

RCTs with Interference in Matching Systems

Mohit Karnani
MIT*

Abstract

This paper studies the estimation of treatment effects using data from randomized controlled trials (RCTs) in the context of matching systems, such as centralized school choice systems. If treated units change their inputs (e.g. reported preferences), they can affect the allocation outcomes determined by the matching system in equilibrium. When the outcome of interest is determined in equilibrium (e.g. being matched to a target school), conventional estimators that compare a treatment group against a control group fail to account for violations of the stable unit treatment value assumption (SUTVA) induced by the matching system. Therefore, they yield biased and inconsistent estimates of the average treatment effect on the treated (ATT), even under perfect randomization designs. I propose an estimator that decomposes the observed average difference between the outcomes of the treatment and control units, and consistently estimates the ATT, along with the residual spillover effect. This nonparametric estimator relies on the matching mechanism satisfying a version of the equal treatment of equals (ETE) property. Moreover, the ATT is identified for any counterfactual fraction of treated units, which allows for the estimation of treatment effects under scaled interventions. I illustrate the applicability of this estimator using data on the Chilean centralized school choice system and the Ecuadorian teacher labor market.

1 Introduction

At least a third of the countries around the world use centralized matching mechanisms to assign students to educational institutions or teachers to their jobs (Neilson, 2024). In these school choice systems, applicants submit rank-ordered lists to a platform that matches them to schools following some predetermined set of rules (Abdulkadiroğlu and Sönmez, 2003). Researchers often conduct large-scale randomized controlled trials (RCTs) where individuals participating in these matching systems are exposed to treatments affecting their application behavior. Examples include informing applicants about nearby high-value-added schools (Ainsworth et al., 2023) or highlighting the non-assignment risk implied by their application (Arteaga et al., 2022).

These RCTs allow researchers to estimate treatment effects on partial equilibrium outcomes using standard methods that compare treated units against control units (Duflo et al., 2007), such as comparing the share of high-value-added schools listed by treatment and control units. However, when the outcome of interest is determined in equilibrium, such as the share of applicants *assigned* to a high-value-added school, conventional estimators fail to account for violations of the stable unit treatment value assumption (SUTVA) induced by the matching system.

This paper studies the estimation of treatment effects when RCTs are conducted with individuals participating in matching systems. We show how standard techniques that compare the equilibrium outcomes of treatment and control units can yield biased and inconsistent estimates of the average treatment effect on the treated (ATT), even under perfect randomization designs. Intuitively, an intervention that induces the treatment group to apply to high-value-added schools not only affects the probability of the treated students of being admitted into these schools, but might also have a *spillover effect* on the outcomes of the control group if schools have fixed capacities and the matching system has to resort to some form of tie-breaking rule.

*Email: karnani@mit.edu. I thank Esther Duflo and Parag Pathak for their guidance and support in this project. I also thank Alberto Abadie, Nikhil Agarwal, Isaiah Andrews, Josh Angrist, Anna Mikusheva and the participants of the 2022 SDSCon at MIT, the 2023 GPR Workshop at Oxford, the 2023 APSA Annual Meeting and the MIT Development, Labor, IO and Applied Microeconomics Lunches for their feedback.

If this spillover effect changes the outcomes of the control units, then they no longer resemble the counterfactual situation where no treatment took place. Therefore, any direct comparison between the treatment and control outcomes, such as comparing the fraction matched to a high-value-added school in each group, yields an invalid estimate of the causal effect of the treatment on the outcome of interest. This problem has been studied in other contexts, such as evaluating the displacement effects of labor market policies (Crépon et al., 2013).¹

Indeed, in the presence of interference, the potential outcomes (Neyman, 1990; Rubin, 1974) of a unit are not only a function of the treatment status of that unit, but can also be a function of the treatment status of all other units. For example, in a binary treatment case, instead of having only two potential outcomes per unit, as in the case without interference, the set of potential outcomes with interference scales to 2^{units} per unit. Therefore, any procedure that aims to estimate treatment effects with interference is required to reduce the dimensionality of the set of potential outcomes to make the problem tractable. Our approach achieves this by exploiting properties of matching systems.

Matching systems are interesting to study the problem of interference for at least two reasons. First, matching systems are widely used in practice to allocate scarce resources in high-stakes situations: from school choice systems, to medical residency matches, to cadet branching (Roth and Sotomayor, 1992). Because of this, policymakers and researchers are increasingly more likely to experimentally evaluate interventions in the context of matching systems to inform policy decisions, making the problem of interference more prevalent. Secondly, as the assignment mechanism is determined by an explicit set of rules described by a matching algorithm, researchers can leverage these rules to characterize the interference, without having to rely on design-based solutions. In a sense, matching systems provide information about the data-generating process of the outcomes, and allow researchers to “correct” their estimates using data from an RCT.

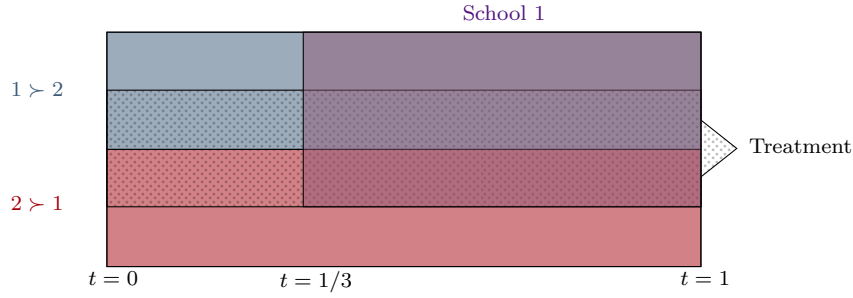
We propose an estimator that decomposes the observed average difference between the outcomes of the treatment and control units, and consistently estimates the ATT, along with the residual spillover effect derived from the matching interference. This nonparametric estimator relies on the matching mechanism satisfying a version of the *equal treatment of equals* property. Moreover, the ATT is identified for any counterfactual fraction of treated units, which allows for the estimation of treatment effects under scaled interventions, targeting a rich set of causal estimands. We also propose an extension to this estimator that accommodates multiple treatment arms and clustered trials.

Finally, we provide two empirical applications for this new estimator. The first application is a simulated experiment in the Chilean centralized school choice system and the second one is a real experiment conducted in the Ecuadorian teacher labor market match. In both cases the implied spillover effects are large, in the latter even to the extent that the estimated ATT cannot be statistically distinguished from zero, i.e. we do not reject that the observed average difference between treatment and control teachers is purely driven by interference in the matching system.

The rest of the paper is organized as follows. In section 2 we illustrate the nature of the interference induced by the matching system and formally introduce the general model we study in the paper. Then, section 3 shows how conventional estimators are invalid in the presence of matching interference and proposes an estimator that solves this problem. We also prove relevant properties of this estimator in section 3, and then propose extensions to it in section 4. Finally, section 5 shows the results of applying our new estimator in two empirical settings, and section 6 concludes.

¹Previously proposed solutions to this problem are usually “design based”, such as designing a two-stage cluster randomized controlled trial, where clusters are randomly assigned to different treatment intensities (i.e. share of treated units), and then units within each cluster are randomized into their treatment status such that they match the treatment intensity of the cluster. This paper proposes a methodological solution that does not rely on special randomization designs, but is specific to RCTs conducted in matching systems. There are other examples where statistical corrections to interference can be achieved without relying on specific randomization designs, such as interference in networks (Leung, 2020).

Figure 1: Illustrative Example of Interference



Note: This figure illustrates the simplified model described by Azevedo and Leshno (2016). The original preferences are colored blue for applicants who prefer school 1 over 2 and red for applicants who prefer school 2 over 1. The shaded area in the middle denotes the treatment population and the overlaid purple rectangle indicates the mass of applicants that are admitted into school 1 once the treatment takes place.

2 The Model

2.1 An Illustrative Example

Consider the following simplified model inspired by Azevedo and Leshno (2016). There are two schools, 1 and 2, and a continuum of applicants seeking admission into these schools. Applicants are indexed by a uniform tie-breaker $t \in [0, 1]$. Each school can take up to 50% of the applicants in the market. Suppose that half of the mass of applicants prefer school 1 over school 2 ($1 \succ 2$) and the other half prefers school 2 over school 1 ($2 \succ 1$). Schools' "preferences" give priority to applicants with higher tie-breakers. Let there be a matching mechanism that assigns applicants to schools by running the student-proposing deferred acceptance (DA) algorithm (Gale and Shapley, 1962). If applicants report their true preferences to the matching mechanism, everyone is assigned to their most preferred option, implying a tie-breaking cutoff of $\bar{t}_0 = 0$, as there is no excess demand for seats at any school in the very first step of the DA algorithm. Let this be the counterfactual scenario in absence of any treatment.

Suppose we do not observe this counterfactual, and conduct an RCT where half of the applicants are randomly selected to be treated with a behavior-changing intervention. Suppose that the intervention effectively modifies the preferences of the treatment group such that they all prefer school 1 over school 2 ($1 \succ 2$), no matter what their counterfactual preferences are in absence of the treatment. In this situation, we have a mass of $3/4$ of the applicants ranking school 1 over school 2; $1/2$ coming from the treatment group and $1/4$ coming from the control group.

Because of the excess demand for seats in school 1, there is a binding tie-breaker induced by the DA matching mechanism, $\bar{t}_1 = 1/3$. This implies that not everyone ranking $1 \succ 2$ is matched to their top listed option. If the outcome of interest is an indicator for being matched to school 1, we observe $2/3$ of the treatment group being matched to school 1 and $1/3$ of the control group being matched to school 1. This is depicted in Figure 1.

Given the observed fractions matched to school 1 in each group, one can (incorrectly) conclude that the average treatment effect of the intervention on the treatment group is an increase of $2/3 - 1/3 = 1/3$ in the probability of admission into school 1. This is incorrect because we know that, in the counterfactual situation described above, half of the treatment group would have been matched to school 1. Therefore, the true average treatment effect on the treated is $2/3 - 1/2 = 1/6$, i.e. only 50% of the observed difference between the treatment and control group outcomes.

This difference arises because of the binding tie-breaker cutoff. Originally, in absence of any intervention, everyone is assigned to their most preferred school, so the unconditional probability of being matched to school 1 is $1/2$. However, the treatment not only affects the treatment group by changing their preferences, but it also interferes with the outcomes of the control group by shifting the cutoff score from $\bar{t}_0 = 0$ to $\bar{t}_1 = 1/3$. This constitutes a violation of the stable unit treatment

value assumption (SUTVA), due to interference in the matching system (Cox, 1958).

We now extend the underlying idea in this simple model to a more general framework that allows us to study treatment effects with interference in matching systems. The rest of the discussion is still framed as a school choice problem, but the results hold in other more general matching systems.

2.2 The Market: Notation and Definitions

Consider a finite set of applicants I and a finite set of schools S with *relative* capacities $q_s \in (0, 1]$, $\forall s \in S$, collected in vector $q \in (0, 1)^{|S|}$. These relative capacities translate into a vector of nq openings, where $n := |I|$. Among these schools, there is an *outside option*, denoted by $\emptyset$, with relative capacity $q_\emptyset = 1$. This null school represents being unmatched.

Applicants have unobserved *types* θ^* that lie on a set of types Θ . For example, if individual i 's type can be described by preferences for schools $\succ_i$ and test scores t_i , then $\theta_i^* := (\succ_i, t_i)$.² Schools do not have types.

Applicants report their types, observed as the profile $\theta := \times_{i \in I} \theta_i$, to a direct stochastic mechanism $\varphi : \theta \rightarrow [0, 1]^{|I| \times |S|}$, which induces a lottery on matchings $\mu(i) \in S$, $\forall i \in I$.³ The reported types θ need not be the same as the true types θ^* , i.e. we do not take a stance on the strategy-proofness of φ .⁴ We also allow agents to misreport their types due to other behavioral factors, such as heterogeneous levels of sophistication (Pathak and Sönmez, 2008) or subjective beliefs (Kapor et al., 2024).

The entries of matrix $\varphi(\theta)$ are the probabilities of assignment of each applicant i to each school s under profile θ , i.e. $\varphi(\theta)[i, s] := \mathbb{P}[\mu(i) = s | \theta]$.⁵ For any profile of reported types θ , each row $\varphi(\theta)[i, :]$ sums to 1, and each column $\varphi(\theta)[:, s]$ adds up to at most nq_s .⁶

We define a *market* as the collection $\mathcal{M} := \{I, S, \theta, q, \varphi\}$. We also define a *large market* as $\lim_{n \rightarrow \infty} \mathcal{M}$. This means that in the large market the options S and relative capacities q_s stay fixed, but the number of applicants of each type grows, i.e. $|I| \rightarrow \infty$ (and we have an infinite mass of each type θ) while $|S|$ is still finite.⁷ The vector of total capacities nq also grows.

2.3 Experimental Design

In a finite market, individuals I participate in an RCT and are randomly assigned to either a *treatment group* $T \subset I$ or a *control group* $C := I \setminus T$. For example, the treatment group can be a random subset of I that receives information about high-value-added schools in their area, while the control group does not receive any such information. Let the fraction of treated individuals be a constant $r := |T|/|I|$, such that the approximate distribution of treatment indicators is $d_i \stackrel{iid}{\sim} \text{Bernoulli}(r)$.

Define $\theta_i^d \in \Theta$ to be the **potential report** of applicant i under treatment status $d \in \{0, 1\}$, where $d = 1$ denotes being treated and $d = 0$ being untreated. This notation conveys that each applicant can be in one of two different states. In our running example, these states are with information ($d_i = 1$) or without information ($d_i = 0$). Agents have a corresponding reported type θ_i^d under each state, even though only one is observed (agents cannot be informed and uninformed simultaneously).

