UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

LOTTERY DESIGN IN SCHOOL CHOICE: USING OBSERVABLE STUDENT CHARACTERISTICS TO IMPROVE EFFICIENCY

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN ECONOMÍA APLICADA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

MARTÍN ANDRÉS VALDEVENITO BELMAR

PROFESOR GUÍA:
JUAN ESCOBAR CASTRO

MIEMBROS DE LA COMISIÓN:
JOSÉ RAFAEL CORREA HAEUSSLER
ALEJANDRO CORVALÁN AGUILAR

Este trabajo ha sido parcialmente financiado por el
Instituto Milenio para el Estudio de Imperfecciones de Mercado y Políticas Públicas.

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE MAGÍSTER EN ECONOMÍA APLICADA
POR: MARTÍN ANDRÉS VALDEVENITO BELMAR
FECHA: 2022
PROF. GUÍA: JUAN ESCOBAR CASTRO

DISEÑO DE LOTERÍAS EN EL SISTEMA DE ADMISIÓN ESCOLAR: UTILIZANDO CARACTERÍSTICAS OBSERVABLES PARA MEJORAR LA EFICIENCIA

El problema de asignación escolar ha sido teorizado como un problema de emparejamiento de dos lados con una particularidad: a priori, los colegios están indiferentes entre grandes grupos de estudiantes. Esto significa que el mecanismo debe especificar cómo desempatar entre estudiantes. Al considerar emparejamientos estables (como aquel que resulta del Algoritmo de Aceptación Diferida), la forma en que se realizan dichos desempates tiene un efecto importante en la eficiencia del resultado.

Comúnmente, como una forma de garantizar equidad ex-ante para los estudiantes, los desempates son definidos de manera aleatoria. Sin embargo, el tipo de lotería utilizada puede generar pérdidas importantes en eficiencia ya que las estabilidad impone restricciones sobre los emparejamientos. Conceptualmente, los mecanismos estables son altamente competitivos y esta competencia excesiva, aunque asegura estabilidad, genera ineficiencias. Dicho esto, pensamos que el diseño de loterías puede ser utilizado para limitar el exceso de competencia al correlacionar preferencias entre ambos lados del mercado: intuitivamente, los colegios otorgan una alta prioridad a estudiantes que a su vez clasifican altamente a ese colegio.

Para correlacionar preferencias en un mecanismo a prueba de estrategias, el diseñador de política se ve de algún modo forzado a predecir preferencias. Definimos las políticas basadas en datos como aquellas políticas que asignan prioridad de acuerdo a la afinidad esperada de un estudiante según una característica observable. Mostramos que cuando esta característica es un buen predictor de las preferencias, las políticas basadas en datos son más eficientes que los diseños aleatorios comunes (MTB y STB). Por el contrario, cuando son un mal predictor de las preferencias, estas políticas pueden llegar a ser más ineficientes dado su carácter determinista.

Complementamos los hallazgos teóricos con simulaciones basadas en datos del Sistema de Admisión Escolar 2021 en Chile. Utilizando las distancias a los colegios como característica observable, mostramos que la política basada en datos reduce sustancialmente los estudiantes pertenecientes a pares de mejora y aumenta la eficiencia en términos de que asigna a más estudiantes a sus primeras 3 preferencias. Además, las simulaciones sugieren que, en contraste con STB, este diseño evita una menor asignación en rankings más bajos y no parece afectar la cantidad de estudiantes sin asignar. Atribuimos este efecto al hecho de implementar políticas que explotan la heterogeneidad de las preferencias en vez de imponer un diseño 'dictatorial'.

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE MAGÍSTER EN ECONOMÍA APLICADA
POR: MARTÍN ANDRÉS VALDEVENITO BELMAR
FECHA: 2022
PROF. GUÍA: JUAN ESCOBAR CASTRO

LOTTERY DESIGN IN SCHOOL CHOICE: USING OBSERVABLE STUDENT CHARACTERISTICS TO IMPROVE EFFICIENCY

The school choice problem is theoretically conceived as a many-to-one two-sided matching market with one particularity: a priori, schools are indifferent over large groups of students. This means that the mechanism must specify how ties are broken within indifference groups. When considering stable mechanisms (such as the one induced by the Deferred Acceptance algorithm) tie-breaking has a significant effect on the efficiency of the allocation (Abdulkadiroglu, Pathak, Roth, & Sonmez, 2006; Erdil & Ergin, 2008).

Commonly, ties are broken randomly for each school as a way to ensure ex-ante equality for students. However, the type of lottery used can generate efficiency loss on the allocation since it imposes artificial stability constraints that must be met. Conceptually, stable mechanisms are highly competitive (Che & Tercieux, 2018) and excess competition, although it grants stability, is generally inefficient. Therefore, we state that lottery design might be a useful tool to limit excess competition by correlating preferences cross-market. Intuitively, schools grant high priority to students that likewise rank that school highly.

A first mechanism that comes to mind is correlating preferences perfectly by waiting for students to submit their preferences before generating lotteries. However, this mechanism would certainly not be strategy-proof. Thus, the policymaker is somehow forced to predict preferences. This is where we state that an observable student characteristic can come in handy as an informational proxy to guide lottery design.

We define data-driven lotteries as a policy that grants priorities according to the likely affinity a student has through the observed informational proxy. We show that when the characteristic is a powerful predictor of preferences, a data-driven policy is more efficient than traditional multiple tie-breaking and single tie-breaking policies. Conversely, when the proxy is a weak predictor, it can perform worse than MTB and STB due to its deterministic nature.

We complement theoretical findings with simulations based on Chilean SAE 2021 data. Using distance to schools as our proxy, we show that a data-driven design reduces students belonging to Pareto-improving pairs substantially and enhances efficiency in terms of increasing students in their top 3 preferences. Moreover, the simulations suggest that, in contrast to STB, this design avoids less allocation in lower ranking and unnasignment. We ascribe this result as a benefit of exploiting preference heterogeneity rather than imposing a 'dictatorial' lottery design.

Para Raúl y Roberto,
mis ejemplos a seguir.

Gracias

Agradecimientos

Qué difícil es agradecer a todos quienes fueron parte de este proceso y transmitir lo importantes que son en lo que he logrado y en la persona que soy. Fuera de estos agradecimientos, ustedes saben quiénes son y sepan que valoro enormemente todo lo vivido.

En primer lugar, agradezco a mi familia, por su apoyo incondicional, su cariño constante y por haberme brindado todas las oportunidades que hoy me permiten ser quien soy y seguir trabajando en quien quiero ser. Especialmente, gracias a Raúl, Edith, Magda y Mati. Sin duda nada de esto sería posible sin ustedes.

Agradezco a mi profesor guía Juan Escobar por instruirme en la rigurosidad de un buen investigador, manteniendo el espacio para ser creativo, para proponer ideas y para pensar. Valoro profundamente las conversaciones, el tiempo y la dedicación que invirtió en este proyecto y en mi formación en general. No sería el mismo profesional sin su apoyo.

A mis amigos, por darle color al día a día, por las risas y por el apoyo en los momentos difíciles. Estoy muy agradecido de haberlos conocido y de haber compartido todo lo que compartimos estos años. Me llena de orgullo poder considerarlos mis amigos y saber que en el futuro seguiremos encontrándonos.

Agradezco también a mis compañeros y a todos aquellos con los que compartí durante la carrera. Gran parte de lo que logré aprender y de lo que me cuestioné se lo debo a ustedes.

Agradezco a la FCFM: a los profesores, funcionarios y todos aquellos que aportan diariamente a la universidad. Admiro profundamente el trabajo que realizan y su dedicación. Particularmente, agradezco a los profesores del CEA por la paciencia y la buena disposición para escuchar y para enseñar. Gracias por haberme inspirado a escoger este camino y por contagiarme la pasión y el compromiso con el que se desenvuelven profesionalmente.

Table of Content

Tables Index

Figures Index

Chapter 1

Introduction

Since the early 2000s, many countries around the world have gravitated towards the use of centralized procedures to assign students to schools motivated by the growing influence and effectiveness of market design theory and its applications. Economists have modeled the school choice problem as a many-to-one, two-sided matching market with one particularity: schools' preferences over students (alternatively, students' priorities) are not given, but are determined by the policy-maker. This means that a priori, schools are indifferent over large groups of students and therefore the mechanism must specify the way in which ties are broken. Tie-breaking has an important effect on the efficiency of the allocation (Abdulkadiroğlu, Pathak, & Roth, 2009).

Recently, the more common tie-breaking procedures use lotteries (random priorities) to ensure ex-ante equality for students. However, Erdil and Ergin (2008) shows that the use of lotteries in stable mechanisms generates a significant efficiency loss. Theoretically, schools' preferences induce artificial stability constraints that, when applied to stable mechanisms that allow for limitless competition to ensure stability (Che & Tercieux, 2018), result in welfare losses.

In this article, we state that lottery design can be a useful tool to limit excessive competition by correlating supply and demand preferences. Intuitively, granting higher priority to a student that effectively has a high preference for that particular school reduces competition as fewer application rounds are needed to achieve stability. However, we seek to maintain strategy-proofness through the use of stable mechanisms rather than implementing the Serial Dictator or the Boston mechanism, which means the policy-maker must predict students' preferences before they submit them. Therefore, we propose the use of observable student characteristics to guide lottery design (what we call data-driven policies) as a way to improve efficiency when implementing stable mechanisms. Observable characteristics might help predict student's preferences, which allows the policy-maker to correlate preferences without conceding strategy-proofness.

We use a continuum model with a stable matching as in Azevedo and Leshno (2016) to evaluate data-driven lottery policies and compare the outcome efficiency with traditional multiple tie-breaking (MTB) and single tie-breaking (STB) procedures. We show that data-driven policies achieve higher efficiency standards compared to MTB when the observable characteristic is sufficiently informative of students' preferences but performs worse when predictions are poor. More precisely, these policies are bold as they are deterministic and naturally grant some students a lower priority when granting other students a higher one. As a consequence, there is no room for randomness to correct misplaced priorities through multiple applications. However, when observable characteristics provide precise information, the result is highly efficient.

Additionally, we uphold our theoretical findings with Chilean admissions data by using students' distance to schools as a predictor of preferences, motivated by Aramayo (2018) results. The simulations show a significant decrease in the number of Pareto-improving pairs and an improvement in the accumulated rank distribution: we obtain stochastic dominance for the first three ranks and barely fewer allocations in lower rankings. Unlike STB procedures, data-driven policies exploit preference heterogeneity, which seems to avoid fewer allocations in lower rankings and unassignment.

