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: March, 2018

Please click here for the most recent version

Abstract

We analyze the properties of matching estimators when the number of treated observations is fixed and the number of control observations is large. We show that, under standard assumptions, the nearest neighbor matching estimator for the average treatment effect on the treated is asymptotically unbiased. However, the estimator is not consistent, and it is generally not asymptotically normal. Since large-sample inferential techniques are inadequate in our setting, we provide alternative inferential procedures based on the theory of randomization tests under approximate symmetry. These tests are asymptotically valid when the number of treated observations is fixed and the number of control observations goes to infinity. Simulations show that our inference methods provide better size and power when compared to existing alternatives. We explore the validity of matching estimators, and of our inferential methods, in the estimation of the effects of an educational program in Brazil that provides a setting with few treated and many control schools.

Keywords: matching estimator, treatment effect, hypothesis testing, randomization inference
JEL Codes: C12; C13; C21

*The author gratefully acknowledges the comments and suggestions of Luis Alvarez, Lucas Finamor, Sergio Firpo, Ricardo Masini, Cristine Pinto, Ricardo Paes de Barros, Vitor Possebom, Pedro Sant'Anna, and participants of the 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 theoretical 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.

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 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 unobservables.[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. Unlike Abadie and Imbens (2006), in our setting, 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 rely on asymptotic results that assume that the number of treated observations goes to infinity.[3] When the dimensionality of the covariates is low, and we consider matching estimators with few nearest neighbors, our Monte Carlo (MC) simulations suggest that the bias of the matching estimator is close to zero, even when the number of control observations is not large, regardless of the number of treated observations. Increasing the dimensionality of the covariates and/or increasing the number of nearest neighbors implies that we need an increasing number of controls to keep our approximation 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

[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.

[3] The finite sample properties of matching estimators have been evaluated in detail in simulations by Frolich (2004) and Busso et al. (2014). 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.

derived by Abadie and Imbens (2006) should not provide a good approximation when the number of treated observations is small, even if there are many control observations. 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] Therefore, we consider alternative inference methods. We first provide two inference procedures 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 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 is fixed. We also consider the approach suggested by Rothe (2017), which provides valid confidence intervals in finite samples, and a wild bootstrap procedure proposed by Otsu and Rai (2017).[6,7]

When the number of treated observations is small, our simulations show significant over rejection for inference based both on the asymptotic distribution derived by Abadie and Imbens (2006) and on wild bootstrap. In the absence of finite-sample bias, the two randomization inference methods we propose and the method suggested by Rothe (2017) control size well with few treated observations in all scenarios, even when the number of control observations is not large. The randomization inference test based on permutations is the most powerful among these three tests, when treatment effect is homogeneous. However, this test relies on a sharper null hypothesis that, conditional on observables, the potential outcomes when treated and untreated have the same distribution. The randomization inference test based on sign changes, and the test based on Rothe (2017), rely on less stringent null hypotheses, but they have poor power in some scenarios.[8] These tests have correct size even when we consider the possibility of finite-sample bias, as long as the number of nearest neighbors used in the estimation and the dimension of the matching covariates are relatively low. With matching estimators using many nearest neighbors and/or multidimensional covariates,

[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 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] The approach suggested by Rothe (2017) is valid in finite samples if potential outcomes are normally distributed and the bias of the matching estimator is negligible. If the number of treated observations is small but the number of control observations is large, then we show that the bias will be negligible. Also, as explained by Rothe (2017), normality is an "asymptotically irrelevant" assumption.

[7] Otsu and Rai (2017) suggest a weighted bootstrap procedure in which the wild bootstrap is a particular case. We do not consider the non-parametric version of their weighted bootstrap because, with few treated observations, such procedure would likely generate bootstrap samples with no treated observation.

[8] The test based on sign changes has poor power when the number of nearest neighbors used for estimation is large relative to the number of control observations, while the test based on Rothe (2017) has poor power when we use few nearest neighbors in the estimation.

the tests remain valid, but it is necessary to have a larger number of control observations to avoid over-rejection. Taken together, our simulations suggest that the alternatives we propose are more reliable than tests that rely on a large number of treated and control observations. This is true even even when the number of treated observations is not very small, and when the number of control observations is not very large. For example, our permutation test provides more reliable hypothesis testing, relative to existing alternatives, even when 100 observations are equally divided in two groups.

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.[9] We find significant treatment effects for Sao Paulo, and small and insignificant effects for Rio de Janeiro, which is consistent with the 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 our proposed randomization inference procedures and the test based on Rothe (2017) control well for size.

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 for our setting. In Section 5, we evaluate the properties of the matching estimator, and we contrast alternative inferential methods, using MC simulations. We present our empirical application in Section 6. Concluding remarks, including a discussion of potential implications for Synthetic Control applications, are presented in Section 7.

[9]Other 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).

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.[10] 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 6, there are a large number of schools that could potentially receive the treatment, but only a small number of schools actually receive 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

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

[10] We abstract from the case in which components of X_i are discrete because, as argued by Abadie and Imbens (2006), discrete covariates with a finite number of support points can be easily dealt with by analyzing estimation of average treatment effects within subsamples defined by their values.

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.[11] However, the density $f_1(X_i)$ for $i \in \mathcal{I}_1$ can potentially be different from the density $f_0(X_i)$ for $i \in \mathcal{I}_0$. This difference in density 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 $N_0 \to \infty$.

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

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

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 $\mu_w(x) = \mathbb{E}[Y(w)|X=x, W_i=1]$.[12] Under Assumption 2, we have that $\mu(x,0) = \mu_0(x)$. Using this notation, note that the CATT is given by

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

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 $\mathbb{E}[\hat{\tau}|\{X_i\}_{i \in \mathcal{I}_1}] \to \tau(\{X_i\}_{i \in \mathcal{I}_1})$ when $N_0 \to \infty$ and N_1 is fixed.*

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

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

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

[12] 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.

Proof. 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 $N_0 \to \infty$, because 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. See details in Appendix A.1. ■

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 $N_0 \to \infty$. 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.

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 assumption used in part 2 of Proposition 1 implies 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 $X^i_{(m)} \xrightarrow{p} \bar{x}$.

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.

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.

Remark 4 With additional assumptions, we can also guarantee that the bias-corrected matching estimator has the same asymptotic distribution as the matching estimator without bias correction. The intuition again is that $\hat{\mu}_0(X_i) - \hat{\mu}_0(X^i_{(m)})$ converge to zero when $N_0 \to \infty$ because $X^i_{(m)} \xrightarrow{p} X_i$. See details in Appendix A.1.

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) should not provide a good approximation in our setting. We therefore consider alternative inference methods. We propose two tests based on the theory of randomization tests under an approximate symmetry assumption developed by Canay et al. (2017), and we show that they 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 transformations given by sign changes. Then we consider a test based on the confidence intervals for treatment effects under limited overlap derived by Rothe (2017), and a test based on wild bootstrap derived by Otsu and Rai (2017). These tests differ in their underlying assumptions and null hypotheses. Moreover, the size and power of these tests depend crucially on the number of observations in the treatment and control groups, and also on the number of nearest neighbors used in the estimation. In Section 5 we consider the finite sample properties of these tests, and we analyze in detail the conditions under which these tests provide valid size and non-trivial power.

