Matching Estimators with Few Treated and Many Control Observations

Ferman, Bruno

Sao Paulo School of Economics - FGV

4 May 2017

Matching Estimators with Few Treated and Many Control Observations*

Bruno Ferman†

Sao Paulo School of Economics - FGV

First Draft: May, 2017

This Draft: September, 2018

Please click here for the most recent version

Abstract

We analyze the properties of matching estimators when there are few treated, but many control observations. We show that, under standard assumptions, the nearest neighbor matching estimator for the average treatment effect on the treated is asymptotically unbiased in this framework. However, when the number of treated observations is fixed, the estimator is not consistent, and it is generally not asymptotically normal. Since standard inferential techniques are inadequate in this setting, we propose alternative inferential procedures based on the theory of randomization tests under approximate symmetry.

Keywords: matching estimators, treatment effects, hypothesis testing, randomization inference

JEL Codes: C12; C13; C21

*The author gratefully acknowledges the comments and suggestions of Luis Alvarez, Ricardo Paes de Barros, Lucas Finamor, Sergio Firpo, Ricardo Masini, Cristine Pinto, Vitor Possebom, Pedro Sant'Anna, and participants of the 2017 California Econometrics Conference and of the Rio-Sao Paulo Econometrics Conference. Deivis Angeli provided outstanding research assistance.

†bruno.ferman@fgv.br

1 Introduction

Matching estimators have been widely used for the estimation of treatment effects under a conditional independence assumption (CIA).[1] In many cases, matching estimators have been applied in settings where (1) the interest is in the average treatment effect for the treated (ATT), and (2) there is a large reservoir of potential controls (Imbens and Wooldridge (2009)). Abadie and Imbens (2006) study the asymptotic properties of matching estimators when the number of control observations grows at a higher rate than the number of treated observations. However, their asymptotic results still depend on both the number of treated and control observations going to infinity. Therefore, reliance on such asymptotic approximations should be considered with caution when the number of treated observations is small, even if the total number of observations is large.

In this paper, we analyze the properties of matching estimators when the number of treated observations is fixed, while the number of control observations goes to infinity. We first show that the nearest neighbor matching estimator is asymptotically unbiased for the ATT, under standard assumptions used in the literature on estimation of treatment effects under selection on observables.[2] This is consistent with Abadie and Imbens (2006), who show that the conditional bias of the matching estimator can be ignored, provided that the number of control observations increases fast enough, relative to the number of treated observations. In their setting, the matching estimator is consistent and asymptotically normal. In our setting, however, the variance of the matching estimator does not converge to zero, and the estimator will not generally be asymptotically normal. Our theoretical results provide a better approximation to the behavior of the matching estimator relative to Abadie and Imbens (2006) in settings where there is a larger number of control relative to treated observations, but the number of treated observations is not large enough, so that we cannot

[1] See Imbens (2004), Imbens and Wooldridge (2009), and Imbens (2014) for reviews.

[2] This is true whether we consider the average treatment effect on the treated conditional or unconditional on the covariates of the treated observations. Also, this is true whether asymptotic unbiasedness is defined based on the limit of the expected value of the estimator, or based on the expected value of the asymptotic distribution.

rely on asymptotic results that assume that the number of treated observations goes to infinity.[3] We conduct an empirical Monte Carlo (MC) study based on real data, as suggested by Huber et al. (2013). When the dimensionality of the covariates is low, and we consider matching estimators with few nearest neighbors, our simulations suggest that, regardless of the number of treated observations, the bias of the matching estimator is close to zero, even when the number of control observations is not large. Increasing the dimensionality of the covariates and/or increasing the number of nearest neighbors used in the estimation implies that we need an increasing number of controls to keep our approximations reliable.

The fact that the matching estimator is not asymptotically normal, in our setting, poses important challenges when it comes to inference. Inference based on the asymptotic distribution of the matching estimator derived by Abadie and Imbens (2006) should not provide a good approximation when the number of treated observations is very small, even if there are many control observations. The bootstrap procedure proposed by Otsu and Rai (2017) also relies on the number of both treated and control observations going to infinity. For finite samples, Rosenbaum (1984) and Rosenbaum (2002) consider permutation tests for observational studies under strong ignorability. However, these tests rely on restrictive assumptions.[4] Rothe (2017) provides robust confidence intervals for average treatment effects under limited overlap. For the case with continuous covariates, he combines his method with subclassification on the propensity score. However, with few treated observations, it would not be possible to reliably estimate a propensity score. Therefore, we consider two alternative inference methods based on the theory of randomization tests under an approximate symmetry assumption, developed by Canay et al. (2017). One test relies on permutations, while the other relies on group transformations given by sign changes.[5] We derive conditions

[3]The finite sample properties of matching and other related estimators have been evaluated in detail in simulations by, for example, Frolich (2004), Busso et al. (2014), Huber et al. (2013), and Bodory et al. (2018). In contrast to their approach, we provide theoretical and simulation results holding the number of treated observations fixed, but relying on the number of control observations going to infinity.

[4]Rosenbaum (1984) assumes that the propensity score follows a logit model, while Rosenbaum (2002) assumes that observations are matched in pairs such that the probability of treatment assignment is the same conditional on the pair.

[5]A test based on permutations has been studied in the context of an approximate symmetry assumption

under which these tests provide asymptotically valid hypothesis testing when the number of control observations goes to infinity, even when the number of treated observations remains fixed.

The different test procedures we consider present important trade-offs in terms of size distortion, power, and the underlying null hypothesis they rely on. With few treated observations, tests based on the asymptotic distribution derived by Abadie and Imbens (2006) and on the bootstrap procedure proposed by Otsu and Rai (2017) can have important size distortions, while the two randomization inference tests we propose control well for size even when the number of treated observations is very small. However, the randomization inference tests rely on sharper null hypotheses. We show that the size distortion and power for each test depend crucially on the number of treated observations, the number of control observations, and the number of nearest neighbors used in the estimation, providing guidance on how to evaluate the trade-offs among these test procedures in different scenarios.

As an empirical illustration, we consider the "Jovem de Futuro" (*Youth of the Future*) program. This is a program that has been running in Brazil since 2008, aimed at improving the quality of education in public schools by improving management practices and allocating grants to treated schools. In 2010, this program was implemented in a randomized control trial with 15 treated schools in Rio de Janeiro and 39 treated schools in Sao Paulo. We estimate the effects of the program using a matching estimator with the non-experimental sample as the control schools. We take advantage of the fact that there were about 1,000 other public schools in Rio de Janeiro and more than 3,000 other public schools in Sao Paulo that did not participate in the experiment, therefore, providing a setting with few treated and many control observations.[6] We find marginally significant treatment effects for Sao Paulo, and small and insignificant effects for Rio de Janeiro, which is consistent with the

by Canay and Kamat (2018) for regression discontinuity designs, while a test based on sign changes has been studied in the context of an approximate symmetry assumption by Canay et al. (2017) for a series of applications.

[6] Influential papers that evaluate the use of non-experimental methods in empirical applications where a randomized control trial is available include LaLonde (1986), Dehejia and Wahba (1999), and Dehejia and Wahba (2002).

estimates based on the randomized control trial. Moreover, using the experimental control schools as the treated group for the matching estimator (so that we should expect to find no significant results), we provide empirical evidence that inference based on the asymptotic distribution derived by Abadie and Imbens (2006) may lead to over-rejection when there are very few treated observations, while the randomization inference procedures control better for size in this case.

The remainder of this paper proceeds as follows. We present our theoretical setup in Section 2. In Section 3, we derive the asymptotic distribution of the matching estimator and derive conditions under which it is asymptotically unbiased. In Section 4, we consider alternative inference methods that are asymptotically valid when the number of control observations goes to infinity, while the number of treated observations remains fixed. In Section 5, we present an empirical MC simulation based on the "Jovem de Futuro" program, and estimate the effects of this program using a matching estimator. In Section 6, we contrast the different inference procedures in light of the theoretical results presented in Section 4 and the simulations presented in Section 5, providing guidance on which method should be chosen depending on the setting. Concluding remarks, including a discussion of potential implications for Synthetic Control applications, are presented in Section 7.

2 Setting and Notation

We are interested in estimating the effect of a binary treatment on some outcome. Following Rubin (1973), for each unit i we denote the potential outcomes $Y_i(1)$ if observation i receives treatment and $Y_i(0)$ if observation i does not receive treatment. Therefore, the observed outcome for unit i is given by $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$, where variable $W_i \in \{0, 1\}$ indicates the treatment received. In addition to Y_i and W_i, we also observe for each unit i a continuous random vector of pretreatment variables of dimension k in $\mathbb{R}^k$, which we denote by X_i. The case in which components of X_i are discrete, with a finite number of support points, can be easily dealt with by estimating treatment effects within subsamples defined by their values, and then aggregating on such covariates, as argued by Abadie and Imbens (2006). We assume that we observe a sample of N_1 treated (N_0 control) units that consists of i.i.d. observations of units with $W_i = 1$ ($W_i = 0$), and that treated and control observations are independent. Let $\mathcal{I}_w$ denote the set of indexes for observations with $W_i = w$.

Assumption 1 (Sample) *For $w \in \{0, 1\}$, $\{Y_i, X_i\}_{i \in \mathcal{I}_w}$ consists of N_w i.i.d. observations with $W_i = w$. Furthermore, we assume that individuals in the treated and control samples are independent.*

We consider the case in which the number of treated observations (N_1) is fixed, while the number of control observations (N_0) goes to infinity. One possibility is that there is a large set of units that could potentially be treated, but only a finite number of those units actually receive treatment. For example, in the empirical application, to be presented in Section 5, there are a large number of schools that could potentially receive the treatment, but only a small number of schools actually received it. Alternatively, we can imagine that there are a large number of treated units, but we only have data from a small sample of them.

We focus on two distinct estimands. First, we consider the conditional average treatment effect on the treated (CATT):

$$\tau(\{X_i\}_{i \in \mathcal{I}_1}) \equiv \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i, W_i = 1\right] \tag{1}$$

which is, conditional on the realization of $\{X_i\}_{i \in \mathcal{I}_1}$, the expected treatment effect for the treated units with these covariate values. We also consider the unconditional average treatment effect on the treated (UATT), which we denote by

$$\tau' \equiv \mathbb{E}\left[Y_i(1) - Y_i(0) | W_i = 1\right]. \tag{2}$$

In both cases, we focus on estimands related to the treatment effect on the treated because, given our setting with N_1 finite and N_0 large, there is no hope of constructing a counterfactual for the control observations using only a finite set of treated observations. In the framework of Imbens and Rubin (2015), these two estimands are defined based on a super-population.

Assumption 1 does not impose any restriction on how the distribution of $(Y_i(1), Y_i(0), X_i)$ for treated and control observations may differ. The following assumption does restrict the way in which these distributions may differ.

Assumption 2 (Conditional Independence Assumption) *Conditional on X_i, the distribution of $Y_i(0)$ is the same for i in the treated and in the control groups.*

Assumption 2 is equivalent to the conditional independence assumption (CIA). While in Assumption 1 we allow for different distributions of $(Y_i(0), Y_i(1), X_i)$ whether i is treated or control, Assumption 2 restricts that the conditional distribution of $Y_i(0)$ given X_i is the same for both treatment and control observations.[7] However, the density of X_i for the treated observations ($f_1(X_i)$) can potentially be different from the density of X_i for the control observations ($f_0(X_i)$). This is what generates potential bias in a simple comparison of means between treated and control groups, without taking into account that these groups might have different distributions of covariates X_i.

The next assumption states that possible values of X_i for the treated observations are in the support of the distribution of X_i for the control observations.

Assumption 3 (Overlap) $\mathbb{X}_1 \subset \mathbb{X}_0$, *where $\mathbb{X}_w$ is the support of $f_w(X_i)$, for $w \in \{0, 1\}$*

Assumption 3 replaces the standard assumption that $Pr(W = 1|X = x) < 1 - \eta$ for some $\eta > 0$. This assumption guarantees that, for each i in the treated group, we can find an observation j in the control group with covariates X_j arbitrarily close to X_i when $N_0 \to \infty$.

[7] We do not need to impose such restriction on $Y_i(1)$ because of our focus on average treatment effects on the treated.