In conclusion, observable student characteristics enable a more tailor-made lottery design as it allows the policy-maker to predict students' preferences and therefore generate preference correlation, which reduces the inefficiency inherent to stable mechanisms whilst maintaining strategy-proofness. Correlating preferences enhances efficiency by making use of available preference heterogeneity rather than imposing a more 'dictatorial' procedure. Empirical work shows this seems to avoid unassignment.

Chapter 2

Literature Review

The school choice problem lies within a broader area of economic theory known as matching markets, which is embedded within the market design literature. Matching theory studies markets in which prices are not used as a mechanism to assign objects, as it is a common agreement that either it is unethical to do so, or because it is ineffective. Some of the more popular examples within the literature include the assignment of doctors to hospitals (Roth & Peranson, 1999), organ transplant allocation (Roth, Sönmez, & Ünver, 2004) and school choice.

The first important contribution to matching markets theory was made by Gale and Shapley (1962) who study the famous marriage problem and design the Deferred Acceptance (DA) algorithm to solve it. Although this algorithm is widely used today in multiple market scenarios and specifically in school choice problems, practical applications acquire greater relevance in the 90s. Since then, great efforts have been made toward the design of allocation algorithms (i.e. a solution to the matching problem) in a variety of different contexts.

Generally, there are three main desirable properties when designing an allocation mechanism: *efficiency* (in the sense of Pareto), *stability* (justified envy), and *strategy-proofness* (incentive-compatible mechanisms). Depending on the problem formulation, these properties tend to clash and therefore present a trade-off for the policy-maker to solve. For instance, Che and Tercieux (2018) describe the trade-off between efficiency and stability in large two-sided markets.

Particularly in school choice, Abdulkadiroğlu and Sönmez (2003) propose this framework to tackle the student allocation problem in Boston schools, specifying this kind of problem as a many-to-one, two-sided matching market. During this period, most of the literature evolved around the ongoing debate to decide which mechanism to implement: the Boston-Mechanism (BM) or the Deferred Acceptance algorithm. The former ensures an efficient allocation, but is not stable or strategy-proof, while the latter ensures stability and strategy-proofness but is not efficient. These authors present this trade-off both theoretically and empirically using

Boston City school data.

Economists have been able to state some general properties about these desiderata and the trade-offs mentioned. For instance, Roth shows that the DA is stable, strategy-proof, and is not Pareto-dominated by any other stable mechanism (Roth, 1982). Pathak and Sönmez (2008) describe the effect of manipulable (not strategy-proof) mechanisms in terms of ex-post efficiency. On the other hand, the Top Trading Cycles (TTC) algorithm is efficient and strategy-proof but not stable (Shapley & Scarf, 1974), although it minimizes instability.

However, the school choice problem is not completely captured by the two-sided matching markets framework as it holds one particularity: student's priorities (or schools' preferences over students) are not given but must be designed by the policy-maker. A priori, schools are indifferent over large groups of students and therefore the mechanism must specify a tie-breaking rule to produce the necessary input for the algorithms mentioned above.

In most of the student allocation procedures, tie-breaking is solved by using lotteries, that is, randomly assigning each student a score which then defines their priorities for each school. Lotteries are used for tie-breaking as a way to attain ex-ante student equality (Abdulkadiroglu et al., 2006). This argument played an important role in the transition towards strategy-proof and stable mechanisms such as the DA using random allocation. However, Erdil and Ergin (2008) and Abdulkadiroğlu et al. (2009) show both theoretically and empirically that random, strategy-proof, stable mechanisms are necessarily inefficient and therefore are subject to a significant welfare loss.

With this in mind, lottery design became a relevant topic in school choice. For instance, Abdulkadiroğlu et al. (2009) compare single tie-breaking (each student uses the same lottery ticket for all schools) versus multiple tie-breaking (students obtain a different lottery ticket for each school) and show it has a large effect on ex-post efficiency. Single tie-breaking leaves more students in their first choice but also leaves more students unassigned. In fact, the difference between the raw definitions of stability and efficiency lies in school preferences, as lotteries induce artificial stability constraints that reduce welfare (Erdil & Ergin, 2008).

Other mechanisms have been designed in order to achieve higher efficiency without compromising stability or strategy-proofness (or at least compromising it minimally). Abdulkadiroğlu, Che, and Yasuda (2015) state that the DA algorithm limits students' ability to communicate their preference intensities, which brings ex-ante inefficiency when ties are broken randomly. They design the 'Choice-Augmented' Deferred Acceptance (CADA) algorithm in which students are able to submit their usual preference rank and additionally the name of a 'target' school. Although targeting clearly induces strategic behavior, the algorithm is still stable and strategy-proof with respect to students' rankings and improves efficiency.

With the same objective in mind, the literature has studied lottery design itself, evaluating its effect on allocation efficiency. For example, Çelebi and Flynn (2021) study coarsenings in priority design in both deterministic and stochastic scenarios and show that, when students' preferences are known, only 3 lotteries suffice to obtain any allocation. Echenique et al. (2020) study the effect of previous interactions between the supply and the demand that restructure preferences and enhance overall efficiency in the actual matching process.

This work studies the possibility of using lottery design as a way of limiting excess competition in a stable mechanism and therefore enhancing efficiency. More precisely, we propose using observable student characteristics that grant valuable information of students' preferences and can therefore generate correlation without compromising strategy-proofness.

Chapter 3

Theoretical Analysis

The main objective of our work is to study the effects of applying different lotteries in the school choice problem. For us to achieve this, we take two approaches for our analysis: a theoretical model and empirical work. This chapter describes the former approach and the theoretical results achieved.

3.1. Notation and Definitions

We consider a continuous matching model as in Azevedo and Leshno (2016). In this setting, there is a finite set $C = \{1, 2, 3, \ldots, n\}$ schools to be matched to a continuum mass of students with total mass 1. A student $i \in I$ is defined by the pair $i = (\succ^i, e^i)$ where $\succ^i$ is a strict ordering over schools $c \in C$ and $e^i \in [0,1]^C$ describes school's ordinal preferences for student i. We refer to e^i_c as student i's score in school c. Schools prefer students with higher scores, that is, school c prefers student i over student i' if $e^i_c > e^{i'}_c$. For simplicity, we will assume that all students and schools are acceptable. Let $\mathcal{R}$ be the set of all strict preference ordering over C, we denote $I = \mathcal{R} \times [0,1]^C$ the set of all student types.

A continuum economy is given by $E = [\eta, S]$, where η is a probability measure over I and $S = (S_1, S_2, \ldots, S_n)$ is a vector of strictly positive capacities for each school. We will assume that every school's indifference curve has measure 0, that is, $\forall c \in C$ and $\forall x \in \mathbb{R}$ we have $\eta(\{i : e^i_c = x\}) = 0$. This is equivalent to imposing that schools have strict preferences over students in the discrete model.

In this context, a matching μ is an allocation of students to schools. Formally, a matching is a function $\mu : C \cup I \to 2^I \cup (C \cup I)$ such that:

1. $\forall i \in I$, $\mu(i) \in C \cup \{i\}$. Each student is matched either to a school or to itself, which represents being unmatched.

2. $\forall c \in C$, $\mu(c) \subseteq I$ is measurable and $\eta(\mu(c)) \leq S_c$. Schools are matched to a set of

students with a mass that doesn't exceed the school's capacity.

3. $c = \mu(i)$ iff $i \in \mu(c)$. A matching is consistent in the sense that a student is matched to a school if and only if the school is matched to that student.

4. For any $c \in C$, the set $\{i \in I : \mu(i) \preceq^i c\}$ is open. This is a regularity condition to avoid adding sets of students with measure 0 to a school because it would generate multiple matchings that only differ in sets of measure 0.

A student-school pair (i, c) *blocks* a matching μ in the economy i if the student prefers school over his match ($c \succ^i \mu(i)$) and either (i) school c has spare capacity $(\eta(\mu(c))) < S_c$ or (ii) school c is matched to another student who has a lower score than i ($\exists i' \in \mu(c)$ such that $e^i_c > e^{i'}_c$).

Definition 3.1 (**Stable Matching**) *A matching μ for a continuum economy E is stable if it is not blocked by any student-school pair.*

A stable matching always exists (Azevedo & Leshno, 2016), similar to the discrete case as shown by Gale and Shapley (1962). The notion of stability is commonly interpreted as 'justifiable envy', as student i and school c prefer each other over their match. Additionally, a stable matching does not create incentives for outside agreements as each student (or school) strictly prefers their match over any other school (or student) that prefers him.

The existence of a stable matching in a discrete economy is proven by the use of the Deferred Acceptance (DA) algorithm (Gale & Shapley, 1962). This algorithm is naturally applied in a discrete scenario that is defined similarly. In this case, there is a finite set of students $i \in I = \{1, 2, 3. \dots, I\}$ with a strict ordering over schools $\succ^i$. Schools are defined similarly with a strict ordering $\succ^c$ over I. Therefore, a matching problem $\mathcal{M}$ is defined by $\mathcal{M} = (C, I, \succ^c, \succ^i, S)$. The algorithm is as follows.

As stated before, we will be specifically interested in studying the efficiency of the output allocation. Generally, the literature uses Pareto-efficiency to study matching outcomes. A matching μ Pareto-dominates another matching μ' in a discrete economy if for all students $i \in I$, $\mu(i) \succeq^i \mu'(i)$ and $\mu(i) \succ^i \mu'(i)$ for at least one student.

Similarly, in a continuous setting, a matching μ Pareto-dominates another matching μ' if $\forall i \in I : \mu(i) \succeq^i \mu'(i)$ and there exists a mass of students with positive measure $\hat{I} \subseteq I : \eta(\hat{I}) > 0$ such that $\forall i \in \hat{I} : \mu(i) \succ^i \mu'(i)$.

Definition 3.2 (**Efficient Matching**) *A matching is efficient if it is not Pareto-dominated by any other matching.*

As it is common in the literature, we are interested in measuring efficiency (or rather, inefficiency) rather than simply stating it. With this in mind, we shall consider two main approaches widely used in the literature to measure inefficiency.

Algorithm 1 Deferred Acceptance Algorithm (students propose)

Require: A discrete economy $\mathcal{M} = (C, I, \succ^c, \succ^i, S)$.

Step 1: Each student applies to his first choice. Each school tentatively assigns its seats to its proposers one at a time following their priority order until capacity is reached. Any remaining applying students are rejected.

In general, at...

Step k: Each student who was rejected in the previous step applies to his next choice, if one remains. Each school considers the set consisting of its previously tentatively assigned students and the new applicants and assigns its seats to these students following their priority order until capacity is reached. Any student remaining after all the seats are assigned is rejected.

