools. Thus, technically, school administrators cannot obtain revenues from SEP subsidies[11].

Additionally, SEP schools had to sign a clause of no selection and no add-on fees for priority students. In Chile, by law, schools that receive public funding cannot select students before 7th grade, but as there is little monitoring, there is overwhelming evidence that schools do incur in selection practices. In fact, over 50% of parents who enrolled their child in voucher schools in 2007 claimed that an admission test was one of the enrollment requisites. Regarding the add-on fees or copayment, there is no clear data for the amount that schools charged before 2015,[12] but from the amounts self-reported by schools, two thirds of voucher schools declared charging under US$20, while 60% declare being free of charge in 2007. Thus, for the majority of private-subsidized schools, the complementary voucher would more than cover their expected add-on.

Finally, government oversight was directly linked to standardized test results. From 2009, SEP schools were classified according to their scores; the better they performed, the less constraints they had in the use of resources (Romaguera & Gallegos, 2010).

Given the incentive structure that the SEP policy provided for schools, it is easy to see that there was a clear trade-off between costs and benefits of adhering to the policy. Private subsidized schools that opted into the new voucher scheme would receive more funding per vulnerable student in terms of subsidies, but, at the same time, would lose additional copayment for those same students if the school had add-ons in place. Schools that concentrated a larger proportion of priority students would also receive more funding, giving an incentive to potentially enroll more of these students *conditionally on subscribing to the SEP policy*. The new resources

[10] There were additional ways that a student could be classified as "priority": If his or her family participated of the socioeconomic program "Chile Solidario" or if the student lived in a high-poverty area. These conditions, however, where applied in the absence of a Social Protection Questionnaire, and had to be re-validated the following year (Romaguera & Gallegos, 2010).

[11] During the first years of the implementation of the policy, there was no way of knowing to what extent this clause was enforced, but sources from the Ministry of Education acknowledge the difficulty to monitor the use of such funds.

[12] Schools were not obligated to report it and, in many cases, it varied between students

also came tied to additional oversight from the government and performance expectations, which also involved additional costs for the school.

4.2. Data

Given that the focus of interest is analyzing the effect of the introduction of the SEP policy on school which opted into the new voucher scheme, I use a complete-cases sample of private subsidized schools between 2005 and 2012. This corresponds to 72% of all private subsidized schools in Chile in 2007.[13]. I choose to exclude public schools from the analysis, given that all of them joined the policy in 2008, and do not face the same cost-benefit trade-offs that private subsidized schools do.

Figure 15 compares the evolution in the number of private subsidized schools throughout the years, according to whether they subscribed to the SEP policy or not within the study period. The difference between number of schools maintains mainly constant during the 2005-2012 period, with a slightly higher rate for SEP schools between the years 2006 and 2008.

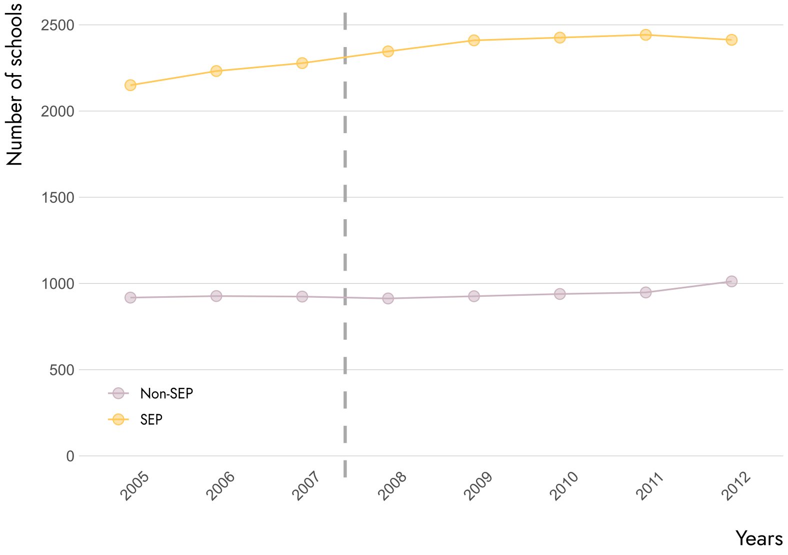


Figure 15: Number of private subsidized schools with primary education by year according to whether they subscribed to the SEP policy between 2008 and 2012 or not

For our sample, we observe that most schools that enrolled in the SEP policy did so during the first year (70%), with few schools enrolling two or three years after.

[13] As a reference, from the 2,833 private subsidized schools that were open in 2007 (used for matching), nearly 90% of them were open since 2005.