The main identification problem arises from the fact that we observe either $Y_i(1)$ or $Y_i(0)$ for each observation i. Note that, if we had two observations, $i \in \mathcal{I}_1$ and $j \in \mathcal{I}_0$, with $X_i = X_j = x$, then, under Assumption 2, $\mathbb{E}[Y_i|W_i = 1, X_i = x] - \mathbb{E}[Y_j|W_j = 0, X_j = x] = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x, W_i = 1]$. The main challenge is that, with a continuous random variable X_i, the probability of finding observations with exactly the same X_i is zero. The idea of the nearest neighbor matching estimator is to input the missing potential outcomes of a treated observation $i \in \mathcal{I}_1$ with observations in the control group $j \in \mathcal{I}_0$ that are as close as possible in terms of covariates X_i. More specifically, for a given metric $d(a, b)$ in $\mathbb{R}^k$, let $\mathcal{J}_M(i)$ be the set of M nearest neighbors in the control group of observation $i \in \mathcal{I}_1$. Then the matching estimator is given by

$$\hat{\tau} = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left[Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j \right]. \tag{3}$$

3 Asymptotic Unbiasedness and Asymptotic Distribution

For $w \in \{0, 1\}$, we define $\mu(x, w) = \mathbb{E}[Y|X = x, W = w]$ and $\epsilon_i = Y_i - \mu(X_i, W_i)$. Since we are focusing on the average treatment effect on the treated, we also define $\mu_w(x) = \mathbb{E}[Y(w)|X = x, W_i = 1]$.[8] Under Assumption 2, we have that $\mu(x, 0) = \mu_0(x)$. Using this notation, note that the CATT is given by

$$\tau(\{X_i\}_{i \in \mathcal{I}_1}) = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} [\mu_1(X_i) - \mu_0(X_i)] \tag{4}$$

[8]Note that Abadie and Imbens (2006) define $\mu_w(x) = \mathbb{E}[Y(w)|X = x]$. We use a slightly different definition because we focus on the average treatment effects on the treated.

and

$$\hat{\tau} = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left[\left(\mu_1(X_i) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \mu_0(X_j) \right) + \left(\epsilon_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \epsilon_j \right) \right]. \tag{5}$$

We first show that $\hat{\tau}$ is an asymptotically unbiased estimator for the CATT when the number of treated observations is fixed and the number of control observations grows, and we derive its asymptotic distribution in this setting.

Proposition 1 *Under Assumptions 1, 2, and 3,*

1. *If $\mu_0(x)$ is continuous and bounded, then $\mathbb{E}[\hat{\tau}|\{X_i\}_{i\in\mathcal{I}_1}] \rightarrow \tau(\{X_i\}_{i\in\mathcal{I}_1})$ when $N_0 \rightarrow \infty$ and N_1 is fixed.*

2. *If $\tilde{h}(x) = \mathbb{E}[h(Y(0))|X = x]$ is continuous and bounded for any $h(y)$ continuous and bounded, then, conditional on $\{X_i\}_{i\in\mathcal{I}_1}$,*

   $$\hat{\tau} \xrightarrow{d} \tau(\{X_i\}_{i\in\mathcal{I}_1}) + \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left(\epsilon_i - \frac{1}{M} \sum_{m=1}^{M} \epsilon_m(X_i) \right) \text{ when } N_0 \rightarrow \infty \text{ and } N_1 \text{ is fixed}$$

   where $\epsilon_m(X_i) \stackrel{d}{=} Y_i(0)|X_i - \mu_0(X_i)$ for $i \in \mathcal{I}_1$, and $\epsilon_m(X_i)$ is independent across m and i.

Proof. See details in Appendix A.1.1. ■

Let $X^i_{(m)}$ be the covariate value of the m-closest match to observation i. The main intuition for the results in Proposition 1 is that, for a fixed $X_i = \bar{x}$, $X^i_{(m)} \xrightarrow{p} \bar{x}$ when $N_0 \rightarrow \infty$, because, holding M fixed, we will always be able to find M observations in the control group that are arbitrarily close to $\bar{x}$. Independence of $\epsilon_m(X_i)$ across m and i follows from the fact that the probability of two treated observations sharing the same nearest neighbor converges to zero.

Proposition 1 shows that, conditional on the realization of $\{X_i\}_{i\in\mathcal{I}_1}$, the expected value of the matching estimator converges to $\tau(\{X_i\}_{i\in\mathcal{I}_1}) = \frac{1}{N_1}\sum_{i\in\mathcal{I}_1}(\mu_1(X_i) - \mu_0(X_i))$ when $N_0 \rightarrow$

∞. We also derive the asymptotic distribution of the matching estimator conditional on $\{X_i\}_{i\in\mathcal{I}_1}$, which is centered on $\tau(\{X_i\}_{i\in\mathcal{I}_1})$. This is important for the construction of the inference methods we propose in Section 4. These results are valid for any fixed value of N_1, including the case with $N_1 = 1$.

Remark 1 The condition that $\mu_0(x)$ is continuous and bounded would be satisfied if we assume that $\mu_0(x)$ is continuous and $\mathbb{X}_0$ is compact, as is assumed by Abadie and Imbens (2006). The intuition behind the assumption used in part 2 of Proposition 1 is that the conditional distribution of $Y(0)$ given $X = x$ changes "smoothly" with x. This guarantees that the outcome of the m-closest match to treated observation i, $Y^i_{(m)}$, converges in distribution to $Y_i(0)|X_i = \bar{x}$ when $X^i_{(m)} \xrightarrow{p} \bar{x}$. In Appendix A.1.2, we show that this condition is satisfied if, for example, $Y(0)|X = x \sim N(\theta(x), \sigma(x))$, where $\theta(x)$ and $\sigma(x)$ are continuous functions of x.

Remark 2 We focus on the properties of the matching estimator conditional on $\{X_i\}_{i\in\mathcal{I}_1}$. We might be interested, however, in the unconditional properties of the matching estimator. Under the assumptions from part 1 of Proposition 1, $\mathbb{E}[\hat{\tau}] = \mathbb{E}\left\{\mathbb{E}[\hat{\tau}|\{X_i\}_{i\in\mathcal{I}_1}]\right\}$ converges to τ', which is the UATT. See details in Appendix A.1.3.

Remark 3 With N_1 fixed, the estimator is not consistent. This happens because, with a fixed number of treated observations, we cannot apply a law of large numbers to the average of the error of the treated observations. For the same reason, the matching estimator will not be asymptotically normal, unless we assume that the error ϵ_i is normal. These conclusions are similar to the ones derived by Conley and Taber (2011) for differences-in-differences estimators with few treated groups.

Remark 4 Consider a bias-corrected estimator given by

$$\hat{\tau}_{biasadj} = \frac{1}{N_1}\sum_{i\in\mathcal{I}_1}\left[Y_i - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}(Y_j + \hat{\mu}_0(X_i) - \hat{\mu}_0(X_j))\right], \tag{6}$$

where $\hat{\mu}_0(X_i)$ is an estimator for $\mu_0(X_i)$. With additional assumptions, we can also guarantee that $\hat{\tau}_{biasadj}$ has the same asymptotic distribution as $\hat{\tau}$. The intuition is that $\hat{\mu}_0(X_i)-\hat{\mu}_0(X^i_{(m)})$ converges in probability to zero when $N_0 \to \infty$, because $X^i_{(m)} \xrightarrow{p} X_i$. See details in Appendix A.1.4.

Remark 5 We consider an asymptotic framework in which M is held fixed, while $N_0 \to \infty$, which is similar to what Abadie and Imbens (2006) call fixed-M asymptotics in their setting.[9] As argued by Abadie and Imbens (2006), the motivation for such fixed-M asymptotics is to provide an approximation to the sampling distribution of matching estimators with a small number of matches. Matching estimators using few matches have been widely used in applied work (see Abadie and Imbens (2006)). Moreover, Imbens and Rubin (2015) argue against using matching estimators with many matches, as this would tends to increase the bias of the resulting estimator, while the marginal gains in precision of increasing the number of matches are limited.

4 Inference

The fact that the matching estimator is not generally asymptotically normal when N_1 is fixed and $N_0 \to \infty$ poses an important challenge when it comes to inference. In particular, inference based on the asymptotically normal distribution derived by Abadie and Imbens (2006), or on the bootstrap procedure suggested by Otsu and Rai (2017), should not provide a good approximation in our setting, as the asymptotic theory behind these methods rely on both N_1 and N_0 going to infinity. We therefore consider alternative inference methods based on the theory of randomization tests under an approximate symmetry assumption, developed by Canay et al. (2017). We derive conditions under which these methods are asymptotically valid when $N_0 \to \infty$, even with fixed N_1. The first test is based on group transformations given by permutations, while the second test is based on group transforma-

[9]The difference relative to the framework considered by Abadie and Imbens (2006) is that we also hold N_1 fixed.

tions given by sign changes. An important caveat is that these different tests differ in their underlying assumptions and null hypotheses. Moreover, the size and power of these tests depend crucially on the number of treated and control observations, and also on the number of nearest neighbors used in the estimation. In Section 6, we contrast the different inference procedures, providing guidance on how to evaluate these trade-offs in different settings.

4.1 Randomization Inference Test Based on Permutations

Consider a function of the data given by

$$\tilde{S}_{N_0} = \left(\tilde{S}^0_{N_0,1}, \tilde{S}^1_{N_0,1}, ..., \tilde{S}^M_{N_0,1}, ..., \tilde{S}^0_{N_0,N_1}, \tilde{S}^1_{N_0,N_1}, ..., \tilde{S}^M_{N_0,N_1}\right)' \tag{7}$$

where $\tilde{S}^0_{N_0,i} = Y_i$ and $\tilde{S}^m_{N_0,i} = Y^i_{(m)}$ for $m = 1, ..., M$. That is, $\tilde{S}_{N_0}$ is a vector containing the outcomes of the treated observations and of their M-nearest neighbors. The distribution of $\tilde{S}_{N_0}$ depends on N_0, because the quality of the matches will depend on N_0. In this notation, the matching estimator is given by

$$\hat{\tau} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(\tilde{S}^0_{N_0,i} - \frac{1}{M}\sum_{j=1}^{M}\tilde{S}^j_{N_0,i}\right). \tag{8}$$

Let $\tilde{G}_i$ be the set of all permutations $\pi_i = (\pi_i(0), ..., \pi_i(M))$ of $\{0, 1, ..., M\}$, $\pi = \otimes_{i=1}^{N_1}\pi_i$, and $\tilde{\mathbf{G}} = \otimes_{i=1}^{N_1}\tilde{G}_i$. Note that $\tilde{\mathbf{G}}$ is the set of all permutations that reassign the treatment status conditional on having exactly one treated observation for each group of treated observation i and its M nearest neighbors. For a given $\pi \in \tilde{\mathbf{G}}$, consider $\tilde{S}^\pi_{N_0} = \left(\tilde{S}^{\pi_1(0)}_{N_0,1}, \tilde{S}^{\pi_1(1)}_{N_0,1}, ..., \tilde{S}^{\pi_1(M)}_{N_0,1}, ..., \tilde{S}^{\pi_{N_1}(0)}_{N_0,N_1}, \tilde{S}^{\pi_{N_1}(1)}_{N_0,N_1}, ..., \tilde{S}^{\pi_{N_1}(M)}_{N_0,N_1}\right)'$.

Let $\tilde{K} = |\tilde{\mathbf{G}}|$ and denote by

$$\tilde{T}^{(1)}(\tilde{S}_{N_0}) \leq \tilde{T}^{(2)}(\tilde{S}_{N_0}) \leq ... \leq \tilde{T}^{(\tilde{K})}(\tilde{S}_{N_0}) \tag{9}$$

the ordered values of $\{\tilde{T}(\tilde{S}^{\pi}_{N_0}) : \pi \in \tilde{\mathbf{G}}\}$, where

$$\tilde{T}(\tilde{S}^{\pi}_{N_0}) = \left[\frac{1}{N_1} \sum_{i=1}^{N_1} \left(\tilde{S}^{\pi_i(0)}_{N_0,i} - \frac{1}{M} \sum_{j=1}^{M} \tilde{S}^{\pi_i(j)}_{N_0,i} \right) \right]^2 . \tag{10}$$

We set $\tilde{k} = \lceil \tilde{K}(1-\alpha) \rceil$, where α is the significance level of the test, and define the decision rule of the test as

$$\tilde{\phi}(S_{N_0}) = \begin{cases} 1 & \text{if } \tilde{T}(\tilde{S}_{N_1}) > \tilde{T}^{(\tilde{k})}(\tilde{S}_{N_1}) \\ 0 & \text{if } \tilde{T}(\tilde{S}_{N_1}) \leq \tilde{T}^{(\tilde{k})}(\tilde{S}_{N_1}). \end{cases} \tag{11}$$

In words, we calculate the test statistic $\tilde{T}(\tilde{S}^{\pi}_{N_0})$ for all possible permutations in $\tilde{\mathbf{G}}$, and then we reject the null if the actual test statistic $\tilde{T}(\tilde{S}_{N_0})$ is large relative to the distribution given by these permutations.