The algorithm terminates when no student is rejected at step k. All tentative assignments are then materialized.

1. Firstly, the number of Pareto improving pairs, which is simply defined as a student pair (i_1, i_2) such that $\mu(i_2) \succ^{i_1} \mu(i_1)$ and $\mu(i_1) \succ^{i_2} \mu(i_2)$, this is, each student prefers the other student's match (Ashlagi & Nikzad, 2020).

2. Secondly, the cumulative rank distribution. Formally, a cumulative rank distribution is $F(k) = \sum_{q=1}^{k} R(q)$, where $R(q)$ is the mass of students assigned to its q-th favorite school. Under this measure, a matching μ is more efficient than another matching μ' if there is stochastic dominance between the cumulative rank distributions, i.e. $F_{\mu}(k) \geq F_{\mu'}(k)$ $\forall k \in \{1, 2, 3, \ldots, n\}$.

3.2. The Model

We study the value of a limited amount of information over students' preferences to guide lottery design. More specifically, our idea consists in using an observable student characteristic (for example, distance from schools) as a predictor of what the student's ranking will look like. Theoretically, a standard lottery procedure assumes the policy maker has no information over student preferences, which leads to efficiency loss. We seek to exploit any degree of information that will allow smarter lottery design in the sense of improving efficiency. Intuitively, correlating students' and schools' preferences will control excessive competition and therefore reduce efficiency loss by limiting the necessary application rounds to achieve stability. We think of higher correlation as granting high priority to a student that likewise grants that school a high ranking.

The school allocation procedure timeline is as follows.

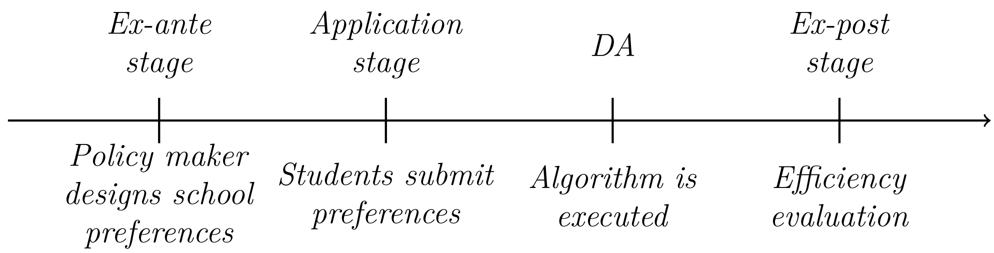

Figure 3.1: Model Timeline

One would argue that we can obtain an efficient outcome simply by designing school preferences once students submit their own by correlating them perfectly. If we would generate each school's preference list by ranking each student according to the position they ranked that school (i.e. first select all students that select that school as their top preference, then students that rank that school second, etc.) this would indeed be efficient. However, this mechanism would not be strategy-proof, as students will be conscious of the effect of each school's ranking position. This is why the policy-maker needs to predict students' rankings before they are submitted.

3.2.1. General Setting

Our model considers a continuous market with a set of two schools $C = \{C_0, C_1\}$ and a continuous mass of students uniformly distributed in $[0, 1]$ with total mass 1. Schools have capacities S_0, S_1 which we will assume are restrictive, i.e. $S_0 + S_1 < 1$. The preference each student has over schools is captured by the following utility function.

$$u_c^i = -d_c^i + \epsilon_c^i \tag{3.1}$$

where d_c^i is the distance between student i and school c, and ϵ_i^c is unobservable and therefore random from the policy-maker's point of view. This utility function structure expresses that the policy-maker has some information over student preferences through the proxy d_c^i, but information is not perfect as ϵ_c^i is unobservable.

In our setting, school C_0 is located at 0 and school C_1 is located at 1. Under these specifications, student utility functions are as follows.

$$u_{C_0}^i = -i + \epsilon_{C_0}^i \qquad u_{C_1}^i = -(1-i) + \epsilon_{C_1}^i$$

We shall state that $\epsilon_{C_1}^i - \epsilon_{C_0}^i \overset{\text{i.i.d}}{\sim} F$, a continuous distribution with mean $\mu_F = 0$. By setting $\mu_F = 0$ we are assuming that there is no idiosyncratic preference for one school. Consequently, the probability that student i prefers school C_0 over school C_1, and vice-versa, is given by

$$\mathbb{P}(u^i_{C_0} > u^i_{C_1}) = \mathbb{P}(-i + \epsilon^i_{C_0} > -(1-i) + \epsilon^i_{C_1}) = \mathbb{P}(1 - 2i > \epsilon^i_{C_1} - \epsilon^i_{C_0}) = F(1-2i)$$
$$\mathbb{P}(u^i_{C_0} < u^i_{C_1}) = 1 - F(1-2i)$$

Visually, the model is as follows.

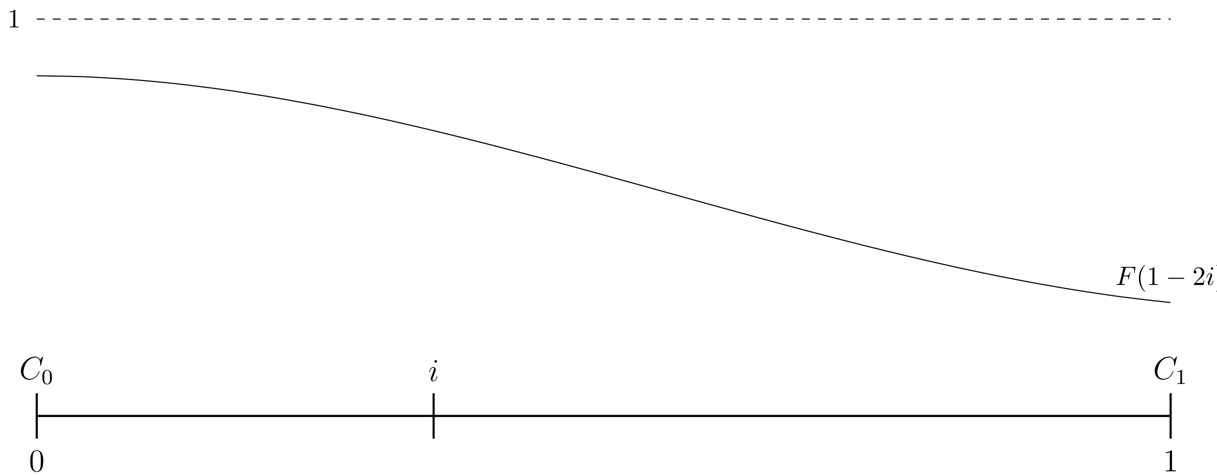

Figure 3.2: Model Visual

In this context, we seek to evaluate different lottery policies in terms of the outcome efficiency.

Definition 3.3 (**Lottery Policy**) *A lottery policy is a family of functions* $\{G^i_c\}^{i \in I}_{c \in C}$ *such that each function*

$$\begin{aligned} G^i_c : [0,1] &\longrightarrow [0,1] \\ p &\longmapsto G^i_c(p) = \mathbb{P}(i \textit{ is rejected at } c \textit{ with equilibrium cutoff } p) \end{aligned} \tag{3.2}$$

assigns each student $i \in [0,1]$ *a probability distribution through which the student obtains his score* e^i_c *at school c. More simply,* $e^i_c \sim G^i_c$.

The timing of the model is the following:

1. The policy-maker observes all of the student's characteristics d^i_c and knows school's capacities S_0, S_1. The policy-maker does not observe students' preferences over schools but has a belief of each student's ranking according to the distribution $F(1-2i)$.
2. The policy-maker designs a lottery policy $\{G^i_c\}^{i \in I}_{c \in C}$.
3. Student's preferences over schools ($\succ^i$) are revealed.

4. Using student's preferences $\succ^i$ and lottery policy $\{G_c^i\}_{c\in C}^{i\in I}$, a stable matching μ is obtained.

In order to understand alterations in the model and compute comparative statistics, we shall use the supply and demand framework proposed by Azevedo and Leshno (2016) when studying large markets. They state an equivalence between a matching μ and a cutoff score that solves a set of market-clearing equations.

A cutoff is a minimal score $P_c \in [0,1]$ required for admission at school c. A student i can afford school c (or school c would accept i) if $P_c \leq e_c^i$. A student's demand given a vector of cutoffs $D^i(P)$ is his favorite school among all those he can afford. In this way, we define aggregate demand for school c as the mass of students who demand it.

$$D_c(p) = \eta(\{i : D^i(P) = c\})$$

Definition 3.4 (**Market Clearing Cutoff**) *A vector of cutoffs P is a market clearing cutoff if it satisfies the following market clearing equations.*

$$D_c(P) \leq S_c \qquad \forall c \in C$$

and

$$D_c(P) = S_c \qquad \forall c \in C : P_c > 0 \tag{3.3}$$

Hence, there is a natural one-to-one correspondence between stable matchings and market clearing cutoffs. This correspondence is described by the following operators.

- Given a market clearing cutoff P, define the associated matching $\mu = \mathcal{M}P$ with the demand function $\mu(i) = D^i(P)$.
- Conversely, given a stable matching μ, define the associated cutoff $P = \mathcal{P}\mu$ by the score of students matched to each school $P_c = \inf_{i\in\mu(c)} e_c^i$.

The authors show that if μ is a market clearing, then $\mathcal{P}\mu$ is a market clearing cutoff. If P is a market clearing cutoff, then $\mathcal{M}P$ is a stable matching ($\mathcal{P}$ and $\mathcal{M}$ are inverses of each other). Moreover, they state that under a set of regularity conditions, a sequence of discrete economies converges to a continuum economy as the market size grows, which then again supports the use of a continuum setting to model a real-case scenario of school choice.

Using this framework, we seek to compare policies in terms of efficiency and ultimately characterize optimal policies for this scenario. We shall measure efficiency using the accumulated rank distribution approximation. In a market with two schools ($|C| = 2$), this is equivalent to comparing the mass of students assigned to their top preference. This also ensures stochastic dominance of the cumulative rank distributions as the mass of unassigned students remains constant.