²Types can be more general and incorporate other objects. For example, if applicants also had a priority vector $\rho_i := (\rho_{is})_{s \in S}$ describing the priorities they have at each school, then types would be $\theta_i^* := (\succ_i, t_i, \rho_i)$. Our general definition of types differs from the usual formulation in the literature studying this class of models (Abdulkadiroğlu et al., 2017). We do this so types contain all relevant variables that can change in an experiment (e.g. a tutoring intervention can change the test scores of treated applicants, even though test scores are usually not included in types). Moreover, in some cases test scores are coarsely measured, and many applicants share the exact same test score. In this situation, test scores resemble priorities and our formulation of type is equivalent to that of the literature.

³In situations where types include a continuous variable, such as a continuous test score that determines admissions, the lottery is degenerate and the mechanism is deterministic.

⁴We impose structure on reports, but this is unrelated to the properties of φ .

⁵To be precise about the notion of probability here, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triplet, where $\Omega := S^{|I|}$ is the sample space with possible match outcomes, $\mathcal{F} := 2^\Omega$ is a σ -algebra containing all valid collections of match events, and $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ is a probability measure induced by the random mechanism φ .

⁶For deterministic mechanisms, the entries of $\varphi(\theta)$ are either 0 or 1, where each row contains at most one 1-entry and each column contains at most nq_s 1-entries.

⁷This notion of large market is later used in section 3 to study the asymptotic properties of the estimators proposed in this paper.

This definition emulates the idea of potential outcomes in the Neyman–Rubin causal model (Neyman, 1990; Rubin, 1974). Moreover, by utilizing this notation without dependence on the full vector of treatments, we explicitly assume a “consistency rule” on the reported types, which implies that there are no direct spillovers of the treatment on the reported types.

Assumption 1 (No type interference). *Types reported follow*

$$\theta_i = d_i \theta_i^1 + (1 - d_i) \theta_i^0, \quad \forall i \in I.$$

This rules out other forms of interference where the treatment status of agent i affects the reported type of another agent j . For example, if i and j are siblings, then θ_j could in principle be a function of d_i (Altmejd et al., 2021; Basse et al., 2019), but our model does not accommodate this spillover. This paper only focuses on spillovers induced by interference in the matching mechanism, and not on other kinds of spillovers that might affect the reported types (e.g. network interference or spatial spillovers). Additionally, Assumption 1 implies perfect compliance with the treatment: informed individuals report their types as if they were informed, and uninformed individuals report their types as if they were not informed. In other words, for each agent i , there are no other reports but the two potential reports θ_i^1 and θ_i^0 . Formally, Assumption 1 is a stable unit treatment value assumption regarding the reported types.

Let X_i denote a vector with other characteristics of applicant i that are excluded from θ_i , such as demographics. Suppose applicants’ potential reports and characteristics are drawn from the same population distribution $F: \{\theta_i^1, \theta_i^0, X_i\} \sim F$. For notation purposes, subscripts on F denote the random variable drawn from the distribution, e.g. F_{θ^0} is the population distribution of potential types in absence of information.

2.4 Target Estimand

Let the outcome of interest be an indicator variable for being matched to a school in some fixed subset $H \subset S$, i.e. $Y_i := \mathbf{1}[\mu(i) \in H|\theta]$. For example, H could be a set of high value added schools targeted by a policymaker, such that Y_i indicates if applicant i was matched to a high value added school or not. Even though this is our preferred outcome variable, the following results directly extend to more general outcomes Y_i that are a linear combination of the match indicators: $Y_i = \sum_{s \in S} w_s \mathbf{1}[\mu(i) = s|\theta]$. For example, w_s could be a measure of value added for school s , and then Y_i would denote the value added of the school assigned to applicant i . In the case we study above, $w_s = \mathbf{1}[s \in H] \quad \forall s \in S$.⁸

The target estimand in this model is the **average treatment effect on the treated** (ATT), i.e. the change in the probability of admission into H for the treatment group when the reported types of the treatment group change from the counterfactual profile $\theta_T^0 := \times_{i \in T} \theta_i^0$ to the observed profile $\theta_T^1 := \times_{i \in T} \theta_i^1$, holding the reported profile of the control group $\theta_C^0 := \times_{i \in C} \theta_i^0$ fixed. For simplicity, we collect the two untreated profiles in one term $\theta_I^0 := (\theta_T^0, \theta_C^0)$.

Definition 1 (ATT). *The average treatment effect on the treated τ is defined as*

$$\begin{aligned} \tau &:= \mathbb{P}[\mu(i) \in H | i \in T, \theta_T^1, \theta_C^0] - \mathbb{P}[\mu(i) \in H | i \in T, \theta_T^0] \\ &= \mathbb{E}[Y_i | i \in T, \theta_T^1, \theta_C^0] - \mathbb{E}[Y_i | i \in T, \theta_T^0]. \end{aligned} \tag{1}$$

The first expectation in (1) is conditioned not only on the treatment condition of the treatment group, but also explicitly incorporates the untreated condition of the control group. The second expectation indicates that no one is treated. This extensive notation becomes necessary because of potential SUTVA violations (Rubin, 1980). Indeed, the outcome Y_i can depend on the treatment status of other units $j \neq i$ and their reported types. Importantly, τ also depends on the fixed treatment

⁸An essential requirement is that the outcome of interest and target parameters depend only on the matches, rather than other downstream outcomes which could be influenced by the matches, such as test scores. Such outcomes and target parameters are typically studied in short-term experimental evaluations (e.g. how does an information campaign affect the match-outcomes of applicants shortly after), and are the main focus of this paper. We do not consider evaluations that go beyond match-outcomes (e.g. how does an information campaign affect test scores after applicants are matched).

fraction r , so this target estimand is specifically defined for a given experimental design. We focus on other causal estimands where r may vary in section 4.⁹

In a similar fashion, the spillover effect that the treatment group imposes on the control group is the change in the expected outcome of the control group when the reported type profile changes from θ_I^0 to (θ_T^1, θ_C^0) , i.e. when the treatment group reports θ_T^1 instead of θ_T^0 . We formally define the spillover population parameter in this setting as follows.

Definition 2 (Spillover). *The spillover effect σ on the control group is defined as*

$$\sigma := \mathbb{E}[Y_i | i \in C, \theta_T^1, \theta_C^0] - \mathbb{E}[Y_i | i \in C, \theta_T^0]$$

These two estimands are of separate interest because they are indicative of the impact the intervention has on the treatment group relative to the counterfactual scenario where no treatment takes place, and the displacement it causes in the control group. Even in situations where researchers care about the total effect that an intervention has on the treated relative to the untreated (e.g. the impact of unionization on relative wages (Lewis, 1983) is a classical example), distinguishing between τ and σ can be of importance for policy reasons, such as scaling up an intervention¹⁰ or quantifying and correcting the “harm” that a policy imposes on the untreated units. Given these two estimands of interest, τ and σ , we now study the estimation and identification of these population objects using data coming from an RCT.

3 Estimation

3.1 Difference in Means

The identification results below are based in the large market approximation using data from an RCT. With this in mind, we first analyze the standard comparison between treatment and control outcomes in absence of interference.

In absence of any interference, when $\sigma = 0$, we can directly estimate τ by computing the sample means of Y_i for individuals in T and C , and subtracting them, as in the standard analysis of RCT outcomes:

$$\hat{\Delta} := \sum_{i \in T} Y_i / |T| - \sum_{i \in C} Y_i / |C|.$$

Lemma 1. *Without interference, $\hat{\Delta}$ is an unbiased, consistent estimator of τ .*

Proof. The independent Bernoulli treatment assignments d_i are linearly mapped to reported types θ_i (Assumption 1) drawn from a common distribution. As there is no interference in φ , the reported types are independently mapped to a discrete support and linearly transformed to Y_i , which has finite expectation. Thus, the strong law of large numbers holds.

By the strong law of large numbers, in the large market (as $n \rightarrow \infty$) the sample averages converge almost surely to their population expectations, so we have that

$$\begin{aligned} \hat{\Delta} &\rightarrow \mathbb{E}[Y_i | i \in T, \theta_T^1, \theta_C^0] - \mathbb{E}[Y_i | i \in C, \theta_T^1, \theta_C^0] \underbrace{+\sigma}_{=0} \\ &= \mathbb{E}[Y_i | i \in T, \theta_T^1, \theta_C^0] - \underbrace{\mathbb{E}[Y_i | i \in C, \theta_I^0]}_{=\mathbb{E}[Y_i | i \in T, \theta_I^0]} \\ &= \tau, \end{aligned}$$

where we add $\sigma = 0$, implied by no-interference, and $\mathbb{E}[Y_i | i \in C, \theta_I^0] = \mathbb{E}[Y_i | i \in T, \theta_I^0]$ because d_i is randomly assigned. $\square$

⁹In addition, τ depends on the properties of φ and relative capacities q , i.e. on the full structure of the market $\mathcal{M}$, but we assume φ and q to be fixed.

¹⁰We further develop this idea in the extension provided in subsection 4.1.

Thus, in absence of interference, as the assignment into the treatment is random, and $\sigma = 0$ by assumption, the sample averages converge to the population expectations in (1) and $\hat{\Delta}$ is a valid estimator for τ .

In empirical practice, the no-interference assumption is likely to hold whenever there is “slack capacity” in S i.e. when the vector of capacities q is such that there are no binding capacity constraints at any school. For example, in a school choice system with heterogeneous preferences and large schools relative to n (i.e. q_s close enough to 1), agents are unlikely to interfere with each other.

Another situation where no-interference is likely to hold is in *small experiments*, i.e. experiments where r is close to 0. In these experiments, the change from profile θ_I^0 to (θ_T^1, θ_C^0) might not significantly change the equilibrium assignments of agents in the control group. Small experiments need not be exclusive to RCTs, but can also be found in *natural experiments*, such as those induced by regression discontinuity designs (Angrist et al., 2022).¹¹

Now consider the case where $\sigma \neq 0$, i.e. where there is interference induced by the matching mechanism. $\hat{\Delta}$ still estimates the second expectation in (1) with the sample average of Y_i in the control group. However, unlike in Lemma 1, this sample average is not a valid approximation of the outcomes of the treatment units in absence of the treatment. This implies the following lemma.

Lemma 2. *If $r \neq 0$, and $\sigma \neq 0$, $\hat{\Delta}$ is a biased, inconsistent estimator of τ .*

Proof. If $r \neq 0$, then the conditioning sets defined by type profiles (θ_T^1, θ_C^0) and (θ_I^0) are different, because the mass of treatment individuals is strictly positive as $n \rightarrow \infty$ in the large market. Then, even if the law of large numbers holds,

$$\begin{aligned}\hat{\Delta} &\rightarrow \mathbb{E}[Y_i | i \in T, \theta_T^1, \theta_C^0] - \mathbb{E}[Y_i | i \in C, \theta_T^1, \theta_C^0] \pm \mathbb{E}[Y_i | i \in T, \theta_I^0] \\ &= \tau + \underbrace{\mathbb{E}[Y_i | i \in T, \theta_I^0] - \mathbb{E}[Y_i | i \in C, \theta_T^1, \theta_C^0]}_{=-\sigma} \\ &= \tau - \sigma \\ &\neq \tau,\end{aligned}$$

where again $\mathbb{E}[Y_i | i \in C, \theta_I^0] = \mathbb{E}[Y_i | i \in T, \theta_I^0]$ because d_i is randomly assigned. $\square$

As mentioned above, the treatment group affects the assignment probabilities of the control group to schools in H . The treatment shifts the type profile θ_T of the treatment group from θ_T^0 to θ_T^1 , which affects the conditioning set of the target population expectation and thus also the resulting expected outcome for the control group. For example, in subsection 2.1 the nature of this interference can be visualized as a change in the equilibrium cutoffs, which not only affects the admission chances of the treatment group, but also those of the control group. This renders the observed sample average of the control group, $\sum_{i \in C} Y_i / |C|$, an invalid estimator of $\mathbb{E}[Y_i | i \in T, \theta_I^0]$.¹²

Moreover, if the treatment induces demand for H such that $\tau \geq 0$, then it is likely that the interference negatively impacts the control group, i.e. $\sigma \leq 0$. This implies that $\hat{\Delta}$ most likely overestimates τ in empirical practice. In section 5, we use data from Ajzenman et al. (2026a) to show how this can be the case.

This negative result has a series of implications. First, even when conducting an RCT, the difference-in-means estimator $\hat{\Delta}$ can be invalid if certain conditions are met: a large experiment in a matching system with interference. Second, it is likely that the bias arising from this interference in empirical applications is *positive*. This means that results from RCTs in matching systems, testing the effectiveness of demand-inducing interventions on applicants’ likelihood of being matched to target schools, tend to be *overestimated*.

¹¹Theoretically, matching systems can significantly change their equilibrium allocation even upon small changes in the reported type profile. However, in practice, large markets are unlikely to exhibit this behavior. For example, in a large DA market, tiebreaker cutoffs are fixed and small changes in reported types do not change them, so the untreated group remains unaffected.

¹²Note that if the relative mass of the treatment group is negligible in the large market, i.e. if $r = 0$, then $\sigma = 0$ because the conditioning set (θ_T^1, θ_C^0) is the same as θ_I^0 . Thus, if $r = 0$, $\hat{\Delta}$ is a consistent estimator of τ .

Without additional assumptions on φ , Lemma 2 poses a practical challenge to experimental studies that change inputs in matching systems to estimate treatment effects on equilibrium outcomes. To overcome this negative result, and correctly identify τ , we impose some structure on the matching mechanism φ .