We show that, if we consider the null hypothesis

$$H_0 : Y_i(0)|X_i \stackrel{d}{=} Y_i(1)|X_i \text{ for all } i \in \mathcal{I}_1, \tag{12}$$

then such test is asymptotically level α, meaning that probability of rejection under the null converges to a value lower or equal to α when $N_0 \to \infty$.

Proposition 2 *Suppose the assumptions used in part 2 of Proposition 1 are valid, and that the distribution of $Y_i(0)|X_i$ is continuous. If we consider the problem of testing 12, then a test based on the decision rule defined in 11 is asymptotically level α, for any $\alpha \in (0,1)$, when $N_0 \to \infty$ and N_1 is fixed.*

Proof. See details of the proof in Appendix A.1.5. ■

The main intuition of the proof is that, when $N_0 \to \infty$, the limiting distribution of $\tilde{S}_{N_0}$, under the null, is invariant to the transformations in $\tilde{\mathbf{G}}$. From the proof of Proposition 1, note that $S^m_{N_0,i} = Y^i_{(m)} \xrightarrow{d} Y_i(0)|X_i$, for all $m = 1, ..., M$. Therefore, under the null that $Y_i(0)|X_i \overset{d}{=} Y_i(1)|X_i$, we have that $\tilde{S}^j_{N_0,i} \xrightarrow{d} Y_i(0)|X_i$ for all $j = 0, ..., M$. Moreover, asymptotically, $\tilde{S}^j_{N_0,i}$ is independent across i and j, because the probability that two treated units share the same nearest neighbor converges to zero when $N_0 \to \infty$.

Remark 6 Rosenbaum (2002) considers Fisher exact tests in observational studies with matched pairs. He shows that, if the probability of treatment assignment is the same for both observations in each pair, then a permutation test conditional on the pair is valid, even in finite samples. With a finite N_0 and continuous X, however, it is not possible to guarantee this condition, even under Assumption 2, since we will not have, in general, a perfect match in terms of covariates. We show that this condition can be approximately satisfied when $N_0 \to \infty$ and N_1 is fixed using the theory of randomization inference under approximate symmetry developed by Canay et al. (2017).

Remark 7 The null hypothesis 12 implies that $\tau(\{X_i\}_{i\in\mathcal{I}}) = 0$, but the converse is not true. To understand why this is crucial for this test, suppose, for example, that $\mathbb{E}[Y_i(1)|X_i] = \mathbb{E}[Y_i(0)|X_i]$ for all $i \in \mathcal{I}$, but $\mathbb{V}[Y_i(1)|X_i] > \mathbb{V}[Y_i(0)|X_i]$. If $M > 1$, then a permutation that uses control observations in place of treated ones would have a less volatile distribution relative to the distribution of the matching estimator. This would lead to a rejection rate higher than α.[10] This is an important drawback of this test, if the underlying interest is in testing whether $\tau(\{X_i\}_{i\in\mathcal{I}_1}) = 0$, because the test may reject at a rate higher than α even when $\tau(\{X_i\}_{i\in\mathcal{I}_1}) = 0$. However, we need to consider a more stringent null hypothesis to guarantee that the test is valid for any fixed value of N_1 (even for $N_1 = 1$). Tests that rely on this kind of null hypothesis have received increasing attention, particularly in randomized experiment settings (e.g. Young (2015)). In Section 4.2 we present an alternative test, that is also valid with fixed N_1, that rely on a less stringent null hypothesis.

Remark 8 This permutation test is similar in spirit to the test proposed by Conley and Taber (2011) for differences-in-differences with few treated and many control groups. Note

[10] Following the same logic, this also implies that such a test may have a low power if the treatment decreases the variance of the outcome (that is, $\mathbb{V}[Y_i(1)|X_i] < \mathbb{V}[Y_i(0)|X_i]$). See Ferman and Pinto (2018) for a related discussion.

that they assume that errors are i.i.d. across groups. In line with their results, the permutation test we propose would also be asymptotically valid if, for example, we test the null hypothesis $\tau(\{X_i\}_{i\in\mathcal{I}_1}) = 0$ (instead of the null hypothesis 12), but we impose the assumptions that ϵ_i is i.i.d. for all i, and that treatment effect is homogeneous. This highlights the fact that we need to rely on stronger assumptions if we want to construct a test that is valid regardless of the number of treated observations.[11]

Remark 9 As outlined by Bugni et al. (2018), the null hypothesis $Y_i(0)|X_i \overset{d}{=} Y_i(1)|X_i$ is implied by what is sometimes referred to as a "sharp null hypothesis," in which $Y_i(1) = Y_i(0)$ with probability one.

Remark 10 The test would remain valid if we consider the null hypothesis $Y_i(0)|X_i \overset{d}{=} Y_i(1)|X_i + c_i$ for all $i \in \mathcal{I}_1$, for a known vector of constants $c = (c_1, ..., c_{N_1})$, instead of the null hypothesis defined in 12.

Remark 11 Canay et al. (2017) consider a randomized version of the test to deal with cases such that $\tilde{T}(\tilde{S}_{N_1}) = T^{(\tilde{k})}(\tilde{S}_{N_1})$. Their approach guarantees a test with asymptotic size α. We focus on the non-randomized version of the test that rejects the null hypothesis if $\tilde{T}(\tilde{S}_{N_1}) > \tilde{T}^{(\tilde{k})}(\tilde{S}_{N_1})$, which guarantees that the test is asymptotically level α, although it may be conservative. The under rejection will only be relevant if $\tilde{K}$ is very small, where $\tilde{K}$ is a function of N_1 and M.

Remark 12 This test is asymptotically valid, when $N_0 \to \infty$, in part because the probability that different treated observations share the same nearest neighbor goes to zero. In finite samples, however, this may not be the case, and two treated observations will likely

[11] Ferman and Pinto (2018) consider a method similar to the one proposed by Conley and Taber (2011), but that allows for specific forms of heteroskedasticity based on the observed covariates. However, if we consider that all the observable variables that may induce heteroskedasticity are already included as covariates in the matching process, then this would be innocuous. More specifically, in this case, we would already be considering only permutations of observations with similar values of X_i, which is essentially a non-parametric version of Ferman and Pinto (2018). Conley and Taber (2011) also consider alternative methods to allow for heteroskedasticity. However, these alternatives would also allow only for heteroskedasticity based on observables.

share the same nearest neighbor when N_0 is not large enough relative to N_1 and M. To take that into account, we consider a finite sample fix in the permutation test. If a control observation is the nearest neighbor for two or more treated observations, then we restrict to permutations of S_{N_0} such that this control observation is always placed as either treated or control. Since the probability that two treated observations share the same nearest neighbor goes to zero when N_1 is fixed and $N_0 \to \infty$, for a fixed M, this finite sample adjustment is asymptotically irrelevant.[12]

Remark 13 This test is also asymptotically valid for bias-corrected matching estimators, as presented in 6. In this case, we define $\tilde{S}^0_{N_0,i} = Y_i$ and $\tilde{S}^m_{N_0,i} = Y^i_{(m)} + \hat{\mu}_0(X_i) - \hat{\mu}_0(X^i_{(m)})$. The key idea is that, again, $\tilde{S}^m_{N_0,i} = Y^i_{(m)} + \hat{\mu}_0(X_i) - \hat{\mu}_0(X^i_{(m)}) \xrightarrow{d} Y_i(0)|X_i$, for all $m = 1, ..., M$, because $\hat{\mu}_0(X_i) - \hat{\mu}_0(X^i_{(m)}) \xrightarrow{p} 0$.

4.2 Randomization Inference Test Based on Sign Changes

We consider now an alternative function of the data given by

$$S_{N_0} = \left(\hat{\tau}_1^{N_0}, ..., \hat{\tau}_{N_1}^{N_0}\right)' \tag{13}$$

where $\hat{\tau}_i^{N_0} = Y_i - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j$. Each $\hat{\tau}_i^{N_0}$ depends on the M nearest neighbors of observation i, so its distribution depends on N_0.

Following Canay et al. (2017), we consider a test statistic given by

$$T(S_{N_0}) = \frac{|\hat{\tau}|}{\sqrt{\frac{1}{N_1 - 1}\sum_{i=1}^{N_1}(\hat{\tau}_i^{N_0} - \hat{\tau})^2}} \tag{14}$$

where $\hat{\tau} = \frac{1}{N_1}\sum_{i \in \mathcal{I}_1} \hat{\tau}_i^{N_0}$ is the matching estimator for the treatment effects on the treated.

[12]Another alternative would be to consider a matching estimator without replacement. However, this would generate lower quality matches, which implies more bias (Abadie and Imbens (2006)). Moreover, matching without replacement has the disadvantage that the estimator is not invariant to different sorting of the data.

We consider the group of transformations given by $\mathbf{G} = \{-1,1\}^{N_1}$, where $gS_{N_0} = \left(g_1\hat{\tau}_1^{N_0}, ..., g_{N_1}\hat{\tau}_{N_1}^{N_0}\right)'$. Let $K = |\mathbf{G}|$ and denote by

$$T^{(1)}(S_{N_0}) \leq T^{(2)}(S_{N_0}) \leq ... \leq T^{(K)}(S_{N_0}) \tag{15}$$

the ordered values of $\{T(gS_{N_0}) : g \in \mathbf{G}\}$. Let $k = \lceil K(1-\alpha)\rceil$, where α is the significance level of the test. Then the test is given by

$$\phi(S_{N_0}) = \begin{cases} 1 & \text{if } T(S_{N_1}) > T^{(k)}(S_{N_1}) \\ 0 & \text{if } T(S_{N_1}) \leq T^{(k)}(S_{N_1}). \end{cases} \tag{16}$$

In words, we calculate the test statistic $T(gS_{N_0})$ for all possible $gS_{N_0} = \left(g_1\hat{\tau}_1^{N_0}, ..., g_{N_1}\hat{\tau}_{N_1}^{N_0}\right)'$, and then we compare the actual test statistic $T(S_{N_0})$ with the distribution $\{T(gS_{N_0}) : g \in \mathbf{G}\}$.

We show that such test is asymptotically valid if we consider the null hypothesis

$$H_0 : \mu_1(X_i) = \mu_0(X_i) \text{ for all } i \in \mathcal{I}_1 \tag{17}$$

Proposition 3 *Suppose the assumptions used in part 2 of Proposition 1 are valid, and that the distribution of $Y_i(1)|X_i$ ($Y_i(0)|X_i$) is continuous and symmetric around $\mu_1(X_i)$ ($\mu_0(X_i)$) for all $i = 1, ..., N_1$. If we consider the problem of testing 17, then a test based on the decision rule defined in 16 is asymptotically level α for any $\alpha \in (0,1)$ when $N_0 \to \infty$ and N_1 is fixed.*

Proof. See details of the proof in Appendix A.1.6. ■

Again, the main intuition of the proof is that, when $N_0 \to \infty$, the limiting distribution of S_{N_0}, under the null, is invariant to the transformations in $\mathbf{G}$. This is true if, asymptotically, $\hat{\tau}_i^{N_0}$ and $\hat{\tau}_j^{N_0}$ are independent for $i \neq j$, and the distribution of $\hat{\tau}_i^{N_0}$ is symmetric around zero. It is not necessary for $\hat{\tau}_i^{N_0}$ to have the same distribution across i. From Proposition 1, we know that, under the null, the asymptotic distribution of $\hat{\tau}_i^{N_0}$, conditional on $\{X\}_{i\in\mathcal{I}_1}$, is given by

$\epsilon_i - \frac{1}{M}\sum_{m=1}^{M}\epsilon_m(X_i)$. This distribution is symmetric around zero given the assumption that $Y_i(1)|X_i$ and $Y_i(0)|X_i$ are symmetric around zero for all $i = 1, ..., N_1$. Moreover, Proposition 1 also shows that, asymptotically, $\hat{\tau}_i^{N_0}$ are independent across i.

Remark 14 This test relies on a null hypothesis that the average treatment effect is equal to zero, conditional on each covariate value in $\{X_i\}_{i\in\mathcal{I}_1}$. This null allows for different distributions of potential outcomes when treated and control. In particular, it allows for heteroskedasticity, as it may be that $\mathbb{V}[Y_i(1)|X_i] \neq \mathbb{V}[Y_i(0)|X_i]$ under the null. This null hypothesis is implied by more narrowly defined null hypotheses that are usually considered in Fisher-type tests, such as $Y_i(0)|X_i \overset{d}{=} Y_i(1)|X_i$ or $Y_i(0) = Y_i(1)$ with probability one. However, it is still more stringent than the null hypothesis that $\tau(\{X_i\}_{i\in\mathcal{I}}) = 0$.