3.2.2. Optimal Policy

In this setting, we search for an optimal lottery policy. For simplicity, let us assume there exists a vector of market clearing cutoffs (p_0, p_1) such that the stable matching exists. Then, let us define $X^i = G^i_{C_0}(p_0)$ and $Y^i = G^i_{C_1}(p_1)$ the rejection probability for student i at schools C_0, C_1 respectively. The optimal policy problem is defined as follows.

$$\begin{aligned} \max_{X^i,Y^i} \quad & \int_0^1 F(1-2i)(1-X^i)di + \int_0^1 (1-F(1-2i))(1-Y^i)di \\ \text{s.t.} \quad & \int_0^1 F(1-2i)(1-X^i)di + \int_0^1 (1-F(1-2i))Y^i(1-X^i)di = S_0 \\ & \int_0^1 F(1-2i)X^i(1-Y^i)di + \int_0^1 (1-F(1-2i))(1-Y^i)di = S_1 \end{aligned} \tag{3.4}$$

The objective function measures the mass of students assigned to their first preference in each school. The restrictions impose that the allocation must meet the market clearing conditions (supply equals demand). Strictly speaking, the total mass of students assigned to each school must be less than or equal to the school's capacity, but we assume that capacities are restrictive ($S_0 + S_1 < 1$) and so the restrictions are always active in the optimum. Because of this, we can rewrite the following unrestricted problem.

$$\max_{X^i,Y^i} \quad S_0 + S_1 - \int_0^1 (1-F(1-2i))Y^i(1-X^i)di - \int_0^1 F(1-2i)X^i(1-Y^i)di$$

We try to minimize students assigned to their second choice (thereby maximizing those assigned to their first choice). With this formulation, we can characterize the optimal policy.

Proposition 3.1 (**Optimal Policy**) *The optimal lottery policy is* $X^i = Y^i \in \{0,1\}\ \forall i \in [0,1]$. *In particular, the optimum is such that*

$$X^i = Y^i = \begin{cases} 0 & \forall i \in \bar{I} \\ 1 & \forall i \notin \bar{I} \end{cases}$$

where the set $\bar{I} \subseteq I$ *of students must be feasible, that is*

$$\int_{\bar{I}} F(1-2i)di = S_0 \quad \text{and} \quad \int_{\bar{I}} (1-F(1-2i))di = S_1$$

Intuitively, the main source of inefficiency is rejection through competition: students rejected by their first choice apply to their next choice and might get accepted. By setting

$X^i = Y^i \in \{0,1\}$ we eradicate this kind of behavior.

Naturally, this lottery policy mimics the Serial Dictator (SD) mechanism: pick a mass of students and allow them to choose their school freely. This mechanism is clearly Pareto-efficient as no selected students are ever rejected. Particularly, the optimal value is obtained, which means a mass of $S_0 + S_1$ students is assigned to their top preference. Although the result might strike as straightforward, the information over student preferences modeled by F allows us to select the set $\bar{I}$ such that, in expectation, actually no student chooses its second preference in the SD-type procedure. This means that the outcome is not only Pareto-efficient, but it maximizes the mass of students assigned to their top preference.

However, the set $\bar{I}$ need not always exist. In fact, for large noises (take $\delta(\epsilon^i_{C_1} - \epsilon^i_{C_0})$ with large $\delta > 0$), if $S_0 \neq S_1$, then the set $\bar{I}$ does not exist. For instance, when $\delta \to \infty$, then $F(1-2i) = 1/2 \ \forall i$ and the market clearing conditions $\frac{1}{2}\int_{\bar{I}} di = S_0$ and $\frac{1}{2}\int_{\bar{I}} di = S_1$ become infeasible.

For example, consider a discrete scenario with two schools C_0, C_1 with capacities $S_0 = 1$ and $S_1 = 2$ and a total of six students where three prefer C_0 over C_1 (group 1) and the other three prefer C_1 over C_0 (group 2). We would like to select $\bar{I}$ such that $|\bar{I}| = S_0 + S_1 = 3$. If we select $\bar{I}$ as one student from group 1 and two from group 2, the SD-type procedure assigns all three students to their top preference. However, if we select $\bar{I}$ as two students from group 1 and one from group 2, necessarily one student from group 1 will be matched to its second choice.

We have shown there exists an optimal lottery policy under certain specifications, but these lotteries not always are feasible. Additionally, this result escapes a natural tie-breaking framework as schools' preferences are dichotomous and therefore impose an allocation rather than actually breaking ties.

3.2.3. Data-driven Policy

We intend to study lotteries driven by observable student characteristics in a stable mechanism setting. As mentioned before, observable characteristics can be informative of students' preferences before they submit them. This can be exploited to generate preference correlation in a strategy-proof manner.

Definition 3.5 (**Data-driven Policy**) *A lottery policy $\{G^i_c\}^{i\in I}_{c\in C}$ is data-driven if G^c_i are degenerate distributions such that*

$$G(x|i,c) = \begin{cases} 1 & x \geq 1 - d^i_c \\ 0 & x < 1 - d^i_c \end{cases}$$

i.e. the score assigned to student i at school c is deterministic $e^i_c = 1 - d^i_c$.

Hence, the market clearing conditions are the following.

$$\int_0^{p_0} [F(1-2i) + (1 - F(1-2i)) \cdot \mathbb{1}_{i<1-p_1}]di = S_0$$
$$\int_{1-p_1}^{1} [(1 - F(1-2i)) + F(1-2i) \cdot \mathbb{1}_{i>p_0}]di = S_1$$

Or equivalently,

$$\min\{p_0, 1-p_1\} + (\mathbb{1}_{1-p_0<p_1}) \int_{1-p_0}^{p_1} F(1-2i)di = S_0$$
$$(1 - \max\{p_0, 1-p_1\}) + (\mathbb{1}_{1-p_1<p_0}) \int_{1-p_1}^{p_0} (1 - F(1-2i))di = S_1$$

We search for cutoffs p_0, p_1 that satisfy both conditions. However, as we have assumed that $S_0 + S_1 < 1$, this implies that $p_0 < 1 - p_1$, and therefore this system of equations is far simpler. The market clearing equations reflect what we mean by bold lotteries. As scores are deterministic, the policy-maker makes some sort of gamble through the proxy as it grants some students high scores (and naturally others a low score) without full knowledge of their preferences. This naturally induces risk as a consequence of the lack of information.

Definition 3.6 ***(Data-driven Equilibrium)*** *With a data-driven lottery design, the equilibrium is given by $p_0 = 1 - S_0, p_1 = 1 - S_1$, which assigns a total mass of*

$$R_{DB}(1) = \int_0^{S_0} F(1-2i)di + \int_{1-S_1}^{1} (1 - F(1-2i))di$$

students to their top preference.

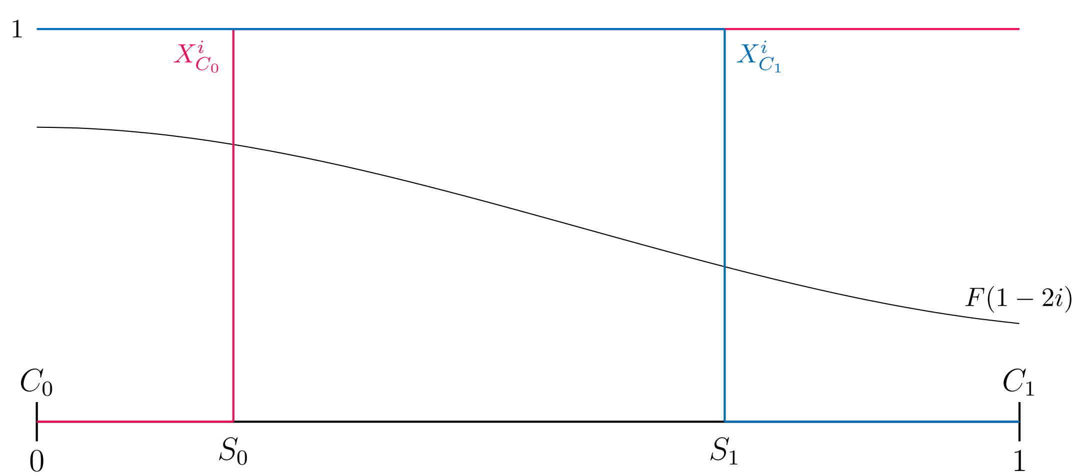

Figure 3.3: Data-driven Equilibrium

The efficiency of the outcome when implementing data-driven policies is determined by F.

When information is perfect (i.e. $\delta(\epsilon^i_{C_1} - \epsilon^i_{C_0})$ with $\delta = 0$) then $F(1-2i) = 1 \ \forall i < 1/2$ and $F(1-2i) = 0 \ \forall i > 1/2$, so the policy-maker is able to predict student's preferences perfectly. In this case, the data-driven policy is efficient as it assigns all students to their first preference ($R_{DB}(1) = S_0 + S_1$). In this case, there is no risk in deterministic scores as the policy-maker has full knowledge of students' preferences through the proxy.

On the other hand, when the policy-maker has no information (i.e. $\delta(\epsilon^i_{C_1} - \epsilon^i_{C_0})$ with $\delta \to \infty$) then $F(1-2i) = 1/2 \ \forall i \in [0,1]$ and the policy performs poorly, assigning $R_{DB}(1) = \frac{S_0+S_1}{2}$ students to their first preference.

3.2.4. Standard Lotteries

With the proposed framework in mind, we would also like to evaluate the more commonly used tie-breaking procedures. We shall study two of the more commonly implemented policies: multiple tie-breaking (MTB) and single tie-breaking (STB). As mentioned above, these policies are random as a way to ensure ex-ante student equality.

Definition 3.7 ***(Multiple Tie-breaking)*** *A multiple tie-breaking lottery is defined by fixing $G^i_c = G_c$, a common distribution $\forall i \in [0,1]$. This is, for each school c, all students have the same probability of being rejected/accepted.*

Through this definition, we can evaluate the efficiency of a MTB procedure with student information. Similar to the optimal policy, for a given cutoff vector (p_0, p_1), we define $X = G_{C_0}(p_0)$ and $Y = G_{C_1}(p_1)$ constant values that reflect the probability of any student being rejected at school C_0 and C_1 respectively. A priori, these probabilities depend on the relative demand for each school, i.e. school's capacities as well as the expected mass of students that prefer each school. Then, the market clearing conditions are the following.

$$\int_0^1 F(1-2i)(1-X)di + \int_0^1 (1-F(1-2i))Y(1-X)di = S_0$$
$$\int_0^1 F(1-2i)X(1-Y)di + \int_0^1 (1-F(1-2i))(1-Y)di = S_1$$

We can solve this 2×2 system of equations and find X, Y.