3.2 Structure of φ

For the main results of this paper, we crucially rely on one assumption about φ . We assume that φ satisfies a *modified* version of the *equal treatment of equals* (ETE) property. We first define the standard ETE property.

Definition 3 (Equal treatment of equals). *φ satisfies equal treatment of equals if for any $i, j \in I$ such that $\theta_i = \theta_j$, the assignment probabilities induced by φ are the same for i and j , i.e. given any fixed type profile θ ,*

$$\theta_i = \theta_j \implies \mathbb{P}[\mu(i) = s | \theta] = \mathbb{P}[\mu(j) = s | \theta], \forall s \in S.$$

Intuitively, ETE induces a “type-stratified RCT” within the matching mechanism (Abdulkadiroğlu et al., 2017), i.e. for any two individuals i, j sharing the same reported type $\theta_i = \theta_j$, their allocation outcomes are drawn from the same distributions $\varphi(\theta)[i, :] = \varphi(\theta)[j, :]$ and any difference in their resulting assignment is as good as random.¹³ In other words, ETE implies that assignments are independent of other applicant characteristics X_i after conditioning on θ_i :

$$\{\mu(i) = s\} \perp X_i | \theta_i \quad \forall s \in S.$$

This independence property holds within a matching system satisfying ETE for any given type profile θ and is unrelated to the RCT-independence in the setting we study. The RCT instead shifts the type profile for the treatment group, where d_i is independent of characteristics X_i and potential types: $\{\theta_i^1, \theta_i^0, X_i\} \perp d_i$.

The importance of ETE is straightforward: it allows researchers to directly map reported types to assignment probabilities, without having to control for additional variables that might be correlated with the assignment outcomes.¹⁴ This ignorability condition has proven to be useful in many contexts, such as estimating value added measures for schools (Abdulkadiroğlu et al., 2017) or the effects of minority-integration in school choice systems (Angrist et al., 2022).

Our identification results also rely on mapping reported types to assignment probabilities. However, instead of the standard ETE assumption, we require a *distributional continuity* assumption on φ to hold in the large market. Specifically, we require that when type *distributions* converge, from where reported types are drawn, then their φ -induced distributions of assignment probabilities also converge.¹⁵

We call this adapted ETE assumption **continuous distributional equal treatment of equals** (CD-ETE).

Assumption 2 (Continuous distributional ETE). *When $n \rightarrow \infty$, the mechanism φ satisfies continuous distributional equal treatment of equals: for any converging sequence of type-distributions F_θ^n approaching a type-distribution F_θ , the corresponding sequence of distributions of match probabilities F_φ^n (induced by φ mapping F_θ^n) also converge to the distribution of match probabilities F_φ (induced by φ mapping F_θ) i.e.*

$$\lim_{n \rightarrow \infty} \sup |F_\theta^n - F_\theta| = 0 \implies \lim_{n \rightarrow \infty} \sup |F_\varphi^n - F_\varphi| = 0.$$

¹³As mentioned above, if types contain a continuous variable (e.g. a test score that determines allocations), then the mechanism becomes deterministic and there is no source of randomness.

¹⁴This is usually done by simulation. As individuals sharing a type also share the same probability distribution of assignments, by simulating the match multiple times, researchers can infer the assignment distribution for each type from the observed empirical measure of simulated assignments. This insight motivates our main estimator below.

¹⁵Recall that each type profile θ is mapped to a bistochastic matrix $\varphi(\theta)$ where the entries correspond to the assignment probabilities of each student to each school, modulo their non-assignment risk. Thus, distributions of random types F_θ are mapped to distributions of random matrices F_φ , where the dependence on θ is suppressed to simplify notation.

The CD-ETE assumption modifies the standard ETE condition. Under ETE, for any given profile θ , individual reports θ_i can be directly mapped to a distribution of allocation outcomes $\varphi(\theta)[i, :]$. Under CD-ETE, this mapping is extended to sequences of profile distributions, such that if profile distributions F_θ^n converge, then assignment-probability distributions F_φ^n also converge. This CD-ETE assumption can also be interpreted as an assumption of distributional uniform convergence under the sup norm.

This assumption allows us to identify our target estimand τ , without relying on other more stringent requirements like a (random serial priority) cutoff structure (Agarwal and Somaini, 2018) or a fully parametric specification of the model (Kapor et al., 2024).¹⁶ In addition, CD-ETE is satisfied by common matching systems in practice, such as those that implement the DA algorithm. Indeed, in a large DA market, cutoffs are fixed and converging sequences of type distributions imply “smoothly” converging cutoff distributions, which then directly map to matching probabilities based on reported types and the equilibrium cutoffs (Azevedo and Leshno, 2016; Abdulkadiroğlu et al., 2017). This applies not only to the DA algorithm, but also to other mechanisms allowing cutoff representations with smoothly converging cutoffs, such as the random serial dictatorship mechanism (Che and Kojima, 2010).

However, CD-ETE rules out other mechanisms that might exhibit “discontinuous” changes in allocations upon “small” changes in the distribution of types. For example, consider a mechanism that grants priority in a school s to a set of applicants $A \subset I$ if at least half of the population I lists that school on top, and priority to another set of applicants $B \neq A$ if fewer than half of I list s as their first choice. Mechanisms like this can satisfy ETE without satisfying CD-ETE: for converging sequences of type distributions with a limit F_θ in which exactly half of the applicants list s as their top choice, there are two limiting distributions of assignment probabilities. As the limit does not exist, this hypothetical mechanism does not satisfy CD-ETE.¹⁷

Provided that the mechanism φ satisfies CD-ETE, there exists an estimator that solves the identification problem introduced in Lemma 2. This alternative estimator is described below.

3.3 Main Results

As noted above, Lemma 2 follows from $\sum_{i \in C} Y_i / |C|$ not being a valid estimator for $\mathbb{E}[Y_i | \theta_I^0]$ in the presence of interference. We first propose a procedure to consistently estimate the population expectation $\mathbb{E}[Y_i | \theta_I^0]$ using RCT data coming from mechanisms that satisfy CD-ETE. Consider the estimator described with pseudo-code in algorithm 1.

Algorithm 1: Consistent estimator of $\mathbb{E}[Y_i | \theta_I^0]$

Data: θ_C, n, H , (large) B

Function: $\varphi : (\theta, seed) \mapsto \mu \in S^n$ satisfying CD-ETE

Init: $\hat{\theta}_I^0 \in \Theta^n, Y^0 \in \{0, 1\}^{n \times B}$

Result: $\bar{Y}^0$

▷ (consistent estimator of $\mathbb{E}[Y_i | \theta_I^0]$)

for $b \in \{1, \dots, B\}$ **do**

$\hat{\theta}_I^0 \leftarrow \text{SampleWithReplacement}(\theta_C, n, seed_b)$

▷ (simulated types)

$\hat{\mu} \leftarrow \varphi(\hat{\theta}_I^0, seed_b)$

▷ (simulated matches)

$Y^0[:, b] \leftarrow \mathbf{1}[\hat{\mu} \in H]$

▷ (simulated outcomes)

$\bar{Y}^0 \leftarrow \sum_{b=1}^B \sum_{i=1}^n Y^0[i, b] / (nB)$

▷ (simulated sample mean)

This is a simulation-based estimator that relies on 4 steps:

¹⁶We do not assume any structure on how d_i affects θ_i , apart from the standard “consistency rule” (Assumption 1), nor on how reported types θ , capacities q or treatment intensity r affect the match.

¹⁷A mechanism satisfying CD-ETE need not satisfy ETE. For example, consider a (dictatorship) mechanism φ in which the individual with index $i = 1$ gets assigned to their top reported choice, while the rest $j \neq 1$ are randomly assigned to their schools, irrespective of their types. This mechanism satisfies CD-ETE because the relative mass of individual $i = 1$ is $1/n$, which decays to 0 as $n \rightarrow \infty$, yielding a constant random assignment mechanism in the large market for any converging sequence F_θ^n and limiting distribution F_θ . Nonetheless, this mechanism does not satisfy ETE, as any other individual $j \neq 1$ with type $\theta_j = \theta_1$ will have assignment probabilities $\varphi[j, :] \neq \varphi[1, :]$ (all other agents have equal, interior assignment probabilities, but agent 1 has a degenerate distribution that guarantees placement in their top choice).

1. Draw n random samples of reported types with replacement from the observed control profile θ_C to create simulated profile $\hat{\theta}_I^0$.
2. Input these simulated types $\hat{\theta}_I^0$ into the matching mechanism φ and save the simulated matching outcomes vector $Y^0[:, b]$.
3. Repeat steps (1)-(2) B times, where B is large.
4. Collapse matrix Y^0 by averaging over all components of all vectors $\{Y^0[:, b]\}_{b=1}^B$ to output the scalar $\bar{Y}^0$.

Intuitively, this estimator uniformly samples types with replacement from the empirical distribution induced by the reported control types. As there is no interference in the reported types (Assumption 1), the distribution of reported types $\hat{\theta}_I^0$ generated by the resampling procedure approaches the population distribution of untreated types θ_I^0 in the large market. Identification of the counterfactual mean then follows from CD-ETE (Assumption 2) and standard results of bootstrap methods.

Lemma 3 (Consistent estimator of counterfactual mean). *If Assumption 1 and Assumption 2 hold, then algorithm 1 yields a simulated sample mean $\bar{Y}^0$ that is a consistent estimator of $\mathbb{E}[Y_i|\theta_I^0]$.*

Proof. As d_i is randomly assigned and Assumption 1 holds, the empirical distribution $F_{\theta_C}^n$ of reported types in the control group θ_C is an independent random sample of size $|C|$ from the population distribution F_{θ^0} .

Then, the random sample $\hat{\theta}_I^0$ comes from independently drawing n random types with replacement from the empirical measure $F_{\theta_C}^n$. This induces the empirical distribution $\hat{F}_{\theta^0}^n$.

When $n \rightarrow \infty$, by the Glivenko-Cantelli theorem,

$$\lim_{n \rightarrow \infty} \sup |\hat{F}_{\theta^0}^n - F_{\theta^0}| = 0,$$

almost surely.

If φ complies with CD-ETE (Assumption 2), then the plug-in empirical measure $\hat{F}_{\varphi}^n$, induced by $\varphi(\hat{\theta}_I^0)$, also converges almost surely to the population distribution F_{φ} :

$$\lim_{n \rightarrow \infty} \sup |\hat{F}_{\varphi}^n - F_{\varphi}| = 0.$$

As φ is a random mechanism, we can draw samples from this consistent estimator of the counterfactual distribution F_{φ} by randomly drawing match-outcomes $\{Y^0[:, b]\}_{b=1}^B$ with a large B , as is standard in bootstrap methods (Efron and Tibshirani, 1994). Almost-sure convergence in the CDF of simulated outcomes resulting from the random matches $\hat{\mu}$ is guaranteed by the Continuous Mapping Theorem because Y_i is a linear function of $\mu(i)$.

Finally, for bounded random variables, if CDFs converge, then their moments also converge. Thus, the sample average $\bar{Y}^0$ converges almost surely to the population expectation $\mathbb{E}[Y_i|\theta_I^0]$, i.e.

$$\lim_{B \rightarrow \infty} \lim_{n \rightarrow \infty} \sum_{b=1}^B \sum_{i=1}^n Y^0[i, b]/(nB) = \mathbb{E}[Y_i|\theta_I^0],$$

almost surely. □

The result in Lemma 3, directly implies identification of τ and σ , as shown below.

Theorem 1 (Consistent estimator of τ and σ). *If Assumption 1 and Assumption 2 hold, then $\hat{\tau} := \sum_{i \in T} Y_i/|T| - \bar{Y}^0$ is a consistent estimator of τ , and $\hat{\sigma} := \sum_{i \in C} Y_i/|C| - \bar{Y}^0$ is a consistent estimator of σ , where $\bar{Y}^0$ is the simulated sample mean obtained from algorithm 1.*

Proof. As d_i is randomly assigned, Lemma 3 implies that

$$\bar{Y}_0 \rightarrow \mathbb{E}[Y_i|\theta_I^0] = \mathbb{E}[Y_i|i \in T, \theta_I^0] = \mathbb{E}[Y_i|i \in C, \theta_I^0],$$

almost surely.

