ECARES

Self-Selection in School Choice

Li Chen
SBS-EM, ECARES, Université libre de Bruxelles

Juan Sebastián Pereyra
SBS-EM, ECARES, Université libre de Bruxelles

December 2015

ECARES working paper 2015-52

ECARES
ULB - CP 114/04
50, F.D. Roosevelt Ave., B-1050 Brussels BELGIUM
www.ecares.org

SELF-SELECTION IN SCHOOL CHOICE

LI CHEN AND JUAN SEBASTIÁN PEREYRA

Abstract. We study self-selection in centralized school choice, a strategic behavior that takes place when students submit preferences before knowing their priorities at schools. A student self-selects if she decides not to apply to some schools despite being desirable. We give a theoretical explanation for this behavior: if a student believes her chances of being assigned to some schools are zero, she may not rank them even when the mechanism is strategy-proof. Using data from Mexico City high school match, we find evidence that self-selection exists, and has negative consequences for low income students as a result of their strategic mistakes.

December, 2015

Key words: School choice, Incomplete Information, Self-selection, Serial Dictatorship Mechanism, Strategy-proofness

JEL Classification: C40, C78, D47, D63, I20, I21, I24

1. Introduction

School choice is aimed at "leveling the playing field": all students should have equal opportunities to choose the school they like, and whenever the number of applicants exceeds a school's capacity, transparent criteria should be employed to decide who will be admitted. To this end, market designers have advocated the use of strategy-proof mechanisms, which guarantees that students can do no better than submitting preferences truthfully (Roth (2008)). However, non-truthful behaviors are not only possible in theory, as truth-telling is a *weakly* dominant strategy, but also well documented in lab (Chen and Sönmez (2006) and Pais and Pintér (2008)).

Our paper ponders the performance of strategy-proof mechanism in a novel situation where students have to submit preferences over schools before knowing their priorities. The introduction

The authors are affiliated with ECARES - Solvay Brussels School of Economics and Management, Université libre de Bruxelles and F.R.S.-FNRS, emails lichen@ulb.ac.be, jpereyra@ulb.ac.be. We especially thank Ana María Aceves Estrada, Roberto Peña Resendiz from COMIPEMS, and Manuel Gil Antón for helping us to access the data, and Estelle Cantillon, Patrick Legros, and Jordi Massó for their insightful suggestions. We also thank Atila Abdulkadiroğlu, Philippe Aghion, Christopher Avery, Péter Biró, David Cantala, Arnaud Dupuy, Álvaro Forteza, Marion Mercier, Dilip Mookherjee, Davy Paindaveine, Francisco Pino, Debraj Ray, Rajiv Sethi, Ran Shorrer, Olivier Tercieux, Alex Teytelboym, Liqiu Zhao, Aiyong Zhu, and seminar participants at ECARES Petit Déjeuner, Economics Department - FCS Uruguay, Renmin University of China, Wuhan University, Barcelona GSE summer forum - Matching in Practice, ThReD 2015 conference, Conference on Economic Design 2015, Public Economic Theory Conference 2015, and the 10th Workshop on Economic Design and Institutions. Financial support from the Belgian National Science Foundation (FNRS) and ERC grant 208535 is gratefully acknowledged.

of uncertainty about schools' priorities is motivated by the high school match in Mexico City, where students are asked to submit a rank ordered list of high schools before taking a standardized exam. The exam score then determines a strict and unique priority order, according to which the serial dictatorship mechanism is used to allocate students to schools. Given that students are not *effectively* constrained by the number of options they can submit, the mechanism in place is **strategy-proof**.[1]

Our contribution is two-fold. First, we present *field* evidence of an overlooked strategic behavior, which we call "self-selection", under a straightforward strategy-proof mechanism. A student self-selects if she does not top rank her most preferred school. Self-selection can be both an equilibrium behavior or the result of strategic mistakes. In the first case, it has not impact on the student assignment, while in the second case it may hurt students once uncertainty is resolved. Our findings highlight that 22% of the students in our selected sample self-select, among which 23% are due to strategic mistakes. Second, we conduct a *counterfactual* analysis to compare the current outcome with the one it is obtained after correcting students' mistakes (which equals the outcome with complete information). We show that in the alternative design students from low income families increase their participation in the best-quality schools. Thus, changing the timing of preferences submission will improve the effectiveness of equal access.