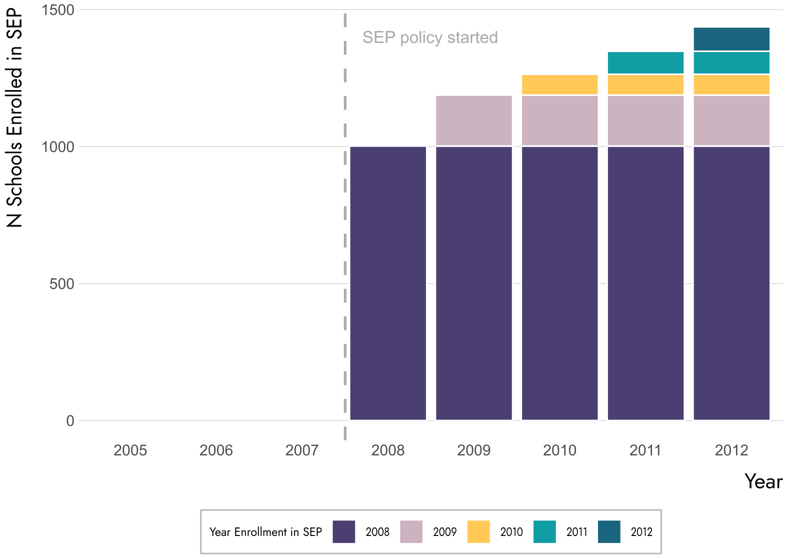


Figure 16: Enrollment in the SEP policy in study sample by year

In terms of schools' characteristics, I use publicly available data from the Chilean Ministry of Education (MINEDUC, 2020) related to the preferential voucher scheme, school subsidies, and other school attributes, such as location and enrollment. Additionally, I use the results from the Education Quality Measurement System (SIMCE), which contains standardized test scores for each school in 4th grade and socioeconomic information from a parents' questionnaire. In terms of outcomes, I focus on average household income at the school level as a socioeconomic component over time, as well as average SIMCE scores for Math and Language to measure academic performance of the school.

In total, my sample consists of a panel of 2032 different schools from 2005 to 2012, where 1437 of them subscribed to the SEP policy during the period of analysis.

4.3. Results

Using Mixed Integer Programming (MIP) Matching (Zubizarreta, 2012), I match exactly on the school's province department,[14] to adjust for geographic location. Additionally, I use restricted mean balance (0.025 SD) on the following characteristics: average grade enrollment, average school subsidy, whether the school charges parents add-ons or not (and the amount of the copayment), vulnerability status of the school, and number of schools in the municipality. These characteristics are likely to capture schools that are subjected to similar competition pressures and funding, which are mainly unchanged during the pre-intervention period.

As it is discussed in Daw & Hatfield (2018a), researchers also need to take into account the specific context of the problem to assess whether it is plausible to assume both populations of treated and control units come from the same population or not. In the case of the SEP policy, given that all private subsidized schools could opt into the program, it is reasonable to assume that both groups stem from the same population, and that the probability of adhering to the intervention depends on pre-intervention characteristics, such as performance and competition. These covariates, however, are practically fixed at the school level, with a serial correlation of over 0.95 in the pre-intervention period for each matching covariate, and over 0.9 if pre- and post-intervention years are included.

[14] Province departments are educational geographic areas that are smaller that geographic provinces, but larger than municipalities.

Using the `designmatch` R package Zubizarreta et al. (2018), I match schools in 2007 (one year before the intervention started) that subscribed to the policy in 2008-2012 to those that did not. I use the last year prior to the introduction of the intervention, because I estimate all post-intervention effects with respect to that particular year. I use cardinality matching to obtain the largest matched sample possible under the balancing constraints that include exact matching for geographic province and restricted mean balance at 0.025 SD for enrollment, average yearly subsidy, whether the school charges parents' a copayment and how much copayment, and number of private subsidized schools in the same municipality.

Under the previous balancing restrictions, I am able to match 578 schools (289 pairs), and compare their outcomes and characteristics between 2005 and 2012 using a matched panel.

4.3.a. *Balance in the matched sample:* In terms of geographic coverage, most matched schools belong to the metropolitan region (59%), while the majority of the other schools are located in other central regions (see Figure 17). The matched sample also covers 16 out of the 53 provincial departments, representing areas that have a larger number of schools.