Definition 3.8 ***(MTB Equilibrium)*** *In a MTB setting, the equilibrium is given by*

$$X = \frac{(\bar{F} - (1-S)(1-\bar{F}) - S_0)}{2\bar{F}} + \frac{\sqrt{((1-S)(1-\bar{F}) + \bar{F} - S_0)^2 + 4\bar{F}(1-S)(1-\bar{F})}}{2\bar{F}}$$

$$Y = \frac{((1-\bar{F}) - S_1 - (1-\bar{F})(1-S)))}{2(1-\bar{F})} + \frac{\sqrt{(-(1-\bar{F}) + S_1 + (1-\bar{F})(1-S))^2 + 4\bar{F}(1-S)(1-\bar{F})}}{2(1-\bar{F})}$$

with $\bar{F} = \int_0^1 F(1-2i)di$ *and* $S = S_0 + S_1$*. This design assigns a total mass of students in their top preference given by*

$$R_{MTB}(1) = \bar{F}(1-X) + (1-\bar{F})(1-Y)$$

The analytical result of the MTB setting reveals there exists an irretrievable efficiency loss when $\bar{F} < S_0$ and $1 - \bar{F} < S_1$, expressed in the equilibrium terms. When the expected mass of students that prefers one school is larger than that school's capacity, there will inevitably exist a mass of students who cannot be assigned to their top preference no matter the lottery policy. Therefore, we can only aim to improve efficiency within a mass of $S_0 + S_1$ students.

Henceforth, we shall assume that schools' capacities are restrictive not only market-wise ($S_0 + S_1 < 1$) but also school-wise ($\bar{F} > S_0$ and $1 - \bar{F} > S_1$). This means we restrict our analysis to the mass of students that can effectively improve efficiency.

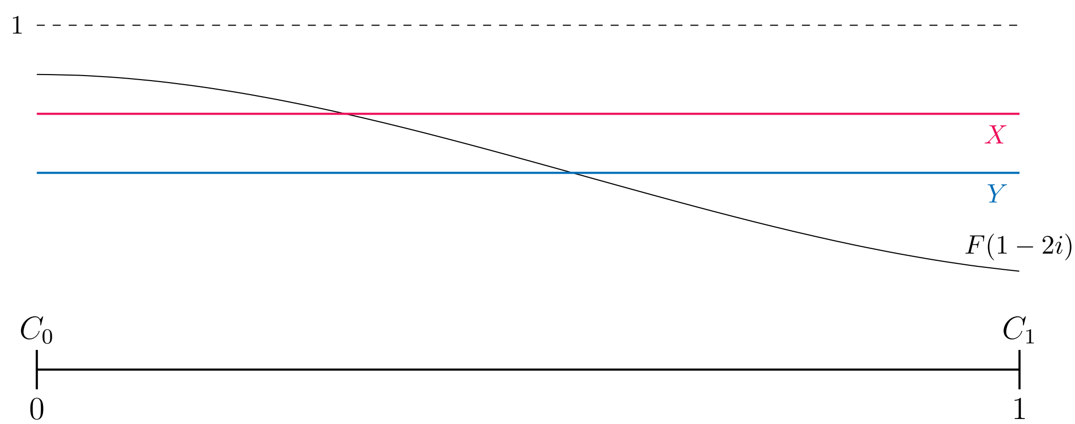

Figure 3.4: Multiple Tie-breaking Equilibrium

Similarly, we can characterize the STB lottery policy in this setting.

Definition 3.9 ***(Single Tie-breaking)*** *A single tie-breaking lottery is defined by fixing* $G_c^i = G$ *a common lottery* $\forall i \in [0,1]$ *and* $\forall c \in C$*. This is, all students obtain the same score for all schools drawn from a common distribution* G*.*

Once again, for a given cutoff vector (p_0, p_1), we define $X = G(p_0), Y = G(p_1)$ constant values. A priori, although the distribution G is the same for both schools and all students, the probability of rejection at each school is not necessarily the same, as it depends on the capacities and student demands. Without loss of generality, let us assume that $\frac{S_0}{\bar{F}} \leq \frac{S_1}{1-\bar{F}}$ so that $p_0 \geq p_1$. Then, the market-clearing equations are the following

$$\int_0^1 F(1-2i)(1-X)di = S_0$$

$$\int_0^1 F(1-2i)(X-Y)di + \int_0^1 (1-F(1-2i))(1-Y)di = S_1$$

We can solve this 2×2 system to find X, Y.

Definition 3.10 ***(STB Equilibrium)*** *In an STB setting, the equilibrium is given by*

$$X = 1 - \frac{S_0}{\bar{F}} \quad , \quad Y = 1 - S_0 - S_1$$

This policy assigns a total mass of students to their top preference given by

$$R_{STB}(1) = S_0 + (1-\bar{F})\left(S_0 + S_1 - \frac{S_0}{\bar{F}}\right)$$

We interpret single tie-breaking as a more 'dictatorial' random policy design. It aims to keep randomization with a similar ex-ante equality goal, but it naturally selects a high-scoring group of students ex-post which is the same for all schools. Consequently, this procedure assigns more schools to their top choice than MTB. Specifically, a highly-demanded school (in this case C_0) is matched only to students which rank that school as their top choice. As we have assumed that capacities are restrictive school-wise and $S_0 < S_1$, this policy induces a kind of cherry-picking for school C_0. However, relatively low-demanded schools (in this case C_1) face applications from both students that have that school as their top choice as well as rejected students from C_0.

3.2.5. Comparing Policies

We would like to provide sufficient conditions to ensure a data-driven policy performs better (or worse) than the MTB policy. Let us assume that $S_0 \neq S_1$. As mentioned before, the key to the policy comparison lies in the magnitude of the error terms, which measures the amount of information the proxy provides over students' preferences.

Proposition 3.2 ***(Data-driven Policy vs MTB)*** *Let us redefine the error term as* $\delta(\epsilon^i_{C_1} - \epsilon^i_{C_0}) \sim F$ *with* $\delta > 0$ *and note* $R^\delta_{DD}(1), R^\delta_{MTB}(1)$ *the mass of students assigned to their top preference with error size* δ *when using data-driven and MTB lottery policies respectively.*

If $S_0 \leq \bar{F}$ and $S_1 \leq 1 - \bar{F}$, then, there exists $\underline{\delta}, \bar{\delta}$ such that:

1. $\forall \delta < \underline{\delta}$: $R^{\delta}_{DD}(1) > R^{\delta}_{MTB}(1)$
2. $\forall \delta > \bar{\delta}$: $R^{\delta}_{DD}(1) < R^{\delta}_{MTB}(1)$

This result shows that an increase in information precision (a decrease in noise) will improve the efficiency of the allocation using a data-driven design. However, if the noise size is too large, the data-driven policy harms efficiency and performs worse than MTB. Then again, this is because data-driven policies are bold in the sense that they are deterministic. When information is scarce, competition in random lottery scenarios can fix the lack of correlation between preferences by multiple application rounds.

For example, take $S_0 = 0.2, S_1 = 0.4$ and $F \sim N(0, \sigma^2)$. Figure 3.5 shows the effect of the variance of F (σ^2) over the number of students assigned to their top preference for each policy through simulations.

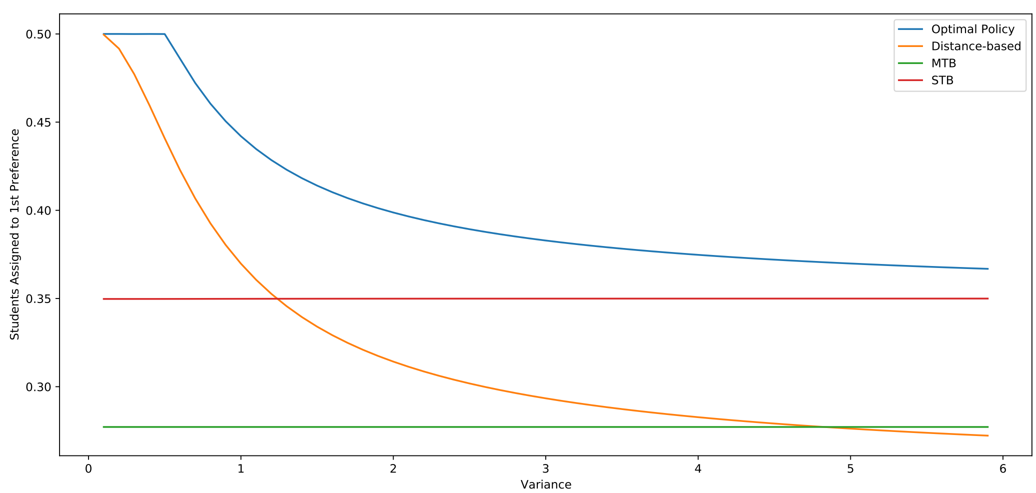

Figure 3.5: Policy Comparison Simulations

As mentioned above, the optimal policy performs efficiently for small errors, assigning $S_0 + S_1$ students to their top preference, but then decreases as the set $\bar{I}$ ceases to exist. The data-driven policy is also efficient for small error magnitudes (a high information setting) but then decreases naturally when less information is available and the gamble is less precise. As stated beforehand, when the error term is too large, the data-driven procedure performs worse than the MTB and STB policies. Naturally, the STB policy always assigns more students to their top preference than the MTB policy.

3.3. Discussion

As Erdil and Ergin (2008) state, random stable mechanisms (such as the DA with randomly generated preferences) are wasteful because randomly generated stability constraints produce efficiency loss. We can think of stable mechanisms as Che and Tercieux (2018) propose: let k be the limit amount of applications each student can perform, then k can be thought of as a measure of competition in matching markets. Then, the Boston Mechanism ($k = 0$) represents minimum competition, and the Deferred Acceptance Mechanism ($k \to \infty$) represents maximum competition. In this sense, excessive competition guarantees stability but damages efficiency. Therefore, one would desire limiting competition but ideally maintaining stability and strategy-proofness. This is where lottery design might be helpful.

One way of enhancing efficiency is the STB lottery design. This policy certainly limits competition but does so in a dictatorial manner by randomly selecting a group of privileged students. This naturally improves top-rankings but has a cost in lower ranking positions and unassigned students.

Our proposition lies in exploiting heterogeneity of students' preferences to enhance efficiency avoiding these costs. Preference heterogeneity would allow the policy-maker to reduce competition without selecting privileged students but by simply granting high priority to students that rank that school highly, i.e. correlating student and school preferences.

Designing schools' preferences based on students' preferences to obtain a high correlation seems natural. The problem is this mechanism would not be strategy-proof, so the timing of the model forces the policy-maker to predict student preferences. This is where student characteristics become handy. We have shown a good proxy of student preferences enables smarter lottery design to obtain high correlation and improve efficiency. If the proxy is verifiable, then the DA is strategy-proof. Similar to the results shown by Echenique et al. (2020), correlating supply and demand preferences leads to increased efficiency in stable mechanisms by limiting excess competition, in their case through a previous interview process. However, in school choice, the particularity of tie-breaking grants an extra degree of freedom in the design of the mechanism that allows for enhanced efficiency without compromising the veracity of student's preferences.