Remark 15 If the null hypothesis 17 is false, but $\tau(\{X_i\}_{i\in\mathcal{I}}) = 0$, then the test would tend to be conservative. The reason is that, in this case, we will have that $\frac{1}{N_1}\sum_{i\in\mathcal{I}_1} g_i(\mu_1(X_i) - \mu_0(X_i)) \neq 0$ for at least some $g_i \neq (1, ..., 1)$, while $\frac{1}{N_1}\sum_{i\in\mathcal{I}_1}(\mu_1(X_i) - \mu_0(X_i)) = 0$. This will tend to generate a distribution for the test statistic given these group transformations that is more volatile than the distribution of the actual test statistic. Therefore, even if we consider a null hypothesis $\tau(\{X_i\}_{i\in\mathcal{I}}) = 0$, we should still expect to have a level α test.

Remark 16 This test can be extended to test null hypotheses of the form $\mu_1(X_i) = \mu_0(X_i) + c_i$ for all $i \in \mathcal{I}_1$, for a known vector of constants $c = (c_1, ..., c_{N_1})$.

Remark 17 Similarly to the point raised in Remark 13, this test is asymptotically valid because the probability that different treated observations share the same nearest neighbor goes to zero, when $N_0 \to \infty$, which implies that $\hat{\tau}_i^{N_0}$ and $\hat{\tau}_{i'}^{N_0}$ are asymptotically uncorrelated for $i \neq i'$. Therefore, we also suggest a finite sample adjustment, in which we restrict to sign changes such that $g_i = g_j$ if i and j share the same nearest neighbor. Similar to the finite sample adjustment used in the test based on permutations, the probability that this modification is relevant converges to zero when $N_0 \to \infty$.

Remark 18 Remark 11 also applies to this test.

Remark 19 This test is also asymptotically valid for bias-corrected matching estimators, as defined in equation 6. In this case, we define $\tilde{\tau}_i^{N_0} = Y_i - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} \left(Y_j - \hat{\mu}_0(X_j) + \hat{\mu}_0(X_i) \right)$.

5 "Jovem de Futuro Program": Monte Carlo Simulations & Empirical Application

We explore the validity of matching estimators and of different inferential methods in the estimation of the effects of an educational program in Brazil called "Jovem de Futuro", that provides a setting with few treated and many control schools. In Section 5.1, we conduct an empirical Monte Carlo (MC) study based on this application (e.g. Huber et al. (2013)), while in Section 5.2 we estimate the effects of the program using matching estimators.

Before we proceed, we start with a brief description of the program, and we present some descriptive statistics (see Barros et al. (2012) for more details). The "Jovem de Futuro" program, an initiative of the "Instituto Unibanco" (Unibanco Institute), aims to improve the quality of education in Brazilian public schools. This is a three-year-long intervention based on two efforts: (i) providing school managers with strategies and instruments to become more efficient and productive, and (ii) providing conditional cash transfers to schools.[13] In 2007, the Unibanco Institute created and implemented the program in three schools in Sao Paulo. Then they implemented a few randomized control trials in the following years to evaluate the impact of the program.

We focus on the 2010 implementation of the program, which took place in Rio de Janeiro and Sao Paulo. Schools in these two states were invited to participate in the program, knowing in advance that they would be randomly assigned to receive the program starting in 2010, or that they would be placed first as a control group and would start the program

[13] The conditions are to improve students' performance on a standardized examination by the Institute at the end of each school year and to implement a participatory budget process in the school.

only in 2013. We use information from the 2007 to 2012 "Exame Nacional do Ensino Médio" (ENEM), a national exam that evaluates high school students in Brazil, as a measure of students' proficiency.[14,15] Focusing on schools with test score information from 2007 to 2012, we have 15 treated schools in Rio de Janeiro and 39 in Sao Paulo, with the same number of control schools in each state.[16]

Column 1 of Table 1 presents the difference in test scores for treated and control experimental schools in Rio de Janeiro, and column 3 shows the same difference for schools in Sao Paulo. Panel A presents this information for 2007 to 2009, which was before the intervention. For Rio de Janeiro, all differences are small and not statistically different from zero, as one would expect given random assignment. For Sao Paulo, however, there are significant differences in test scores in 2007 and 2008, suggesting that there may have been some problems in the assignment of treatment schools. Panel B presents the results for the three years after the implementation of the program. The comparison between treated and control schools suggest a null effect of the program in Rio de Janeiro, and a positive and significant effect in Sao Paulo. We should be careful in interpreting the results for Sao Paulo, however, due to the imbalances in pre-intervention test scores.[17]

[14]It is not possible to identify the schools that participated in the "Jovem de Futuro" experiment using the public-access ENEM microdata before 2007. For this reason, we do not consider earlier implementations of the program in Minas Gerais and Rio Grande do Sul, because we would only have one year of pre-treatment outcome.

[15]For 2007 and 2008, we focus on the score on a 63-question multiple-choice test on various subjects (Portuguese, History, Geography, Math, Physics, Chemistry and Biology). Since 2009, the exam has been composed of 180 multiple-choice questions, equally divided into four areas of knowledge: languages, codes and related technologies; human sciences and related technologies; natural sciences and related technologies; and mathematics and its technologies. In this case, we consider the average score for these four areas. For each year and for each state, we standardize the test scores based on the sample of students from the experimental control schools.

[16]We exclude one control and two treated schools from Sao Paulo because they lack information for at least one of these years.

[17]Rosa (2015) analyzes the "Jovem de Futuro" program using a differences-in-differences approach, exploiting the experimental design of the program. He finds a positive and significant effect of the program for both Rio de Janeiro and Sao Paulo. There are a few differences in our analyses that justify the different results. First, we consider an intention to treat effect, including schools that abandoned the program after its implementation, while Rosa (2015) includes only strata with no attritors (see Ferman and Ponczek (2017) for a discussion on potential bias from the exclusion of strata with attrition problems). Second, Rosa (2015) considers an exam that was administered on the treated and control schools to evaluate this program. We are not able to use this dataset because this information is not available for non-experimental schools. Finally, we aggregate our data at the school level, while Rosa (2015) uses individual-level data.

Columns 2 and 4 of Table 1 present differences in test scores for public schools that did not participate in the experiment and schools in the experimental control group. In Rio de Janeiro, schools that (voluntarily) decided to participate in the experiment had better outcomes prior to the intervention, relative to other schools that did not participate in the experiment. In Sao Paulo, schools in the experimental control group were, on average, worse than the schools that did not participate in the experiment. Interestingly, Rio de Janeiro has 966 and Sao Paulo has 3481 non-experimental public schools, thus providing a setting with few treated and many (non-experimental) control schools.

5.1 Empirical Monte Carlo Study

We consider an empirical MC study based on the "Jovem de Futuro" implementation. We first estimate a probit model using schools' average test scores in the three years prior to the intervention as covariates. We estimate the probit model using the implementation of the program in Sao Paulo, which was a place where the program focused on attending schools with lower test scores, so treatment selection is a more severe problem in this case. We also include private schools to have a larger population for the simulation study.[18] Then we exclude the treated schools and draw placebo treatments for all schools in Brazil with a treatment selection process based on the estimated probit model. We have a population of 20,363 schools for this simulation study. Based on these simulations, we find, on average, a difference of -0.32 points in a standardized test score when we simply compare treated and control schools under this selection process, revealing that schools that participated in this program had, on average, worse test scores relative to other schools.

For each realization of the placebo treatment, we control the number of treated and control observations by selecting a random sample of $N_1 \in \{5, 10, 25, 50\}$ treated and $N_0 \in \{50, 500\}$ control schools. We then estimate the nearest neighbor matching estimator with $M \in \{1, 4, 10\}$, and calculate rejection rates based on the asymptotic distribution derived

[18]Simulation results are similar if we include only public schools. Results available upon request.

by Abadie and Imbens (2006), and based on the randomization inference tests presented in Section 4. For each scenario, we draw 10,000 samples.

Bias and Mean Square Error

Panel A of Table 2 shows the average bias of the nearest-neighbor matching estimator. Columns 1 and 2 has $M = 1$. For $N_0 = 50$, the matching estimator for the treatment effect on the treated has a bias of around 0.01, regardless of the number of observations in the treated group, which reflects the fact that, with a finite N_0, it is impossible to guarantee a perfect match in X for the treated observations and their nearest neighbors. This bias, however, equals only about 3% of the bias of a naive comparison between treated and control observations, suggesting that, in this setting, the matching estimator is very effective in controlling for differences in observables of treated and control schools, even when N_0 is not large. Consistent with Proposition 1, the average bias shrinks to zero when we increase the number of control observations, regardless of the number of treated observations. When the matching estimator has more nearest neighbors, the bias increases, but it remains close to zero when $N_0 = 500$. This happens because, with a limited number of control observations, we end up with poorer matches when considering an estimator with more nearest neighbors. This loss in match quality becomes less relevant when there are many control observations.

Panel B of Table 2 presents the mean square error (MSE) of the matching estimators. While the MSE is always decreasing in N_1 and N_0, two competing forces come into play when M increases. On the one hand, using more nearest neighbors reduces the variance of the matching estimator. On the other hand, this increases the bias of the estimator. With $N_0 = 500$, since increasing M from one to ten has little impact on the bias, using more nearest neighbors — in this range — always reduces the MSE of the matching estimator. However, with smaller N_0 there are some cases in which increasing M actually increases the MSE, exposing the trade-off between bias and variance for the matching estimator.

Appendix Table A.1 presents simulations when the dimensionality of the covariates increases.[19] While the number of covariates does not affect our theoretical results in Proposition 1, these simulations confirm the intuition that, when the dimensionality of the covariates increases, a larger N_0 is required to keep our approximations reliable. Finally, Appendix Table A.2 presents simulations for a bias-corrected estimator, as defined in equation 6.[20] While the average bias is reduced using this procedure, the effects on the MSE are ambiguous. In particular, the bias corrected estimator may lead to higher MSE when N_1 is very small and N_0 is not large. When N_0 is large, the bias correction becomes less relevant, so the bias and MSE of the two estimators become very similar.

Inference: test size

Panels C to E of Table 2 show rejection rates for 5% tests using different inference methods. A superscript "+" indicates a rejection rate greater than 6%, and a superscript "−", a rejection rate lower than 4%.[21] Importantly, while the different test procedures rely on different null hypotheses, all these null hypotheses are valid in the simulations. We discuss in detail the implications of considering tests that rely on different null hypotheses in Section 6.

Panel C of Table 2 presents rejection rates using the test based on Abadie and Imbens (2006).[22] Rejection rates for a 5% test are higher than 13% when $N_1 = 5$, and around 9% when $N_1 = 10$, for all values of N_0 and M. This happens because the asymptotic distribution derived by Abadie and Imbens (2006) relies on $N_1 \rightarrow \infty$, even though it allows N_0 to grow

[19] We generate three additional covariates with the same distributions of the test scores from 2007, 2008, and 2009, but that are independent of all other random variables in the model. Then we estimate the matching estimator including these variables, in addition to the original ones, as covariates. A mismatch in these additional variables would not directly generate bias in the matching estimator. However, the addition of these variables makes it harder to find a good match in terms of relevant covariates, which might lead to higher bias.

[20] We use linear least squares using only the nearest neighbors to estimate $\mu_0(x)$. This is the procedure used in the teffects command in Stata.

[21] While there is an asymmetry in that over-rejection is usually considered a more relevant problem relative to under-rejection, it is also important to highlight cases in which a test under-rejects, as this might imply that the test is under-powered.

[22] We consider in our simulations the default options of the teffect program in Stata, which uses the robust standard errors derived by Abadie and Imbens (2006) with two nearest neighbors for the estimation of the variance.

at a faster rate than N_1. When N_1 increases, rejection rates go down, although they are still marginally higher than 5% even when $N_1 = 50$. The simulations suggest that rejection rates computed using the asymptotic variance derived by Abadie and Imbens (2006) should be considered with caution when the number of treated observations is small.

Panel D of Table 2 shows rejection rates using randomization inference test based on permutations. Rejection rates are close to 5% in most cases. The exceptions are the scenarios with $M = 1/N_1 = 5$, and with $M = 10/N_1 \in \{25, 50\}$, in which the test is conservative. In both cases, the test is conservative because there are relatively few possible permutations.[23] In the first case, there are few possible permutations because the dimension of $\tilde{S}_{N_0}$ is small. Therefore, the test should remain conservative even when we increase N_0 even further. In the second case, the test is conservative because we end up with many shared nearest neighbors (see Remark 13). Therefore, the test would lead to rejection rates closer to 5% if we increase N_0.