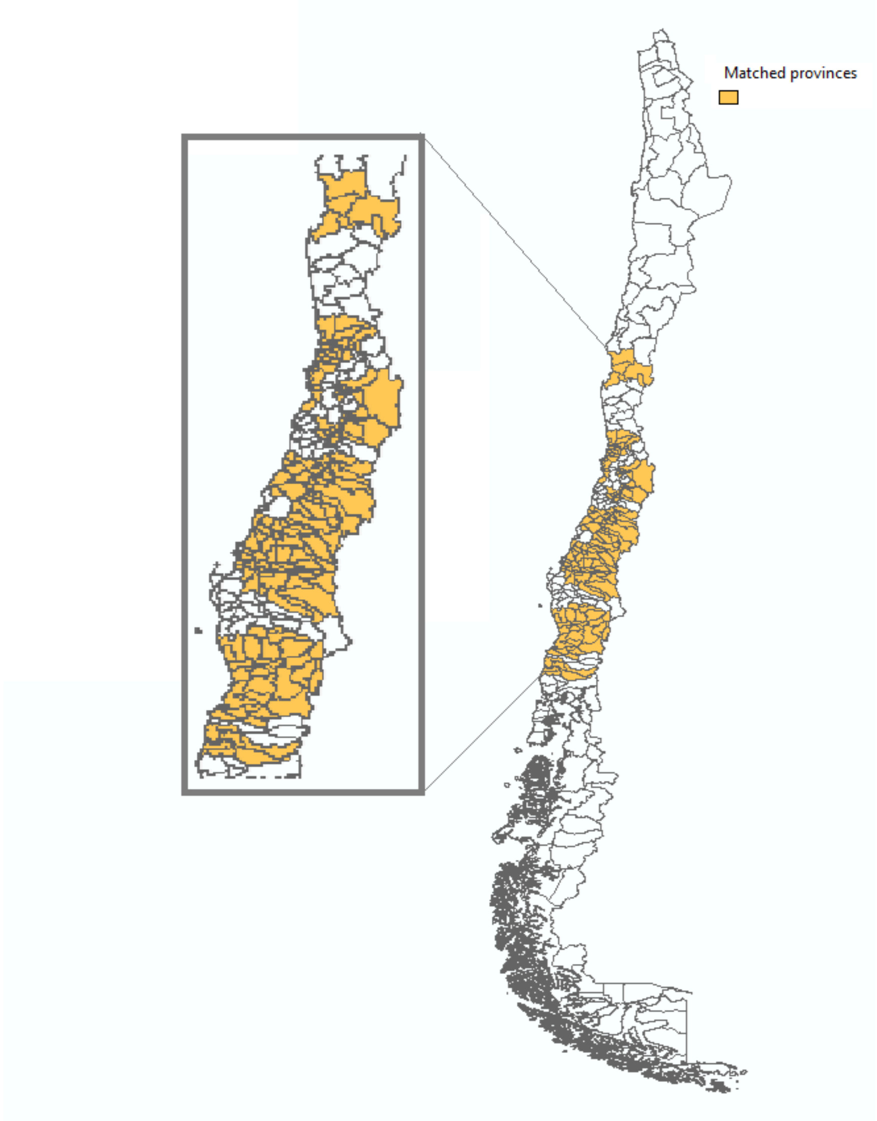


Figure 17: Chilean provinces in the matched sample

Table 5 shows the covariate balance in 2007, both before and after matching. As it can be seen from that previous table, both groups had substantial differences, especially in terms of charging copayment. While 89% of schools that did not subscribe to the policy charged additional add-ons, only 49% of the schools that did enroll in the SEP policy asked for copayment. However, after matching, differences in covariates are substantially reduced, with differences that are even smaller than the original 0.025 SD restriction.

Table 5: Covariate profile of the 2007 sample of private subsidized schools that opted into the SEP policy between 2008-2012 and those that did not

(a) All schools in the sample

Variables	SEP schools (All)	Non-SEP schools (All)	Diff (SD)
Avg. grade enrollment	42.41	50.76	-0.23
Avg. subsidy (1,000 CL$)	15083.53	19873.2	-0.33
Has copayment	0.49	0.89	-0.95
Copayment < 0.5 USE	0.25	0.09	0.43
Copayment \>= 0.5 USE	0.24	0.79	-1.34
Reported copayment (parents)	7853.83	25913.13	-1.39
Num. schools in municipality	49.22	60.01	-0.33
Group A (most vulnerable)	0.17	0.01	0.57
Group B (second most vulnerable)	0.24	0.03	0.66

(b) Matched schools

Variables	SEP school (matched)	Non-SEP schools (matched)	Diff (SD)
Avg. grade enrollment	49.91	49.85	0
Avg. subsidy (1,000 CL$)	18088.95	18018.58	0
Has copayment	0.84	0.84	0
Copayment < 0.5 USE	0.15	0.15	0
Copayment \>= 0.5 USE	0.69	0.69	0
Reported copayment (parents)	16841.56	16971.57	-0.01
Num. schools in municipality	58.86	58.9	0
Group A (most vulnerable)	0.02	0.02	0
Group B (second most vulnerable)	0.06	0.06	0

These differences are also portrayed in Figure 18, where the absolute standardized differences in mean are plotted before and after matching. It can be seen that after-matching differences are well below the 0.025 SD threshold indicated in the balancing constraints.[15]

[15] Even though other pre-intervention years were not used for matching, given the stability of the covariates over time, the differences in 2005 and 2006 between the treatment and control groups are still under 0.025 SD.

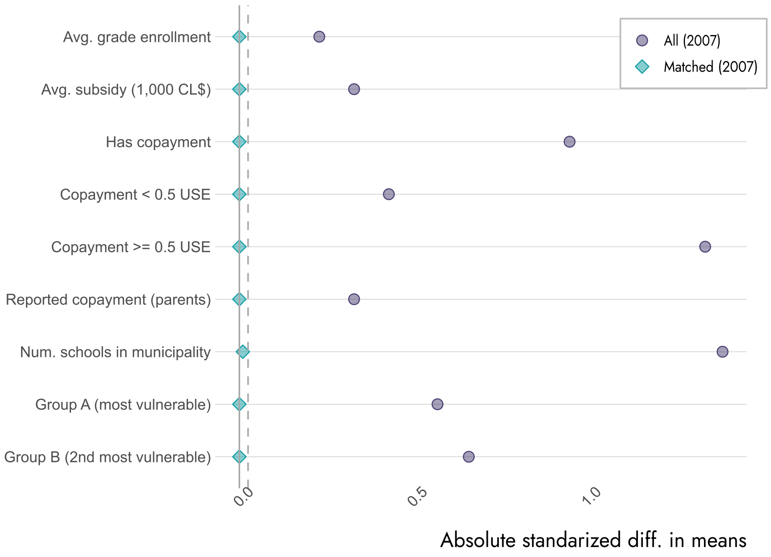


Figure 18: Absolute standardized differences in covariates between 2007 SEP and non-SEP schools before and after matching

4.3.b. *Event study results:* Figure 19 and Figure 20 show the event study estimates for the entire sample and the matched sample for two outcomes: average household income and average SIMCE score. For the pre-intervention period 2005-2007, even though there is suggestive evidence of pre-trends in the entire sample, this issue is minimized when using the matched sample (see Figure 19 (b) and Figure 20 (b)). It is important to note that schools were only matched in 2007 characteristics that are time invariant (or with high temporal correlation) and previous outcomes were not used as covariates, so the fact that the differences are held constant for 2005 and 2006 is not necessarily a mechanical construction of the matching process.