We begin by introducing an incomplete information game to study students' equilibrium strategy in the Mexico City high school match. When facing incomplete information, students' best responses depend on their beliefs. We show that when schools do not behave strategically (priorities are exogenously defined), and students' beliefs are such that **each** profile of other students' preferences and schools' priority has a positive probability to occur, the **unique** equilibrium (in the game induced by the serial dictatorship mechanism) is for students to submit their preferences truthfully. However, when beliefs do not have full support, strategic behaviors may be triggered at equilibrium.

Our theory suggests a discontinuity in students' equilibrium strategies: when every student's beliefs are such that the probability of being assigned to each school is positive, submitting truthfully is the unique equilibrium; but as soon as this probability goes to zero for some schools, strategic behaviors may arise even if the mechanism in place is strategy-proof. Among these behaviors, we focus on self-selection, a strategy by which some students do not top rank their most preferred school.

[1]Although students can rank up to 20 schools, as we show in Section 6.2, only 3% of them use the full list.

Given that both truth-telling and self-selection are possible at equilibrium, why do we expect self-selection to arise in data? The official advice in fact recommends students to take expected priorities into account:

> *"Which is the best option? ... In addition to your preferences, interests and circumstances, it is worth considering other factors before choosing your options. ... it turns out evidently that your chances of getting a place in the option you prefer more will depend on the score of your exam. In this sense,* ***you should be very conscious and objective about your likely performance****."*

Therefore, self-selection is not only possible in theory, but also can happen in practice if such (misleading) advice is followed.

We begin our empirical analysis by asking: *do students self-select when there is uncertainty about priorities, even if the mechanism is strategy-proof?* To approach this question, we consider those students who agree that their most desirable choice belongs to a set of top quality high schools affiliated to *Universidad Nacional Autónoma de México* (UNAM). These students are selected using a survey question about which university they would like to attend in the future. Since the chances of being admitted to UNAM are significantly higher for those that attended a UNAM high school, we argue for those who would like to attend UNAM that their top choice is one of the UNAM high schools. By looking at the submitted choices, we find one fifth of them do not top rank any UNAM high school. This type of self-selection strategy is consistent with our characterization of non-truthful equilibrium strategy when students' beliefs are such that the probability of being assigned to these schools is zero, but it could also be driven by mistakes.

In a second step, we explore which factors affect self-selection. Our probit estimation reveals that secondary school average grade and family income are the most important factors contributing to self-selection, after controlling for other variables (including distance). In particular, an increase of one standard deviation in the secondary school average grade decreases the probability of self-selection by 7.7 percentage points. In addition, the probability of self-selection by students from low income families is 8 percentage points higher than those from high income families.

These results unveil the decision process when students are deciding on which schools to apply to. Average grade from secondary school helps students to form beliefs about their future exam scores. However, given the same grade, their strategies depend on socio-economic backgrounds. Those coming from low income families tend to be more pessimistic about their future performance, resulting in higher probability of self-selection. Moreover, the difference between high and low income families increases as we move to lower grades. For example, a student with a top 10% grade

in secondary school self-selects with almost the same probability independent of her family's income level. However, a student with a bottom 10% performance in secondary school is 13 percentage points more likely to self-select if she is from a low income than from a high income family.

In a third step, we disentangle those cases where self-selection is an equilibrium strategy from those where it is a consequence of strategic mistakes. We say that self-selection is a strategic mistake for a student if she self-selects and finally obtains a score high enough to enter one of the UNAM high schools. The findings show that among self-selected students, for close to 23% of them self-selection is a strategic mistakes which raises further concerns in terms of the stability of the final matching.

Finally, we correct the submissions of those students that self-select due to strategic mistakes. When all students play equilibrium strategies, the matching equals the stable matching under complete information. Compared to the current matching, the participation of students from low socio-economic backgrounds rises, creating more social diversity at UNAM high schools. Indeed, while the increase with respect to the observed matching is 5.2% and 1.6% within low and middle income groups, respectively, the number of high income students is reduced by 0.5%. Therefore, changing the timing of preferences submission to after knowing priorities can prevent strategic mistakes, and therefore improves the access of disadvantaged students to good schools. If the modification is difficult, clear advice encouraging truth-telling without considering the chances of admissions (in contrast to the current advice), will also help low-income students.