However, we have shown that using information to guide lottery design might be risky. The data-driven policies effectively improve efficiency only when (1) the proxy is a powerful predictor which leads to high correlation, and (2) the market shows sufficient heterogeneity in students' preferences. If this is not the case, these policies perform worse than MTB. This is because data-driven policies are bold as they not only grant some students higher priority but because they grant the rest a lower one in a deterministic manner. This is beneficial when student preference prediction is precise, but it becomes harmful when it is not. With poor predictions, competition is in fact efficient as it corrects poor predictions through multiple applications.

Chapter 4

Application - School Choice in Chile

As mentioned before, we shall contrast our theoretical findings with simulated results using school choice data recovered from the chilean SAE 2021 process. This chapter provides a brief summary of the chilean system as well as a general overview of the data and the results obtained through our simulations.

4.1. School Choice in Chile - *Sistema de Admisión Escolar*

Since 2016, the Chilean Government, as well as many other countries or cities around the world (Boston, New York, Amsterdam, Seúl, Ghana, England and others) adopted a centralized procedure to assign students to schools. In May 2015, as part of a major educational reform, a student inclusion bill was passed. In this bill, one of the main aspects included the prohibition of student selection in schools. This led to the foundation of the *Sistema de Asignación Escolar* (SAE), an institution dependent of the Educational Ministry (MINSAL) which is entrusted to assign students to schools. This reform meant great change as the chilean educational system migrated from a traditional decentralized structure towards a centralized school assignment.

Every year, the SAE team has to solve the following task: generate a matching of student to schools considering student preferences and school capacities while meeting a series of restraints specific to the chilean context. Law 20.835 carefully dictates the rules that define the admission process (Biblioteca del Congreso Nacional de Chile, 2015).

In general, the process is based on two main steps: the main application and assignment round, and the complementary round. The former phase presents itself as the more relevant and complex step. In the main round, all elegible school programs are offered to all students, which generates a large allocation problem. Through the SAE webpage, students (truthfully parents or guardians) submit a ranking over desired schools. This serves as the main input to generate the allocation procedure, which is done using the Deferred Acceptance Algorithm. The algorithm generates a student-school matching that assigns each student at most one slot in one school.

The second round procedes similarly with families that either did not participate in the previous round or remained unassigned due to the limiting capacities of their submitted schools. In this round, no student can remain unassigned. A student which remains once again unassigned is simply assigned to the nearest school (geographically) with spare capacity.

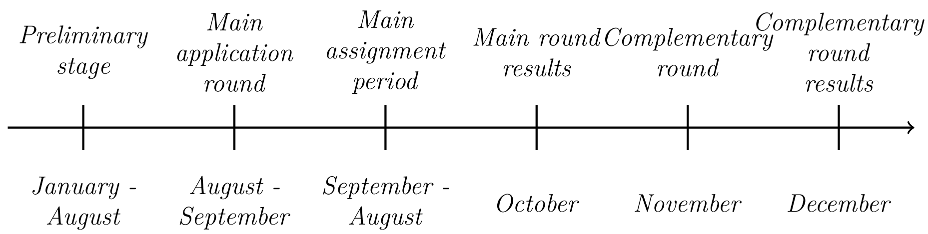

Figure 4.1: General Assignment Proccess Steps

The core of the SAE admission problems lies in the SAE webpage[1]. Through this online platform students are able to review and compare all available school characteristics, they are able to submit and edit their preferences and finally accept or reject once the results are published.

The deployment of the SAE project has been gradual. In 2016 it only included pre-kinder, kinder, primero básico, séptimo básico and primero medio age groups and only considered students from the Magallanes Region. In 2017, the project included all age group levels from 5 regions. From 2020 onwards, the system includes all age groups and all regions of chilean public schools.

4.2. Data Overview

All of the data used is publicly available in the MINEDUC data webpage[2]. The Ministry uploads this dataset every year as part of a governmental transparency policy. The data available is enough to replicate every step of the allocation process, including the main round and complementary round. However, we shall focus only on the main round as it represents the more complex matching problem we intend to tackle.

[1] For further information visit: https://www.sistemadeadmisionescolar.cl/

[2] For more information, visit: http://datos.mineduc.cl/dashboards/20514/descarga-bases-de-datos-sistema-de-admision-escolar/

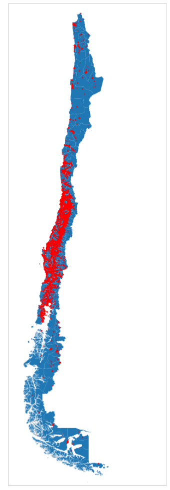

Figure 4.2: Student Map

The data set is divided into 5 files: supply (schools), demand (students), applications, outcome and additional information.

4.2.1. Supply Data

Supply data is described in the file A1. Each of the entries in this file represents a pair (establishment, course) which in theory represents a single 'school' to which a student applies to. The establishment is identified with a unique *RBD*, while the course is identified by the *COD_CURSO* variable.

For each (establishment, course) pair, the file contains the following variables:

1. Educational day format. This is either morning, afternoon or complete.
2. Genre of admissible students. Either masculine, feminine or both.
3. Geographical location of school in lat-long format.
4. A binary variable indicting if the school charges some sort of tuition.
5. Total available seats and vacant seats.

4.2.2. Demand Data

Demand data is described in the file B1. Each of the entries in this file represents a student which is uniquely identified by the *mrun* code, consistently used in MINEDUC data to ensure anonymity. Each *mrun* contains the following variables.

1. Educational level. Integer variable ranging from −1 (pre-kinder) to 12 (4th medio).
2. Genre, which is either masculine, feminine or other.
3. A binary variable indicating if the student is priority-type or not.
4. A binary variable stating if the student is part of the top 20% ranking of his/her school.
5. Georeference of the student's declared location in lat-long format and a quality index representing the quality of the georeferencing.

With respect to georeferencial data, it is important to notice that these are subject to error. When submitting their information through the SAE webpage, students are asked to enter their address manually. This information is then processed by MINEDUC using the *Google Geocoding API*[3] which returns both lat and long variables as well as a quality index associated to the precision of the output for each address.

The quality index ranges from 1 to 5 according to the following scheme. Quality index 1 means that the application was able to obtain a single result with either 'rooftop' or 'range interpolated' quality, which reflects a very precise result. Quality index 2 also reflects a single result but in this case with 'geometric center' or 'approximate' quality, which means that the algorithm identified relevant information in the address and used geometric centers of streets, districts or whatever information was available to obtain the lat-long values. Quality index 3 means that the API obtained multiple results. Quality index 4 means the address corresponds to the municipality address. Quality index 5 means that the student shared his location.

Table 4.1: Georeferential Quality Index Distribution

Quality Index	**Count**	**Percentage**
1	43,355	37.5%
2	23,157	20%
3	0	0%
4	17,190	15%
5	31,860	27.5%

Additionally, for anonymity purposes, the lat-long values of the publicly available SAE data is slightly distortioned using a random procedure. A direction between 0 and 360 degrees is chosen randomly as well as a random distance between 50 and 300 meters. The point is then shifted in that direction. For instance, data is slightly noisy.

Although georeferential data is clearly not perfect, it provides much information of student's preferences and it is the most complete available student characteristic other than their preferences, which is what we seek to exploit. Consequently, as we will argue in the theoretical

[3] Visit https://developers.google.com/maps/documentation/geocoding/overview for additional information.

model, we allow for noisy information when it comes to including student's characteristics.

4.2.3. Applications

The application information is described in the C1 file. Each of the entries corresponds to a student's application to one school program and its associated preference ranking. Therefore, each entry can be uniquely defined by the (*mrun*, *RBD*, *COD_CURSO*, *preference*) tuple, which means student *mrun* applies to school (*RBD,COD_CURSO*) as its preference number *preference*.

Theoretically, students rank all schools as it is common to assume that all schools are acceptable for all students. However, this is not true in practice, as the supply of programs is vast and therefore it is very costly (generally time-wise) for parents or guardians to gather information of all programs. The data shows students apply to around 3 schools on average.

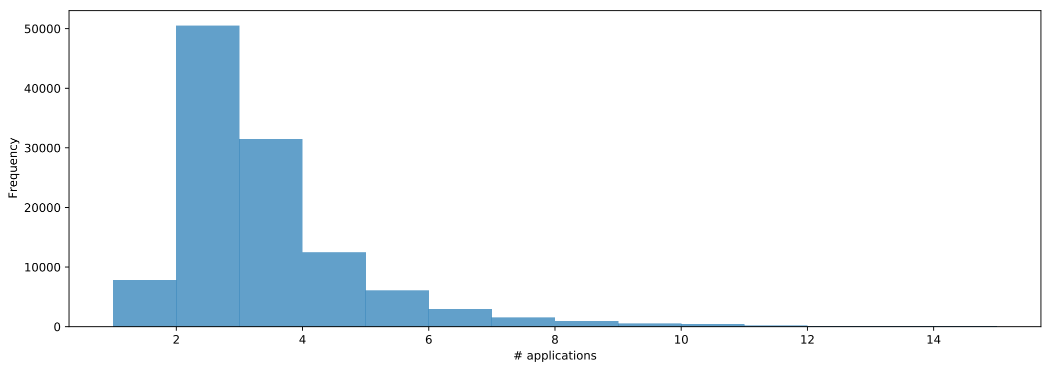

Figure 4.3: Distribution of Student's Number of Applications

Furthermore, with respect to distance data, we can observe that, on average, students apply to schools which are around 3.8km away. The distribution suggests, as Aramayo and Goic (2018) state, that students tend to prefer schools that are close by. This is a key feature in our simulations, as we will use distance as our observable characteristic to construct school's preferences.

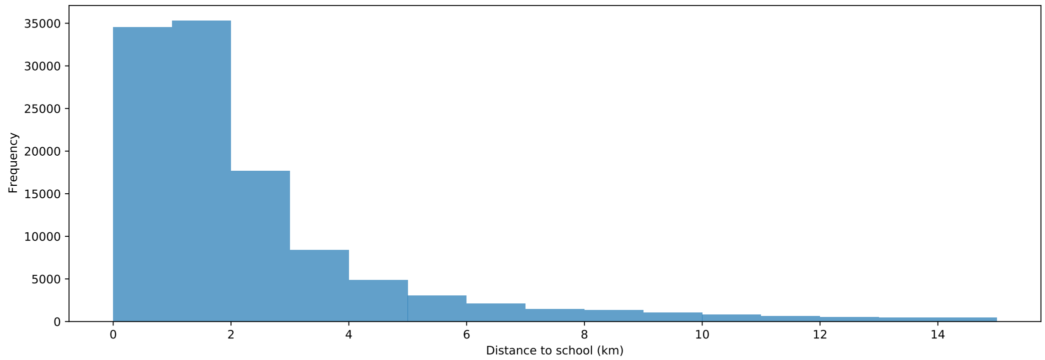

Figure 4.4: Distribution of Student's Distance to Applied Schools

4.2.4. Outcome

The official results obtained in that year's allocation procedure are contained in the D1 file. Each entry correspond to a (*mrun*, *RBD_CURSO*, *COD_CURSO*) tuple, this is, the one-to-one matching obtained for each student-school pair. If the *RBD_CRUSO*,*COD_CURSO* variables are empty, that student was left unassigned in the first round. This file also contains information regarding the posterior student's response to the assigned school as to where the student accepted or rejected the assignment. Although this might also be interesting to exploit, we shall only consider the algorithm's output as the official allocation for this project's purpose.