In terms of the outcomes of interest, Figure 19 shows that after the policy was implemented, schools that opted into the policy enrolled an even larger proportion of lower-income students, increasing socioeconomic differences between schools. On the other hand, Figure 20 shows evidence that schools that opted into the new voucher scheme were able to close the gap between schools that did not (see Figure 21 (b)), suggesting a positive effect of the policy in terms of academic outcomes.

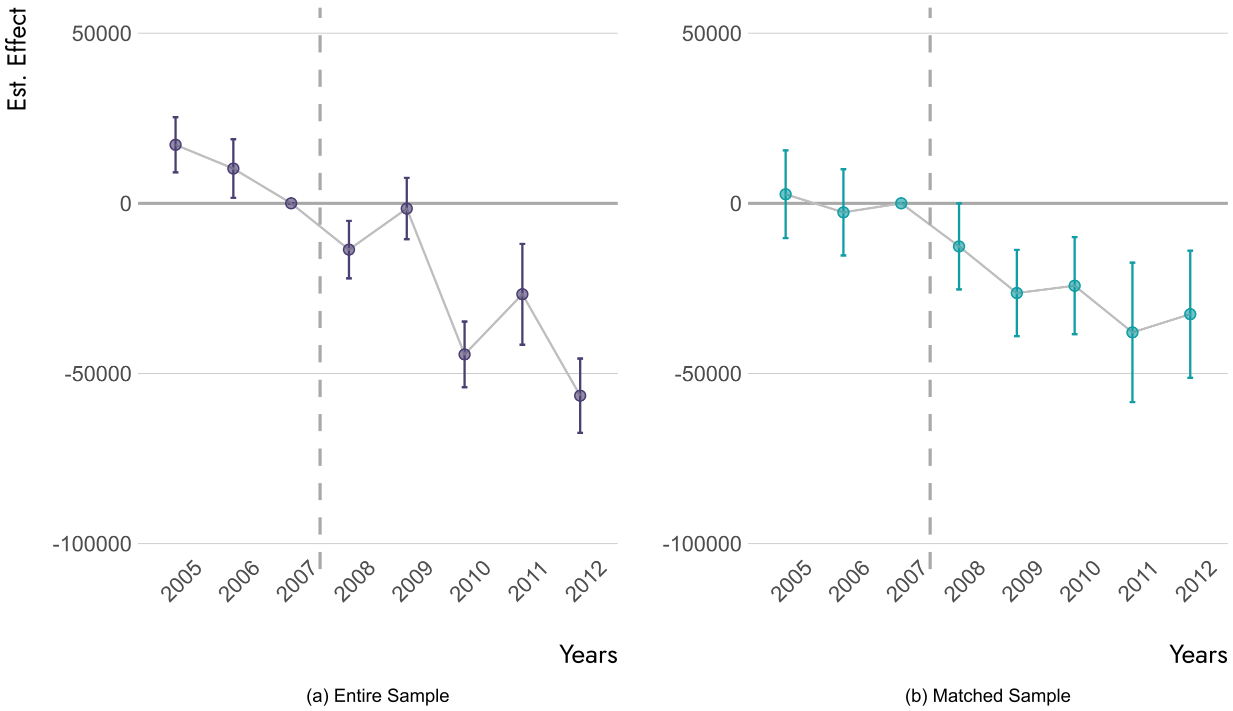


Figure 19: Event study estimates for household income (1,000 CL$)