Panel E of Table 2 shows rejection rates using the randomization inference test based on sign changes, presented in Section 4.2. When the nearest-neighbor matching estimator with $M = 1$ is considered, rejection rates using this test are close to 5%, except when $N_1 = 5$. In this case, few different group transformations exist, which explains why the test is conservative.[24] When we consider matching estimators with $M > 1$ and $N_0 = 50$, the test under-rejects the null hypothesis, even for larger N_1. This happens because increasing M increases the probability that different treated observations share the same nearest neighbors, which in turn reduces the number of group transformations. When $N_0 = 500$, this problem becomes less relevant, and rejection rates approach 5%, when $M = 4$. However, the test is still conservative when $M = 10$. Since this comes from a higher proportion of shared neighbors when $M = 10$, the test would lead to rejection rates closer to 5% if we increase

[23] We use the non-randomized version of the test in which we do not reject the null hypothesis in case of equality. We could guarantee the correct size if we used a randomized version of the test, as explained in Remark 11.

[24] Similar to the case of permutations, this happens because we use the non-randomized version of the test in which we do not reject in case of equality. We could guarantee the correct size if we used a randomized version of the test.

N_0 (except for the case with $N_1 = 5$).

Appendix Table A.1 show some over-rejection for the randomization inference tests when we increase the dimensionality of the covariates, which is explained by the fact that the bias is more relevant in this scenario. Again, such over-rejection does not arise if N_0 is large enough. When a bias-corrected estimator is used, Appendix Table A.2 also show some over-rejection in the permutation test when N_1 and N_0 are small, despite the fact that the bias is smaller. When N_0 is large, there is not much difference in rejection rates between the standard and the bias-corrected matching estimator. Finally, Appendix Table A.3 shows that the bootstrap test proposed by Otsu and Rai (2017) can also lead to over-rejection when N_1 is small. When N_1 and N_0 increases, rejection rates converge to 5%, which was expected given the theoretical results derived by Otsu and Rai (2017).[25]

Inference: test power

Table 3 present rejection rates when we assume a homogeneous treatment of 0.2 standard deviations in the students' outcomes (that is, $Y_i(1) = Y_i(0) + 0.2$ for all i). An important caveat when comparing these different inference procedures is that inference based on the asymptotic distribution derived by Abadie and Imbens (2006) leads to over-rejection under the null, particularly when N_1 is small. Therefore, these results should be considered with caution. As expected the power of these tests are increasing with N_1. The power is also increasing with M, but at decreasing rates, which is expected given the discussion presented by Imbens and Rubin (2015) that M should not be large. Most importantly, the two randomization inference tests present non-trivial power in many settings in which tests that rely on $N_1 \to \infty$ would lead to over-rejection. The only exceptions are the cases in which

[25]We focus on the wild bootstrap implementation of test using the two point distribution suggested by Mammen (1993). Another alternative proposed by Otsu and Rai (2017) would be a nonparametric bootstrap. However, with few treated and many control observations, we would likely generate bootstrap samples with no treated observations. Differently from the other tests we considered, this test must be based on a bias-corrected estimator, and it requires some properties on the estimator for $\mu_0(x)$ (see Otsu and Rai (2017) for details). Following Otsu and Rai (2017), we estimate $\mu_0(x)$ using a linear OLS with all control observations. We also present results using the estimator for $\mu_o(x)$ used by default in the teffects command in Stata, which makes the over-rejection more significant.

there are few possible group changes, so that the test is conservative. Therefore, there are settings in which the randomization inference tests may provide an important alternative to inference methods that rely on $N_1 \to \infty$. In Section 6, we contrast these different inference procedures in more detail, providing guidance on how to evaluate the trade-offs of these methods in different settings.

5.2 Empirical Application

Our idea is to estimate the effects of the program using a matching estimator with the experimental treated schools as treated observations and schools that did not participate in the experiment as control observations, therefore providing a setting with few treated and many control observations. Moreover, we take advantage of the randomized control trial to analyze the validity of the matching estimator and of different inference methods in this setting. More specifically, we consider a matching estimator using the experimental *control* schools as treated observations, and schools that did not participate in the experiment as control observations. Since the experimental control schools did not actually receive the treatment in the analyzed period, we should not expect to find significant effects in this case.

One important caveat in using ENEM test scores is that the treatment may have affected the probability that a student would take the exam. We do not find, however, significant differences in the number of students who took the exam between treated and control schools (see Appendix Table A.4). Moreover, one of our main exercises in this empirical application is to analyze the performance of matching estimators using the experimental *control* schools as the treated observations. Since the experimental control schools were not affected by the treatment, we do not have any reason to believe sample selection should be a problem in this case.

Table 4 shows estimated effects from 2010 to 2012 using the experimental *control* schools as the treated observations in our matching estimators. These schools volunteered to participate in the program, but were not actually treated during this period. Therefore, if the matching estimators are valid, then we should not expect to find significant effects. In addition to the point estimates, p-values are calculated using the asymptotic distribution derived by Abadie and Imbens (2006), and from the two proposed RI tests. We use test scores from 2007 to 2009 as matching variables. Interestingly, estimates for Rio de Janeiro (columns 1 to 4) generally have lower p-values using the test based on Abadie and Imbens (2006), relative to the alternative inference procedures. In particular, a test based on Abadie and Imbens (2006) would reject the null at 10% in two cases, while the other tests would fail to reject the null. This is consistent with our simulations from Section 5.1, that show the test based on Abadie and Imbens (2006) may lead to over-rejection when N_1 is small. The difference in p-values across different methods is less pronounced when we consider estimates for Sao Paulo, which is consistent with having a larger number of "treated" schools in Sao Paulo.

Finally, Table 5 presents estimated effects using the experimental treated schools as the treated observations in our matching estimators. The effects for Rio de Janeiro are small and not significantly different from zero, which is consistent with the experimental results presented in Table 1. For Sao Paulo, some results for 2011 and 2012 are significant, depending on the specification. While positive, the estimates for Sao Paulo are generally smaller than the experimental results presented in Table 1, which is consistent with the imbalances in pre-treatment outcomes for the experimental sample.

6 Choosing Among Alternative Inference Methods

The different test procedures we consider present important trade-offs in terms of size distortion, power, and the underlying null hypothesis they rely on. In light of the theoretical properties derived in Section 4, and of the empirical evidence presented in Section 5, we provide guidance on how to evaluate these trade-offs. First, note that tests based on the asymptotic distribution derived by Abadie and Imbens (2006), and on the bootstrap procedure proposed by Otsu and Rai (2017), are valid to test the null hypothesis $\tau(\{X_i\}_{i\in\mathcal{I}}) = 0$, provided both N_1 and N_0 are large enough so that the asymptotic approximations are reliable. The randomization inference tests, in contrast, rely on more stringent null hypotheses. Therefore, if we believe N_1 is large enough so that this approximation is reasonable, then we should use one of these tests that rely on $N_1 \to \infty$, instead of the randomization inference ones. In our simulations, for example, there is only a slight over-rejection when $N_1 = 50$, so the advantage of using an inference method that is valid to test a less stringent null hypothesis should dominate.

When N_1 is not that large, then the over-rejection of tests that rely on $N_1 \to \infty$ becomes more relevant, so it may be reasonable to consider alternative inference procedures that allow for N_1 fixed. The randomization inference test based on sign changes relies on a slightly more stringent null hypothesis, that the average treatment effect for each value of X_i is equal to zero. However, in light of Remark 15, if $\tau(\{X_i\}_{i\in\mathcal{I}}) = 0$, but the null is false because treatment effects are heterogeneous across X_i, then the test would under-reject. This means that such test would only reject at a rate greater than α when it is actually the case that $\tau(\{X_i\}_{i\in\mathcal{I}}) \neq 0$ (asymptotically, with N_1 fixed and $N_0 \to \infty$). Therefore, if the goal is to test the null $\tau(\{X_i\}_{i\in\mathcal{I}}) = 0$, then we should not expect over-rejection for any value of N_1. This test would have low power if N_1 is very small, or if N_0 is not very large, so that many treated observations share the same nearest neighbors. Since the proportion of shared nearest neighbors is increasing with M, this provides another reason to avoid using matching estimators with large M (see Imbens and Rubin (2015) for other reasons to avoid using large M). In our simulations, the test based on sign changes becomes an attractive alternative when $N_1 \in \{10, 25\}$. In these cases, it has the correct size and non-trivial power, while tests that rely on $N_1 \to \infty$ presented relevant size distortions. When $N_1 = 5$, however, this test is underpowered, so it should not be used.

Finally, the randomization inference test based on permutations is the only one that provides correct size and non-trivial power when N_1 is very small. However, it relies on a very stringent null hypothesis, which implies that we may reject at a rate greater than α even when $\tau(\{X_i\}_{i\in\mathcal{I}}) \neq 0$ (see Remark 7). Therefore, this test should only be used when alternative methods either lead to significant over-rejection or provide trivial power.

7 Conclusion

We consider the asymptotic properties of matching estimators when the number of control observations is large, but the number of treated observations is fixed. In this setting, the nearest neighbor matching estimator is asymptotically unbiased for the ATT under standard assumptions used in the literature on estimation of treatment effects under selection on unobservables. Moreover, we provide tests, based on the theory of randomization under approximate symmetry, that are asymptotically valid when the number of treated observations is fixed and the number of control observations goes to infinity. The different test procedures we consider present important trade-offs in terms of size distortion, power, and the underlying null hypothesis they rely on. We, therefore, provide guidance on on how to evaluate the trade-offs among these different test procedures in specific settings.

Our results are also relevant for Synthetic Control (SC) applications. Following Doudchenko and Imbens (2016), the SC and the matching estimators are nested in a framework in which the estimated counterfactual outcome for the treated observation is a linear combination of the outcomes for the controls. In the framework of Doudchenko and Imbens (2016), if we consider linear combinations of the controls such that the weights given to observations with large discrepancies in pre-treatment outcomes relative to the treated units go to zero, then, following the same arguments as we do for the matching estimator, the estimator is asymptotically unbiased if treatment assignment is "as good as random," conditional on this set of pre-treatment outcomes.[26] Moreover, under these conditions, the randomization

[26] If however, treatment assignment is only "as good as random" conditional on a set of common factors (which allows for some correlation between treatment assignment and post-treatment potential outcomes), then this would not necessarily be true. Abadie et al. (2010) show that, conditional on a perfect pre-treatment match, the bias of the SC estimator is bounded by a function that goes to zero when the number of pre-

inference test based on sign changes remains asymptotically valid when the number of control units goes to infinity. Given recent concerns regarding the validity of the placebo test proposed by Abadie et al. (2010) (see, for example, Ferman and Pinto (2017) and Hahn and Shi (2017)), the randomization inference test based on sign changes may provide a feasible alternative when there are multiple treated units and a large number of control units.[27] The only caveat is that a very large number of control observations are needed when the number of pre-treatment periods is large, so that approximations remain reliable.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statiscal Association*, 105(490):493–505.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.

Barros, R., de Carvalho, M., Franco, S., and Rosalém, A. (2012). Impacto do projeto jovem de futuro. *Estudos em Avaliação Educacional*, 23(51):214–226.

Bodory, H., Camponovo, L., Huber, M., and Lechner, M. (2018). The finite sample performance of inference methods for propensity score matching and weighting estimators. *Journal of Business & Economic Statistics*, 0(ja):1–43.

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*.

treatment periods increases, even if the number of control units is fixed, while Gobillon and Magnac (2016) show provide conditions such that this perfect pre-treatment match is achieved when number of control units and the number of pre-treatment periods go to infinity. See also Ferman and Pinto (2016) for a discussion of the conditions for asymptotic unbiasedness for the SC estimator when the number of control units is fixed, and a perfect pre-treatment match is not assumed.

[27] Kreif et al. (2016) propose a permutation test similar to the one of Abadie et al. (2010) for the case with multiple treated, so it is subject to the same concerns presented by Ferman and Pinto (2017) and Hahn and Shi (2017). Chernozhukov et al. (2017) propose a permutation test based on the timing of the intervention. This test, however, would require a very large number of periods. Instead, our test may be an alternative when the number of periods is not large, but the number of control units is large.

Busso, M., DiNardo, J., and McCrary, J. (2014). New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. *The Review of Economics and Statistics*, 96(5):885–897.