4.1 Randomization Inference Test Based on Permutations

Consider a function of the data given by

$$\tilde{S}_{N_0} = \left(\tilde{S}^0_{N_0,1}, \tilde{S}^1_{N_0,1}, ..., \tilde{S}^M_{N_0,1}, ..., \tilde{S}^0_{N_0,N_1}, \tilde{S}^1_{N_0,N_1}, ..., \tilde{S}^M_{N_0,N_1} \right)' \tag{6}$$

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 the number of control observations. The matching estimator is given by

$$\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N_1} \left(\tilde{S}^0_{N_0,i} - \frac{1}{M} \sum_{j=1}^{M} \tilde{S}^j_{N_0,i} \right). \tag{7}$$

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

$$\tilde{T}^{(1)}(\tilde{S}_{N_0}) \leq \tilde{T}^{(2)}(\tilde{S}_{N_0}) \leq ... \leq \tilde{T}^{(\tilde{K})}(\tilde{S}_{N_0}) \tag{8}$$

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

$$\tilde{T}(\tilde{S}^{\pi}_{N_0}) = \left[\frac{1}{N_1} \sum_{i=1}^{N_1} \left(\tilde{S}^{\pi_i(0)}_{N_0,i} - \frac{1}{M} \sum_{j=1}^{M} \tilde{S}^{\pi_i(j)}_{N_0,i} \right) \right]^2 . \tag{9}$$

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

$$\tilde{\phi}(S_{N_0}) = \begin{cases} 1 & \text{if } \tilde{T}(\tilde{S}_{N_1}) > \tilde{T}^{(\tilde{k})}(\tilde{S}_{N_1}) \\ 0 & \text{if } \tilde{T}(\tilde{S}_{N_1}) \leq \tilde{T}^{(\tilde{k})}(\tilde{S}_{N_1}). \end{cases} \tag{10}$$

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.

Proposition 2 *Under the same assumptions used in part 2 of Proposition 1, and considering a null hypothesis that $Y_i(0)|X_i \overset{d}{=} Y_i(1)|X_i$ for all $i \in \mathcal{I}_1$, a test based on the decision rule defined in 10 is asymptotically level α for any $\alpha \in (0,1)$ when $N_0 \to \infty$.*

Proof.

We apply Theorem 3.1 from Canay et al. (2017). We only need to show 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 $Y^i_{(m)} \xrightarrow{d} Y_i(0)|X_i$. 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 the transformations in $\tilde{\mathbf{G}}$. ■

Remark 5 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 $N_0 \to \infty$ using the theory of randomization inference under approximate symmetry developed by Canay et al. (2017).

Remark 6 An important limitation of this test is that it relies on a null hypothesis that $Y_i(0)|X_i \overset{d}{=} Y_i(1)|X_i$ for all $i \in \mathcal{I}_1$. To understand why this assumption is crucial for this test, suppose, for example, that $\mathbb{E}[Y_i(1)|X_i] = \mathbb{E}[Y_i(0)|X_i]$, 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 leads to a rejection rate higher than α. 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, $\mathbb{V}[Y_i(1)|X_i] < \mathbb{V}[Y_i(0)|X_i]$).

Remark 7 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 8 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. The under rejection will only be relevant if $\tilde{K}$ is very small.

Remark 9 This test is also asymptotically valid for bias-corrected matching estimators. In this case, we define $\tilde{S}^0_{N_0,i} = Y_i - \hat{\mu}_0(X_i)$ and $\tilde{S}^m_{N_0,i} = Y^i_{(m)} - \hat{\mu}_0(X^i_{(m)})$.

4.2 Randomization Inference Test Based on Sign Changes

We consider now an alternative function of the data given by

$$S_{N_0} = (\hat{\tau}_1, ..., \hat{\tau}_{N_1})' \tag{11}$$

where $\hat{\tau}_i = Y_i - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j$. Each $\hat{\tau}_i$ 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

$$T(S_{N_0}) = \frac{|\hat{\tau}|}{\sqrt{\frac{1}{N_1 - 1}\sum_{i=1}^{N_1}(\hat{\tau}_i - \hat{\tau})^2}} \tag{12}$$

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

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

$$T^{(1)}(S_{N_0}) \leq T^{(2)}(S_{N_0}) \leq ... \leq T^{(K)}(S_{N_0}) \tag{13}$$

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

$$\phi(S_{N_0}) = \begin{cases} 1 & \text{if } T(S_{N_1}) > T^{(k)}(S_{N_1}) \\ 0 & \text{if } T(S_{N_1}) \leq T^{(k)}(S_{N_1}). \end{cases} \tag{14}$$

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

Proposition 3 *Under the same assumptions used in part 2 of Proposition 1, if we further assume that ϵ_i is symmetric around zero for all $i = 1, ..., N_1$, and consider a null hypothesis that $\mu_1(X_i) = \mu_0(X_i)$ for all $i \in \mathcal{I}_1$, then a test based on the decision rule defined in 14 is asymptotically level α for any $\alpha \in (0, 1)$ when $N_0 \to \infty$.*

Proof.

Again, we apply Theorem 3.1 from Canay et al. (2017). We only need to show that, when $N_0 \to \infty$, the limiting distribution of S_{N_0} under the null 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 ϵ_i is symmetric around zero for all $i = 1, ..., N_1$. Moreover, Proposition 1 also shows that, asymptotically, $\hat{\tau}_i$ are independent across i. Therefore, the assumptions for Theorem 3.1 of Canay et al. (2017) are satisfied. ■

Remark 10 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 hypothesis is implied by more narrowly defined null

hypotheses that are usually considered in Fisher-type tests, such as $Y_i(0)|X_i \stackrel{d}{=} Y_i(1)|X_i$ or $Y_i(0) = Y_i(1)$ with probability one. 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.

Remark 11 Remark 8 also applies to this test.

4.3 Test based on Rothe (2017)

Rothe (2017) constructs robust confidence intervals for treatment effects estimators under limited overlap. The main idea of his approach is that, under limited overlap, "local sample sizes" can be effectively very small in applications, so that approximations based on asymptotic theory would not be reliable. Instead, he constructs confidence intervals based on classical approaches to small sample inference. He shows that inference for the matching estimator can be considered as a generalized version of the Behrens-Fisher problem, where the test statistic is a studentized version of a linear combination of independent means. In the case in which X is discrete and can take J different values, the matching estimator for the ATT is a linear combination of $J+1$ sample means.[13] Under the assumption that outcomes are normally distributed, he constructs a confidence interval that guarantees coverage greater than or equal to $1-\alpha$ (Proposition 2 of Rothe (2017)). With continuous covariates, Rothe (2017) considers a partition of the data based on an estimated propensity score. He shows that, if the bias is negligible, then the conclusion based on discrete covariates is still valid.

We consider a slightly different way to partition the data, based on the nearest neighbors of the treated observations. More specifically, we consider a partition in which a treated observation i is joint with its M nearest neighbors. Therefore, if treated observations i and i' share at least one nearest neighbor, then they belong to the same partition. Suppose we end up with J partitions, and let $S_j(i) = 1$ if observation i belongs to partition j. Then the estimator for the average treatment effect on the treated would be given by

$$\hat{\tau}' = \hat{\mu}_1 - \sum_{j=1}^{J} \hat{\mu}_0(j)\hat{f}_1(j) \tag{15}$$

where $\hat{\mu}_1 = \frac{1}{N_1}\sum_{i\in\mathcal{I}_1} Y_i$ is the average of the treated observations, $\hat{\mu}_0(j) = \frac{1}{\sum_{i\in\mathcal{I}_0} S_j(i)}\sum_{i\in\mathcal{I}_0} S_j(i)Y_i$ is the average of the control observations in partition j, and $\hat{f}_1(j) = \frac{\sum_{i\in\mathcal{I}_1} S_j(i)}{N_1}$ is the proportion of the treated observations that belong to partition j. Since the probability that two treated observations share the same

[13] One for the treated observations, and J for the control observations with each $X = x$.

nearest neighbor goes to zero when N_1 is fixed and $N_0 \to \infty$, note that, for a fixed M, the estimators $\hat{\tau}$ and $\hat{\tau}'$ are asymptotically equivalent. Importantly, this estimator is a linear combination of independent sample means, so the insights of Rothe (2017) apply to this case. If we assume that the finite sample bias of the matching estimator is negligible, then we can construct a test statistic and calculate a critical value that guarantee a rejection rate of at most α for an α-level test if $Y_i|X_i$ is normally distributed. This is valid even in finite samples.