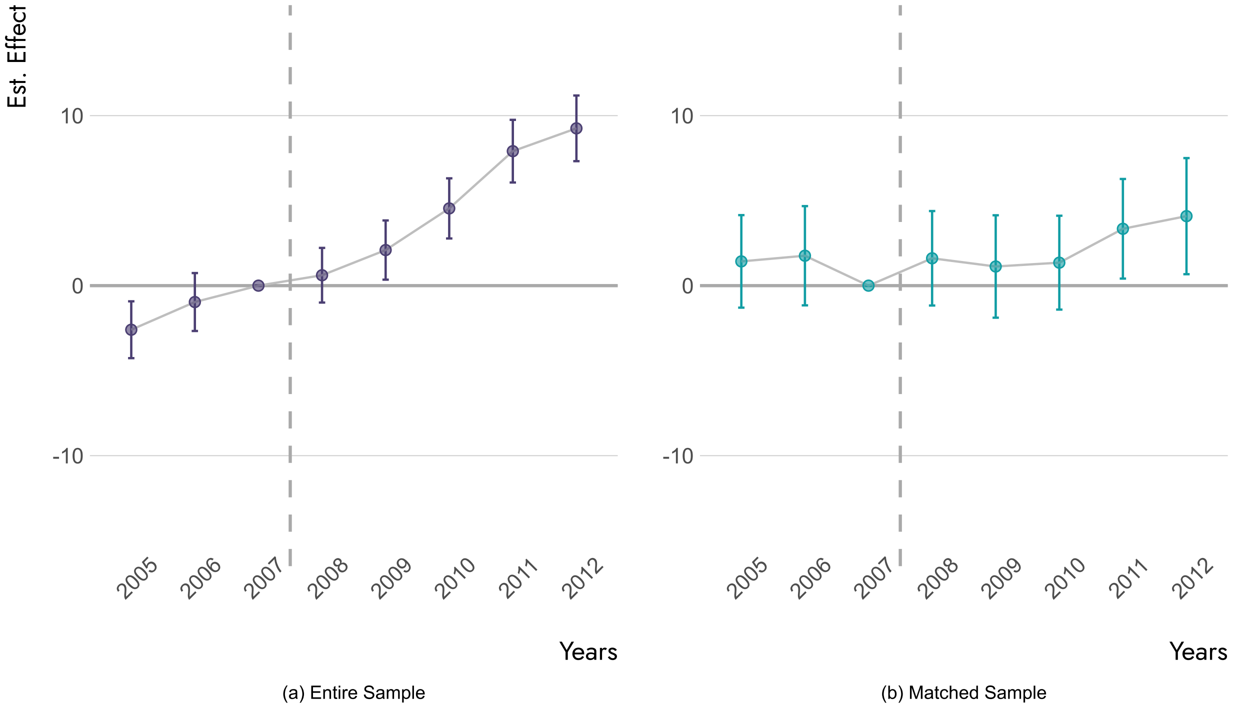
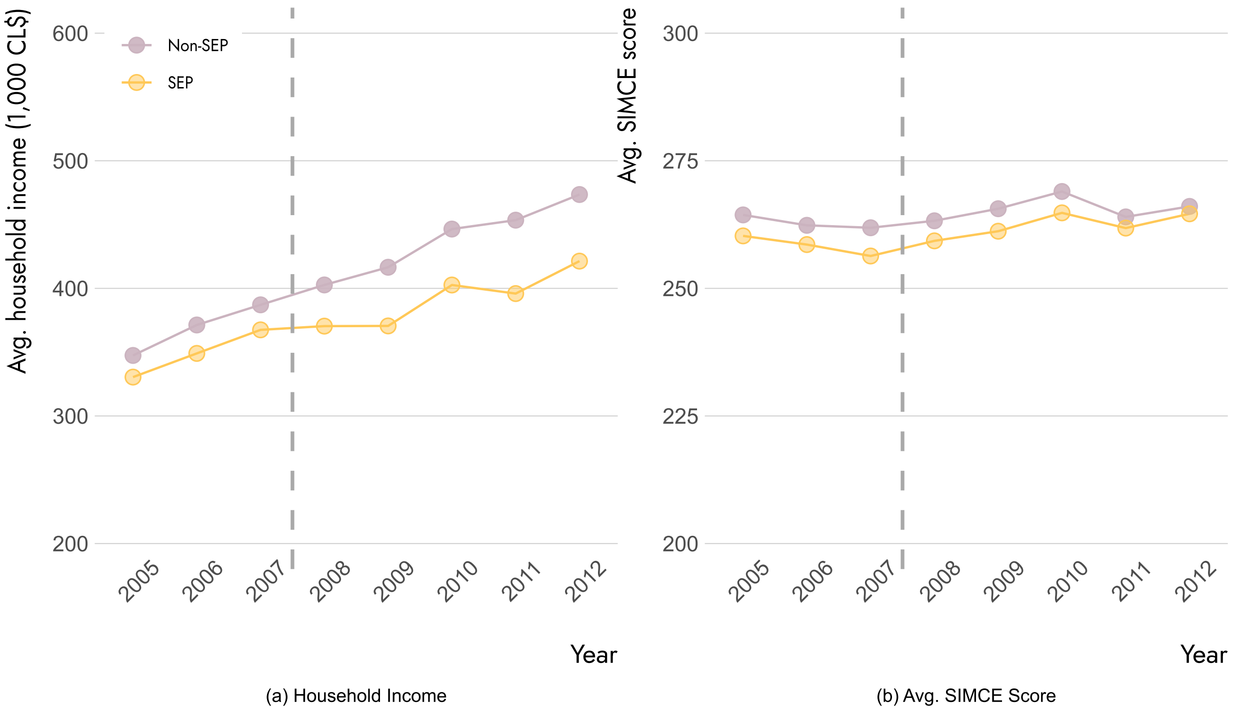


Figure 20: Event study estimates for average SIMCE score



(a) Household Income

(b) Avg. SIMCE Score

Figure 21: Outcome evolution over time for SEP and Non-SEP scools - Matched Sample

Table 6 shows the point estimates and 95% confidence intervals for all the post-intervention periods with respect to 2007. From 2009 to 2011, all differences between SEP and non-SEP schools are statistically significant from the same differences in 2007. Panel A from the same table shows that SEP schools increasingly attracted students of lower income households (compared to non-SEP school), with a sizable effect of 0.09 SD in 2011. Panel B, on the other hand, also shows significant and positive effects in terms of academic performance.