Canay, I. A. and Kamat, V. (2018). *The Review of Economic Studies*. Forthcoming.

Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3):1013–1030.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2017). An exact and robust conformal inference method for counterfactual and synthetic controls.

Conley, T. G. and Taber, C. R. (2011). Inference with Difference in Differences with a Small Number of Policy Changes. *The Review of Economics and Statistics*, 93(1):113–125.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Dehejia, R. H. and Wahba, S. (2002). Propensity Score-Matching Methods For Nonexperimental Causal Studies. *The Review of Economics and Statistics*, 84(1):151–161.

Doudchenko, N. and Imbens, G. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis.

Ferman, B. and Pinto, C. (2016). Revisiting the Synthetic Control Estimator. MPRA Paper 73982, University Library of Munich, Germany.

Ferman, B. and Pinto, C. (2017). Placebo Tests for Synthetic Controls. MPRA Paper 78079, University Library of Munich, Germany.

Ferman, B. and Pinto, C. (2018). Inference in differences-in-differences with few treated groups and heteroskedasticity. *The Review of Economics and Statistics, forthcoming.*

Ferman, B. and Ponczek, V. (2017). Should we drop covariate cells with attrition problems? Mpra paper, University Library of Munich, Germany.

Frolich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *The Review of Economics and Statistics*, 86(1):77–90.

Gobillon, L. and Magnac, T. (2016). Regional Policy Evaluation: Interative Fixed Effects and Synthetic Controls. *Review of Economics and Statistics*. Forthcoming.

Hahn, J. and Shi, R. (2017). Synthetic control and inference. *Econometrics*, 5(4).

Huber, M., Lechner, M., and Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, 175(1):1 – 21.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*.

Imbens, G. (2014). Matching Methods in Practice: Three Examples. NBER Working Papers 19959, National Bureau of Economic Research, Inc.

Imbens, G. and Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, USA.

Kreif, N., Grieve, R., Hangartner, D., Turner, A. J., Nikolova, S., and Sutton, M. (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Economics*, 25(12):1514–1528.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–20.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Statist.*, 21(1):255–285.

Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520):1720–1732.

Rosa, L. (2015). Avaliação de impacto do programa jovem de futuro.

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.*, 17(3):286–327.

Rothe, C. (2017). Robust confidence intervals for average treatment effects under limited overlap. *Econometrica*, 85(2):645–660.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183.

Young, A. (2015). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results.

Table 1: **"Jovem de Futuro": Summary Statistics**

	Rio de Janeiro		Sao Paulo	
	Exp. Treated - Exp. Control	Nonexp. Control - Exp. Control	Exp. Treated - Exp. Control	Nonexp. Control - Exp. Control
	(1)	(2)	(3)	(4)
	Panel A: Before treatment			
2007	0.040	-0.091	0.116***	0.117***
	(0.111)	(0.082)	(0.042)	(0.034)
2008	0.006	-0.136**	0.091**	0.061
	(0.098)	(0.059)	(0.041)	(0.046)
2009	0.026	-0.122	0.030	0.096**
	(0.111)	(0.079)	(0.053)	(0.045)
	Panel B: After treatment			
2010	-0.063	-0.197***	0.097*	0.070*
	(0.124)	(0.073)	(0.057)	(0.042)
2011	0.065	-0.086	0.142***	0.112***
	(0.101)	(0.059)	(0.048)	(0.039)
2012	0.016	-0.121**	0.129**	0.093**
	(0.102)	(0.050)	(0.054)	(0.041)
# of Schools				
Exp. Treated	15		39	
Exp. Control	15		39	
Nonexp. Control	966		3481	

Note: Columns 1 and 3 present differences in test scores between experimental treated and control schools, calculated using a regression with strata fixed effects, for Rio de Janeiro and Sao Paulo respectively. Columns 2 and 4 present differences between non-experimental public schools and experimental control schools, for Rio de Janeiro and Sao Paulo respectively. Test scores are normalized such that students in the experimental control group have zero mean and variance one for each year. From 2009 to 2012 there are separate test scores for math, Portuguese, natural sciences, and human sciences, so we use the average of these four scores. Robust standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2: **Empirical Monte Carlo Simulation**

	$M=1$		$M=4$		$M=10$	
	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: $\lvert$average bias $\times$ 100$\rvert$			
$N_1=5$	1.143	0.338	1.618	0.673	2.156	0.936
$N_1=10$	1.112	0.465	1.585	0.711	2.085	0.706
$N_1=25$	0.883	0.369	1.547	0.576	2.148	0.833
$N_1=50$	1.030	0.466	1.608	0.635	2.137	0.771
			Panel B: mean squared error ($\times$100)			
$N_1=5$	2.587	2.481	1.822	1.591	1.733	1.440
$N_1=10$	1.425	1.286	1.005	0.816	0.989	0.739
$N_1=25$	0.677	0.516	0.515	0.344	0.522	0.315
$N_1=50$	0.453	0.268	0.357	0.186	0.365	0.167
			Panel C: rejection rates based on AI (2006)			
$N_1=5$	0.139^{+}	0.148^{+}	0.140^{+}	0.145^{+}	0.133^{+}	0.144^{+}
$N_1=10$	0.093^{+}	0.098^{+}	0.084^{+}	0.089^{+}	0.090^{+}	0.090^{+}
$N_1=25$	0.068^{+}	0.065^{+}	0.067^{+}	0.064^{+}	0.077^{+}	0.071^{+}
$N_1=50$	0.063^{+}	0.055	0.071^{+}	0.062^{+}	0.082^{+}	0.064^{+}
			Panel D: test based on RI, permutations			
$N_1=5$	0.009^{-}	0.016^{-}	0.050	0.052	0.040^{-}	0.056
$N_1=10$	0.049	0.053	0.045	0.052	0.024^{-}	0.051
$N_1=25$	0.053	0.049	0.025^{-}	0.045	0.009^{-}	0.035^{-}
$N_1=50$	0.054	0.046	0.016^{-}	0.040^{-}	0.007^{-}	0.022^{-}
			Panel E: test based on RI, sign changes			
$N_1=5$	0.009^{-}	0.015^{-}	0.000^{-}	0.012^{-}	0.000^{-}	0.005^{-}
$N_1=10$	0.049	0.053	0.000^{-}	0.046	0.000^{-}	0.024^{-}
$N_1=25$	0.052	0.052	0.000^{-}	0.049	0.000^{-}	0.023^{-}
$N_1=50$	0.053	0.050	0.000^{-}	0.052	0.000^{-}	0.004^{-}

Note: This table presents simulation results from the empirical MC study described in Section 5.1. Panel A reports the average bias (multiplied by 100), while Panel B reports the mean squared error (multiplied by 100) of the matching estimator. Panel C presents rejection rates based on the asymptotic distribution derived by Abadie and Imbens (2006). Panel D presents rejection rates for the randomization inference test based on permutations, proposed in Section 4.2, while Panel E presents rejection rates for the randomization inference test based on sign changes, proposed in Section 4.1. We include a superscript "+" when rejection rate is greater than 6% and a superscript "−" when rejection rate is lower than 4%. For each combination (N_1, N_0), we run 10,000 simulations.

Table 3: **Empirical Monte Carlo Simulation: Test Power**

	$M=1$		$M=4$		$M=10$	
	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: rejection rates based on AI (2006)			
$N_1=5$	0.386^{+}	0.409^{+}	0.464^{+}	0.514^{+}	0.459^{+}	0.545^{+}
$N_1=10$	0.471^{+}	0.516^{+}	0.558^{+}	0.655^{+}	0.561^{+}	0.701^{+}
$N_1=25$	0.683^{+}	0.790^{+}	0.767^{+}	0.911^{+}	0.765^{+}	0.925^{+}
$N_1=50$	0.821^{+}	0.957^{+}	0.883^{+}	0.993^{+}	0.882^{+}	0.996^{+}
			Panel B: test based on RI, permutations			
$N_1=5$	0.034^{-}	0.051	0.268^{+}	0.341^{+}	0.209^{+}	0.372^{+}
$N_1=10$	0.314^{+}	0.372^{+}	0.407^{+}	0.562^{+}	0.262^{+}	0.582^{+}
$N_1=25$	0.608^{+}	0.737^{+}	0.560^{+}	0.880^{+}	0.301^{+}	0.877^{+}
$N_1=50$	0.785^{+}	0.945^{+}	0.633^{+}	0.984^{+}	0.308^{+}	0.976^{+}
			Panel C: test based on RI, sign changes			
$N_1=5$	0.036^{-}	0.050	0.001^{-}	0.053	0.000^{-}	0.025^{-}
$N_1=10$	0.312^{+}	0.374^{+}	0.003^{-}	0.467^{+}	0.000^{-}	0.257^{+}
$N_1=25$	0.606^{+}	0.739^{+}	0.000^{-}	0.843^{+}	0.000^{-}	0.312^{+}
$N_1=50$	0.786^{+}	0.943^{+}	0.000^{-}	0.962^{+}	0.000^{-}	0.045

Note: Note: This table presents simulation results from the empirical MC study described in Section 5.1, when we consider a homogeneous treatment effect of 0.20 standard deviations in the individual-level test scores. Panel A presents rejection rates based on the asymptotic distribution derived by Abadie and Imbens (2006). Panel B presents rejection rates for the randomization inference test based on permutations, proposed in Section 4.2, while Panel C presents rejection rates for the randomization inference test based on sign changes, proposed in Section 4.1. We include a superscript "+" when rejection rate is greater than 6% and a superscript "−" when rejection rate is lower than 4%. For each combination (N_1, N_0), we run 10,000 simulations.

Table 4: **Non-experimental Results, Experimental Control Schools as Treated Observations**

	Rio de Janeiro			Sao Paulo		
	$M = 1$	$M = 4$	$M = 10$	$M = 1$	$M = 4$	$M = 10$
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment effects in 2010						
Point Estimate	0,087	-0,003	0,046	0,000	0,018	0,004
p-values:						
AI (2006)	0,091	0,941	0,086	0,995	0,601	0,924
RI-permutation	0,124	0,960	0,449	0,996	0,679	0,933
RI-sign changes	0,123	0,938	0,179	0,996	0,609	0,917
Treatment effects in 2011						
Point Estimate	0,043	-0,032	0,000	-0,019	-0,027	-0,013
p-values:						
AI (2006)	0,566	0,396	0,997	0,746	0,475	0,692
RI-permutation	0,659	0,626	0,999	0,771	0,553	0,783
RI-sign changes	0,662	0,438	0,997	0,734	0,496	0,693
Treatment effects in 2012						
Point Estimate	0,070	-0,019	0,006	-0,072	-0,034	-0,019
p-values:						
AI (2006)	0,263	0,522	0,885	0,169	0,383	0,616
RI-permutation	0,295	0,742	0,918	0,189	0,453	0,665
RI-sign changes	0,306	0,576	0,896	0,185	0,382	0,495

Note: This table presents non-experimental results using a matching estimator with experimental control schools as treated observations and non-experimental schools as control observations. Columns 1 to 3 present results for Rio de Janeiro using 1, 4, or 10 nearest neighbors in the estimation, while columns 4 to 6 present results for Sao Paulo. We present the estimated effects separately for 2010, 2011, and 2012. For each estimate, we present p-values calculated based on the asymptotic distribution derived by Abadie and Imbens (2006), and based on the randomization inference procedures described in Section 4.

Table 5: **Non-experimental Results, Experimental Treated Schools as Treated Observations**

	Rio de Janeiro			Sao Paulo		
	$M = 1$	$M = 4$	$M = 10$	$M = 1$	$M = 4$	$M = 10$
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment effects in 2010						
Point Estimate	-0,056	-0,012	0,017	0,039	0,025	0,051
p-values:						
AI (2006)	0,319	0,736	0,596	0,412	0,516	0,119
RI-permutation	0,344	0,851	0,816	0,429	0,587	0,305
RI-sign changes	0,349	0,766	0,624	0,427	0,538	0,111
Treatment effects in 2011						
Point Estimate	-0,100	0,045	0,033	0,040	0,070	0,055
p-values:						
AI (2006)	0,247	0,415	0,545	0,318	0,080	0,123
RI-permutation	0,280	0,600	0,667	0,321	0,098	0,181
RI-sign changes	0,273	0,551	0,537	0,346	0,116	0,213
Treatment effects in 2012						
Point Estimate	0,023	0,030	0,044	0,054	0,089	0,063
p-values:						
AI (2006)	0,719	0,516	0,293	0,312	0,032	0,090
RI-permutation	0,739	0,694	0,543	0,311	0,052	0,142
RI-sign changes	0,720	0,566	0,257	0,337	0,043	0,122

Note: This table replicates the results from Table 4 using the experimental treated schools as treated observations for the matching estimators.