Remark 12 Calculating critical values for this method requires at least two control observations for each partition of the data.

Remark 13 Unlike the tests presented in Sections 4.1 and 4.2, the null hypothesis in this case is that the sample or the population ATT equals zero.

4.4 Test based on wild bootstrap

We also consider a bootstrap procedure based on the work of Otsu and Rai (2017). As explained by Abadie and Imbens (2008), naive bootstrap procedures are not valid for matching estimators because they fail to reproduce the distribution of the number of times each observation is used as a match. Otsu and Rai (2017) overcome this problem by considering bootstrap procedures that treat the number of times an observation is used for a match as one of the characteristics of the sample. More specifically, let $\tilde{\tau}$ be a bias-corrected estimator for the average treatment effect on the treated using $\hat{\mu}_0(x)$ as an estimator for $\mu_0(x)$. Otsu and Rai (2017) note that[14]

$$\tilde{\tau} = \frac{1}{N_1}\sum_{i=1}^{N}\left[W_i(Y_i - \hat{\mu}_0(X_i)) - (1-W_i)\frac{K_M(i)}{M}(Y_i - \hat{\mu}_0(X_i))\right] = \frac{1}{N_1}\sum_{i=1}^{N}\tilde{\tau}_i \tag{16}$$

where $K_M(i)$ is the number of times a control observation i is used as a match and $\tilde{\tau}_i = W_i(Y_i - \hat{\mu}_0(X_i)) - (1-W_i)\frac{K_M(i)}{M}(Y_i - \hat{\mu}_0(X_i))$. The weighted bootstrap counterpart for $\sqrt{N_1}(\tilde{\tau} - \tau)$ is obtained as

$$\sqrt{N_1}T^* = \sum_{i=1}^{N} e_i^*(\tilde{\tau}_i - W_i\tilde{\tau}) \tag{17}$$

where e_i^* are random variables satisfying specific conditions explained by Otsu and Rai (2017). Two particular cases that are encompassed in this model are nonparametric bootstrap (Efron (1979)) and wild bootstrap

[14] We use a different notation than Otsu and Rai (2017).

(Mammen (1993)). Otsu and Rai (2017) show that such bootstrap procedures are asymptotically valid when $N \to \infty$, even if we consider a setting in which N_0 grows at a faster rate than N_1. Importantly, the asymptotic theory of Otsu and Rai (2017) also relies on $N_1 \to \infty$, as does that of Abadie and Imbens (2006).

5 Monte Carlo Simulations

We use a data generating process (DGP) similar to the one used by Frolich (2004) and Busso et al. (2014) in our Monte Carlo (MC) simulations. Following Busso et al. (2014), these DGPs can be expressed as

$$\begin{aligned} Y_i(0) &= m(Z_i) + \sigma\epsilon_i \\ W_i^* &= \alpha + \beta Z_i - U_i \end{aligned} \tag{18}$$

where $Z_i = \Lambda(\sqrt{2}X_i)$, $\Lambda(.)$ is the logistic function, and X_i is a normal covariate; the error term U_i is i.i.d. standard uniform and is independent of ϵ_i and X_i; and W_i^* is the latent variable corresponding to treatment ($W_i = 1$ if $W_i^* > 0$). Since we want to consider the case in which N_1 is finite while N_0 is large, we generate a large population based on this DGP, and then we sample a small number N_1 of treated observations and a large number N_0 of control observations.[15] Frolich (2004) considers five combinations of (α, β). For clarity, we focus on the combination of (α, β) used in design 1 of Frolich (2004), which sets $\alpha = 0$ and $\beta = 1$. This is the design that induces the highest correlation between treatment assignment and covariate X among the parameters considered.

We start presenting in Section 5.1 a simpler case in which $m(.) = 0$, and ϵ_i is normally distributed and independent of X, so that there is no selection on observables. This way can focus on the size and power of the different inferential procedures, without the finite sample bias of matching estimators. In this case, all assumptions of Rothe (2017) are satisfied. In Section 5.2, we consider a functional form $m(.)$ of Frolich (2004), so that the matching estimator is biased in finite samples.[16] This way, we can analyze how different specifications affect the finite sample bias of the matching estimator and the rejection rates for the different test procedures. For each scenario, we draw 10,000 samples for MC simulations.

[15] We use the program available in the supplemental appendix of Busso et al. (2014).

[16] We focus on specification 1 from Frolich (2004). Results using alternative specifications are similar. Results available upon request.

5.1 Simulations with no selection on observables

Test size

We start with a simple case in which $Y_i(0)|X_i \sim N(0,1)$ and $Y_i(0) = Y_i(1)$. In this case, the matching estimator is unbiased even in finite samples. Table 1 shows rejection rates for 5% tests using different inference methods for combinations of (N_1, N_0) where $N_1 \in \{5, 10, 25, 50\}$ and $N_0 \in \{50, 500\}$. A superscript "+" indicates a rejection rate greater than 6%, and a superscript "−", a rejection rate lower than 4%.

Panel A of Table 1 presents rejection rates using the test based on Abadie and Imbens (2006) for different matching estimators for the ATT, varying the number of nearest neighbors, $M \in \{1, 2, 4, 10\}$.[17] Rejection rates for a 5% test are higher than 13% when $N_1 = 5$, for all values of N_0 and M. This happens because the asymptotic distribution derived by Abadie and Imbens (2006) relies on $N_1 \to \infty$, even though it allows N_0 to grow at a faster rate than N_1. When N_1 increases, rejection rates go down. However, except for the case in which $N_0 = 500$, rejection rates do not approach 5%, even when we increase N_1. For example, with $N_0 = N_1 = 50$, the rejection rate is still greater than 7.3% for most specifications.

The simulations suggest that rejection rates computed using the asymptotic variance derived by Abadie and Imbens (2006) may not be reliable when the number of treated observations is small. Panel B of Table 1 therefore shows rejection rates using randomization inference test based on permutations. As discussed in Section 4.1, 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. To take that into account, we consider permutations of treatment status in partitions of the sample as discussed in Section 4.3. The probability that this finite sample adjustment is relevant goes to zero when $N_0 \to \infty$.[18] Rejection rates are remarkably close to 5% in all cases. The only exception is when $N_1 = 5$ and $M = 1$, in which case the test is overly conservative because there are relatively few possible permutations.[19]

Panel C of Table 1 presents rejection rates using the randomization inference test based on sign changes, presented in Section 4.2. As in the previous test, a key feature of the test based on sign changes is that $\hat{\tau}_i$ become asymptotically independent, because the probability that two treated observations share the same

[17] 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.

[18] 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.

[19] 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. See Canay et al. (2017) for details.