Table 6: Difference-in-difference estimates for post-intervention periods (relative to 2007) for matched sample

	2008	2009	2010	2011	2012
(a) Avg. HH Income (1,000 CL$)					
Point Est.	-12.66**	-26.37***	-24.22***	-37.93***	-32.59***
	[-25.31,-0.02]	[-39.07,-13.66]	[-38.49,-9.95]	[-58.45,-17.42]	[-51.27,-13.9]
Non-SEP mean	347.43	371.31	402.7	416.55	446.49
(b) Avg. SIMCE score					
Point Est.	1.61	1.13	1.35	3.34**	4.09**
	[-1.17,4.39]	[-1.88,4.14]	[-1.41,4.11]	[0.42,6.27]	[0.67,7.5]
Non-SEP mean	264.39	262.36	263.24	265.61	268.98
N Obs	578	578	578	578	578

Notes: 95% CI in brackets.

As a robustness check, I conduct a sensitivity analysis (Rambachan & Roth, 2023) to assess whether findings are robust to different magnitudes of violations of parallel trends. I find that while the point estimate using matching DD is smaller in magnitude than the DD estimate for the complete sample, findings are still more robust and would hold even for violations of 50% of the magnitude of the pre-intervention period trends (Figure 22). In

Figure 23 we observe a different story, with a statistically significant point estimate, but that is highly sensitive to small parallel trend violations when using a matching DD approach.

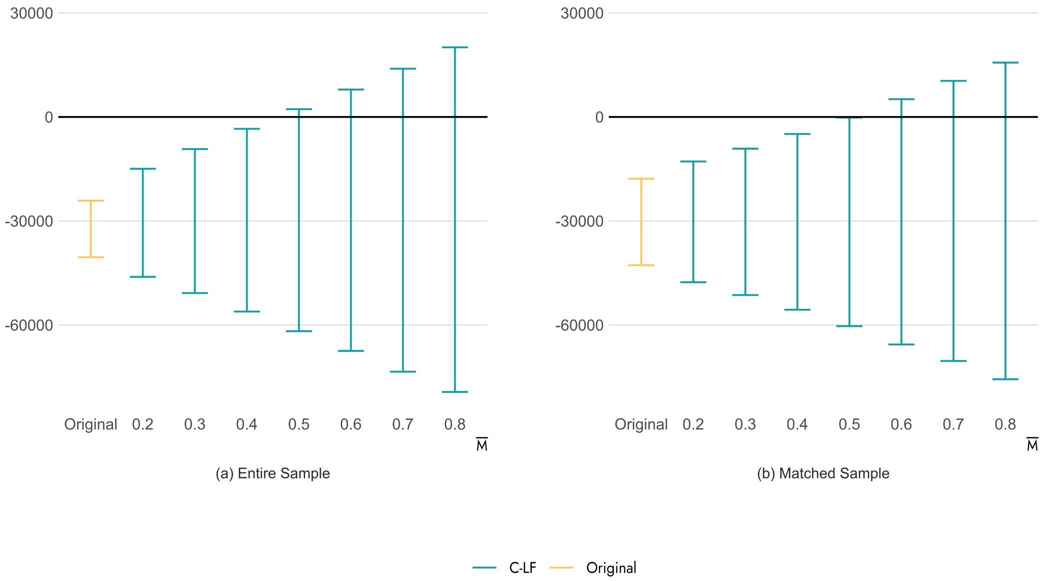


Figure 22: Bounds on Relative Magnitudes for Household Income Event Study Estimates

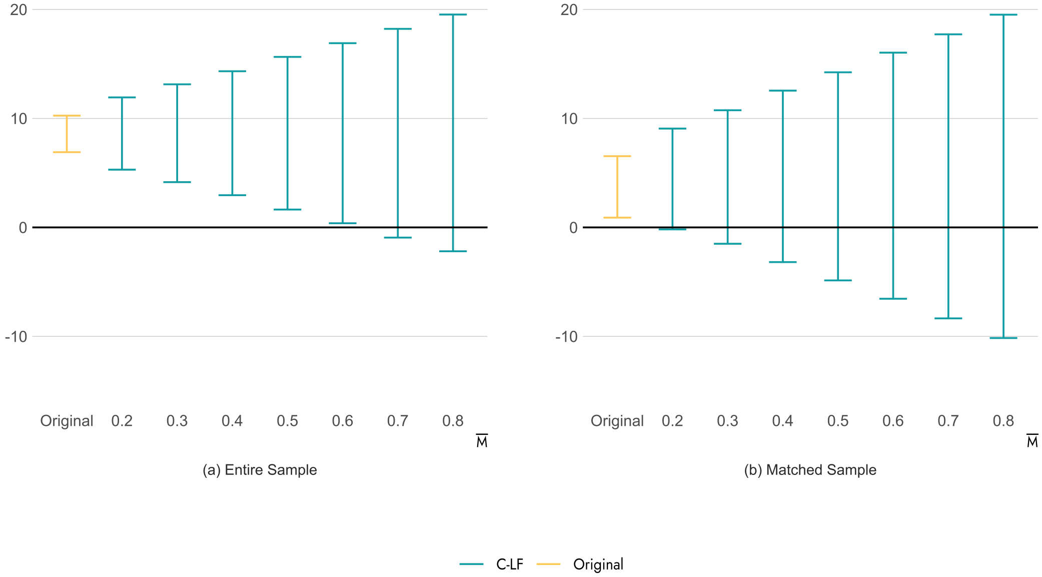


Figure 23: Bounds on Relative Magnitudes for Avg. SIMCE Event Study Estimates

I also conduct a sensitivity analysis to hidden bias on the matched ATT following Keele, Small, Hsu, & Fogarty (2019), finding that an unobserved confounder should affect the probability of assignment to treatment or control by 32.7% in order to explain away my findings, making these results moderately sensitive to hidden biases.[16]

[16] The estimate for average SIMCE is not statistically significant at conventional levels when testing the paired differences, so a sensitivity analysis is not conducted for this outcome.