There is an ongoing debate in Mexico on the access of low-income students to higher education and in particular to UNAM, the most prestigious university of the country. Our results suggest that the problem of low access may not root in the admission to UNAM, but in the assignment of students to its high schools. As students from UNAM high schools have priority over other students to get a seat in UNAM, the participation for self-selected low-income students is impeded.

Our paper offers two main insights beyond the specific matching problem studied.

First, self-selection points out the caveat of treating submitted preferences as true preferences under strategy-proof mechanisms. By doing so, if some students of low socio-economic backgrounds do not apply to good schools, one might conclude that they do it simply because they prefer their submitted choices over good schools. Moreover, policy interventions on the market design such as giving complete information to students when submitting preferences, may indeed have no impact. However, our findings suggest that this conclusion is wrong if students are facing uncertainty about priorities. In fact, strategic behaviors, and self-selection in particular, are triggered by uncertainty.

Therefore, we should expect changes in the application patterns of some students when the timing of preference submission is modified.

Second, our findings can apply to a more general context where both sides of the market are strategic. This is the case of college admissions in the US, where colleges have preferences which are unknown to students, and students may use private information to predict their likelihood of admission. As in the market studied here, some students may self-select and not apply to selective colleges based on their predicted probability of being assigned to them. If after the uncertainty is resolved, it turns out that some students self-selected due to strategic mistakes, these non-truthful strategies may harm students in terms of their final assignment.[2]

The rest of the paper is organized as follows. Section 2 discusses our results in the perspective of the related literature. Section 3 describes the high school public system in Mexico City. Section 4 introduces the theoretical model. Section 5 describes our data, based on which, Section 6 presents a first piece of evidence of self-selection. Section 7 includes the results from the probit regression analysis, and Section 8 studies students' mistakes and simulates the outcome when students play equilibrium strategies. Finally, Section 9 concludes.

2. Related Literature

The literature on centralized school choice is initiated by Balinski and Sönmez (1999) and Abdulkadiroglu and Sönmez (2003). In the last decade, many advances have been made in the theory and practice of the design of school choice systems, and in particular, in the advantages and drawbacks of popular mechanisms. Special attention has been paid to strategy-proof mechanisms, so that their use is nowadays advocated by many researchers. However, experimental papers have found evidence of non-truthful behaviors under strategy-proof mechanisms. In one of the first experiments on school choice, Chen and Sönmez (2006) find that 28% of the subjects do not report their preferences truthfully under the deferred acceptance mechanism (which is strategy-proof and equivalent to the serial dictatorship studied in our framework). In a different experiment, Pais and Pintér (2008) assess the influence of information in students' strategies. As students have more information on schools' priorities and other students' preferences, the percentage of subjects that do not submit truthfully increases. Featherstone and Niederle (2013) analyze the performance of the same mechanism when preferences are aligned, and show that 20% of the subjects do not reveal

[2]As we discuss in the following section, recent papers have pointed out the existence of "missing applicants" in this market. That is, students with high ability that do not apply to selective colleges. Although the phenomenon is similar to self-selection, its driving forces are different.

their true preferences. The common goal of these studies is to compare the performance of the deferred acceptance and Boston mechanism, so they do not explore the reasons behind non-truthful behaviors. Moreover, they are based on experimental evidence, and then the presence of these strategies in the field remains as an open question. In a recent paper, Fack, Grenet, and He (2015) propose a method to estimate students' preferences under deferred acceptance mechanism without assuming truth-telling. Different from our approach, application cost of submitting preferences is their main concern for non-truth-telling behavior.

We contribute to this literature by giving *field* evidence of non-truthful behaviors when a strategy-proof mechanism is in place. We explore a new channel from which these behaviors arise: given the uncertainty about schools' priorities, students form priors to decide about their strategies, which may trigger non-truthful submissions at equilibrium. Thus, our paper gives a new account from both theoretical and empirical perspective of these strategic behaviors when students apply to schools without knowing their priorities.

The framework used in the paper is a game with incomplete information, which has been little-studied in the matching literature. The vast majority of the existing papers assumes complete information. In school choice this implies that students know at the time of submission, other students' preferences and schools' priorities. Roth (1989) is one of the exceptions, followed by recent studies by Ehlers and Massó (2007), Chakraborty, Citanna, and Ostrovsky (2010), Liu, Mailath, Postlewaite, and Samuelson (2014), and Ehlers and Massó (2015) who further extend the model to an incomplete information environment. There are little empirical studies trying to identify strategic behaviors in the presence of unknown priorities. A remarkable exception is Budish and Cantillon (2012) who study the allocation of elective courses to MBA students by a non-strategy-proof mechanism where students have to anticipate which courses will be popular to find their equilibrium strategies.