Table 1: **Test Sizes - No Selection on Observable**

	$M=1$		$M=2$		$M=4$		$M=10$	
	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)
Panel A: test based on AI (2006)								
$N_1=5$	0.146^+	0.147^+	0.141^+	0.146^+	0.133^+	0.148^+	0.134^+	0.148^+
$N_1=10$	0.086^+	0.096^+	0.086^+	0.093^+	0.083^+	0.093^+	0.084^+	0.092^+
$N_1=25$	0.071^+	0.067^+	0.075^+	0.065^+	0.073^+	0.066^+	0.068^+	0.064^+
$N_1=50$	0.075^+	0.055	0.081^+	0.057	0.078^+	0.055	0.067^+	0.057
Panel B: test based on RI, permutation								
$N_1=5$	0.020^-	0.016^-	0.047	0.047	0.048	0.049	0.052	0.049
$N_1=10$	0.049	0.049	0.046	0.050	0.050	0.052	0.051	0.049
$N_1=25$	0.048	0.052	0.051	0.048	0.052	0.049	0.051	0.051
$N_1=50$	0.049	0.049	0.050	0.047	0.052	0.049	0.051	0.051
Panel C: test based on RI, sign changes								
$N_1=5$	0.007^-	0.013^-	0.003^-	0.012^-	0.000^-	0.009^-	0.000^-	0.006^-
$N_1=10$	0.038^-	0.051	0.023^-	0.050	0.004^-	0.046	0.000^-	0.034^-
$N_1=25$	0.048	0.050	0.041	0.052	0.010^-	0.048	0.000^-	0.045
$N_1=50$	0.050	0.050	0.045	0.050	0.005^-	0.046	0.000^-	0.046
Panel D: test based on Rothe (2017)								
$N_1=5$	-	-	0.002^-	0.000^-	0.023^-	0.020^-	0.037^-	0.037^-
$N_1=10$	-	-	0.001^-	0.000^-	0.027^-	0.022^-	0.043	0.042
$N_1=25$	-	-	0.002^-	0.000^-	0.038^-	0.027^-	0.048	0.046
$N_1=50$	-	-	0.007^-	0.000^-	0.046	0.030^-	0.047	0.050
Panel E: test based on wild bootstrap								
$N_1=5$	0.072^+	0.058	0.085^+	0.080^+	0.100^+	0.101^+	0.118^+	0.128^+
$N_1=10$	0.063^+	0.051	0.072^+	0.066^+	0.082^+	0.078^+	0.090^+	0.088^+
$N_1=25$	0.073^+	0.051	0.073^+	0.055	0.080^+	0.061^+	0.083^+	0.066^+
$N_1=50$	0.084^+	0.051	0.084^+	0.053	0.090^+	0.056	0.092^+	0.054

Note: This table presents simulation results using Design 1 from Frolich (2004) and Busso et al. (2014). Potential outcomes are normally distributed with mean zero and variance one. Panel A presents rejection rates under the null based on the asymptotic distribution of the matching estimator derived by Abadie and Imbens (2006) (AI). Panel B presents rejection rates under the null for the randomization inference test based on permutations, proposed in Section 4.2 (RI, permutation). Panel C presents rejection rates under the null for the randomization inference test based on sign changes, proposed in Section 4.1 (RI, sign changes). Panel D presents rejection rates under the null for the test based on the robust confidence intervals derived by Rothe (2017). Finally, Panel E presents rejection rates under the null for the test based on wild bootstrap. 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.

nearest neighbor converges to zero. For finite N_0, however, there is a positive probability that $\hat{\tau}_i$ is correlated across i, as different treated observations may share the same nearest neighbor. For this reason, we consider a slight modification of our test, 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 $N_0 \to \infty$. 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.[20] 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%.

Panel D of Table 1 presents rejection rates for the test based on Rothe (2017), as described in Section 4.3. As explained in Remark 12, this test is not well-defined for the case of $M = 1$. While the test is well defined for $M = 2$, note that rejection rates virtually equal zero in this case. Therefore, while it is possible to guarantee that this test is level α even in finite samples, it is overly conservative for the case with very few nearest neighbors. With more nearest neighbors, rejection rates approach 5%. Finally, Panel E of Table 1 presents rejection rate using the bootstrap test based on Otsu and Rai (2017). We focus on the wild bootstrap implementation of test, because a nonparametric bootstrap with few treated observations would likely generate bootstrap samples with no treated observations. Following Otsu and Rai (2017), we estimate $\mu_0(x)$ using a linear regression with all control observations, and the two point distribution suggested by Mammen (1993).[21] Rejection rates are generally higher than 5%, except when $N_1 = 50$ and $N_0 = 500$. This is consistent with the fact that the test relies on an asymptotic approximation with $N_1 \to \infty$, even though it allows N_1 to grow at a slower rate relative to N_0.

Test power

Given that the two randomization inference tests and the test based on Rothe (2017) have correct size in all scenarios (although, in some cases, they may be conservative), we consider the power of these tests. Table 2 presents rejection rates when $Y_i(1) = Y_i(0) + 0.5$ for these three tests. In most scenarios, the randomization inference test based on permutations has the highest power. The randomization inference test based on sign changes has good power when M is small relative to the number of control observations, but poor power otherwise. This is not surprising, given that this test is overly conservative when there are few control observations relative to the number of nearest neighbors used in the estimation. Finally, the test based on Rothe (2017) has good power when M is large, but poor power otherwise. Again, this is consistent with the fact that the test based on Rothe (2017) is overly conservative when a matching estimator has few nearest

[20] 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.

[21] That is, we assign $e_i^* = (\sqrt{5}-1)/2$ with probability $(\sqrt{5}+1)/2\sqrt{5}$ and $e_i^* = (\sqrt{5}+1)/2$ with probability $(\sqrt{5}-1)/2\sqrt{5}$.

neighbors.

Table 2: **Test Power - No Selection on Observable**

	$M=1$		$M=2$		$M=4$		$M=10$	
	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)
Panel A: test based on RI, permutation								
$N_1=5$	0.033^-	0.030^-	0.105^+	0.119^+	0.134^+	0.150^+	0.155^+	0.170^+
$N_1=10$	0.128^+	0.155^+	0.179^+	0.218^+	0.215^+	0.266^+	0.249^+	0.307^+
$N_1=25$	0.222^+	0.351^+	0.295^+	0.453^+	0.365^+	0.541^+	0.441^+	0.610^+
$N_1=50$	0.286^+	0.569^+	0.408^+	0.707^+	0.524^+	0.795^+	0.637^+	0.854^+
Panel B: test based on RI, sign changes								
$N_1=5$	0.011^-	0.025^-	0.006^-	0.025^-	0.002^-	0.021^-	0.000^-	0.011^-
$N_1=10$	0.103^+	0.156^+	0.067^+	0.183^+	0.012^-	0.193^+	0.000^-	0.146^+
$N_1=25$	0.233^+	0.353^+	0.202^+	0.431^+	0.035^-	0.471^+	0.000^-	0.427^+
$N_1=50$	0.322^+	0.578^+	0.294^+	0.684^+	0.023^-	0.719^+	0.000^-	0.621^+
Panel C: test based on Rothe (2017)								
$N_1=5$	-	-	0.004^-	0.001^-	0.071^+	0.073^+	0.108^+	0.120^+
$N_1=10$	-	-	0.008^-	0.001^-	0.151^+	0.172^+	0.217^+	0.259^+
$N_1=25$	-	-	0.029^-	0.001^-	0.320^+	0.448^+	0.444^+	0.593^+
$N_1=50$	-	-	0.103^+	0.003^-	0.506^+	0.737^+	0.646^+	0.853^+

Note: This table presents simulation results using Design 1 from Frolich (2004) and Busso et al. (2014). Potential outcomes are normally distributed with mean zero and variance one. For the treated observations, we add a treatment effect of 0.5. Panel A presents rejection rates for the randomization inference test based on permutations, proposed in Section 4.2 (RI, permutation). Panel B presents rejection rates for the randomization inference test based on sign changes, proposed in Section 4.1 (RI, sign changes). Panel D presents rejection rates for the test based on the robust confidence intervals derived by Rothe (2017), 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.