Then, by the Strong Law of Large Numbers,

$$\begin{aligned} \hat{\tau} &= \sum_{i \in T} Y_i / |T| - \bar{Y}^0 \\ &\rightarrow \mathbb{E}[Y_i|i \in T, \theta_T^1, \theta_C^0] - \mathbb{E}[Y_i|i \in T, \theta_I^0] \\ &= \tau \end{aligned}$$

and

$$\begin{aligned} \hat{\sigma} &= \sum_{i \in C} Y_i / |C| - \bar{Y}^0 \\ &\rightarrow \mathbb{E}[Y_i|i \in C, \theta_T^1, \theta_C^0] - \mathbb{E}[Y_i|i \in C, \theta_I^0] \\ &= \sigma \end{aligned}$$

almost surely. □

3.4 Inference

We now briefly turn to the matter of statistical inference. First, we focus on constructing confidence intervals for our point estimate. Assumption 1 and Assumption 2 not only imply that the counterfactual expectation $\mathbb{E}[Y_i|\theta_I^0]$ is identified (Lemma 3), but also that the counterfactual distribution of matches $F_\varphi|\theta_I^0$ is identified. By adjusting the procedure described in algorithm 1, under Assumption 1 and Assumption 2, we can also identify the distribution of matches $F_\varphi|(\theta_T^1, \theta_C^0)$. With these two objects, we can construct bootstrap confidence intervals for $\hat{\tau}$. Bootstrap confidence intervals will have correct coverage in large samples provided suitable regularity conditions hold (Horowitz, 2001). A valid procedure is described below:

1. Implement algorithm 1 and average the components of each column in Y^0 to form the set of bootstrapped empirical probabilities $\{Y_b^0\}_{b=1}^B$. As shown in the proof of Lemma 3, $\hat{F}_0(t) := \sum_{b=1}^B \mathbf{1}[Y_b^0 < t]/B$ is a consistent estimator of $F_\varphi|\theta_I^0$ mapped to outcomes Y_i .
2. Implement a variant of algorithm 1 but randomly sampling reported types with replacement from the treatment and control group, instead of only the control group (this variant is described below in algorithm 2). Average the treatment components (rows $i : \hat{\theta}_i \sim \theta_T^1$) of each column in Y^1 to form the set of bootstrapped empirical probabilities $\{Y_b^1\}_{b=1}^B$. Construct $\hat{F}_1(t) := \sum_{b=1}^B \mathbf{1}[Y_b^1 < t]/B$, which is a consistent estimator of $F_\varphi|T, (\theta_T^1, \theta_C^0)$ mapped to outcomes Y_i .
3. Randomly draw samples with replacement from $\hat{F}_1$ to compute the average $\tilde{Y}^1$ and sample from $\hat{F}_0$ to compute the average $\tilde{Y}^0$. Store the difference $\tau_k = \tilde{Y}_k^1 - \tilde{Y}_k^0$ and repeat the process for $k = 1, \dots, K$, where K is large.
4. As in standard bootstrap applications (Efron and Tibshirani, 1994), use the percentile method to estimate the confidence interval by computing the required quantiles of the empirical distribution of $\{\tau_k\}_{k=1}^K$ for the desired confidence level (e.g. 5th and 95th percentile for a 90% confidence interval).

Similarly, we can use the bootstrapped sample $\{\tau_k\}_{k=1}^K$ to estimate standard errors by computing $\hat{se} := \sqrt{\sum_{k=1}^K (\tau_k - \bar{\tau})^2 / (K - 1)}$, where $\bar{\tau} = \sum_{k=1}^K \tau_k / K$ is the bootstrap sample mean.

Alternatively, we can directly test the hypothesis of zero treatment effect $H_0 : \tau = 0$ by using the independent samples of empirical assignment probabilities $\{\tilde{Y}_k^1, \tilde{Y}_k^0\}_{k=1}^K$. Formally, we can compute the bootstrap test statistic for equality of means Efron and Tibshirani (1994):

1. Compute and store the t -statistic $\hat{t} = \bar{\tau} / \sqrt{(\hat{s}e_1^2 + \hat{s}e_0^2)/K}$
2. Recenter the samples by computing $\bar{y}^1 = \sum_{k=1}^K \bar{Y}_k^1 / K$, $\bar{y}^0 = \sum_{k=1}^K \bar{Y}_k^0 / K$ and then generating $y_k^1 := \bar{Y}_k^1 - (\bar{y}^1 - \bar{y}^0)/2$, $y_k^0 := \bar{Y}_k^0 - (\bar{y}^0 - \bar{y}^1)/2 \forall k = 1, \dots, K$. Now both recentered groups share the same sample mean (the pooled sample mean).
3. Form M datasets $\{\mathbf{y}^1, \mathbf{y}^0\}_{m=1}^M$ by drawing K random samples with replacement from each of the recentered samples in each case (so each dataset m has $2K$ observations).
4. Compute the t -statistic $\hat{t}_m$ for each of the M datasets.
5. Compute the empirical probability $\hat{p} := \sum_{m=1}^M \mathbf{1}[\hat{t}_m \geq |\hat{t}|] / M$.

These standard nonparametric bootstrap methods are valid in large markets, as CD-ETE (Assumption 2) along with the random assignment of d_i and Assumption 1 imply that we can construct two independent data generating processes $\hat{F}_1$ and $\hat{F}_0$ that approximate the distributions of outcomes Y_i arbitrarily well under profiles (θ_T^1, θ_C^0) and θ_I^0 (i.e. they converge almost surely). As the match-dependent outcomes are bounded and have finite moments, the procedures described above are valid estimators for confidence intervals, standard errors, and p -values associated to the null $H_0 : \tau = 0$.

4 Extensions

This section introduces extensions that build on algorithm 1. In each case, we provide identification results for different causal parameters under different identification assumptions and research designs.

4.1 Scaled Interventions

In many situations, researchers are interested in estimating the impacts of treatments or policies at scale, even though they only observe data coming from a small-scale intervention targeting a limited sample of the population (Allende et al., 2019). Specifically, they target estimands such as

$$\tau(r) := \mathbb{E}[Y_i | i \in T, \theta_T^1, \theta_C^0, r] - \mathbb{E}[Y_i | i \in T, \theta_I^0],$$

which is simply Equation 1, but conditioning the first expectation on having a fraction r of the population being treated. A common value of interest is $r = 1$, i.e the ATT $\tau(1)$ under the scenario where a policy is scaled to the whole population. However, other interior values of $r \in (0, 1)$ can also be justified, such as when budget constraints do not allow for a full roll out of the intervention or they call for a downscale of the program.

An adaptation of algorithm 1 allows us to consistently estimate $\mathbb{E}[Y_i | i \in T, \theta_T^1, \theta_C^0, r]$ for any fraction $r \in (0, 1]$. This is formalized in algorithm 2.

Algorithm 2: Consistent estimator of $\mathbb{E}[Y_i | i \in T, \theta_T^1, \theta_C^0, r]$

Data: $\theta_T, \theta_C, n, r, H$, (large) B

Function: $\varphi : (\theta, seed) \mapsto \mu \in S^n$ satisfying CD-ETE

Init: $\hat{\theta}_T^1 \in \Theta^{nr}, \hat{\theta}_C^0 \in \Theta^{n(1-r)}, Y^0 \in \{0, 1\}^{n \times B}$

Result: $\bar{Y}^1$ $\triangleright$ (consistent estimator of $\mathbb{E}[Y_i | i \in T, \theta_T^1, \theta_C^0, r]$)

for $b \in \{1, \dots, B\}$ **do**

$\hat{\theta}_T^1 \leftarrow \text{SampleWithReplacement}(\theta_T, nr, seed_b)$

$\hat{\theta}_C^0 \leftarrow \text{SampleWithReplacement}(\theta_C, n(1-r), seed_b)$

$\hat{\theta}_I^r \leftarrow \text{Append}(\hat{\theta}_T^1, \hat{\theta}_C^0)$

$\triangleright$ (simulated types)

$\hat{\mu} \leftarrow \varphi(\hat{\theta}_I^r, seed_b)$

$\triangleright$ (simulated matches)

$Y^1[:, b] \leftarrow \mathbf{1}[\hat{\mu} \in H][: nr]$

$\triangleright$ (simulated outcomes)

$\bar{Y}^1(r) \leftarrow \sum_{b=1}^B \sum_{i=1}^{nr} Y^1[i, b] / (nrB)$

$\triangleright$ (simulated sample mean)

Intuitively, this variant of algorithm 1 resamples a fraction r of types from the treatment group and a fraction $1 - r$ from the control group to form the simulated types $\hat{\theta}_I^r$. This allows us to compute

the estimate $\bar{Y}^1(r)$, which approximates a counterfactual situation where rn units are treated and the remaining are untreated.

Thus, provided that we can compute $\bar{Y}^1(r)$ using algorithm 2 and $\bar{Y}^0$ using algorithm 1, the difference between these two terms yields a consistent estimator of $\tau(r)$ as a function of the treatment fraction r , as long as r is orthogonal to the potential reported types $\{\theta^0, \theta^1\}$.

Corollary 1. *Let $r \in (0, 1]$ and $\theta_i^d \perp r$ for $d \in \{0, 1\}$. If Assumption 1 and Assumption 2 hold, then $\hat{\tau}(r) := \bar{Y}^1(r) - \bar{Y}^0$ is a consistent estimator of $\tau(r)$, where $\bar{Y}^1(r)$ is defined as the output of algorithm 2 and $\bar{Y}^0$ is the output of algorithm 1.*

Proof. See Appendix C. □

The resulting estimates $\hat{\tau}(r)$ for different values of r allow researchers to trace out how their treatment effects are expected to vary under different scale-up or scale-down scenarios vis-à-vis the baseline scenario where no intervention took place, even if they only conducted a single RCT with a fixed r . This can be a policy-relevant tool, especially when it comes to evaluating the expansion of high-cost interventions that might induce applicants to significantly interfere with each other, such as providing free tuition to college applicants (Bucarey, 2018). Having an estimate of $\tau(r)$ allows policymakers to optimize the scale of an intervention (e.g. considering the cost of expanding the intervention, and some welfare metric that depends on $\tau(r)$) and target this optimal r^* after conducting a smaller experiment.¹⁸

4.2 Congestion Effects

As the fraction of treated units grows, the treatment can induce congestion effects within the treatment group. For example, in our running example, the impact of providing information about high value added schools can decrease as the number of treated applicants increases, due to potential displacements of other treated applicants that would have been matched to a school in the target set H had the treatment been provided to fewer applicants. As opposed to the spillovers *between* experimental arms (where treated units affect the outcomes of control units), these congestion effects arise *within* the treatment group, i.e. not only can treated units affect control units, but they can also affect other treated units.

The congestion effect $c(r_0, r_1)$ stemming from an increase in the fraction of treated applicants from r_0 to r_1 is the change in the expected outcome for treated units when shifting from r_0 to r_1 :

$$c(r_0, r_1) := \mathbb{E}[Y_i | i \in T, \theta_T^1, \theta_C^0, r_1] - \mathbb{E}[Y_i | i \in T, \theta_T^1, \theta_C^0, r_0].$$

As shown above, we can consistently estimate $\mathbb{E}[Y_i | i \in T, \theta_T^1, \theta_C^0, r]$ for any $r \in (0, 1]$. Thus, $c(r_0, r_1)$ can be estimated using algorithm 2.

Corollary 2. *Let $r_0, r_1 \in (0, 1]$ and $\theta_i^d \perp r_0, r_1$ for $d \in \{0, 1\}$. If Assumption 1 and Assumption 2 hold, then $\hat{c}(r_0, r_1) := \bar{Y}^1(r_1) - \bar{Y}^1(r_0)$ is a consistent estimator of $c(r_0, r_1)$, where $\bar{Y}^1(r)$ is defined as the output of algorithm 2.*

The congestion effect $c(r_0, r_1)$ can be equivalently expressed as a difference in treatment effects, as $c(r_0, r_1) = \tau(r_1) - \tau(r_0)$. A direct implication from this analysis is that researchers can identify the values of r for which *marginal* congestion effects are stronger by tracing the estimates $\hat{\tau}(r)$ for a granular set of values of $r \in (0, 1)$ and finding the regions with steeper slopes.¹⁹ While regions with slopes close to 0 can be interpreted as evidence of insignificant marginal congestion within the treatment group ($c(r_0, r_0 + \epsilon) \approx 0$), they do not rule out the existence of spillovers σ .

¹⁸For example, if $\hat{\tau}(r)$ is a monetary measure of the ATT (e.g. computed using estimates of value added in future income), and $c(r)$ is the average cost of providing the intervention to a share r of the population, then a policymaker interested in the welfare of the treatment group can solve $r^* := \arg \max_{r \in [0, 1]} nr[\hat{\tau}(r) - c(r)]$. An analogous estimate of $\hat{\sigma}(r)$ must be incorporated to account for the welfare of the control group.

¹⁹Alternatively, the same procedure can be implemented with $\bar{Y}^1(r)$ instead of $\hat{\tau}(r)$.

4.3 Multiple Treatment Arms and Clustered Trials

Consider an experiment with multiple treatment arms indexed by $d = 0, \dots, D$, where $d = 0$ denotes the “pure control” and $D \geq 1$ is the number of treatment arms. In this situation, just as in the standard case, the different treatments can affect the outcomes of the control group. However, these treatments can also interfere among each other in potentially complex ways (e.g. treatments that induce demand for similar schools can interfere more than treatments that promote disjoint sets of schools).

Indeed, there are $2^D - 1$ counterfactual situations of interest, considering all the combinations where each treatment either takes place or not (minus the observed “factual” situation where all treatments take place). Similar to algorithm 1 and algorithm 2, we can simulate all these $2^D - 1$ assignment probabilities, creating a set of estimates $\bar{Y}^a$, where the binary vector $a \in \{0, 1\}^D$ denotes the combination of “active” treatment arms in each estimate. For each estimate, the reported types in the inactive treatment groups are replaced by random samples with replacement from the control group ($d = 0$) and the match is simulated multiple times.

An analogous procedure can be applied to clustered trials, defined by clusters $k = 1, \dots, K$, where each cluster has a control group and at least one treatment group. In this case, we can define the treatment arms in each cluster with the index $d_k = 0, \dots, D_k$ and repeat the previous procedure. This yields a set of $2^{\mathbf{D}} - 1$ estimates $\bar{Y}^{\mathbf{a}}$, where $\mathbf{D} := \sum_{k=1}^K D_k$ and $\mathbf{a} \in \{0, 1\}^{\mathbf{D}}$ denotes the concatenation of active treatments in each cluster. Importantly, for each active treatment arm, the resampling of control units should be done exclusively with units belonging to the cluster of the corresponding treatment group.