A Supplemental Appendix for "Matching Estimators with Few Treated and Many Control Observations

A.1 Proof of Main Results

A.1.1 Proof of Proposition 1

Proof. For a given realization of $X_i = \bar{x}$ for an observation in the treated group and for a given $\epsilon > 0$, consider the probability that the M-closest realizations of $\{X_j\}_{j \in \mathcal{I}_0}$ are such that $d(X_j, \bar{x}) < \epsilon$. Let $X^i_{(M)}$ be the M-closest match of observation i. Then,

$$\begin{aligned}
\Pr\left(d(X^i_{(M)}, \bar{x}) > \epsilon\right) &= \sum_{m=0}^{M-1} \Pr\left(d(X_j, \bar{x}) < \epsilon \text{ for exactly } m \text{ observations}\right) \\
&= \sum_{m=0}^{M-1} \binom{N_0}{m} [\Pr(d(X_j, \bar{x}) < \epsilon)]^m [\Pr(d(X_j, \bar{x}) > \epsilon)]^{N_0 - m}. \quad (18)
\end{aligned}$$

Since $\bar{x} \in \mathbb{X}_0$, we have that $Pr(d(X_j, \bar{x}) < \epsilon) > 0$, which implies that $\Pr(d(X_j, \bar{x}) > \epsilon) < 1$. Therefore, we have that $\Pr\left(d(X^i_{(M)}, \bar{x}) > \epsilon\right) \to 0$. By analogy, the m-nearest neighbor of i for $m < M$ also converges in probability to $\bar{x}$.

Now consider

$$\mathbb{E}[\hat{\tau} | \{X_i\}_{i \in \mathcal{I}_1}] = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left(\mu_1(X_i) - \mathbb{E}\left[\frac{1}{M} \sum_{m=1}^{M} \mu_0(X^i_{(m)}) \right] \right). \quad (19)$$

Since $\mu_0(x)$ is continuous and bounded and $X^i_{(m)} \xrightarrow{p} X_i$, then we have that $\mathbb{E}[\mu_0(X^i_{(m)}) | X_i] \to \mu_0(X_i)$, which proves part 1 of Proposition 1.

For part 2, we assume that $\tilde{h}(x) = \mathbb{E}[h(Y(0)) | X = x]$ is continuous and bounded for any $h : \mathbb{R} \to \mathbb{R}$ continuous and bounded. Let $Y^i_{(m)}$ be the outcome of the m-nearest neighbor of treated observation i. Therefore, for any $h(y)$ continuous and bounded, and for a given $X_i = \bar{x}$, we have that

$$\mathbb{E}[h(Y^i_{(m)})] = \mathbb{E}\left\{ \mathbb{E}[h(Y^i_{(m)}) | X^i_{(m)}] \right\} = \mathbb{E}\left\{ \tilde{h}(X^i_{(m)}) \right\} \to \tilde{h}(\bar{x}) = \mathbb{E}[h(Y(0)) | X = \bar{x}]. \quad (20)$$

By the Portmanteau Lemma, we have that $Y^i_{(m)} \xrightarrow{d} Y(0) | \{X = \bar{x}\}$. Under Assumption 2, $Y^i_{(m)} \xrightarrow{d} \mu_0(X_i) + \epsilon_m(X_i)$, where $\epsilon_m(X_i) \overset{d}{=} Y_i(0) | X_i - \mu_0(X_i)$. Therefore, conditional on $\{X_i\}_{i \in \mathcal{I}_1}$,

$$\hat{\tau} = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left[Y_i - \frac{1}{M} \sum_{m=1}^{M} Y^i_{(m)} \right] \xrightarrow{d} \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left[(\mu_1(X_i) - \mu_0(X_i)) + \left(\epsilon_i - \frac{1}{M} \sum_{m=1}^{M} \epsilon_m(X_i) \right) \right]. \quad (21)$$

Now we just have to show that $\epsilon_m(X_i)$ is independent across m and i. Since X_i is a continuous random variable, then $X_i \neq X_j$ with probability one for $i \neq j$ with $i, j \in \mathcal{I}_1$.

Since there is a finite number of treated observations, then it must be that, conditional on $\{X_i\}_{i=1}^{N_1}$, there is an $\eta > 0$ such that $d(X_i, X_j) > \eta$ for all $i, j \in \mathcal{I}_1$ with $i \neq j$. However, we know that $Pr(d(X_i, X^i_{(m)}) > \epsilon) \to 0$ for all $\epsilon > 0$. Therefore, the probability that $k \in \mathcal{I}_0$ belongs to $\mathcal{J}_M(i)$ and $\mathcal{J}_M(j)$ converges to zero. Under the assumption that the errors ϵ_i are independent across i (which is guaranteed from Assumption 1), we have that $\epsilon_m(X_i)$ is independent across m and i. ■

A.1.2 Particular case: $Y(0)|X$ is normally distributed

Let $Y \sim N(\theta, \sigma^2)$. We first want to show that $\tilde{h}(\theta, \sigma) = \mathbb{E}[h(Y)|\theta, \sigma]$ is continuous and bounded for any $h()$ continuous and bounded. In this case,

$$\tilde{h}(\theta, \sigma) = \int h(y) \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2} dy. \tag{22}$$

Let $g(y, \theta, \sigma) = h(y)\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}e^{-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2}$. Since $h(y)$ is continuous and bounded, $g(y, \theta, \sigma)$ is integrable for all (θ, σ), and, for all $y \in \mathbb{R}$, $g(y, \theta, \sigma)$ is continuous in (θ, σ). We now show that there is a neighborhood of (θ, σ) and an integrable function $q : \mathbb{R} \to \mathbb{R}_+$ such that, for all (θ, σ) in this neighborhood, $|g(y, \theta, \sigma)| \leq q(y)$.

Consider the neighborhood of (θ, σ) given by $(\theta - \delta, \theta + \delta) \times (\sigma - \delta, \sigma + \delta)$ (where δ is sufficiently small so that $\sigma - \delta > 0$), and define

$$q(y) = \begin{cases} |h(y)|\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma-\delta}e^{-\frac{1}{2}\left(\frac{y-(\theta+\delta)}{\sigma+\delta}\right)^2}, \text{ if } y > \theta + \delta \\ |h(y)|\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma-\delta}, \text{ if } y \in [\theta - \delta, \theta + \delta] \\ |h(y)|\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma-\delta}e^{-\frac{1}{2}\left(\frac{y-(\theta-\delta)}{\sigma+\delta}\right)^2}, \text{ if } y < \theta - \delta \end{cases} \tag{23}$$

For any $(\theta, \sigma) \in (\theta - \delta, \theta + \delta) \times (\sigma - \delta, \sigma + \delta)$, $|g(y, \theta, \sigma)| \leq q(y)$, and $q(y)$ is integrable. Therefore, $h(\theta, \sigma)$ is continuous at any point (θ, σ). Moreover, since $h(y)$ is bounded, $\tilde{h}(\theta, \sigma)$ is also bounded.

Now let $Y(0)|X = x \sim N(\theta(x), \sigma(x))$. Since compositions of continuous functions are continuous, it follows that $\tilde{h}(x) = \int h(y)\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma(x)}e^{-\frac{1}{2}\left(\frac{y-\theta(x)}{\sigma(x)}\right)^2} dy$ is bounded and continuous in x.

A.1.3 Unconditional Expectation

Now we consider the unconditional expectation of $\hat{\tau}$:

$$\mathbb{E}[\hat{\tau}] = \mathbb{E}\{\mathbb{E}[\hat{\tau}|\{X_i\}_{i \in \mathcal{I}_1}]\} = \frac{1}{N_1}\sum_{i \in \mathcal{I}_1} \mathbb{E}\left[\mu_1(X_i) - \frac{1}{M}\sum_{m=1}^{M} \mu_0(X^i_{(m)})\right]. \tag{24}$$

We need that $\mathbb{E}[\mu_0(X^i_{(m)})] \to \mathbb{E}[\mu_0(X_i)]$. We know that $\mathbb{E}[\mu_0(X^i_{(m)})|X_i] \to \mu_0(X_i)$ for all X_i. Again using the fact that $\mu_0(x)$ is continuous and bounded, we have that $\mathbb{E}[\mu_0(X^i_{(m)})] =$

$\mathbb{E}\{\mathbb{E}[\mu_0(X^i_{(m)})|X_i]\} \to \mathbb{E}[\mu_0(X_i)]$. Therefore,

$$\mathbb{E}[\hat{\tau}] \to \mathbb{E}\left[\mu_1(X_i) - \mu_0(X_i)\right] \tag{25}$$

where this expectation is taken according to $f_1(x)$, the density function of the treated units.

A.1.4 Bias-corrected Matching Estimator

We consider the bias-corrected matching estimator using linear least squares on the nearest neighbors to estimate $\mu_0(x)$. This is the procedure used in the teffects command in Stata. Considering, for simplicity, the case with $k = 1$, note that

$$\hat{\tau}_{biasadj} = \hat{\tau} + \frac{1}{N_1}\frac{1}{M}\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\hat{\beta}\left(X^i_{(m)} - X_i\right) \tag{26}$$

where $\hat{\beta} = \frac{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\left(X^i_{(m)}-\bar{X}_1\right)Y^i_{(m)}}{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\left(X^i_{(m)}-\bar{X}_1\right)^2}$ and $\bar{X} = \frac{1}{N_1}\frac{1}{M}\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}X^i_{(m)}$. We assume that $Y_i(0)|X_i = x$ is uniformly bounded for almost all $x \in \mathbb{X}_0$ and that X_i is bounded.[28] Define $\mathcal{X} = \sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}(X^i_{(m)} - \bar{X}_1)^2$. If we have at least two treated observations, then $\exists C_1 > 0$ such that $\Pr\left(\mathcal{X} < C_1\right) \to 0$. Therefore,

$$\begin{aligned}
\Pr\left(|\hat{\beta}| \geq c\right) &= \Pr\left(\left|\frac{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\left(X^i_{(m)} - \bar{X}_1\right)Y^i_{(m)}}{\mathcal{X}}\right| \geq c\right) \\
&\leq \Pr\left(\frac{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\left|X^i_{(m)} - \bar{X}_1\right|\left|Y^i_{(m)}\right|}{\mathcal{X}} \geq c\right) \\
&\leq \Pr\left(\frac{C_2\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\left|Y^i_{(m)}\right|}{\mathcal{X}} \geq c \mid \mathcal{X} < C_1\right)\Pr\left(\mathcal{X} < C_1\right) \\
&\quad +\Pr\left(\frac{C_2\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\left|Y^i_{(m)}\right|}{C_1} \geq c \mid \mathcal{X} > C_1\right)\Pr\left(\mathcal{X} > C_1\right).
\end{aligned}$$

Since $\Pr\left(\mathcal{X} < C_1\right) \to 0$, the first term converges to zero. Since we assume that $Y_i(0)|X_i = x$ is uniformly bounded for almost all $x \in \mathbb{X}_0$, we can always find c such that the second term is lower than any $\eta > 0$, which implies that $\hat{\beta} = O_p(1)$. Since $X^i_{(m)} - X_i = o_p(1)$ for all i and m, $\frac{1}{N_1}\frac{1}{M}\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\hat{\beta}\left(X^i_{(m)} - X_i\right) = o_p(1)$, so $|\hat{\tau}_{biasadj} - \hat{\tau}| = o_p(1)$.[29]

[28] These assumptions are weaker than the assumptions of Abadie and Imbens (2011).

[29] The proof would be easier if we used all control observations to estimate $\mu_0(x)$ using linear least squares. In this case, $\hat{\beta}$ would converge to the population OLS coefficients.

A.1.5 Proof of Proposition 2

Proof. We apply Theorem 3.1 from Canay et al. (2017). We first show that, when $N_0 \to \infty$, the limiting distribution of $\tilde{S}_{N_0}$ under the null, $\tilde{S}$, is invariant to transformations in $\tilde{\mathbf{G}}$. From the proof of Proposition 1, note that $Y^i_{(m)} \xrightarrow{d} Y_i(0)|X_i$. Therefore, under the null that $Y_i(0)|X_i \overset{d}{=} Y_i(1)|X_i$, we have that $\tilde{S}^j_{N_0,i} \xrightarrow{d} Y_i(0)|X_i$ for all $j = 0, ..., M$. Moreover, asymptotically, $\tilde{S}^j_{N_0,i}$ is independent across i and j because the probability that two treated units share the same nearest neighbor converges to zero when $N_0 \to \infty$. Therefore, the asymptotic distribution of $\tilde{S}_{N_0}$ is invariant to transformations in $\tilde{\mathbf{G}}$.