The dominance of the randomization inference test based on permutations in these simulations relies crucially on the use of alternatives with homogeneous treatment effect. If $\mathbb{V}[Y_i(1)|X_i] < \mathbb{V}[Y_i(0)|X_i]$, then the test based on permutations would likely have lower power than alternative tests in some scenarios (see Remark 6).

5.2 Simulations with selection on observables

In Section 5.1 we consider a simplified DGP, such that potential outcomes are unrelated with covariates that determine treatment assignment. This simplification enables analysis of different inference methods without finite N_0 bias of the matching estimator. Now, we consider a case in which potential outcomes are correlated with X, so the matching estimator is biased when N_0 is finite. We consider the first conditional expectation

function $m(.)$ used by Frolich (2004), and set $\sigma = \sqrt{0.1}$.[22]

Panel A of Table 3 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 equals about 5% of the bias of a naive comparison between treated and control observations. 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 rises significantly when $N_0 = 50$, 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 is less relevant when there are many control observations, which explains why bias increases only slightly when $N_0 = 500$.

Panel B of Table 3 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.

Finally, Panels C-E of Table 3 presents rejection rates for alternative inference methods under selection on observables. With $M \in \{1, 2\}$, the test based on permutations still controls well for size. This happens because the finite N_0 bias of the matching estimator is negligible, relative to the variance of the matching estimator, so it does not generate strong size distortions. With $M \in \{4, 10\}$, however, strong size distortions appear when $N_0 = 50$. This happens because both the bias increases and the variance of the estimator decreases, so the finite N_0 bias of the matching estimator becomes more relevant and generates larger size distortions. With $N_0 = 500$, however, rejection rates are close to 5% even when $M \in \{4, 10\}$.[23] The test based on sign changes never over-rejects. However, it is overly conservative (and has poor power) in the

[22] Simulations using the other specifications are qualitatively the same. Results available upon request.

[23] The over-rejection is more relevant if $\sigma = \sqrt{0.01}$ instead of $\sigma = \sqrt{0.1}$. This is expected, because decreasing the variance of ϵ_i reduces the variance of the matching estimator, but it does not affect the average finite N_0 bias. Still, rejection rates remain close to 5% for for the permutation test when $N_0 = 500$, except when $M = 10$. See Appendix Table A.1.

Table 3: **MC Results with Selection on Observable**

	$M=1$		$M=2$		$M=4$		$M=10$	
	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)
				Panel A: average $\|bias \times 1000\|$				
$N_1=5$	10.30	1.42	12.68	1.84	22.07	2.57	49.16	4.69
$N_1=10$	9.41	0.58	12.99	0.63	22.60	1.54	49.34	4.30
$N_1=25$	8.47	0.15	12.41	0.95	22.42	2.19	49.05	5.38
$N_1=50$	6.74	0.21	12.23	0.31	22.16	1.16	49.50	4.30
				Panel B: mean squared error ($\times 1000$)				
$N_1=5$	48.07	40.81	36.25	30.69	30.93	25.58	29.37	22.20
$N_1=10$	28.61	20.67	21.86	15.50	18.80	12.83	18.36	11.24
$N_1=25$	17.76	9.20	13.46	6.93	11.69	5.81	11.53	5.05
$N_1=50$	13.49	5.17	10.35	3.94	8.87	3.31	9.22	2.94
				Panel C: test based on RI, permutation				
$N_1=5$	0.021^-	0.016^-	0.046	0.045	0.050	0.045	0.062^+	0.046
$N_1=10$	0.048	0.047	0.050	0.045	0.056	0.046	0.081^+	0.045
$N_1=25$	0.054	0.051	0.052	0.052	0.068^+	0.049	0.247^+	0.051
$N_1=50$	0.051	0.049	0.056	0.050	0.132^+	0.047	0.529^+	0.058
				Panel D: test based on RI, sign changes				
$N_1=5$	0.007^-	0.014^-	0.003^-	0.014^-	0.001^-	0.009^-	0.000^-	0.004^-
$N_1=10$	0.037^-	0.047	0.023^-	0.045	0.004^-	0.044	0.000^-	0.028^-
$N_1=25$	0.050	0.051	0.043	0.049	0.008^-	0.047	0.000^-	0.044
$N_1=50$	0.048	0.045	0.043	0.046	0.005^-	0.050	0.000^-	0.050
				Panel E: test based on Rothe (2017)				
$N_1=5$	-	-	0.002^-	0.000^-	0.024^-	0.022^-	0.047	0.039^-
$N_1=10$	-	-	0.002^-	0.000^-	0.031^-	0.017^-	0.062^+	0.032^-
$N_1=25$	-	-	0.004^-	0.000^-	0.047	0.018^-	0.229^+	0.031^-
$N_1=50$	-	-	0.010^-	0.000^-	0.117^+	0.022^-	0.533^+	0.039^-

Note: This table presents simulation results using Design 1 and the conditional expectation function 1 from Frolich (2004) and Busso et al. (2014). Panel A reports the average bias (multiplied by 1000), while Panel B reports the mean squared error (multiplied by 1000) of the matching estimator. Panel C presents rejection rates for the randomization inference test based on permutations, proposed in Section 4.2 (RI, permutation). Panel D presents rejection rates for the randomization inference test based on sign changes, proposed in Section 4.1 (RI, sign changes). Panel E presents rejection rates for the test based on the robust confidence intervals derived in Rothe (2017), 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.

settings in which the bias would be largest. At the other extreme, the test based on Rothe (2017) has good size and non-trivial power in specific scenarios with many nearest neighbors and many control observations.

5.3 Multidimensional covariates

The MC simulations in Section 5.2 use a unidimensional covariate X_i. As stressed by Abadie and Imbens (2006), the bias of the matching estimator converges to zero at a lower rate when X_i is multidimensional. While this does not affect our theoretical results in Proposition 1, it can have important effects on the finite N_0 behavior of the matching estimator. In order to evaluate the implications of a multidimensional X_i on simulations comparable to those in Section 5.2, we include a marginal modification in the DGP. We generate $k-1$ new random variables $\widetilde{X}_{2i}, ..., \widetilde{X}_{ki}$, with the same distribution as X_i, that are independent of all other random variables in the model. Then we estimate the matching estimator using $\widetilde{X}_i = (X_i, \widetilde{X}_{2i}, ..., \widetilde{X}_{ki})'$ as covariates. A mismatch in $\widetilde{X}_{k'i}$ for $k' = 2, ..., k$ would not directly generate bias in the matching estimator. However, the addition of these variables makes it more difficult to find a good match in terms of X_i, which might lead to higher bias.

The MC results for the case with $k = 2$ appears in Table 4. For a given (N_1, N_0), the average bias of the matching estimator is higher when compared to the case of $k = 1$. Nevertheless, the average bias goes to zero with N_0 for any given N_1, which is consistent with our Proposition 1. Rejection rates using our randomization inference test based on permutations remain close to 5% when $M = 1$. With more nearest neighbors, rejection rates stay close to 5%, provided a large number of control observations. The results in Appendix Table A.2 show that even larger N_0 is required to keep the bias under control and rejection rates close to 5% when $k = 4$.

Overall, our conclusions remain valid with multidimensional covariates. However, our asymptotic approximations require an increasing number of control observations to be reliable when the number of covariates increases.