For example, consider the design in Ajzenman et al. (2026a), where they describe a clustered RCT in the Chilean centralized college application system. They study the effectiveness of humans and chatbots motivating high school students to apply for education degrees when submitting their college applications. In the first cluster ($k = 1$) they targeted high school students who had expressed interest in education degrees with both human and chatbot promoters (so $D_1 = 2$), whereas in the second cluster ($k = 2$) they targeted students interested in social sciences only with chatbots (i.e. $D_2 = 1$). In this case, there are $2^{(2+1)} - 1 = 7$ target parameters that can be estimated with this extension of algorithm 1.²⁰

5 Empirical Applications

We now show how these tools can be applied in empirical settings with two examples. The first example uses data from the Chilean centralized school choice system and a simulated experiment changing the preferences of applicants in the system. The second example uses data from Ajzenman et al. (2026b). This experiment took place in the Ecuadorian centralized teacher labor market, and made teachers more likely to apply for jobs at hard-to-staff schools.

5.1 Simulated Experiment: The Chilean School Choice System

5.1.1 Setting

In 2016, Chile initiated the implementation of a centralized school choice system for all PreK-12 public schools. The rollout of the system started with some levels in the Magallanes (southernmost) region in 2016, and ended with all levels across all regions in the country in 2020. In this application, we focus on the 2016 Magallanes-region implementation. Specifically, we simulate an experiment with the 2016 applicants seeking admission to 9th grade²¹ in 2017.

In this centralized school choice system, applicants can rank as many schools as they want and then submit this rank-ordered-list (ROL) to an online platform. Only schools that offer their corresponding grade (9th grade in this case) can be listed. The reported types θ are the preferences submitted to

²⁰To be explicit about the notation above, in this case the vector $\mathbf{a} \in \{0, 1\}^3$ represents the active treatment indicators (human-education, bot-education, bot-social science) $\in \{0, 1\}^3$.

²¹In Chile, the 9th grade of instruction is the first grade of high school and is called *primero medio*.

the platform. In this specific cohort of 9th grade applicants in 2016, the matching algorithm φ is equivalent to the Student-Proposing Deferred Acceptance (DA) Algorithm (Gale and Shapley, 1962) with a single random tie-breaker, which satisfies CD-ETE.²² The student-proposing DA algorithm with a single tiebreaker is described below:

- Step 0 Generate random tiebreakers $u_i \sim U[0, 1]$, such that schools have strict preferences over applicants.
- Step 1 All students apply to their favorite school, and all schools *tentatively* accept their most preferred applicants up to their capacity nq_s . All other applicants are rejected.
- Step k All students rejected in step $k - 1$ apply to their next preferred school. All schools *tentatively* accept their most preferred applicants among the pool of new applicants and previously accepted applicants, up to their capacity nq_s . All other applicants are rejected.
- Stop When there are no more rejections, the final allocation is the tentative assignment in the last step.

5.1.2 Data

This implementation of the school choice match was the smallest in the history of the Chilean school choice system: for the 9th grade level there were $|I| = 1040$ applicants that could rank up to $|S| = 24$ schools offering that level. For each applicant, we observe data on the individually submitted types θ (including the realized random tie-breaking numbers) and the resulting DA allocation $\mu(i), \forall i \in I$. We also observe the school capacities $nq_s, \forall s \in S$.

We report summary statistics of this data in Table 1. Even though applicants can list up to 24 schools, the average student listed only 3.7 schools, and the longest list had 15 schools ranked. Moreover, the average school has 53.67 vacant seats, which implies that there is excess supply of vacancies in this system (1288 seats available for 1040 applicants). If schools have heterogeneous qualities and some high-quality schools have slack capacity, inducing demand for these schools could be welfare-enhancing (Ainsworth et al., 2023).

Using a rich set of controls from the Ministry of Education (MINEDUC) and test results from Chile’s national 10th-grade math exam (SIMCE), we estimate a fixed-effects regression model with the 2018 (future) SIMCE test scores as a dependent variable:

$$score_i = \sum_{s \in S} \alpha_s \mathbf{1}[\mu(i) = s] + \beta' X_i + \varepsilon_i,$$

where $score_i$ denotes the math test score of student i , X_i contains a set of controls (schooling records, including grades and attendance for the previous 10 years of schooling), and ε_i is assumed to be uncorrelated with the regressors. Under unconfoundedness, the set of fixed effects $\{\alpha_s\}_{s \in S}$ is a value-added measure for schools.

We plot these coefficients in Figure 2, along with the fraction of ROLs that list each school. This figure depicts how demand for schools is *negatively* correlated with the estimated value added. One mechanism that could explain this empirical finding is the presence of uninformed applicants, which has both efficiency implications (e.g. if seats at high-value-added schools remain unfilled) and distributive implications (e.g. if low-income families are more prone to being uninformed).²³

The aforementioned descriptive facts motivate our simulated experiment. Consider an intervention where 20% of the applicants ($r = 0.2$) randomly change their preferences to follow the estimated value-added profile in Figure 2, i.e. all students in T list the highest value-added school first, the second

²²There is a second round held for individuals that remain unmatched after the DA allocation, but we only study the first round of the system. Moreover, for other grade-levels and other years of implementation, the algorithm considers other factors, such as priorities based on siblings or test scores at some selective schools. For more details on how the Chilean matching algorithm works, see Correa et al. (2019).

²³Even though this correlation is not causal, and there might be other mechanisms explaining the pattern (e.g. students might not apply to high-value-added schools because they offer few seats), we use it to justify the simulated experiment below in the spirit of Allende et al. (2019).

Table 1: Summary Statistics

Variable	Mean	St. Dev.	Min.	P25	Median	P75	Max.
<i>Panel A: Schools (N=24)</i>							
Vacancies	53.67	55.78	0	5	31	97	173
Copayment Indicator	0.7917	0.4149	0	1	1	1	1
Coed. Indicator	0.4583	0.5090	0	0	0	1	1
<i>Panel B: Students (N=1,040)</i>							
Female Indicator	0.4769	0.4997	0	0	0	1	1
Low-SES Indicator	0.5221	0.4998	0	0	1	1	1
(Younger) Sibling Indicator	0.0837	0.2770	0	0	0	0	1
<i>Panel C: Applications (N=3,846*)</i>							
Priority Indicator	0.0941	0.2920	0	0	0	0	1
List Length	3.70	1.58	2	3	3	5	15
<i>Panel D: Allocation (N=966)</i>							
1st pref. Indicator	0.6605	0.4738	0	0	1	1	1
2nd pref. Indicator	0.1594	0.3663	0	0	0	0	1
3rd pref. Indicator	0.1046	0.3061	0	0	0	0	1

Note: This table reports summary statistics using data from the Chilean centralized school choice system. (*List length is collapsed)

highest value-added next and so on. The types of the remaining students are unchanged. Suppose that the outcome of interest Y_i is an indicator equal to 1 if student i is matched with a top-5 VA school.

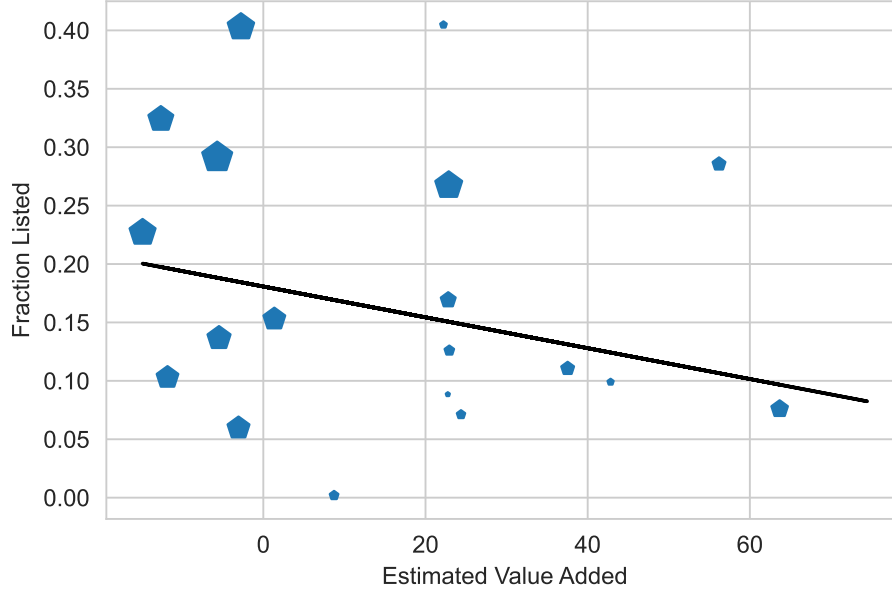
5.1.3 Results

Table 2 reports the results of conducting two estimation procedures on the experimentally modified data. The first column reports the estimates from the regression model described by

$$Y_i = \alpha + \beta \text{Treatment}_i + \varepsilon_i,$$

where the estimate $\hat{\beta}$ is equivalent to the difference-in-means estimator $\hat{\Delta}$. The second column shows the resulting estimates from applying our consistent estimator described in Theorem 1.

Figure 2: Value added is negatively correlated with share listed



Note: This figure plots the estimated value added for each school and the fraction of applications that list each school. Pentagon sizes are proportional to available slots.

Table 2: Difference-in-means vs. Consistent estimator

	(1) Mean comparison	(2) Consistent estimator
Treatment Effect	0.0635*** (0.0174)	0.0462* (0.0180)
Control Mean	0.0558*** (0.0123)	0.0730*** (0.0127)
Observations	1040	1040

Standard errors in parentheses

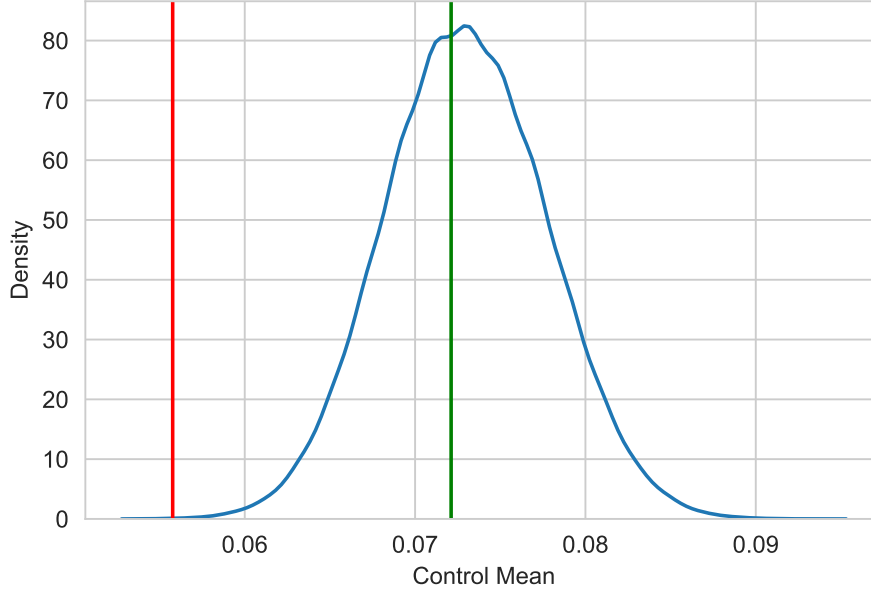
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Column (1) reports the estimates of a simple linear regression model with the treatment indicator as the independent variable. Column (2) reports the results of the consistent estimator derived from applying algorithm 1 to the simulated data. The t -statistic associated with the null hypothesis of equal treatment effects is 0.689 ($p = 0.491$).

The standard regression model estimate implies a 6.4 percentage point increase in the probability of being admitted into a top-5 school. On the contrary, the result of applying our consistent estimator to the data suggests an increase of only 4.6 percentage points. The former overestimates the latter by over 37%, albeit the difference in the estimated coefficients is not statistically significant at conventional levels.

Recall that in this simulated experiment, we artificially modified the data. Therefore, we can observe the true “counterfactual” distribution of the data, in absence of any treatment. Figure 3 shows how the consistent estimator correctly recenters the control mean by comparing the simulated control means with the true counterfactual probability of assignment into a top-5 school. Indeed, the red line shows the observed control average after the experiment, without using algorithm 1. This

Figure 3: Consistent Estimation of the Control Mean



Note: The red line denotes the observed control average of $Y_i, \forall i \in C$ in the experiment, without any corrections. The green line denotes the original control mean of Y_i , in absence of the experimental manipulation, i.e. the true counterfactual mean. The blue function is a smoothed density estimate of the simulated control means derived from applying algorithm 1 to the experimental data.

observed average significantly underestimates the control mean in absence of the treatment, suggesting that the treatment had spillover effects on the control group that reduced their likelihood of being assigned to a top-5 school. Thus, because the control mean is being underestimated, $\sigma < 0$, the treatment effect is being overestimated, i.e. $\hat{\Delta} \rightarrow \tau - \sigma > \tau$.

Finally, we apply the extension proposed in subsection 4.1 to estimate $\tau(r)$ under different fractions of treatment population $r \in \{0.1, 0.2, \dots, 0.8, 0.9\}$. The results are plotted in Figure 4. These results show a strong decay in $\tau(r)$ for low values of r , with a flatter curve after $r = 0.5$. This could be the case if for high values of r the target schools fill their capacities such that the observed differences between treatment and control units are driven less by τ and more by spillovers σ .

This simulated experiment highlights the empirical relevance of the theoretical results shown above. In this controlled simulation environment, we can observe how recentering the control mean using algorithm 1 to account for SUTVA violations can change the conclusions drawn from an experiment. We now also show how this can be the case in a *real* experiment.