In terms of socioeconomic characteristics of the schools, these results are consistent with the incentive structure provided by the SEP policy: Schools that had a higher concentration of lower SES students received an additional concentration bonus. For instance, Figure 24 shows the discontinuous jump of 4.7 percentage points (SE = 0.024; p-val = 0.05) in the probability of schools being SEP in 2009 at the 60% level of concentration of priority students in the prior year (2008); in this case, 60% represents the threshold that the SEP policy established for schools to receive the largest concentration bonus.

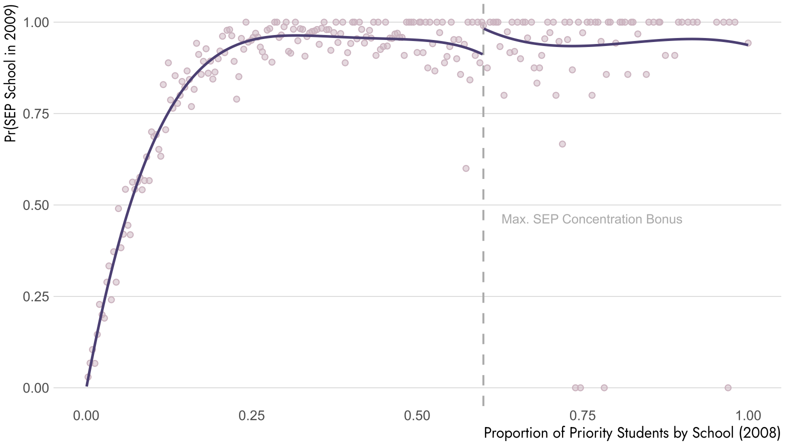


Figure 24: Probability of schools being subscribed to the SEP policy in 2009 as a function of their concentration of priority students in 2008

Evidence of the positive impact of the SEP policy on test scores is actually mixed. While some studies have found positive effects (see, for example, Navarro-Palau (2017), Correa, Parro, & Reyes (2014), Mizala & Torche (2013), and Neilson (2021)), others like Feigenberg, Yan, & Rivkin (2019)[17] and Aguirre (2022) found little to no significant effects of the policy in terms of academic performance. Conclusions seem to depend on the sample and the methodology used to identify these effects. In this case, the sample I focus the analysis on corresponds only to private subsidized schools which are more likely to respond to the policy. Additionally, because of the balancing constraints I am using for cardinality matching, the matched sample represents schools that belong to more densely populated areas with larger education markets.

To a certain extent, the ambiguity of this evidence can be reconciled by comparing the 2010 test score results using a regression discontinuity design (RDD) on students that belong to the matched schools. In this way, I partially recreate the methodology used by Aguirre (2022), but on my sample of analysis, shedding some light to whether the differences potentially stem from a different sample selection. Using this approach, I find a 4-point difference (p-value = 0.061) in math test scores for students in the margin of SEP eligibility (40% vulnerability score cutoff) that belong to the previously matched schools. However, it is important to note that these effects are not comparable: If the respective assumptions hold, RDD identifies a local average treatment effect for *students* at the margin of eligibility, while a difference-in-differences approach at the *school level* identifies an average treatment effect for the matched population.

5. Conclusions

[17] Feigenberg, Yan, & Rivkin (2019) are able to replicate the gain scores cited in the literature, but attribute this effect to the shift in socioeconomic characteristics in low SES students' households, and not to the policy itself.

Difference-in-differences can be a useful tool for causal inference and, given the current advances in the literature, particularly regarding sensitivity analysis, we now have a better framework to understand the robustness of our findings. However, it is important to understand that bias still plays an important role in any sort of observational study, and researchers should take additional precautions to make sure that findings from a study hold under a variety of conditions. In that sense, coupling matching on time-invariant characteristics with DD can help not only reduce or eliminate bias that stems from differential trends, but also provide more accurate sensitivity analyses for our findings even in the presence of violations to the main identification assumption.

This study contributes to the growing literature on DD and causal inference by emphasizing the interplay between design-based and model-based methods. The integration of matching into DD designs offers a promising path for researchers aiming to mitigate biases while preserving the interpretability and transparency of their findings. However, these results also call for careful consideration of the context-specific challenges associated with matching, such as the potential loss of sample size and its implications for statistical power.

Future research can extend this framework by exploring potential bounds on causal estimates in the presence not only of differential associations of time-invariant variables between groups, but also on *time-varying* covariates, which pose a direct violation of the main identification assumption for these studies.

References

Aguirre, J. (2022). How Can Progressive Vouchers Help the Poor Benefit from School Choice? Evidence from the Chilean Voucher System. *Journal of Human Resources*, *57*(3), 956—997.