5.4 Bias-corrected matching estimator

We also consider a bias-corrected estimator, as proposed by Abadie and Imbens (2011). 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. Then the bias-corrected matching estimator is given by

$$\tilde{\tau} = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left[(Y_i - \hat{\mu}_0(X_i)) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j - \hat{\mu}_0(X_j)) \right]. \tag{19}$$

This bias-corrected matching estimator is featured in Table 5. While the average bias is reduced using

Table 4: **MC Results with Selection on Observable,** $k = 2$

	$M = 1$		$M = 2$		$M = 4$		$M = 10$	
	$N_0 = 50$	$N_0 = 500$	$N_0 = 50$	$N_0 = 500$	$N_0 = 50$	$N_0 = 500$	$N_0 = 50$	$N_0 = 500$
	(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)
				Panel A: average $\lvert bias \times 1000 \rvert$				
$N_1 = 5$	22.61	6.16	31.42	7.28	45.11	9.48	75.37	17.07
$N_1 = 10$	25.83	4.07	33.29	5.37	46.77	8.00	77.48	15.49
$N_1 = 25$	26.83	6.16	34.21	8.19	48.36	11.23	78.61	18.48
$N_1 = 50$	28.18	5.19	36.27	7.05	50.45	10.21	81.23	18.11
				Panel B: mean squared error ($\times 1000$)				
$N_1 = 5$	45.58	39.61	35.80	29.95	31.03	24.98	31.84	22.24
$N_1 = 10$	26.46	20.66	20.90	15.44	19.06	12.89	20.98	11.61
$N_1 = 25$	14.71	8.91	12.23	6.74	11.70	5.70	14.31	5.24
$N_1 = 50$	11.17	4.88	9.57	3.82	9.51	3.32	12.68	3.11
				Panel C: test based on RI, permutation				
$N_1 = 5$	0.020^-	0.014^-	0.047	0.046	0.051	0.048	0.076^+	0.047
$N_1 = 10$	0.054	0.049	0.053	0.047	0.066^+	0.047	0.145^+	0.050
$N_1 = 25$	0.051	0.051	0.072^+	0.051	0.174^+	0.051	0.376^+	0.070^+
$N_1 = 50$	0.055	0.050	0.134^+	0.050	0.403^+	0.057	0.634^+	0.179^+
				Panel D: test based on RI, sign changes				
$N_1 = 5$	0.008^-	0.014^-	0.003^-	0.012^-	0.000^-	0.009^-	0.000^-	0.004^-
$N_1 = 10$	0.045	0.049	0.020^-	0.046	0.001^-	0.042	0.000^-	0.021^-
$N_1 = 25$	0.048	0.052	0.031^-	0.048	0.000^-	0.048	0.000^-	0.030^-
$N_1 = 50$	0.053	0.049	0.025^-	0.048	0.000^-	0.051	0.000^-	0.027^-
				Panel E: test based on Rothe (2017)				
$N_1 = 5$	-	-	0.003^-	0.000^-	0.030^-	0.023^-	0.059	0.041
$N_1 = 10$	-	-	0.003^-	0.000^-	0.044	0.019^-	0.121^+	0.034^-
$N_1 = 25$	-	-	0.011^-	0.000^-	0.156^+	0.022^-	0.366^+	0.049
$N_1 = 50$	-	-	0.062^+	0.000^-	0.404^+	0.035^-	0.645^+	0.159^+

Note: This table replicates the simulations presented in Table 3 with the difference that it considers a matching estimator on X and $\tilde{X}_2$, where $\tilde{X}_2$ is a random variable independent of all other random variables in the model.

this procedure, it generally comes at a cost of a higher MSE. The MSE is significantly higher when N_1 is very small, because in this case $\hat{\mu}_0(x)$ is estimated using very few observations. This is the bias-corrected matching estimator employed in the teffects command in Stata, so care should be taken when using this bias correction with few treated observations. Interestingly, despite having a lower average bias, in some cases rejection rates are higher when we use the bias-adjusted estimator. This happens because the bias adjustment $\hat{\mu}_0(X_i)$ is chosen to fit Y_i for the control observations, so in finite samples we under-estimate the variation generated by the control observations.

Table 5: **MC Results with Selection on Observable, Bias-corrected Estimator**

	$M=1$		$M=2$		$M=4$		$M=10$	
	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)
				Panel A: average $\lvert bias \times 1000 \rvert$				
$N_1=5$	9.11	1.87	13.61	2.64	14.60	4.09	15.18	7.21
$N_1=10$	10.77	2.78	14.01	3.91	15.22	5.18	16.23	7.97
$N_1=25$	10.51	3.30	13.27	3.71	14.80	4.68	16.10	6.97
$N_1=50$	12.80	3.62	14.12	4.38	15.64	5.72	16.18	8.14
				Panel B: mean squared error $(\times 1000)$				
$N_1=5$	97.30	41.23	98.11	31.08	42.48	25.90	32.60	22.62
$N_1=10$	34.23	20.84	26.27	15.70	22.17	13.03	19.76	11.48
$N_1=25$	21.09	9.34	16.18	7.07	14.04	5.97	12.12	5.25
$N_1=50$	16.79	5.27	13.04	4.05	11.16	3.43	9.58	3.11
				Panel C: test based on RI, permutation				
$N_1=5$	0.019^-	0.016^-	0.047	0.046	0.061^+	0.046	0.074^+	0.048
$N_1=10$	0.053	0.049	0.056	0.049	0.066^+	0.046	0.080^+	0.044
$N_1=25$	0.058	0.052	0.064^+	0.051	0.077^+	0.051	0.122^+	0.053
$N_1=50$	0.062^+	0.050	0.073^+	0.050	0.101^+	0.052	0.178^+	0.062^+
				Panel D: test based on RI, sign changes				
$N_1=5$	0.006^-	0.014^-	0.003^-	0.014^-	0.001^-	0.009^-	0.000^-	0.004^-
$N_1=10$	0.036^-	0.046	0.022^-	0.045	0.003^-	0.044	0.000^-	0.028^-
$N_1=25$	0.050	0.050	0.042	0.050	0.008^-	0.047	0.000^-	0.045
$N_1=50$	0.047	0.046	0.042	0.045	0.004^-	0.050	0.000^-	0.048

Note: This table replicates the simulations presented in Table 3 with the difference that it considers a bias-corrected matching estimator suggested by Abadie and Imbens (2011).

Overall, it might be possible to construct a bias-corrected matching estimator if we have a large number of control observations. In this case, we could use, for example, non-parametric estimation and have a good approximation to $\mu_0(0)$. However, with a fixed number of treated and many control observations, the matching estimator without correction would also work well in terms of bias and the randomization inference tests would provide good size and power. Therefore, it is unclear whether such bias correction would be warranted. When N_0 is not large, the bias correction can potentially do more harm than good.

6 Empirical Application: "Jovem de Futuro" Program

In this section, we explore the validity of matching estimators and of our inferential methods in the estimation of the effects of an educational program in Brazil. This setting has few treated and many control schools.

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.[24] 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 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.[25,26] 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.[27] 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.3). Moreover, one

[24] 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 (see Barros et al. (2012) for details).

[25] 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.

[26] 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.

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

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 6: **"Jovem de Futuro": Summary Statistics**

	Rio de Janeiro		Sao Paulo	
	Exp. Treated - Exp. Control (1)	Nonexp. Control - Exp. Control (2)	Exp. Treated - Exp. Control (3)	Nonexp. Control - Exp. Control (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 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%.