5.2 Real Experiment: The Ecuadorian Teacher Labor Market

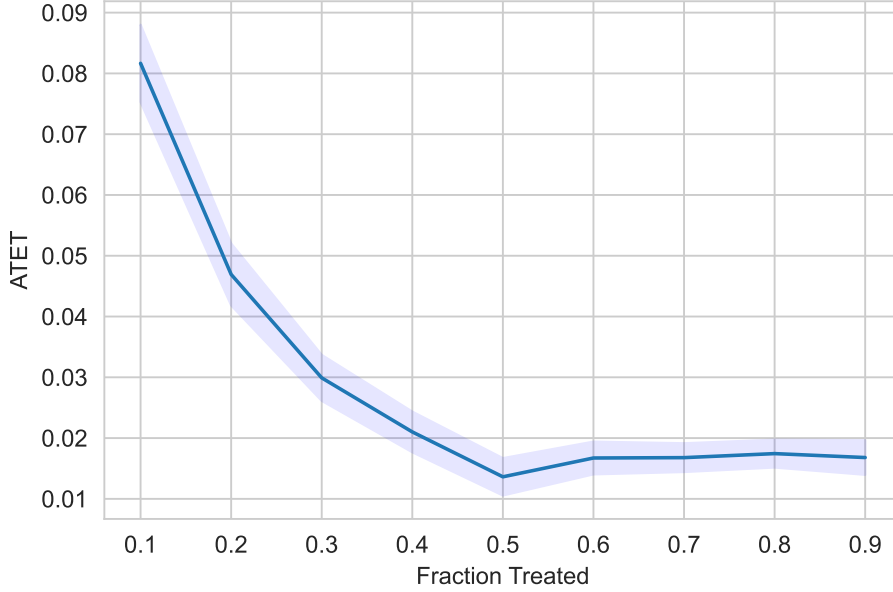
5.2.1 Setting

The Ecuadorian Ministry of Education uses a centralized matching system to assign teachers to jobs in public schools. The matching system is called *Quiero Ser Maestro*²⁴ (QSM) and has been in place since 2013. This empirical application focuses on a nationwide experiment conducted in 2019 by the Ministry of Education in collaboration with the Interamerican Development Bank (IDB). More details about this experiment can be found in Ajzenman et al. (2026b).

The objective of this experiment was twofold. First, the Ministry identified a set of schools which they classified as “hard to staff”, and decided to experimentally evaluate an intervention to better

²⁴This is “I want to be a teacher” in Spanish.

Figure 4: Consistent Estimates of $\tau(r)$



Note: The blue line plots the point estimates of $\tau(r)$ for different values of r . The shaded regions denote the 95% confidence intervals around the point estimates.

staff these schools. Hard-to-staff schools, which were categorized as such prior to the experiment, tend to be located in low-income areas and exhibit low average test scores when compared to other schools (see Figure 6 and Figure 7). Importantly, teaching jobs at these hard-to-staff schools constitute a fixed subset of jobs $H \subset S$ and the Ministry cared about the “staffing indicator” $Y_i = \mathbf{1}[\mu(i) = H]$, which is 1 when a teacher i is matched to a hard-to-staff position.

Second, the Ministry and the IDB aimed to evaluate a zero-cost intervention to test for “order effects” in this matching system. Specifically, a third of the teachers applying for jobs were randomly assigned into a treatment group T and saw hard-to-staff schools sorted at the top of the list of options in the application platform ($d_i = 1$). The control group C saw schools listed in alphabetical order ($d_i = 0$). Figure 5 depicts how the application platform looked for treatment and control teachers. They intended to test if highlighting hard-to-staff schools at the beginning of the applicants’ list of options would translate into a higher likelihood of applying and being matched to these target schools. In other words, they wanted to test how reports θ_i vary, and how staffing outcomes Y_i change as a result.

When teachers apply in the online portal, they first filter schools by *parroquia*²⁵ and then the system automatically shows all the job openings for their teaching specialty (e.g. high school math) in the selected parroquia. Some parroquias did not have hard-to-staff schools, while others only had hard-to-staff schools. This implies that teachers who only saw these parroquias did not observe any variation induced by the treatment. The IDB excluded teachers who were not subject to any sorting variation by the treatment due to their parroquia-specialty combination.

In addition, there was a second treatment arm targeting another third of the applicants. This second treatment was not deployed correctly, and thus the IDB also excluded this third of the sample from the analysis. The remaining applicants constitute the “restricted sample” defined by Ajzenman et al. (2026b), and the experimental results are valid for this subset of applicants.

Once the application period closes, the QSM platform runs the teacher-proposing DA algorithm to assign teachers to jobs in public schools. Teacher preferences for schools are determined by the rank-

²⁵*Parroquias* (parishes) are geographic administrative units used in Ecuador. They are similar to municipalities in other countries.

Figure 5: Screenshots of application platform (Ajzenman et al., 2026b)

Código AMIE	Institución educativa	Zonas	Provincias	Dirección de la institución educativa	Especialidad: EDUCACIÓN GENERAL BÁSICA (EGB) DE 200 A 7MO		Postular
					# de vacantes ofertadas	# de postulantes hasta 2020-05-18 10:29:05	
24H00091	AURELIO CARRERA CALVO	Zona: 5 Distrito: 24D01	Provincia: SANTA ELENA Cantón: SANTA ELENA Parroquia: COLONCHE	COMUNIDAD BAMBIL COLLAO - DIAGONAL AL DISPENSARIO	1	5	☐
24H00096	CARLOS JULIO AROSEMENA MONRROY	Zona: 5 Distrito: 24D01	Provincia: SANTA ELENA Cantón: SANTA ELENA Parroquia: COLONCHE	COMUNA AGUADITA	1	2	☐
24H00104	ESCUELA DE EDUCACION BASICA 24 DE MAYO	Zona: 5 Distrito: 24D01	Provincia: SANTA ELENA Cantón: SANTA ELENA Parroquia: COLONCHE	RECINTO EL COROZO VIA PREZA SAN VICENTE	2	4	☐
24H00092	ESCUELA DE EDUCACION BASICA CASIMIRO SORIANO BORBOR	Zona: 5 Distrito: 24D01	Provincia: SANTA ELENA Cantón: SANTA ELENA Parroquia: COLONCHE	FRENTE A LA CAPILLA SEÑOR DE LA MISERICORDIA	1	5	☐
				RCTO			

→

Código AMIE	Institución educativa	Zonas	Provincias	Dirección de la institución educativa	Especialidad: EDUCACIÓN GENERAL BÁSICA (EGB) DE 200 A 7MO		Postular
					# de vacantes ofertadas	# de postulantes hasta 2020-05-18 10:24:43	
24H00260	UNIDAD EDUCATIVA AYANGUE	Zona: 5 Distrito: 24D01	Provincia: SANTA ELENA Cantón: SANTA ELENA Parroquia: COLONCHE	COMUNA AYANGUE BARRIO SAN FRANCISCO VIA LABORATORIO AGUALAB	1	6	☐
24H00104	ESCUELA DE EDUCACION BASICA 24 DE MAYO	Zona: 5 Distrito: 24D01	Provincia: SANTA ELENA Cantón: SANTA ELENA Parroquia: COLONCHE	RECINTO EL COROZO VIA PREZA SAN VICENTE	2	3	☐
24H00106	UNIDAD EDUCATIVA SAN MARCOS	Zona: 5 Distrito: 24D01	Provincia: SANTA ELENA Cantón: SANTA ELENA Parroquia: COLONCHE	VIA A GUANGALA - FRENTE A LA IGLESIA SAN MARCOS	1	6	☐
24H00107	ESCUELA DE EDUCACION BASICA PRESIDENTE LIZARDO GARCIA	Zona: 5 Distrito: 24D01	Provincia: SANTA ELENA Cantón: SANTA ELENA Parroquia: COLONCHE	BARRIO FRANCISCO PIZARRO CALLE PRESBITERO CARVAJAL	2	11	☐

(a) Control Screen
(b) Treatment Screen

Note: The left panel shows a screenshot of the application platform for a teacher in the control group and the right panel shows a screenshot of the application platform for a teacher in the treatment group. On the left screen schools are sorted alphabetically (“Aurelio...”, then “Carlos...”, then “Escuela”), while on the right screen the schools at the top of the list are all hard-to-staff schools (highlighted with a circular logo in the first column).

ordered lists submitted in the online platform, which are limited to up to 5 jobs listed per applicant. School priorities for teachers are based on a composite score computed for each teacher at each job. Each teacher knows their scores in advance, as well as the number of vacancies for each job (e.g. teacher i knows that school s is hiring 2 high school math teachers and their composite score is 88.8 at that job).

Tiebreakers in this system give priority to teachers who ranked an option higher. In the event a tie persists, a (tentative) winner is picked at random. A more detailed description of the algorithm is presented in Elacqua et al. (2021), and summarized below:

Step 0 Generate tiebreakers giving priority to teachers that ranked jobs higher and a (dominated) random component. For example, a valid tiebreaker is

$$tiebreaker_{is} = \epsilon \times (6 - rank_{is} + u_i) \quad u_i \sim U[0, 1],$$

where $rank_{is} \in \{1, 2, 3, 4, 5\}$ is the position where teacher i ranked school s and $\epsilon > 0$ is a small enough scaling factor such that the original order induced by test scores is unaltered when adding $score_{is} + tiebreaker_{is}$, modulo the set of tied scores. After incorporating this tiebreaker, all preferences in the system are strict.

Step 1 All teachers apply to their favorite job, and all jobs *tentatively* accept their top-scoring applicants up to their capacity nq_s . All other applicants are rejected.

Step k All teachers rejected in step $k - 1$ apply to their next preferred job. All jobs *tentatively* accept their top-scoring applicants among the pool of new applicants and previously accepted applicants, up to their capacity nq_s . All other applicants are rejected.

Stop When there are no more rejections, the final allocation is the tentative assignment in the last step.

Because of this specific tie-breaking rule and the constraint on the length of the rank-ordered lists, this variant of the DA mechanism is not strategy-proof for teachers. However, this mechanism satisfies CD-ETE, as it has a cutoff representation where cutoffs converge in the large market. In fact, the QSM mechanism is a variant of the Cambridge mechanism described in Agarwal and Somaini (2018).

5.2.2 Data

In 2019, there were $|I| = 27,235$ qualified teachers²⁶ that applied for a teaching job at public schools through the QSM platform. There were 9,074 teachers assigned to a *failed* treatment arm, and they were excluded from the analysis by the IDB. Out of the 18,161 remaining teachers, 12,379 were not subject to any treatment variation because of their parroquia-specialty combination. Thus, the restricted sample as defined by Ajzenman et al. (2026b) contains only 5,782 applicants.²⁷

Table 3 shows descriptive statistics for baseline characteristics under different sample specifications. In the unrestricted sample, about 72% of applicants are women, 51% are single and 10% belong to ethnic minorities. On average, they score about 66/100 points and have around 3.4 years of experience. The restricted sample is slightly different: there are 5pp more females, they 3pp more likely to be single, they score about 1.3 more points on average, they have about 0.4 more years of experience and they are 0.7pp more likely to belong to an ethnic minority. In all cases, the treatment and control groups are balanced in their baseline characteristics.

These applicants chose from $|S| = 10,067$ jobs when applying through the online platform. On average, jobs offered 1.44 openings, the median being 1 opening. The 75th percentile was 2 openings, the 95th percentile was 3 openings, the 99th percentile was 6 openings, and the maximum was 16 openings. As most jobs offer few vacancies, and the total number of vacancies is smaller than the number of applicants n , it is unlikely that SUTVA holds in this setting.

5.2.3 Results

The first set of results describes the effects of the treatment on the reported types θ . Specifically, we focus on the length of the reported lists and the presence of hard-to-staff schools in them. For these “partial equilibrium” outcomes, we consider Assumption 1 holds and report linear regression estimates.

Table 4 reports the results of estimating the parameters of the following regression model:

$$L_i = \alpha + \beta \text{Treatment}_i + \gamma' X_i + \varepsilon_i, \quad (2)$$

where L_i is either the length of the list submitted to the system, the share of hard-to-staff schools listed, an indicator for listing a hard-to-staff school or an indicator for listing one as a first choice. The vector X_i contains the same set of controls variables as Ajzenman et al. (2026b): composite score, indicators for being female, being single, belonging to an ethnic minority, indicators for years of experience and the applicant’s district.

Column (1) in Table 4 shows no significant difference in the length of the lists reported by treatment and control teachers. However, columns (2)-(4) indicate a significant change in the composition of the lists submitted by the treatment group. Treated teachers increase their share of hard-to-staff schools listed by 1.3 percentage points (a 3% increase) and are 1.5 percentage points more likely to list at least one hard-to-staff school (a 1.7% increase). Moreover, treated applicants are 5.1 percentage points more

²⁶The Ministry of Education does not allow low-scoring teachers to apply for jobs in public schools. Qualified teachers are those who cleared the screening process comprised by a psychological test and a knowledge test developed by the Ministry. According to Ajzenman et al. (2026b): “only 27% of the 129,114 candidates who registered for the teacher selection process passed the eligibility phase”.

²⁷There is a minor difference in the number of applicants in this analysis when compared to Ajzenman et al. (2026b). This is due to 28 applicants not having an ID number in the data provided by the Ministry. In our analysis, we include these 28 applicants, as opposed to Ajzenman et al. (2026b), where they are dropped. We do this to replicate the results of the matching process as closely as possible. With our full sample, we are able to correctly assign 99.8% of applicants to their assigned jobs by running the teacher-proposing DA algorithm. The difference is likely due to issues in the raw input data (e.g. some applicants list the same job more than once in their rank-ordered lists).