Arkhangelsky, D., Athey, S., Hirshberg, D., Imbens, G., & Wager, S. (2021). Synthetic Difference-in-Differences. *American Economic Review*, *111*(12), 4088—4118.

Arteaga, F., Paredes, V., & Paredes, R. (2016). School Segregation in Chile: Residence, Co-Payment, or Preferences?. *Working Paper Pontificia Universidad Católica De Chile*.

Basu, P., & Small, D. S. (2020). Constructing a More Closely Matched Control Group in a Difference-in-Differences Analysis: Its Effect on History Interacting with Group Bias. *Observational Studies*, *6*(1), .

Bennett, M., Vielma, J., & Zubizarreta, J. (2020). Building Representative Matched Samples with Multi-valued Treatments in Large Observational Studies. *Journal of Computational and Graphical Statistics*, *29*(4), 744—757.

Chabé-Ferret, S. (2015). Analysis of the Bias of Matching and Difference-In-Difference under Alternative Earnings and Selection Processes. *Journal of Econometrics*, *185*(1), 110—123.

Chabé-Ferret, S. (2017). Should We Combine Difference In Differences with Conditioning on Pre-Treatment Outcomes?. *Working Paper, Toulouse School of Economics*.

Correa, J., Parro, F., & Reyes, L. (2014). The effects of vouchers on school results: evidence from Chile's targeted voucher program. *Journal of Human Capital*, *8*(4), 351—398.

Daw, J. R., & Hatfield, L. A. (2018b). Matching and Regression to the Mean in Difference-in-Differences Analysis. *Health Services Research*, *53*(6), 4138—4156.

Daw, J. R., & Hatfield, L. A. (2018a). Matching in Difference-in-Differences: Between a Rock and a Hard Place. *Health Services Research*.

Epple, D., & Romano, R. (2008). Educational Voucher and Cream Skimming. *International Economic Review*, *49*(4).

Epple, D., Newlon, E., & Romano, R. (2002). Ability Tracking, School Competition, and the Distribution of Educational Benefits. *Journal of Public Economics*, *83*, 1—48.

Feigenberg, B., Yan, R., & Rivkin, S. (2019). Illusory gains from Chile's targeted school voucher experiment. *The Economic Journal*, *129*(623), 2805—2832.

Ham, D. W., & Miratrix, L. (2024). Benefits and costs of matching prior to a Difference in Differences analysis when parallel trends does not hold. *Arxiv Working Paper*, *56*, .

Hsieh, C., & Urquiola, M. (2006). The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program. *Journal of Public Economics*, *90*, 1477—1503.

Keele, L., Small, D., Hsu, J., & Fogarty, C. (2019). Patterns of Effects and Sensitivity Analysis for Differences-in-Differences. *Arxiv Working Paper*, *0*, .

MINEDUC. (2020). *Datos Abiertos*.

Mizala, A., & Torche, F. (2013). Logra la subvención escolar preferencial igualar los resultados educativos?. *Documento De Referencia, Espacio Público*, *9*.

Navarro-Palau, P. (2017). Effects of differentiated school vouchers: Evidence from a policy change and date of birth cutoffs. *Economics of Education Review*, *0*(58), 86—107.

Neilson, C. (2021). Targeted Vouchers, Competition Among Schools, and the Academic Achievement of Poor Students. *Working Paper, Yale University*.

OECD. (2014). Latin American Economic Outlook 2015: Education, Skills, and Innovation for Development. *OECD Publishing*.

Rambachan, A., & Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, *90*, 2555—2591.

Romaguera, P., & Gallegos, S. (2010). Financiando la Educación de Grupos Vulnerables: La Subvención Escolar Preferencial. In O. Larrañaga & D. Contreras, *Las Nuevas Políticas de la Protección Social en Chile*. UNDP.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41—55.

Roth, J., Sant'Anna, P., Bilinski, A., & Poe, J. (2023). What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. *Journal of Econometrics*, *235*(2), 22218—22244.

Ryan, A. M., Burgess, J. F., & Dimick, J. B. (2015). Why We Should Not be Indifferent to Specification Choices for Difference-in-Differences. *Health Services Research*, *50*(4), 1211—1235.

Sant'Anna, P., & Zhao, J. (2020). Doubly Robust Difference-in-Differences Estimators. *Journal of Econometrics*, *219*(1), 101—122.

Urquiola, M. (2016). Competition among schools: traditional public and private schools. In E. A. Ed. Hanushek, S. Machin, & L. Woessman, *Handbook of the Economics of Education*. Elsevier.

World Bank. (2015). World Development Indicators. *World Bank*.

Zeldow, B., & Hatfield, L. (2021). Confounding and Regression Adjustment in Difference-in-Differences Studies. *Health Services Research*, *56*(5), 932—941.

Zubizarreta, J. R. (2012). Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure After Surgery. *Journal of the American Statistical Association*, *107*(500), 1360—1371.

Zubizarreta, J. R., Kilcioglu, C., & Vielma, J. P. (2018). designmatch: Matched samples that are balanced and representative by design. *R Package Version 0.3*, *0*.

Zubizarreta, J. R., Paredes, R. D., & Rosenbaum, P. R. (2014). Matching for Balance, Pairing for Heterogeneity in an Observational Study of the Effectiveness of For-profit and Not-for-profit High Schools in Chile. *Annals of Applied Statistics*, *8*, 204—231.