Column 1 of Table 6 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.[28]

Columns 2 and 4 of Table 6 present differences in test scores for 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 non-experimental schools and Sao Paulo has 3481 non-experimental schools, thus providing a setting with few treated schools and many (non-experimental) control schools. We, therefore, consider the use of matching estimators with outcomes from 2007 to 2009 as matching variables.

Table 7 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), the two proposed RI tests, and the test based on the confidence intervals derived by Rothe (2017). 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 in two cases reject the null at 10%, while the other tests would fail to reject the null. This is consistent with our MC simulations in Section 5, 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.

[28] 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.

Table 7: **Non-experimental Results, Experimental Control Schools as Treated Observations**

	Rio de Janeiro				Sao Paulo			
	$M=1$	$M=2$	$M=4$	$M=10$	$M=1$	$M=2$	$M=4$	$M=10$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment effects in 2010								
Point Estimate	0.087	0.038	-0.003	0.046	0.000	0.019	0.018	0.004
p-values:								
AI (2006)	0.091	0.478	0.941	0.086	0.995	0.624	0.601	0.924
RI-permutation	0.114	0.591	0.954	0.459	0.993	0.621	0.663	0.941
RI-sign changes	0.105	0.547	0.919	0.169	0.996	0.613	0.619	0.917
Rothe (2017)	-	0.599	0.889	0.513	-	0.737	0.703	0.945
Treatment effects in 2011								
Point Estimate	0.043	0.017	-0.032	0.000	-0.019	0.005	-0.027	-0.013
AI (2006)	0.566	0.740	0.396	0.997	0.746	0.915	0.475	0.692
RI-permutation	0.676	0.784	0.573	0.996	0.748	0.937	0.585	0.741
RI-sign changes	0.649	0.769	0.451	0.997	0.767	0.921	0.485	0.687
Rothe (2017)	-	0.896	0.655	0.969	-	0.907	0.531	0.772
Treatment effects in 2012								
Point Estimate	0.070	0.027	-0.019	0.006	-0.072	-0.052	-0.034	-0.019
p-values:								
AI (2006)	0.263	0.525	0.522	0.885	0.169	0.267	0.383	0.616
RI-permutation	0.294	0.682	0.708	0.920	0.176	0.309	0.437	0.661
RI-sign changes	0.278	0.522	0.561	0.894	0.160	0.259	0.401	0.535
Rothe (2017)	-	0.603	0.837	0.869	-	0.372	0.447	0.658

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 4 present results for Rio de Janeiro using 1, 2, 4, or 10 nearest neighbors in the estimation, while columns 5 to 8 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), the randomization inference procedures based on permutations and based on sign changes, and based on Rothe (2017).

Finally, Table 8 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 6. 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 6, which is consistent with the imbalances in pre-treatment outcomes for the experimental sample.

Table 8: **Non-experimental Results, Experimental Treated Schools as Treated Observations**

	Rio de Janeiro				Sao Paulo			
	$M=1$	$M=2$	$M=4$	$M=10$	$M=1$	$M=2$	$M=4$	$M=10$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment effects in 2010								
Point Estimate	-0.056	-0.003	-0.012	0.017	0.039	0.024	0.025	0.051
p-values:								
AI (2006)	0.319	0.947	0.736	0.596	0.412	0.571	0.516	0.119
RI-permutation	0.331	0.943	0.850	0.781	0.402	0.619	0.533	0.293
RI-sign changes	0.363	0.937	0.764	0.624	0.422	0.566	0.542	0.124
Rothe (2017)	-	0.915	0.871	0.837	-	0.705	0.657	0.317
Treatment effects in 2011								
Point Estimate	-0.100	-0.009	0.045	0.033	0.040	0.066	0.070	0.055
p-values:								
AI (2006)	0.247	0.900	0.415	0.545	0.318	0.119	0.080	0.123
RI-permutation	0.272	0.924	0.579	0.681	0.330	0.117	0.072	0.193
RI-sign changes	0.285	0.890	0.564	0.566	0.330	0.136	0.089	0.214
Rothe (2017)	-	0.884	0.547	0.772	-	0.341	0.260	0.331
Treatment effects in 2012								
Point Estimate	0.023	0.022	0.030	0.044	0.054	0.087	0.089	0.063
p-values:								
AI (2006)	0.719	0.711	0.516	0.293	0.312	0.054	0.032	0.090
RI-permutation	0.730	0.796	0.673	0.495	0.343	0.065	0.054	0.128
RI-sign changes	0.745	0.736	0.563	0.294	0.330	0.071	0.031	0.117
Rothe (2017)	-	0.809	0.650	0.609	-	0.281	0.171	0.275

Note: This table replicates the results from Table 7 using the experimental treated schools as treated observations for the matching estimators.

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 control observations goes to infinity. Our theory provides a better approximation to the behavior of the matching estimator and more reliable hypothesis testing compared to Abadie and Imbens (2006), in settings in which not only there is a much larger number of control observations relative to treated observations, but also the number of treated observations is too small to allow reliance on asymptotic results that rely on

the number of treated observations going to infinity. Our MC simulations and empirical application confirm that, in settings with few treated observations, our inference methods may be more reliable than existing inference methods.

Finally, 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 asymptotic bias goes to zero if treatment assignment is "as good as random," conditional on this set of pre-treatment outcomes.[29] Moreover, under these conditions, the randomization 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.[30] 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.

[29] 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. Gobillon and Magnac (2016) show that the SC estimator can be asymptotically unbiased if the number of control units and the number of pre-treatment periods go to infinity, while 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-treatment periods increases, even if the number of control units is fixed. 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.

[30] 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.

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. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

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.

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

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

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.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26.

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 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).

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.

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

A.1 Proofs

Proposition 1

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,

$$\begin{aligned} \Pr\left(d(X^i_{(M)}, \bar{x}) > \epsilon\right) &= \sum_{m=0}^{M-1} \Pr\left(d(X_j, \bar{x}) < \epsilon \text{ for exactly } m \text{ observations}\right) \\ &= \sum_{m=0}^{M-1} \left(\begin{array}{c} N_0 \\ m \end{array} \right) [\Pr(d(X_j, \bar{x}) < \epsilon)]^m [\Pr(d(X_j, \bar{x}) > \epsilon)]^{N_0 - m}. \end{aligned} \tag{20}$$

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

$$\mathbb{E}[\hat{\tau} | \{X_i\}_{i \in \mathcal{I}_1}] = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left(\mu_1(X_i) - \mathbb{E}\left[\frac{1}{M} \sum_{m=1}^{M} \mu_0(X^i_{(m)}) \right] \right). \tag{21}$$

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

$$\mathbb{E}[f(Y^i_{(m)})] = E\left\{ \mathbb{E}[f(Y^i_{(m)}) | X^i_{(m)}] \right\} = E\left\{ \tilde{f}(X^i_{(m)}) \right\} \to \tilde{f}(\bar{x}) = E[f(Y(0)) | X = \bar{x}]. \tag{22}$$

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

$$\hat{\tau} = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left[Y_i - \frac{1}{M} \sum_{m=1}^{M} Y^i_{(m)} \right] \xrightarrow{d} \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left[(\mu_1(X_i) - \mu_0(X_i)) + \left(\epsilon_i - \frac{1}{M} \sum_{m=1}^{M} \epsilon_m(X_i) \right) \right]. \tag{23}$$

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.

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.