Table 3: Baseline characteristics and sample balance tests

Variable	(1) Mean (sd)	(2) Control (sd)	(3) Treatment (sd)	(4) Difference (p-value)
<i>Panel A: Full Sample</i>				
Female	0.715 (0.451)	0.714 (0.452)	0.718 (0.450)	0.004 (0.480)
Single	0.513 (0.500)	0.511 (0.500)	0.517 (0.500)	0.006 (0.325)
Comp. Score	66.033 (9.914)	66.042 (9.954)	66.016 (9.833)	-0.026 (0.840)
Experience	3.375 (3.278)	3.390 (3.291)	3.346 (3.253)	-0.044 (0.297)
Minority	0.098 (0.298)	0.097 (0.296)	0.101 (0.301)	0.004 (0.278)
Observations	27,235	18,176	9,059	27,235
<i>Panel B: Excluding Failed Treatment Arm</i>				
Female	0.716 (0.451)	0.715 (0.452)	0.718 (0.450)	0.003 (0.648)
Single	0.512 (0.500)	0.507 (0.500)	0.517 (0.500)	0.010 (0.167)
Comp. Score	66.016 (9.910)	66.016 (9.987)	66.016 (9.833)	-0.000 (0.998)
Experience	3.359 (3.268)	3.371 (3.284)	3.346 (3.253)	-0.025 (0.600)
Minority	0.099 (0.298)	0.096 (0.295)	0.101 (0.301)	0.005 (0.252)
Observations	18,161	9,102	9,059	18,161
<i>Panel C: Final Restricted Sample</i>				
Female	0.766 (0.423)	0.768 (0.422)	0.765 (0.424)	-0.003 (0.785)
Single	0.540 (0.498)	0.532 (0.499)	0.549 (0.498)	0.017 (0.204)
Comp. Score	67.385 (9.393)	67.416 (9.467)	67.353 (9.319)	-0.064 (0.797)
Experience	3.737 (3.275)	3.739 (3.275)	3.734 (3.276)	-0.005 (0.954)
Minority	0.106 (0.308)	0.111 (0.314)	0.102 (0.302)	-0.010 (0.241)
Observations	5,782	2,918	2,864	5,782

Note: This table shows summary statistics for baseline characteristics in different subsamples. Female is an indicator equal to 1 if the applicant is female. Single is an indicator equal to 1 if the applicant is single. Comp. Score is the base composite score of the applicant, excluding application-specific bonuses. Experience is the years of experience of the applicant. Minority is an indicator equal to 1 if the applicant belongs to an ethnic minority. Panel A contains all applicants in the data. Panel B excludes applicants that were assigned to the failed treatment arm. In addition, Panel C also excludes applicants that were not subject to treatment variation because of their parroquia-specialty combination. The final column reports differences in means between the treatment and control groups, along with the associated p-value when testing the null hypothesis of zero difference in means.

Table 4: Treatment effects on application behavior

Outcome:	(1) List Length	(2) Share of HTS	(3) Ranked HTS	(4) First is HTS
Treatment	-0.005 (0.011)	0.013 (0.006)	0.015 (0.008)	0.051 (0.012)
Observations	5765	5765	5765	5765
Control Mean	4.908	.431	.865	.402
p -value	.65	.044	.048	0

Robust standard errors, clustered at the district level, in parentheses.

Treatment p -values reported in the last row.

Table 5: Treatment effects on match outcome

Outcome:	(1) Matched to HTS (Inconsistent)	(2) Matched to HTS (Consistent)
Treatment	0.034 (0.012)	-0.009 (0.012)
Observations	5765	5765
Control Mean	.27	.313
p -value	.005	.448

Robust standard errors, clustered at the district level, in parentheses.

Treatment (bootstrapped) p -values reported in the last row.

likely to list a hard-to-staff school as their first choice (a 12.7% increase). Results for other samples are comparable to those in Table 4 and are reported in the Appendix (Table 6 and Table 7).

It is important to notice that 86.5% of the control group ranked at least one hard-to-staff school, as shown in column (3) of Table 4. If hard-to-staff schools are “small”, and both experimental groups have high application rates to these schools, treatment applicants might crowd out control applicants from hard-to-staff schools when the treatment induces additional demand for these schools.

We now turn to the estimation of the treatment effect that the intervention had on the matching outcome Y_i . Table 5 shows the estimated treatment effects from computing the difference-in-means estimator $\hat{\Delta}$ and the consistent estimator $\hat{\tau}$ introduced in Theorem 1. Column (1) shows the results of computing the treatment effect using a standard linear regression (Equation 2). This estimate implies that teachers who saw hard-to-staff schools at the top of their list were 3.4 percentage points more likely to be assigned to a hard-to-staff school than the control group. The point estimate is significantly different from 0 and accounts for a 12.8% increase in the odds of being assigned to a hard-to-staff school when compared to the control mean.

However, in column (2), we account for SUTVA violations using algorithm 1 to adjust the control mean. Our consistent estimator suggests that the impact of the policy is significantly different from $\hat{\Delta}$, and cannot be statistically distinguished from zero. Indeed, the simulated control mean is about 4 percentage points higher than the observed control mean in the sample. This implies that the observed difference in means, which can be decomposed as $\hat{\tau} - \hat{\sigma}$, is mostly driven by spillovers of the treatment group on the control group, i.e. treated teachers were likely to displace control teachers in the experiment.

6 Conclusion

In this paper, we study the estimation of treatment effects in an RCT setting where the experimental units participate in a matching system. In this setting, the no-interference assumption is unlikely to hold when measuring match-dependant outcomes, such as being matched to a hard-to-staff school

in the Ecuadorian teacher labor market. This is because treated units might affect the equilibrium placement of control units, as each individual match depends on the system-wide profile of reported types. For example, if an intervention induces applicants to list hard-to-staff schools when applying for a job, this additional demand for the target schools might displace control units that would have listed hard-to-staff schools in the absence of any treatment.

Interference poses a challenge to the estimation of causal treatment effects, even if the data comes from a well-executed RCT. We first show how conventional estimators that compare the average outcomes for treatment units against control units are biased and inconsistent whenever there is interference in a large experiment (i.e. when the fraction of treated units $r \not\rightarrow 0$). Indeed, this observed difference not only accounts for the ATT, but also contains the spillover effect, which we define as the “treatment effect” on the control group. Moreover, in empirical settings where the treatment induces demand for a set of options and the outcome depends on being matched to these options, the estimates obtained from the standard difference-in-means approach are likely to overestimate the true ATT. We show how this is the case in two empirical applications.

To solve the interference issue, we propose a nonparametric, simulation-based estimator that consistently estimates the target causal parameter under a modified version of the equal treatment of equals (ETE) condition. This condition on the matching algorithm, which we call continuous distributional equal treatment of equals (CD-ETE), allows researchers to decompose the observed average difference between treatment and control units into an estimator for the ATT, and a residual estimator for the spillover effect. We show the validity of this estimator when SUTVA holds at the report level (Assumption 1), without requiring it to hold for equilibrium outcomes.

Moreover, we propose extensions to this estimator, such that we can estimate treatment effects under different counterfactual intensities of treatment (different fractions of treatment shares r) and accommodate RCTs with multiple treatment arms and clusters. The former is a useful tool for policy recommendations, such as estimating treatment effects under an expansion or contraction of the intervention, and the latter fits popular RCT designs that are implemented in practice.

We apply our estimator using data from the Chilean school choice system, and the Ecuadorian teacher labor market match. In both cases, as expected, we find lower point estimates for the ATT parameter when correcting for the interference induced in the matching system. In the case of Ecuador, we cannot distinguish our estimated treatment effect from a statistical zero, implying that the observed difference between the treatment and control groups in that experiment is likely driven by control units being displaced by treatment units. This result highlights the importance of our contribution, as it pertains to experiments that are conducted for policy evaluations in settings with high-stake outcomes (e.g. jobs of teachers in a whole country).

Finally, in this paper, we specifically focus on RCTs conducted in matching systems that comply with CD-ETE. Future research can explore quasi-experimental settings where no RCT takes place, such as estimating treatment effects with interference using a difference-in-difference research design, or studying local interference where a change in reported types occurs in a regression discontinuity setting. More research related to estimating treatment effects on other downstream outcomes (e.g. future test scores) that can be a function of match outcomes is also warranted. Additionally, further research is needed to study the behavior of matching systems that either do not comply with CD-ETE (e.g. that comply with ETE, but not CD-ETE), or are part of a specific subset of mechanisms that provide additional structure that researchers can exploit to improve on our results (e.g. Agarwal and Somaini (2018) show how simulations of RSP+C mechanisms, which comply with CD-ETE, can be used to construct estimators of match probabilities that are also asymptotically normal).

References

- Abdulkadiroğlu, A., Angrist, J. D., Narita, Y., and Pathak, P. A. (2017). Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation. Econometrica, 85(5):1373–1432.
- Abdulkadiroğlu, A. and Sönmez, T. (2003). School Choice: A Mechanism Design Approach. American Economic Review, 93(3):729–747.
- Agarwal, N. and Somaini, P. (2018). Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism. Econometrica, 86(2):391–444.
- Ainsworth, R., Dehejia, R., Pop-Eleches, C., and Urquiola, M. (2023). Why do households leave school value added on the table? the roles of information and preferences. American Economic Review, 113(4):1049–82.
- Ajzenman, N., Elacqua, G., Jaimovich, A., and Pérez-Núñez, G. (2026a). Humans Versus Chatbots: Scaling-Up Behavioral Interventions to Reduce Teacher Shortages. Journal of Political Economy: Microeconomics.
- Ajzenman, N., Elacqua, G., Marotta, L., and Olsen, A. S. (2026b). Order Effects and Employment Decisions: Experimental Evidence from a Nation-Wide RCT in Ecuador. American Economic Journal: Applied Economics.
- Allende, C., Gallego, F., and Neilson, C. (2019). Approximating the Equilibrium Effects of Informed School Choice. Working Paper, (1100623).
- Altmejd, A., Barrios-Fernández, A., Drlje, M., Goodman, J., Hurwitz, M., Kovac, D., Mulhern, C., Neilson, C., and Smith, J. (2021). O Brother, Where Start Thou? Sibling Spillovers on College and Major Choice in Four Countries. The Quarterly Journal of Economics, 136(3):1831–1886.
- Angrist, J., Gray-Lobe, G., Idoux, C., and Pathak, P. A. (2022). Still Worth the Trip? School Busing Effects in Boston and New York. SSRN Electronic Journal, (July).
- Arteaga, F., Kapor, A. J., Neilson, C. A., and Zimmerman, S. D. (2022). Smart Matching Platforms and Heterogeneous Beliefs in Centralized School Choice. Quarterly Journal of Economics, 137(3):1791–1848.
- Azevedo, E. M. and Leshno, J. D. (2016). A Supply and Demand Framework for Two-Sided Matching Markets. Journal of Political Economy, 124(5):1235–1268.
- Basse, G. W., Feller, A., and Toulis, P. (2019). Randomization tests of causal effects under interference. Biometrika, 106(2):487–494.
- Bucarey, A. (2018). Who Pays for Free College? Crowding Out on Campus. Job Market Paper, pages 1–71.
- Che, Y.-K. and Kojima, F. (2010). Asymptotic Equivalence of Probabilistic Serial and Random Priority Mechanisms. Econometrica, 78(5):1625–1672.
- Correa, J., Epstein, R., Escobar, J., Rios, I., Bahamondes, B., Bonet, C., Epstein, N., Aramayo, N., Castillo, M., Cristi, A., and Epstein, B. (2019). School choice in Chile. ACM EC 2019 - Proceedings of the 2019 ACM Conference on Economics and Computation, pages 325–343.
- Cox, D. (1958). Planning of Experiments. John Wiley & Sons, Inc., New York.
- Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., and Zamora, P. (2013). Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment. The Quarterly Journal of Economics, 128(2):531–580.

- Duflo, E., Glennerster, R., and Kremer, M. (2007). Chapter 61 Using Randomization in Development Economics Research: A Toolkit. In Handbook of Development Economics, volume 4, pages 3895–3962.
- Efron, B. and Tibshirani, R. (1994). An Introduction to the Bootstrap. Number 3. Chapman and Hall/CRC.
- Elacqua, G., Westh Olsen, A. S., and Velez-Ferro, S. (2021). The Market Design Approach to Teacher Assignment: Evidence from Ecuador. Technical report, Inter-American Development Bank, Washington, D. C.
- Gale, D. and Shapley, L. S. (1962). College Admissions and the Stability of Marriage. The American Mathematical Monthly, 69(1):9–15.
- Horowitz, J. L. (2001). The Bootstrap. In Handbook of Econometrics, volume 5, pages 3159–3228.
- Kapor, A., Karnani, M., and Neilson, C. (2024). Aftermarket frictions and the cost of off-platform options in centralized assignment mechanisms. Journal of Political Economy, 132(7):2346–2395.
- Leung, M. P. (2020). Treatment and Spillover Effects Under Network Interference. The Review of Economics and Statistics, 102(2):368–380.
- Lewis, H. G. (1983). Union Relative Wage Effects: A Survey of Macro Estimates. Journal of Labor Economics, 1(1):1–27.
- Neilson, C. A. (2024). The rise of coordinated choice and assignment systems in education markets around the world. pages 1–24.
- Neyman, J. (1990). On the Application of Probability Theory to Agricultural Experiments. Statistical Science, 5(4):465–472.
- Pathak, P. A. and Sönmez, T. (2008). Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism. American Economic Review, 98(4):1636–1652.
- Roth, A. E. and Sotomayor, M. (1992). Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis. Econometric Society Monographs.
- Rubin, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701.
- Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. Journal of the American Statistical Association, 75(371):591.