Our model builds on the framework introduced by Ehlers and Massó (2015). Their paper analyzes a two-sided matching model with firms and workers, and shows a connection between Nash equilibrium under complete information and Ordinal Bayesian Nash equilibrium under incomplete information. We show that there exists a unique equilibrium in the case where one side of the market, schools, is not strategic (as it is generally the case in school choice) and students' priors have full support.

Our paper links with the new literature on "under-matching" in college admission in the US. Avery, Hoxby, Jackson, Burek, Pope, and Raman (2006) are among the first to identify the existence of "missing applicants", that is, students with high ability who do not apply to selective colleges. This

fact is particularly strong among students from low-income families, who are called high-achieving low-income students (see also Dillon and Smith (2013), and Pallais (2013)). Self-selection is similar to under-matching: some students do not apply to good high schools, despite the fact that they are very likely to be admitted by these schools. Nonetheless, the driving force behind under-matching is difficulty in accessing information, different from the reason for self-selection. Hoxby and Turner (2015) show that when students receive information about the cost of college and financing options, the availability of the curricula and peers, and the different types of colleges available to them, they have a higher probability to apply to selective colleges. Moreover, those high-achieving low-income students who do not apply to selective institutions come from small districts which are geographically isolated, supporting the argument of difficulty in accessing valuable information (Hoxby and Avery, 2012).

We are compelled to search for reasons beyond the channel of accessing information due to the features of Mexico City high school match. First and foremost, the match we study is coordinated through a central clearinghouse, simpler than the decentralized market of college admission in the US. This centralized match aims at providing students with equal access to information. The clearinghouse publishes every year detailed information of all the available options (and the cut-offs of previous years). Students and their parents can read the information through the hand-outs or via Internet. Second, students in our study are geographically concentrated in Mexico City and its metropolitan area, not in geographically isolated regions. Moreover, the application fee is low, not depending on the number of applications, and all participating high schools are public and free.[3] Then, it is not surprising to see that students' participation is not an issue in our case: the coverage rate of high school education in Mexico City was 78% of total population between 15 and 17 years old for the year we consider in our data (INEE (2012)). Finally, an important information that students have when applying to colleges in the US is their scores, which serves as a proxy of the probability of acceptance to college. In the Mexico City high school match, students have to decide their applications based on their priors on future scores. Therefore, our explanation for self-selection is grounded in the uncertainty about information, not the difficulty in accessing information.

3. Public High School System in Mexico City

School education in Mexico is compulsory from age 6 to 17. Elementary school covers age 6 to 11, secondary school covers age 12 to 14, and high school age 15 to 17. While most families choose free public schools for their children, there is an increasing number of private schools, especially at high

[3]In 2015 the application fee was approximately USD 23.

school level where students have to pay a significant amount for tuition fees.[4] Nonetheless, private schools have remained a relatively small share as compared to public high schools. In the academic year 2010, public high schools cover 81% of the total number of school-age students (INEE, 2011).

We consider those students leaving secondary schools and applying to high school. Admissions to public high schools are often decentralized with the exception of a few metropolitan areas. One of them being Mexico City together with 22 municipalities in the neighboring state of Mexico. *Comisión Metropolitana de Instituciones Públicas de Educación Media Superior* (COMIPEMS) has been the organizing body since 1996.

There are three types of high schools in the match: *Bachillerato General*, which provides general academic courses; *Bachillerato Tecnológico* where in addition to the academic courses, students take extra classes to obtain a professional certificate, in fields such as computer technology; finally, *Carrera Técnica* offers no general academic courses and students finish with their vocational certificate in, for example, nursing or graphic design. Only the first two types of programs prepare students for access to universities or colleges. Each high school is managed by a public institution, and there are 9 such public institutions. Schools managed by the *Universidad Nacional Autónoma de México* (UNAM) are generally viewed as "elite" high schools. In addition, *Instituto Politécnico Nacional* (IPN) is the other institution that operates the technically-oriented "elite" schools.

The procedure of the match is arranged as follows (Figure 1).

In *late January*, COMIPEMS hands out brochures disseminating information about available options, and instructions of the process. In addition, information about past assignments is available to students.