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.[31] 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 $\Pr(\mathcal{X} < C_1) \to 0$. Therefore,

$$\begin{aligned}
\Pr\left(|\hat{\beta}| \geq c\right) &= \Pr\left(\left|\frac{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\left(X^i_{(m)} - \bar{X}_1\right)Y^i_{(m)}}{\mathcal{X}}\right| \geq c\right) \leq \Pr\left(\frac{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\left|X^i_{(m)} - \bar{X}_1\right|\left|Y^i_{(m)}\right|}{\mathcal{X}} \geq c\right) \\
&\leq \Pr\left(\frac{C_2\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\left|Y^i_{(m)}\right|}{\mathcal{X}} \geq c \mid \mathcal{X} < C_1\right)\Pr(\mathcal{X} < C_1) \\
&\quad + \Pr\left(\frac{C_2\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_1}\left|Y^i_{(m)}\right|}{C_1} \geq c \mid \mathcal{X} > C_1\right)\Pr(\mathcal{X} > C_1).
\end{aligned} \tag{27}$$

Since $\Pr(\mathcal{X} < C_1) \to 0$, 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 $|\hat{\tau}_{biasadj} - \hat{\tau}| = o_p(1)$.[32]

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

[32] 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.2 Appendix Tables and Figures

Table A.1: **MC Results with Selection on Observable,** $\sigma = \sqrt{0.01}$

	$M = 1$		$M = 2$		$M = 4$		$M = 10$	
	$N_0 = 50$	$N_0 = 500$	$N_0 = 50$	$N_0 = 500$	$N_0 = 50$	$N_0 = 500$	$N_0 = 50$	$N_0 = 500$
	(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)
				Panel A: average $\lvert bias \times 1000 \rvert$				
$N_1 = 5$	9.63	1.08	13.54	1.56	22.79	2.47	49.45	5.16
$N_1 = 10$	9.50	0.80	13.83	1.14	23.14	2.11	49.81	4.98
$N_1 = 25$	9.04	0.66	13.45	1.24	22.91	2.29	49.42	5.28
$N_1 = 50$	8.57	0.54	13.53	1.03	23.00	1.96	49.75	4.95
				Panel B: mean squared error ($\times 1000$)				
$N_1 = 5$	5.09	4.09	4.04	3.08	3.96	2.57	6.02	2.28
$N_1 = 10$	3.08	2.07	2.52	1.55	2.63	1.29	4.65	1.16
$N_1 = 25$	1.94	0.92	1.62	0.70	1.82	0.59	3.71	0.54
$N_1 = 50$	1.47	0.52	1.28	0.40	1.50	0.34	3.43	0.32
				Panel C: test based on RI, permutation				
$N_1 = 5$	0.019^-	0.019^-	0.047	0.048	0.057	0.045	0.107^+	0.046
$N_1 = 10$	0.048	0.048	0.052	0.049	0.070^+	0.047	0.244^+	0.046
$N_1 = 25$	0.054	0.051	0.063^+	0.050	0.172^+	0.049	0.755^+	0.062^+
$N_1 = 50$	0.052	0.047	0.091^+	0.051	0.479^+	0.050	0.977^+	0.117^+
				Panel D: test based on RI, sign changes				
$N_1 = 5$	0.007^-	0.013^-	0.003^-	0.014^-	0.001^-	0.010^-	0.000^-	0.004^-
$N_1 = 10$	0.037^-	0.047	0.023^-	0.045	0.004^-	0.044	0.000^-	0.028^-
$N_1 = 25$	0.048	0.051	0.043	0.049	0.008^-	0.046	0.000^-	0.045
$N_1 = 50$	0.049	0.045	0.043	0.047	0.005^-	0.050	0.000^-	0.051
				Panel E: test based on Rothe (2017)				
$N_1 = 5$	-	-	0.001^-	0.000^-	0.006^-	0.003^-	0.032^-	0.006^-
$N_1 = 10$	-	-	0.001^-	0.000^-	0.012^-	0.002^-	0.117^+	0.002^-
$N_1 = 25$	-	-	0.003^-	0.000^-	0.082^+	0.002^-	0.710^+	0.003^-
$N_1 = 50$	-	-	0.015^-	0.000^-	0.399^+	0.002^-	0.976^+	0.014^-

Note: This table replicates the simulations presented in Table 3 with the difference that it considers a DGP with $\sigma = \sqrt{0.01}$.

Table A.2: **MC Results with Selection on Observable,** $k = 4$

	$M = 1$		$M = 2$		$M = 4$		$M = 10$	
	$N_0 = 50$	$N_0 = 500$	$N_0 = 50$	$N_0 = 500$	$N_0 = 50$	$N_0 = 500$	$N_0 = 50$	$N_0 = 500$
	(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)
				Panel A: average $\lvert bias \times 1000 \rvert$				
$N_1 = 5$	51.10	22.54	62.37	26.94	77.16	32.27	104.55	44.16
$N_1 = 10$	55.18	18.76	65.77	24.03	80.80	30.73	107.66	42.80
$N_1 = 25$	58.69	22.64	68.84	27.38	83.08	33.87	109.58	46.23
$N_1 = 50$	60.85	23.36	71.78	27.87	86.42	34.43	113.23	46.78
				Panel B: mean squared error ($\times 1000$)				
$N_1 = 5$	47.79	41.26	38.57	31.25	35.50	26.02	37.59	24.20
$N_1 = 10$	27.62	20.86	23.87	15.85	23.29	13.76	26.83	13.11
$N_1 = 25$	16.27	9.00	14.99	7.24	15.89	6.64	20.13	6.94
$N_1 = 50$	12.56	5.25	12.55	4.39	14.09	4.28	18.85	4.93
				Panel C: test based on RI, permutation				
$N_1 = 5$	0.022^{-}	0.017^{-}	0.056	0.049	0.068^{+}	0.050	0.099^{+}	0.050
$N_1 = 10$	0.058	0.047	0.073^{+}	0.046	0.103^{+}	0.052	0.189^{+}	0.064^{+}
$N_1 = 25$	0.074^{+}	0.055	0.134^{+}	0.062^{+}	0.276^{+}	0.076^{+}	0.437^{+}	0.147^{+}
$N_1 = 50$	0.093^{+}	0.061^{+}	0.297^{+}	0.079^{+}	0.530^{+}	0.137^{+}	0.685^{+}	0.410^{+}
				Panel D: test based on RI, sign changes				
$N_1 = 5$	0.012^{-}	0.013^{-}	0.003^{-}	0.013^{-}	0.000^{-}	0.008^{-}	0.000^{-}	0.002^{-}
$N_1 = 10$	0.056	0.046	0.024^{-}	0.046	0.000^{-}	0.039^{-}	0.000^{-}	0.011^{-}
$N_1 = 25$	0.073^{+}	0.055	0.023^{-}	0.059	0.000^{-}	0.059	0.000^{-}	0.003^{-}
$N_1 = 50$	0.094^{+}	0.063^{+}	0.005^{-}	0.070^{+}	0.000^{-}	0.063^{+}	0.000^{-}	0.000^{-}
				Panel E: test based on Rothe (2017)				
$N_1 = 5$	-	-	0.002^{-}	0.001^{-}	0.041	0.023^{-}	0.072^{+}	0.044
$N_1 = 10$	-	-	0.006^{-}	0.000^{-}	0.080^{+}	0.024^{-}	0.157^{+}	0.045
$N_1 = 25$	-	-	0.042	0.000^{-}	0.268^{+}	0.042	0.431^{+}	0.127^{+}
$N_1 = 50$	-	-	0.217^{+}	0.000^{-}	0.535^{+}	0.106^{+}	0.695^{+}	0.400^{+}

Note: This table replicates the simulations presented in Table 3 with the difference that it considers a matching estimator on X, $\tilde{X}_2$, $\tilde{X}_3$ and $\tilde{X}_4$.

Table A.3: **"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%.