4.3. Simulated Results

Our simulations only consider the pre-kinder age group ($cod\_nivel = -1$). This is because it represents the larger age group market without saturating computation time and it simplifies calculations of vacancies which depend on continuing students and other criteria.

As mentioned before, we shall use the distance between the student and the school as our information source/proxy over student preferences in a data-driven policy design, as defined in our theoretical approach. This is mainly motivated by the results shown by Aramayo (2018), who shows empirically that the distance to schools has a significant effect over student's ranking decisions. Moreover, as mentioned in the general overview, the distance from schools is the only completely available student characteristic.

In the 2021 admissions process, the market size is as follows.

Table 4.2: Market Size

Market side	**Count**
Supply (schools)	5,221 (total capacity 167,980)
Demand (students)	115,562

4.3.1. Pareto-improving Pairs

Firstly, we evaluate the allocation from a Pareto-improving pairs point of view. As we mentioned before, this metric is widely used in the literature, probably due to it's simple intuition: we can swap two student's schools and Pareto-improve the allocation.

When generating data-driven lotteries, we observe a significant reduction in the amount of students that belong to Pareto-improving pairs.

Table 4.3: Simulation of Pareto-improving pairs

Lottery Policy	**2021**
MTB lottery	12,641 (10.94%)
Data-driven lottery	4,315 (3.73%)

This is a first empirical result that supports the theoretical idea of improving efficiency by using student information. Intuitively, using data-driven policies reduces the probability of a school granting higher priority to a student that does not rank that school highly versus a student that does. This kind of randomly-induced result that allows for Pareto-improving pairs to emerge is less likely if information is considered.

4.3.2. Cumulative Rank Distribution

Secondly, we evaluate the allocation from a cumulative rank distribution point of view. This metric grants a more holistic perspective as it specifies student's rankings and the amount of students that remain unassigned. Figure 4.5 shows the simulated results for the three analyzed policies: STB, MTB and data-driven (based on student's distance to schools).

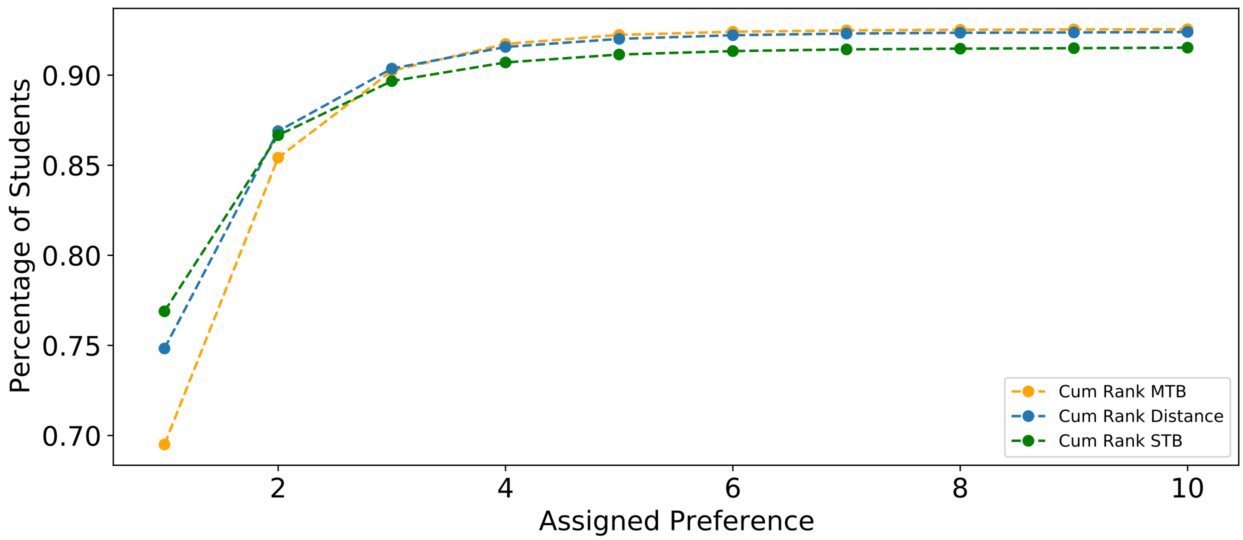

Figure 4.5: Cumulative Rank Distributions

As discussed by Abdulkadiroğlu et al. (2009), the single tie-breaking (STB) lottery policy assigns more student to their top preference than the MTB policy, although it leaves more students unassigned. Moreover, there is no stochastic dominance between the rank distributions as the STB policy assigns less students from their third preference onwards. Students that obtain high lottery tickets have a good chance of being assigned to their top schools, but the students that get lower lottery tickets find themselves rejected repeatedly with higher probability. This generates a sort of 'snowball effect' that results in a greater number of students unassigned.

On the other hand, the data-driven lottery policy assigns more students to their top three preferences than MTB but not as much as STB. In contrast to STB, although we cannot ensure stochastic dominance, the data-driven lottery performs quite similarly to the MTB in lower rankings and the amount of unassigned students remains practically constant. Data-driven lotteries, similar to STB, assigns more students to their top preferences but seems to avoid the 'snowball effect' that generates a reduction in lower rank assignments and unassigned students. This design improves efficiency by exploiting student heterogeneity rather than forcing a more dictatorial kind of procedure.

Table 4.4: Policy Comparison - Unassigned Students

Lottery Policy	**Unassigned Students**	**Percentage**
MTB	8,609	7.45%
STB	9,626	8.33%
Data-driven	8,786	7.60%

However, using the distance from schools as a proxy comes at a cost. When using distance to schools to define priorities we observe a significant increase in segregation. This is probably due to the fact that Santiago (and other chilean cities) show important geographical segregation. Therefore, the data-driven design limits priority student's chances for admission in better schools which are usually located outside of their districts. As shown by the Duncan Index (Duncan & Duncan, 1955) (a commonly used metric to evaluate segregation) the simulations show that segregation nearly doubles.

Lottery Policy	**Duncan Index**
MTB	0.298
STB	0.534
Data-driven	0.566

However, this cost is specific to this observable characteristic. Optimistically, there might be other available student characteristics that are simultaneously effective predictors of preferences and not segregators, avoiding these kind of costs.

Chapter 5

Conclusion

Recently, many centralized school allocation markets worldwide have gravitated towards the use of the DA algorithm with randomly generated tie-breaking procedures as a way of granting equality ex-ante. Erdil and Ergin (2008) show this mechanism damages efficiency as lotteries impose artificial stability constraints. Stable mechanism are highly competitive for students (Che & Tercieux, 2018) which is the main cause of efficiency loss.

We propose designing lotteries by correlating student's and school's preferences to reduce excessive competition and therefore improve efficiency. To maintain strategy-proofness through the DA, we can use observable student characteristics to predict student preferences and, if the proxy is informative, effectively correlate preferences.

Theoretically, we have shown that data-driven policies perform better than MTB policies in terms of efficiency when observable characteristics are good predictors, but perform worse when they are not. Additionally, enhancing efficiency by correlating preferences seems less costly than doing so by using STB procedures. This is because data-driven policies exploit preference heterogeneity rather than imposing a more 'dictatorial' lottery design.

The theoretical results are consistent with empirical simulations using chilean SAE 2021 admissions data. Using distance to schools as our observable characteristic, the number of students in Pareto-improving pairs is greatly reduced. Furthermore, data-driven policies cumulative rank distributions stochastically dominate MTB in the top three preferences and, unlike STB, assigns almost the same mass of students in lower rankings, leaving the same amount of students unassigned.

Bibliography

Abdulkadiroglu, A., Pathak, P., Roth, A., & Sonmez, T. (2006, 01). Changing the boston school choice mechanism: Strategy-proofness as equal access.

Abdulkadiroğlu, A., Che, Y.-K., & Yasuda, Y. (2015). Expanding "choice" in school choice. *American Economic Journal: Microeconomics*, *7*, 1–42.

Abdulkadiroğlu, A., Pathak, P., & Roth, A. (2009). Strategy-proofness versus efficiency in matching with indifferences: Redesigning the nyc high school match. *American Economic Review*, *99*, 1974–1978.

Abdulkadiroğlu, A., & Sönmez, T. (2003). School choice: A mechanism design approach. *American Economic Review*, *93*, 729–747.

Aramayo, N. (2018). Implementation of the matching mechanism for the new school admissions system and modeling of the school choice for chilean families. *Universidad de Chile*.

Ashlagi, I., & Nikzad, A. (2020). What matters in school choice tie-breaking? how competition guides design. *Journal of Economic Theory*, *190*, 105–120.

Azevedo, E., & Leshno, J. (2016). A supply and demand framework for two-sided matching markets. *Journal of Political Economy*, *124*.

Biblioteca del Congreso Nacional de Chile. (2015). *Ley 20.835.*

Che, Y.-K., & Tercieux, O. (2018). Efficiency and stability in large matching markets. *Journal of Political Economy*, *127*, 0–12.

Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, *20*, 210.

Echenique, F., Gonzalez, R., Wilson, A., & Yariv, L. (2020, 02). Top of the batch: Interviews and the match. *SSRN Electronic Journal*.

Erdil, A., & Ergin, V. (2008). What's the matter with tie-breaking? improving efficiency in school choice. *American Economic Review*, *98*, 669–689.

Gale, D., & Shapley, L. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, *69*, 9–15.

Pathak, P. A., & Sönmez, T. (2008). Leveling the playing field: Sincere and sophisticated players in the boston mechanism. *American Economic Review*, *98*, 1636-52.

Roth, A. E. (1982). The economics of matching: Stability and incentives. *Mathematics of Operations Research*, *7*, 617-628.

Roth, A. E., & Peranson, E. (1999). The redesign of the matching market for american physicians: Some engineering aspects of economic design. *American Economic Review*, *89*, 748-780.