In *March*, after registering with COMIPEMS, students submit their preferences over up to 20 options. A high school can offer one option (for example, all high schools affiliated to UNAM), or multiple options, as is the case of technical high schools. Additionally, students have to fill out a survey questionnaire.

In *June*, all students simultaneously take a standardized exam. The final score is used to determine student's priority in the match. A minimum score of 31 out of 128 is required for eligibility.[5] Additionally, if a student wants to apply to UNAM high schools, a minimum average grade of 7 in the secondary school (out of a scale of 10) is also required.

In *mid July*, students have to provide their secondary school certificate to be eligible for the match.

[4]This is the case, for example, of private schools with international curriculum that aim to prepare students for studying in universities aboard.

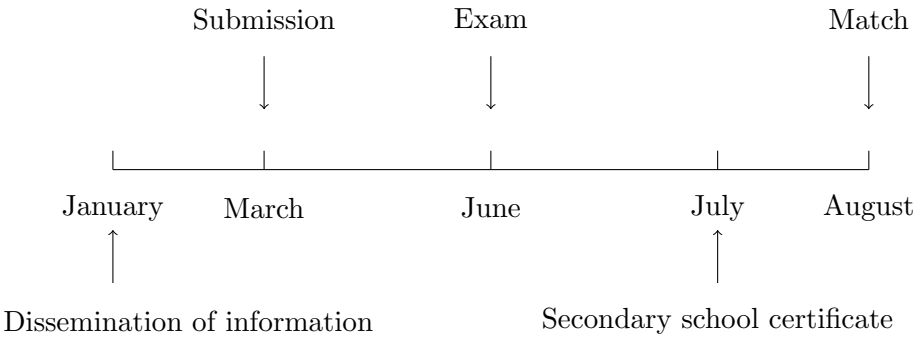
[5]This minimum score requirement was abolished as of year 2013 as an endeavor to extend compulsory education to high school level.

In *late July*, the match starts. It takes two phases:

(1) Main Phase: Based on the priority defined by exam scores, COMIPEMS allocates students using the *serial dictatorship* (SD) mechanism.[6] When students tie for the last seat at an option, COMIPEMS consults the school either to take all or reject all tied students.[7]

(2) Second Phase: Those students that remain unassigned because all their submitted options are full, are allowed to register in options with excess capacity. This process is decentralized.

FIGURE 1. The timing of COMIPEMS match



4. THEORETICAL FRAMEWORK

In this section we develop a school choice model with incomplete information. Each student has a prior over other students' preferences and schools' priority, based on which she chooses the preferences to be revealed. Our main result shows that when students' priors have full support, the unique equilibrium is to submit truthfully. Then, we study other equilibrium strategies when the full-support hypothesis does not hold. In particular, we introduce an assumption that allows us to derive precise implications on the equilibrium strategy that we may expect, linking in this way the theoretical model with our empirical analysis. Finally, we introduce the concept of self-selection.

Let $S = \{s_1, \ldots, s_m\}$ denote the set of schools, and I the finite set of students.[8] The capacity of school s_j is q_j, and $q = (q_1, \ldots, q_m)$ is the vector with each school's capacity. Assume that $\sum_{j=1}^{m} q_j \leq |I|$. Additionally, there is a null school, denoted by s_0, which is used to assign no school to students; and without loss of generality we suppose that s_0 is not scarce.

[6]The SD mechanism works as the following: it assigns the student with the highest priority her most favorite choice, then proceeds to the student with the next highest priority and assigns her the most favorite choice which is still available, and so forth.

[7]This way of breaking ties makes the mechanism manipulable in theory. However, the manipulation is extremely difficult in practice as it requires complete knowledge of future priorities and other students' preferences.

[8]To keep the convention in the literature, we use the term schools. In our empirical part, however we will use the term options, to accommodate the fact that some schools can host multiple options.

A matching is a function $\mu : I \to S \cup \{s_0\}$, satisfying $|\mu^{-1}(s_j)| \leq q_j$ for every $j \in \{1, \ldots, m\}$, and $\mathcal{M}$ is the set of matchings.

Students have strict preferences over the set of schools. Let P_i denote student i's preferences over $S \cup \{s_0\}$, and $\mathcal{P}_i$ the set of all possible preferences of student i. The notation sP_is' means that student i prefers school s over s', and when $s \equiv s_0$, that school s' is not acceptable for student i. Let R_i be the weak preferences associated with P_i. A preference profile is a vector in $\times_{i \in I}\mathcal{P}_i$.