We also have that the test statistic function $\tilde{T}(\tilde{S})$ is continuous. Finally, we show that, for two distinct elements $\pi \in \tilde{G}$ and $\pi' \in \tilde{G}$, either $\tilde{T}(\tilde{S}^\pi) = \tilde{T}(\tilde{S}^{\pi'})$ for all possible realizations of $\tilde{S}$, or $Pr(\tilde{T}(\tilde{S}^\pi) \neq \tilde{T}(\tilde{S}^{\pi'})) = 1$. Suppose $M > 1$. Then, if π and π' are such that $\pi_i(0) = \pi'_i(0)$ for all $i \in \mathcal{I}_1$, then we will have $\tilde{T}(\tilde{S}^\pi) = \tilde{T}(\tilde{S}^{\pi'})$ for all possible realizations of $\tilde{S}$. If π and π' are such that $\pi_i(0) \neq \pi'_i(0)$ for at least one $i \in \mathcal{I}_1$, then the probability that $\tilde{T}(\tilde{S}^\pi) = \tilde{T}(\tilde{S}^{\pi'})$ would be equal to zero, because $\tilde{S}$ is a continuous random variable. For the case $M = 1$, we would have $\tilde{T}(\tilde{S}^\pi) = \tilde{T}(\tilde{S}^{\pi'})$ for all possible realizations of $\tilde{S}$ if π and π' are such that $\pi_i(0) = \pi'_i(0)$ for all i, or $\pi_i(0) = \pi'_i(1)$ for all i. Otherwise, $Pr(\tilde{T}(\tilde{S}^\pi) \neq \tilde{T}(\tilde{S}^{\pi'})) = 1$. ■

A.1.6 Proof of Proposition 3

Proof. Again, we apply Theorem 3.1 from Canay et al. (2017). We first show that, when $N_0 \to \infty$, the limiting distribution of S_{N_0} under the null, S, is invariant to sign changes. This is true if, asymptotically, $\hat{\tau}_i$ and $\hat{\tau}_j$ are independent for $i \neq j$, and the distribution of $\hat{\tau}_i$ is symmetric around zero. It is not necessary for $\hat{\tau}_i$ to have the same distribution across i. From Proposition 1, we know that, under the null, the asymptotic distribution of $\hat{\tau}_i$ conditional on $\{X\}_{i \in \mathcal{I}_1}$ is given by $\epsilon_i - \frac{1}{M}\sum_{m=1}^{M} \epsilon_m(X_i)$. This distribution is symmetric around zero given the assumption that $Y_i(1)|X_i$ and $Y_i(0)|X_i$ are symmetric around their mean for all $i = 1, ..., N_1$. Moreover, Proposition 1 also shows that, asymptotically, $\hat{\tau}_i$ are independent across i.

We also have that the test statistic function $T(S)$ is continuous. Finally, we show that, for two distinct elements $g \in G$ and $g' \in G$, either $T(gS) = T(g'S)$ for all possible realizations of S, or $Pr(T(gS) \neq T(g'S)) = 1$. If g and g' are such that $g_i = g'_i$ for all i, or $g_i = -g'_i$ for all i, then $T(gS) = T(g'S)$ for all possible realizations of S. Otherwise, given that S is a continuous random variable, $Pr(T(gS) \neq T(g'S)) = 1$. ■

A.2 Appendix Tables and Figures

Table A.1: **Empirical Monte Carlo Simulation: More Covariates**

	$M=1$		$M=4$		$M=10$	
	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: $\lvert$*average bias* $\times 100\rvert$			
$N_1=5$	3,181	1,703	3,999	2,280	4,625	2,509
$N_1=10$	2,822	1,717	3,776	2,201	4,889	2,951
$N_1=25$	3,005	1,744	3,656	2,196	4,538	2,657
$N_1=50$	2,657	1,657	3,476	2,138	4,294	2,644
			Panel B: mean squared error ($\times 100$)			
$N_1=5$	3,190	2,679	2,228	1,720	2,282	1,645
$N_1=10$	1,682	1,299	1,261	0,887	1,393	0,875
$N_1=25$	0,820	0,557	0,718	0,403	0,764	0,389
$N_1=50$	0,543	0,299	0,493	0,227	0,588	0,240
			Panel C: rejection rates based on AI (2006)			
$N_1=5$	$0{,}154^{+}$	$0{,}156^{+}$	$0{,}144^{+}$	$0{,}154^{+}$	$0{,}154^{+}$	$0{,}156^{+}$
$N_1=10$	$0{,}091^{+}$	$0{,}096^{+}$	$0{,}099^{+}$	$0{,}098^{+}$	$0{,}122^{+}$	$0{,}108^{+}$
$N_1=25$	$0{,}073^{+}$	$0{,}072^{+}$	$0{,}095^{+}$	$0{,}082^{+}$	$0{,}108^{+}$	$0{,}094^{+}$
$N_1=50$	$0{,}072^{+}$	$0{,}067^{+}$	$0{,}093^{+}$	$0{,}084^{+}$	$0{,}129^{+}$	$0{,}107^{+}$
			Panel D: test based on RI, permutations			
$N_1=5$	$0{,}013^{-}$	$0{,}015^{-}$	0,049	0,050	$0{,}038^{-}$	0,056
$N_1=10$	0,051	0,051	0,047	0,058	$0{,}032^{-}$	0,057
$N_1=25$	$0{,}064^{+}$	0,055	0,043	0,059	$0{,}022^{-}$	0,044
$N_1=50$	$0{,}069^{+}$	0,059	$0{,}028^{-}$	0,057	$0{,}021^{-}$	0,045
			Panel E: test based on RI, sign changes			
$N_1=5$	$0{,}012^{-}$	$0{,}015^{-}$	$0{,}000^{-}$	$0{,}011^{-}$	$0{,}000^{-}$	$0{,}004^{-}$
$N_1=10$	0,049	0,051	$0{,}000^{-}$	0,048	$0{,}000^{-}$	$0{,}015^{-}$
$N_1=25$	$0{,}064^{+}$	0,055	$0{,}000^{-}$	$0{,}061^{+}$	$0{,}000^{-}$	$0{,}002^{-}$
$N_1=50$	$0{,}069^{+}$	0,059	$0{,}000^{-}$	0,059	$0{,}000^{-}$	$0{,}000^{-}$

Note: This table replicates the simulations presented in Table 2 with the difference that we add three additional covariates that are uncorrelated with the potential outcomes.

Table A.2: **Empirical Monte Carlo Simulation: Bias-corrected Estimator**

	$M=1$		$M=4$		$M=10$	
	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: $\lvert$*average bias* $\times$ 100$\rvert$					
$N_1=5$	13,181	0,141	0,212	0,014	0,495	0,064
$N_1=10$	0,042	0,027	0,357	0,062	0,684	0,151
$N_1=25$	0,452	0,125	0,452	0,105	0,620	0,255
$N_1=50$	0,285	0,007	0,403	0,081	0,614	0,130
	Panel B: mean squared error (×100)					
$N_1=5$	>100	5,704	2,229	1,669	1,856	1,477
$N_1=10$	1,889	1,349	1,119	0,814	1,009	0,737
$N_1=25$	0,739	0,518	0,551	0,340	0,508	0,297
$N_1=50$	0,491	0,266	0,365	0,181	0,347	0,162
	Panel C: rejection rates based on AI (2006)					
$N_1=5$	$0{,}146^+$	$0{,}145^+$	$0{,}138^+$	$0{,}144^+$	$0{,}137^+$	$0{,}142^+$
$N_1=10$	$0{,}094^+$	$0{,}098^+$	$0{,}087^+$	$0{,}089^+$	$0{,}087^+$	$0{,}091^+$
$N_1=25$	$0{,}071^+$	$0{,}064^+$	$0{,}071^+$	$0{,}063^+$	$0{,}071^+$	$0{,}064^+$
$N_1=50$	$0{,}070^+$	0,054	$0{,}068^+$	$0{,}061^+$	$0{,}075^+$	$0{,}061^+$
	Panel D: test based on RI, permutations					
$N_1=5$	$0{,}010^-$	$0{,}016^-$	$0{,}085^+$	$0{,}065^+$	$0{,}060^+$	$0{,}066^+$
$N_1=10$	0,049	0,051	$0{,}064^+$	0,055	$0{,}031^-$	0,055
$N_1=25$	0,054	0,050	$0{,}032^-$	0,047	$0{,}010^-$	$0{,}032^-$
$N_1=50$	0,056	0,048	$0{,}016^-$	$0{,}039^-$	$0{,}005^-$	$0{,}017^-$
	Panel E: test based on RI, sign changes					
$N_1=5$	$0{,}009^-$	$0{,}014^-$	$0{,}000^-$	$0{,}011^-$	$0{,}000^-$	$0{,}005^-$
$N_1=10$	0,049	0,052	$0{,}000^-$	0,045	$0{,}000^-$	$0{,}028^-$
$N_1=25$	0,052	0,050	$0{,}000^-$	0,048	$0{,}000^-$	$0{,}023^-$
$N_1=50$	0,058	0,048	$0{,}000^-$	0,049	$0{,}000^-$	$0{,}004^-$

Note: This table replicates the simulations presented in Table 2 with the difference that it considers a bias-corrected matching estimator suggested by Abadie and Imbens (2011). The large bias reported for the case with $N_1 = 5$ and $N_0 = 50$ comes from the fact that the variance of the bias-corrected estimator is very large. With 10,000 simulations, it is not possible to reject that the average bias is equal to zero.

Table A.3: **Empirical Monte Carlo Simulation: Wild Bootstrap**

	$M = 1$		$M = 4$		$M = 10$	
	$N_0 = 50$	$N_0 = 500$	$N_0 = 50$	$N_0 = 500$	$N_0 = 50$	$N_0 = 500$
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: estimate $\mu_0(x)$ using all observations						
$N_1 = 5$	0.062^+	0.052	0.095^+	0.089^+	0.105^+	0.111^+
$N_1 = 10$	0.055	0.049	0.070^+	0.068^+	0.076^+	0.080^+
$N_1 = 25$	0.059	0.046	0.061^+	0.056	0.060	0.059
$N_1 = 50$	0.061^+	0.043	0.058	0.050	0.067^+	0.052
Panel B: estimate $\mu_0(x)$ using anly nearest neighbors						
$N_1 = 5$	0.157^+	0.140^+	0.119^+	0.104^+	0.123^+	0.118^+
$N_1 = 10$	0.107^+	0.082^+	0.085^+	0.068^+	0.085^+	0.083^+
$N_1 = 25$	0.088^+	0.059	0.075^+	0.060	0.073^+	0.060
$N_1 = 50$	0.093^+	0.054	0.078^+	0.053	0.080^+	0.054

Note: This table presents rejection rates for the same simulations presented in Table 2 using the wild bootstrap procedure proposed by Otsu and Rai (2017). Panel A estimates $\mu_0(x)$ using linear OLS for the full sample of controls, as done by Otsu and Rai (2017) in their simulations. Panel B estimates $\mu_0(x)$ using linear OLS only for the sample of nearest neighbors, which is how the method is implemented by default in Stata.

Table A.4: **"Jovem de Futuro": Effects of the Treatment on ENEM Enrollment**

	Rio de Janeiro		Sao Paulo	
	Control	Treated - Control	Control	Treated - Control
	(1)	(2)	(3)	(4)
		Panel A: Before treatment		
2007	129.467	-8.200	72.872	6.162
	[67.567]	(34.314)	[29.119]	(11.797)
2008	146.667	-21.667	76.256	3.579
	[61.586]	(20.350)	[31.212]	(9.625)
2009	123.600	-13.800	46.103	7.368
	[61.814]	(21.177)	[19.820]	(8.451)
		Panel B: After treatment		
2010	153.267	-32.067	55.154	11.737
	[71.611]	(20.282)	[26.298]	(9.957)
2011	157.400	-21.133	72.154	-0.947
	[79.469]	(25.025)	[34.159]	(9.993)
2012	210.800	-7.933	83.641	5.737
	[92.378]	(47.515)	[41.161]	(11.407)

Note: Column 1 presents the number of students that took the ENEM exam in control schools in Rio de Janeiro, while column 4 presents this information for control schools in Sao Paulo. Standard deviation in brackets. Columns 2 and 4 present differences between experimental treated and control schools, calculated using a regression with strata fixed effects. * significant at 10%; ** significant at 5%; *** significant at 1%.