A Additional Tables

Table 6: Effects on applications - Full sample

Outcome:	(1) List Length	(2) Share of HTS	(3) Ranked HTS	(4) First is HTS
Treatment	0.001 (0.011)	0.007 (0.003)	0.012 (0.006)	0.025 (0.006)
Observations	27164	27164	27164	27164
Control Mean	4.908	.431	.865	.402
<i>p</i> -value	.897	.012	.034	0

Robust standard errors, clustered at the district level, in parentheses.

Treatment *p* -values reported in the last row.

Table 7: Effects on applications - Without failed treatment arm

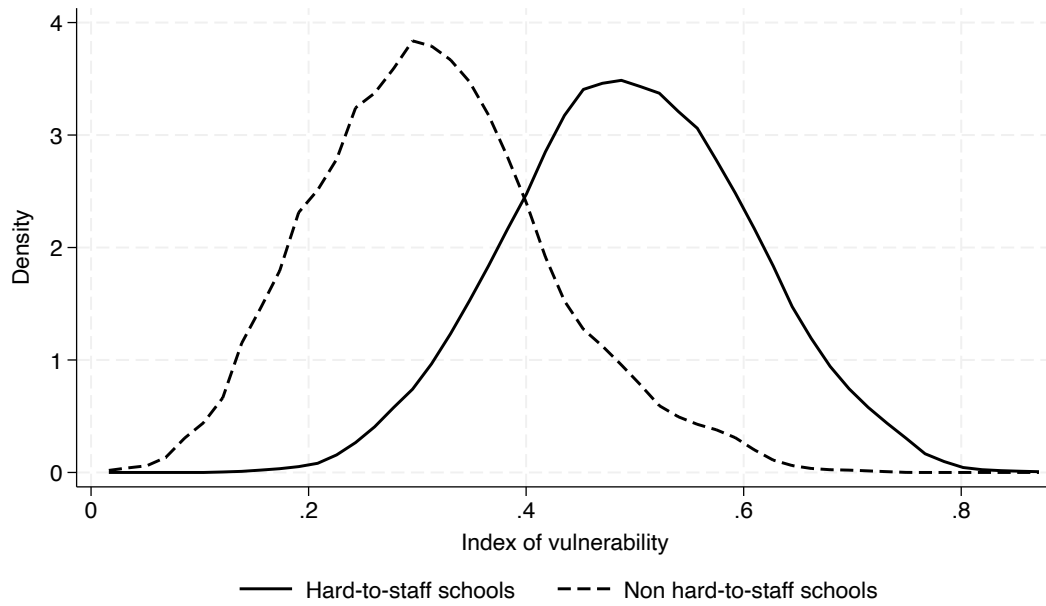
Outcome:	(1) List Length	(2) Share of HTS	(3) Ranked HTS	(4) First is HTS
Treatment	0.003 (0.012)	0.006 (0.003)	0.011 (0.007)	0.026 (0.007)
Observations	18108	18108	18108	18108
Control Mean	4.908	.431	.865	.402
<i>p</i> -value	.811	.06	.102	0

Robust standard errors, clustered at the district level, in parentheses.

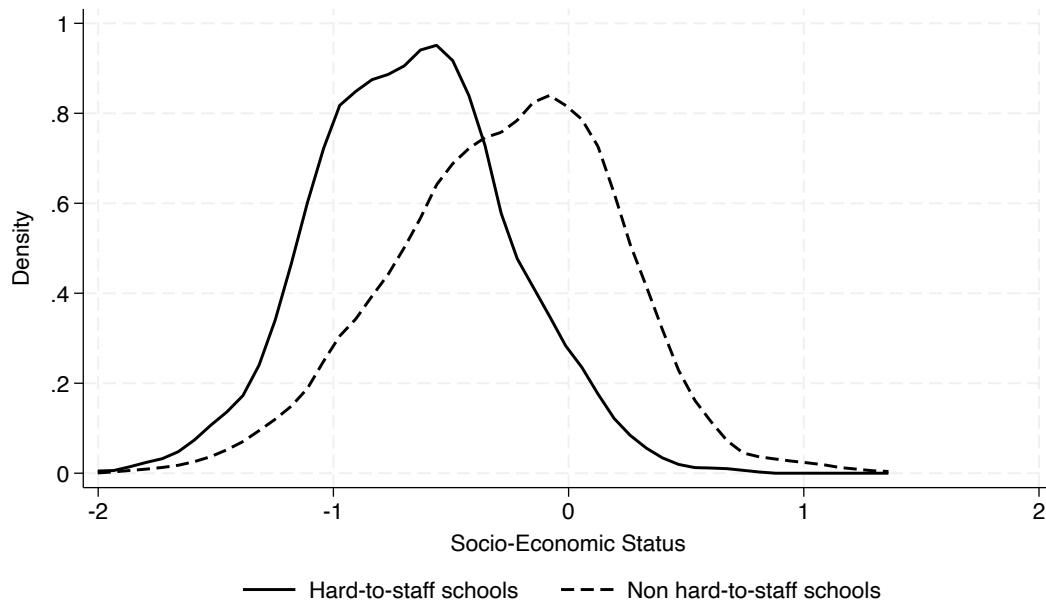
Treatment *p* -values reported in the last row.

B Additional Figures

Figure 6: School Characteristics - Socio-Economic Variables



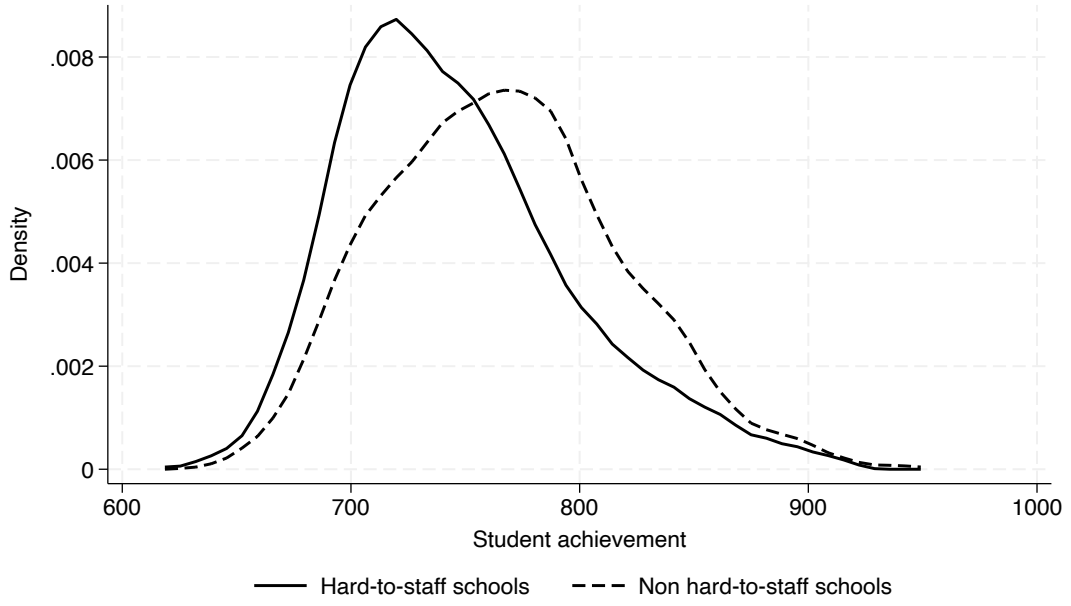
(a) School Vulnerability Index (Poverty Rate)



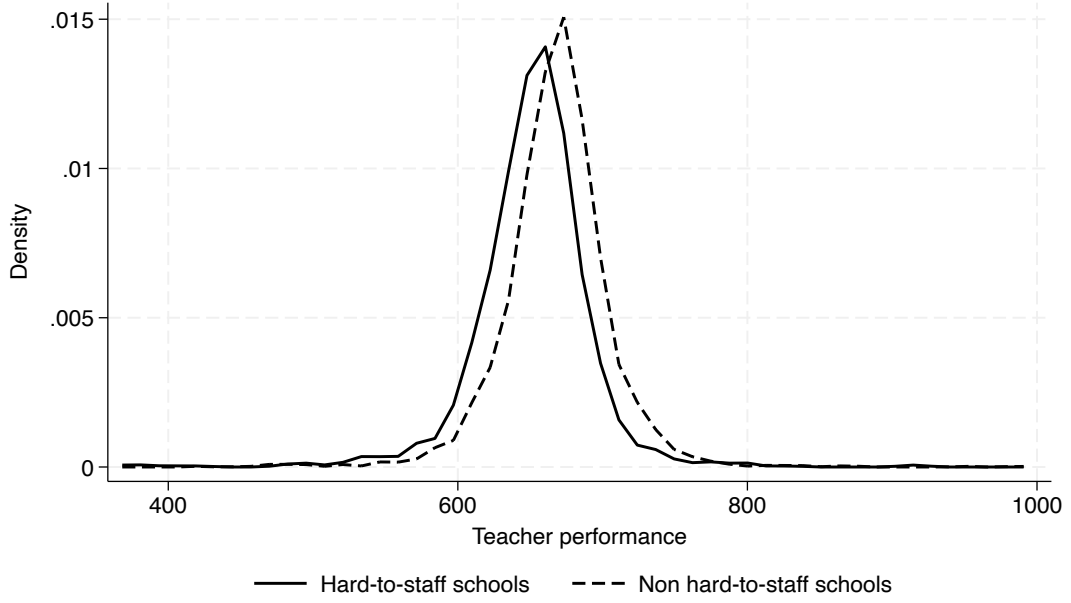
(b) Student SES Index

Note: This figure is based on Ajzenman et al. (2026b) and describes socioeconomic characteristics of hard-to-staff schools and other schools. Panel (a) shows the estimated densities of the school-level poverty rates, while panel (b) plots the estimated densities of an index for the socioeconomic status of students.

Figure 7: School Characteristics - Achievement Variables



(a) Student Test Scores



(b) Teacher Test Scores

Note: This figure is based on Ajzenman et al. (2026b) and describes academic achievement variables of hard-to-staff schools and other schools. Panel (a) shows the estimated densities of the student test scores at schools, while panel (b) plots the estimated densities of teachers' base composite scores.

C Additional Proof

Let $r \in (0, 1]$ and $\theta_i^d \perp r$ for $d \in \{0, 1\}$. If Assumption 1 and Assumption 2 hold, then $\hat{\tau}(r) := \bar{Y}^1(r) - \bar{Y}^0$ is a consistent estimator of $\tau(r)$, where $\bar{Y}^1(r)$ is defined as the output of algorithm 2 and $\bar{Y}^0$ is the output of algorithm 1.

Proof. As d_i is randomly assigned, Lemma 3 implies that

$$\bar{Y}_0 \rightarrow \mathbb{E}[Y_i | \theta_I^0] = \mathbb{E}[Y_i | i \in T, \theta_I^0],$$

almost surely.

Similar to the proof of Lemma 3, the empirical distribution $F_{\theta_T}^n$ of reported types in the treatment group θ_T is an independent random sample of size $|T|$ from the population distribution F_{θ^1} . Analogously, $F_{\theta_C}^n$ is an independent random sample of size $|C|$ from the population distribution F_{θ^0} . Thus, $\hat{\theta}_I^r$ comes from independently drawing nr types from the empirical measure $F_{\theta_T}^n$ and $n(1-r)$ types from the empirical measure $F_{\theta_C}^n$. This induces the empirical measure $\hat{F}_{(\theta^1, \theta^0), r}^n$.

By the Glivenko-Cantelli theorem,

$$\lim_{n \rightarrow \infty} \sup |\hat{F}_{(\theta^1, \theta^0), r}^n - F_{(\theta^1, \theta^0), r}| = 0,$$

almost surely.

As φ complies with CD-ETE (Assumption 2), then the plug-in empirical measure $\hat{F}_{\varphi, r}^n$, induced by $\varphi(\hat{\theta}_I^r)$, also converges almost surely to the population distribution $F_{\varphi, r}$:

$$\lim_{n \rightarrow \infty} \sup |\hat{F}_{\varphi, r}^n - F_{\varphi, r}| = 0.$$

As φ is a random mechanism, we can draw empirical bootstrap samples from this consistent estimator of the counterfactual distribution $F_{\varphi, r}$ by randomly drawing match-outcomes of size nr for treatment individuals in the simulation $\{Y^1[:, b]\}_{b=1}^B$ with a large B , as is standard in bootstrap methods. Convergence in the simulated outcomes resulting from the random matches $\hat{\mu}|T$ is guaranteed by the Continuous Mapping Theorem because Y_i is a linear function of $\mu(i)|T$.

Finally, as the empirical distribution of simulated bootstrap outcomes for the treatment group comes from a consistent estimator of the counterfactual distribution $F_{\varphi, r}|T$, the continuous (linear) functional of this bootstrap distribution $\bar{Y}^1(r)$ is a sample average that converges almost surely to the population expectation $\mathbb{E}[Y_i|i \in T, \theta_T^1, \theta_C^0, r]$:

$$\lim_{B \rightarrow \infty} \lim_{n \rightarrow \infty} \sum_{b=1}^B \sum_{i=1}^{nr} Y^1[i, b] / (nrB) = \mathbb{E}[Y_i|i \in T, \theta_T^1, \theta_C^0, r],$$

almost surely.

Then,

$$\begin{aligned} \hat{\tau}(r) &= \bar{Y}^1(r) - \bar{Y}^0 \\ &\rightarrow \mathbb{E}[Y_i|i \in T, \theta_T^1, \theta_C^0, r] - \mathbb{E}[Y_i|i \in T, \theta_I^0] \\ &= \tau(r) \end{aligned}$$

almost surely. □