The exam scores form a common priority ranking of all students, and high scores are preferred over low scores. In this sense, our problem is the student placement problem (Balinski and Sönmez, 1999) with the specificity that the same priority ranking is used for all schools.

Let P_s denote a strict priority order over I, and $\mathcal{P}_s$ the set of all possible priority orders. The set of all preference profiles and priorities is denoted by $\mathcal{P} = (\times_{i \in I}\mathcal{P}_i) \times \mathcal{P}_s$. Let $\mathcal{P}_{-i} = (\times_{j \in I \setminus \{i\}}\mathcal{P}_j) \times \mathcal{P}_s$.

We fix throughout this paper I, S and q, thus a mechanism is a function that maps the set of all preference profiles and priorities to the set of all possible matchings: $\phi : \mathcal{P} \to \mathcal{M}$. For each $P \in \mathcal{P}$, and $i \in I$, let $\phi_i[P]$ denote the school assigned to student i by the mechanism ϕ when preferences and priority order are P.

A matching is **stable** if no student prefers being unassigned to her assigned school, and whenever a student prefers another school to her current assignment, either she has lower priority than the assigned students, or there is no empty seat at that school. When schools share the same priority (as in our framework), the set of stable matchings is a singleton. A stable mechanism is a mechanism that associates a stable matching to every preference profile and priority.

An important property with respect to the mechanism itself is strategy-proofness. A mechanism ϕ is **strategy-proof** if truth-telling is a weakly dominant strategy for all students. Strategy-proofness is desirable because it reduces costly and risky strategic behavior of students by rewarding truth-telling students a no-worse outcome than if they had adopted any other strategy.

We consider the serial dictatorship (SD) mechanism which is outcome equivalent to the DA mechanism (Gale and Shapley, 1962) when schools have strict and identical preferences over students. Like DA, SD finds the unique stable matching, and it is strategy-proof.

4.1. **Incomplete Information.** The standard school choice model assumes complete information. That is, at the time of submission, students know other students' preferences and schools' priorities. The Mexican City high school match clearly departs from this assumption: students do not observe other students' preferences, and the timing of submission creates uncertainty about the exact priorities, which makes incomplete information a natural choice for our analysis. Moreover, the use of incomplete information framework brings two other advantages. First, it helps us to pin

down the issue of multiple equilibria in strategy-proof mechanisms, and second it offers a direct implication when truth-telling is not the unique equilibrium.

Under incomplete information, students form priors about other students' preferences and priorities at schools. The key element is the support of the priors. The intuition is that as long as every student assigns a positive probability to every priority order and profile of students' preferences, there is a **unique** equilibrium at which students reveal their preferences truthfully. On the contrary, when a student has a prior with non-full support, other strategies may emerge in equilibrium.

We follow the definitions introduced by Ehlers and Massó (2015). Each student's private information is her preferences over schools, so the set of all possible types of student i is $\mathcal{P}_i$. Additionally, each student i has a private prior (a probability distribution) $\tilde{P}^i$ over the set of all preferences profiles and priorities $\mathcal{P}$. The support of a prior is the set of all elements of $\mathcal{P}$ on which the prior puts positive probability. This prior may vary across students, and, in particular, it may have different support. When $\tilde{P}^i$ is such that $Pr\{\tilde{P}^i = P\} > 0$ for every $P \in \mathcal{P}$, we say that $\tilde{P}^i$ has full support. A profile of priors is a vector $\tilde{P} = (\tilde{P}^i)_{i \in I}$ specifying a private prior for each student.

A strategy of a student i defines for each type P_i the preference that i submits to the central clearinghouse, that is, a strategy is a function $r_i : \mathcal{P}_i \to \mathcal{P}_i$. A strategy profile is a vector $r = (r_i)_{i \in I}$ that specifies for each preference profile in $\times_{i \in I} \mathcal{P}_i$ a profile to be submitted. We denote as r_{-i} a vector of strategies for all students different from i.

Given a private prior $\tilde{P}^i$ and $P_i \in \mathcal{P}_i$, let $\tilde{P}^i_{-i} \,|P_i$ denote the probability distribution which $\tilde{P}^i$ induces over $\mathcal{P}_{-i}$ conditional on P_i. This conditional probability describes the uncertainty of student i with type P_i about other students preferences and the priority order.