Roth, A. E., Sönmez, T., & Ünver, U. (2004). Kidney exchange. *The Quarterly Journal of Economics*, *119*, 457–488.

Shapley, L., & Scarf, H. (1974). On cores and indivisibility. *Journal of Mathematical Economics*, *1*, 23-37.

Çelebi, O., & Flynn, J. P. (2021). Priority Design in Centralized Matching Markets. *The Review of Economic Studies*, *89*(3), 1245-1277.

Annex

Optimal Policy Proof

As mentioned above, the optimal policy problem for a setting with 2 schools is the following.

$$\begin{aligned}
\max_{X^i,Y^i} \quad & \int_0^1 F(1-2i)(1-X^i)di + \int_0^1 (1-F(1-2i))(1-Y^i)di \\
\text{s.t.} \quad & \int_0^1 F(1-2i)(1-X^i)di + \int_0^1 (1-F(1-2i))Y^i(1-X^i)di = S_1 \\
& \int_0^1 F(1-2i)X^i(1-Y^i)di + \int_0^1 (1-F(1-2i))(1-Y^i)di = S_2
\end{aligned} \tag{.1}$$

As we assume capacities are restrictive, we can rewrite the problem in an unrestricted manner.

$$\max_{X^i,Y^i} \quad S_1 + S_2 - \int_0^1 (1-F(1-2i))Y^i(1-X^i)di - \int_0^1 F(1-2i)X^i(1-Y^i)di$$

We would like to show that the optimal policy is given by $X^i = Y^i \in \{0,1\}$ $\forall i \in [0,1]$ and in particular

$$X^i = Y^i = \begin{cases} 0 & \forall i \in \bar{I} \\ 1 & \forall i \notin \bar{I} \end{cases}$$

where the set $\bar{I} \subseteq I$ of students must be feasible, that is

$$\int_{\bar{I}} F(1-2i)di = S_1 \quad \text{and} \quad \int_{\bar{I}} (1-F(1-2i))di = S_2$$

For this, we can argue in 2 different ways.

- Let $R^*(1)$ be the optimal value in the unrestricted problem, the objective function has a straightforward bound $R^*(1) \leq S_1 + S_2$ because we know that $\int_0^1 (1-F(1-2i))(1-Y^i)di \geq 0$ and $\int_0^1 F(1-2i)X^i(1-Y^i)di \geq 0$ for any policy X^i, Y^i. Then, using $X^i = Y^i$ such that $\bar{I}$ is feasible, the upper bound is obtained and therefore it is optimal.

- Using Euler-Lagrange, we can solve the optimization problem for a fixed $i \in [0,1]$. Each problem is as follows.

$$\max_{X,Y} \quad S_1 + S_2 - (1 - F(1-2i))Y^i(1 - X^i) - F(1-2i)X^i(1 - Y^i)$$

We shall note $R_i(1)$ as the optimum for this problem.

Then, the derivatives for X, Y are:

$$\frac{\partial R_i(1)}{\partial X} = (1 - F(1-2i))Y - F(1-2i)(1-Y) = Y - F(1-2i)$$
$$\frac{\partial R_i(1)}{\partial Y} = -(1 - F(1-2i))(1-X) + F(1-2i)X = -1 + X + F(1-2i)$$

As the derivatives are constant with respect to the variable in question, we will find corner solutions. We can check that the proposed optimum meets these conditions.

For X:

  - If $Y = 0$, then $\frac{\partial R_i(1)}{\partial X} = -F(1-2i) < 0$ and therefore $X = 0$ is optimal.
  - If $Y = 1$, then $\frac{\partial R_i(1)}{\partial X} = 1 - F(1-2i) > 0$ and therefore $X = 1$ is optimal.

For Y:

  - If $X = 0$, then $\frac{\partial R_i(1)}{\partial X} = -1 + F(1-2i) < 0$ and therefore $Y = 0$ is optimal.
  - If $X = 1$, then $\frac{\partial R_i(1)}{\partial X} = F(1-2i) > 0$ and therefore $Y = 1$ is optimal.

Additionally, we can discard any interior solution. In contradiction, let X, Y be such that

$$\frac{\partial R_i(1)}{\partial X} = Y - F(1-2i) = 0 \implies Y = F(1-2i)$$
$$\frac{\partial R_i(1)}{\partial Y} = -1 + X + F(1-2i) = 0 \implies X = 1 - F(1-2i)$$

Then, at this point, the objective function is

$$R_i(1) = S_1 + S_2 + F(1-2i)^2 - F(1-2i) < S_1 + S_2$$

which is not optimal.

MTB Equilibrium

The market clearing conditions for a MTB equilibrium are:

$$\int_0^1 F(1-2i)(1-X)di + \int_0^1 (1-F(1-2i))Y(1-X)di = S_1$$
$$\int_0^1 F(1-2i)X(1-Y)di + \int_0^1 (1-F(1-2i))(1-Y)di = S_2$$

Let $\bar{F} = \int_0^1 F(1-2i)di$, the conditions define the following 2×2 system.

$$\bar{F}(1-X) + (1-\bar{F})Y(1-X) = S_1$$

$$\bar{F}X(1-Y) + (1-\bar{F})(1-Y) = S_2$$

Subtracting both equations we get

$$XY = 1 - S$$

where $S = S_1 + S_2$. If we replace this condition in the market clearing condition for school C_0 we get the following quadratic equation.

$$\bar{F}(1-X) + \frac{1-S}{X}(1-X)(1-\bar{F}) = S_0$$

$$\implies -X^2\bar{F} + X[+\bar{F} - (1-S)(1-\bar{F}) - S_0)] + (1-S)(1-\bar{F}) = 0$$

Which has solution

$$X^* = \frac{(\bar{F} - (1-S)(1-\bar{F}) - S_0)}{2\bar{F}} + \frac{\sqrt{((1-S)(1-\bar{F}) + \bar{F} - S_0)^2 + 4\bar{F}(1-S)(1-\bar{F})}}{2\bar{F}}$$

Similarly, replacing our condition in the market clearing condition for school C_1 we get

$$\bar{F}(1-Y)\frac{1-S}{Y} + Y(1-Y)(1-\bar{F}) = S_1$$

$$\implies -Y^2(1-\bar{F}) + Y[-\bar{F}(1-S) + (1-\bar{F}) - S_1] + \bar{F}(1-S) = 0$$

Which has solution

$$Y^* = \frac{((1-\bar{F}) - S_1 - (1-\bar{F})(1-S)))}{2(1-\bar{F})} + \frac{\sqrt{(-(1-\bar{F}) + S_1 + (1-\bar{F})(1-S))^2 + 4\bar{F}(1-S)(1-\bar{F})}}{2(1-\bar{F})}$$

STB Equilibrium

The market clearing conditions for a STB equilibrium are the following.

$$\int_0^1 F(1-2i)(1-X)di = S_0$$
$$\int_0^1 F(1-2i)(X-Y)di + \int_0^1 (1-F(1-2i))(1-Y)di = S_1$$

The difference in the market clearing equation from school C_0 comes from the fact that, as $\frac{S_0}{\bar{F}} \leq \frac{S_1}{1-\bar{F}}$ then $p_0 \geq p_1$. Therefore, as the score for each student i is identical, the conditional probability of being accepted at school C_0 when previously rejected at school C_1 is 0. Then, the solution is the following.

$$X = 1 - \frac{S_0}{\bar{F}}$$
$$Y = 1 - S_0 - S_1$$

Then, the total mass of students assigned to their top preference is:

$$R_{STB}(1) = S_0 + (1-\bar{F})(S_0 + S_1 - \frac{S_0}{\bar{F}})$$

Data-driven Policy vs MTB Proof

Let $\epsilon^i_{C_1} - \epsilon^i_{C_0} \sim F$ and $\underline{\delta} \in \mathbb{R}^+$ be such that the data-driven policy assigns a mass of students to their top preference given by

$$R^{\underline{\delta}}_{DB}(1) = \int_0^{S_1} F\left(\frac{1-2i}{\underline{\delta}}\right) di + \int_0^{S_1} \left(1 - F\left(\frac{1-2i}{\underline{\delta}}\right)\right) di$$

and particularly $R^{\underline{\delta}}_{DB}(1) > R^{\underline{\delta}}_{MTB}(1)$. Consider $\delta < \underline{\delta}$, this is $\delta = \alpha \cdot \underline{\delta}$ with $\alpha < 1$. Because F is an increasing function, it is true that $\forall i < \frac{1}{2}$, $F\left(\frac{1-2i}{\alpha \cdot \underline{\delta}}\right) > F\left(\frac{1-2i}{\underline{\delta}}\right)$ and $\forall i > \frac{1}{2}$, $F\left(\frac{1-2i}{\alpha \cdot \underline{\delta}}\right) < F\left(\frac{1-2i}{\underline{\delta}}\right)$. Then, as $S_1, S_2 < 1/2$ we have that

- $\forall i < \frac{1}{2}$: $1-2i > 0$ and then $F\left(\frac{1-2i}{\alpha \cdot \underline{\delta}}\right) > F\left(\frac{1-2i}{\underline{\delta}}\right) \implies \int_0^{S_1} F\left(\frac{1-2i}{\underline{\delta}}\right) di < \int_0^{S_1} F\left(\frac{1-2i}{\alpha \cdot \underline{\delta}}\right) di$
- $\forall i > \frac{1}{2}$: $1-2i < 0$ and then $F\left(\frac{1-2i}{\alpha \cdot \underline{\delta}}\right) < F\left(\frac{1-2i}{\underline{\delta}}\right) \implies \int_0^{S_1} \left(1 - F\left(\frac{1-2i}{\underline{\delta}}\right)\right) di < \int_0^{S_1} \left(1 - F\left(\frac{1-2i}{\alpha \cdot \underline{\delta}}\right)\right) di$

and therefore $R^{\delta}_{DB}(1) > R^{\underline{\delta}}_{DB}(1) > R_{MTB}(1)$, which proves the result. Similarly, if there exists $\bar{\delta} = \alpha \cdot \bar{\delta}$ with $\alpha > 1$, then the converse is true. $\forall i < \frac{1}{2}$ $F\left(\frac{1-2i}{\alpha \cdot \bar{\delta}}\right) < F\left(\frac{1-2i}{\bar{\delta}}\right)$ and $\forall i > \frac{1}{2}$ $F\left(\frac{1-2i}{\alpha \cdot \bar{\delta}}\right) > F\left(\frac{1-2i}{\bar{\delta}}\right)$. Therefore $R^{\delta}_{DB}(1) < R^{\bar{\delta}}_{DB}(1) < R_{MTB}(1)$.