A random matching $\tilde{\eta}$ is a probability distribution over the set of matchings $\mathcal{M}$. Given $\mu \in \mathcal{M}$, let $Pr\{\tilde{\eta} = \mu\}$ be the probability that $\tilde{\eta}$ assigns to μ. The random matching $\tilde{\eta}$ induces for each student a probability distribution over $S \cup \{s_0\}$ that represents the probability with which the student is assigned to each school by $\tilde{\eta}$. That is, for each $i \in I$, $\tilde{\eta}(i)$ is defined as:

$$Pr\{\tilde{\eta}(i) = s\} = \sum_{\mu \in \mathcal{M} \text{ s.t. } \mu(i) = s} Pr\{\tilde{\eta} = \mu\}, \text{ for each } s \in S \cup \{s_0\}$$

Given two random matchings $\tilde{\eta}$ and $\tilde{\eta}'$, for $i \in I$ and $P_i \in \mathcal{P}_i$ we say that $\tilde{\eta}(i)$ *first-order stochastically* P_i *dominates* $\tilde{\eta}'(i)$, denoted by $\tilde{\eta}(i) \succcurlyeq \tilde{\eta}'(i)$, if for all $s \in S \cup \{s_0\}$

$$\sum_{s' \in S \cup \{s_0\} : s' R_i s} Pr\{\tilde{\eta}(i) = s'\} \quad \geq \quad \sum_{s' \in S \cup \{s_0\} : s' R_i s} Pr\{\tilde{\eta}'(i) = s'\}$$

Given a mechanism ϕ and a prior $\tilde{P}^i$, a strategy profile r induces a random matching $\phi[r(\tilde{P}^i)]$ in the following way: for each $\mu \in \mathcal{M}$

$$Pr\{\phi[r(\tilde{P}^i)] = \mu\} = \sum_{P \in \mathcal{P}: \phi[r(P)] = \mu} Pr\{\tilde{P}^i = P\}.$$

Using Bayesian updating, the relevant random matching for student i, given her type P_i and a strategy profile r, is $\phi[r_i(P_i), r_{-i}(\tilde{P}^i_{-i}|P_i)]$, where $r_{-i}(\tilde{P}^i_{-i}|P_i)$ is the probability distribution induced by r_{-i} and $\tilde{P}^i$ over $\mathcal{P}_{-i}$ conditional on P_i. Further, given a student i with private prior $\tilde{P}^i$, let $\tilde{P}^i_i$ denote the marginal distribution of $\tilde{P}^i$ over $\mathcal{P}_i$.

Definition 1 (**Equilibrium Concept**)**.** Let $\tilde{P}$ be a profile of priors. A strategy profile r is an *ordinal Bayesian Nash equilibrium* (OBNE) in the revelation game induced by a mechanism ϕ under incomplete information $\tilde{P}$, if for all $i \in I$, and all $P_i \in \mathcal{P}_i$ such that $\Pr\{\tilde{P}^i_i = P_i\} > 0$,

$$\phi_i[r_i(P_i), r_{-i}(\tilde{P}^i_{-i}|P_i)] \succcurlyeq \phi_i[P'_i, r_{-i}(\tilde{P}^i_{-i}|P_i)], \quad \text{for all } P'_i \in \mathcal{P}_i.$$

The adoption of OBNE is appropriate for school choice where students are typically required to submit a ordered preferences over schools and not a specific utility representation of their preferences. Moreover, this equilibrium concept is not restrictive. In fact, as Ehlers and Massó (2007) point out, an OBNE is equivalent to a Bayesian Nash equilibrium for every von Neumann–Morgenstern utility representation of the preference order.

A truth-telling OBNE is an OBNE strategy profile such that for every $i \in I$, $r_i(P_i) = P_i$ for every $P_i \in \mathcal{P}_i$ such that $\Pr\{\tilde{P}^i_i = P_i\} > 0$.

To show truth-telling is indeed an OBNE in our setting, we need the following result which connects Nash equilibrium under complete information and OBNE under incomplete information.

Proposition 1 (**Equilibrium Strategies**)**.** *A strategy profile r is an OBNE in the revelation game induced by the SD mechanism under incomplete information $\tilde{P}$ if and only if for all $i \in I$ and any $P \in \mathcal{P}$ in the support of $\tilde{P}^i$, $r_i(P_i)$ is a best response to $r_{-i}(P_{-i})$ in the revelation game induced by the SD under complete information P.*

Proposition 1 is implied by Theorem 1 of Ehlers and Massó (2015). Their result relates the concept of OBNE in an incomplete information setting to Nash equilibrium under complete information for any stable mechanism, and we consider the SD mechanism, which is stable under complete information.

Thus, by Proposition 1, truth-telling is indeed an OBNE, because truth-telling is a weakly dominant strategy under complete information.

4.2. **Truth-telling as the unique OBNE.** Our main theoretical result assumes that the private prior of each student has full support. We then argue that there exists a unique OBNE at which each student submits truthfully: if the strategy of a student i is such that $r_i(P_i) \neq P_i$, we can always construct some preferences for the other students and a priority order such that if i submits P_i she is assigned to a school which is preferred to her assignment with $r_i(P_i)$. Note that the assumption of full support is crucial in this argument because when we construct the preferences and the priority that are needed, we know that every student plays at any OBNE a best response to the strategy of other students.

We introduce the following notation. Let $\pi_j(P_i)$ be the school in position j at P_i.

Theorem 1 (**Uniqueness**). *Let $\tilde{P}$ be a profile of priors such that $\tilde{P}^i_{-i}$ $|P_i$ has full support for each $i \in I$ and $P_i \in \mathcal{P}_i$ with $Pr\{\tilde{P}^i_i = P_i\} > 0$. Then, there exists a unique OBNE, which is the truth-telling OBNE.*

Proof. Consider an OBNE r. The proof is by induction. We show first that in any OBNE students should submit truthfully their first and second options, and then we finish the proof using induction in the number of options. In fact, we only need to prove the statement for the first option, including the second option is to illustrate the construction of the general argument.

Claim 1: At any OBNE, each student submits truthfully the first option of each preferences profile: $\pi_1(P_i) = \pi_1(r_i(P_i))$ for every $i \in I$.

Suppose this is not the case, and consider a student i and preferences $P_i = (s_l, \ldots)$ such that $r_i(P_i) = (s_{j\neq l}, \ldots)$. Given that $\tilde{P}^i$ has full support, we can consider an order of students P_s where i is ranked the first, and for any $r_{-i}(P_{-i})$, student i will not be assigned to s_l, her true first choice under $r_i(P_i)$, because she will be assigned to s_j. This contradicts that $(r_i(P_i), r_{-i}(P_{-i}))$ is a Nash equilibrium under complete information, and then r is not an OBNE. Thus, we should have every student submitting truthfully the first option at any OBNE.

Claim 2: At any OBNE, each student submits truthfully the first two schools: $\pi_j(P_i) = \pi_j(r_i(P_i))$ for every $i \in I$ and $j = 1, 2$.

Let the first two most preferred options of a given preference profile of student i be $P_i = (s_h, s_l, \ldots)$. Consider a profile for other students such that their most preferred school is s_h, and a priority order P_s such that student i is ranked in the $(q_h + 1)$-th position. Given Claim 1, we know that at r every student submits truthfully her first option. Then, the best response of i implies that she should submit truthfully her second most preferred school.

Induction step: Suppose that all students submit truthfully their first $k-1$ options, and consider a preference order for student i, $P_i = (s_1, \ldots, s_{k-1}, s_k, \ldots)$. Construct other students

preference profile such that the first $k-1$ options are the same than the first $k-1$ options of i (but possibly in a different order). Consider a priority order P_s where i ranks as the $\left(\sum_{j=1,\dots,k-1} q_j + 1\right)$-th student. A best response of i to $(r_{-i}(P_{-i}), P_s)$ implies that $\pi_k(P_i) = \pi_k(r_i(P_i))$. Then student i should submit her k-th choice truthfully. □

Theorem 1 says when each student's prior about other students' preferences and schools priority has full support, there exists a unique OBNE at which students submit truthfully. However, the converse does not hold. Following example demonstrates this.

Example 1. Consider a market with two schools $S = \{s_1, s_2\}$, each with a capacity of one seat, and two students $I = \{i_1, i_2\}$. Student i_2's prior has full support, so she submits truthfully at every OBNE. For student i_1 assume that $\tilde{P}^1$ is such that $\text{Prob}\{\tilde{P}^1_1 = (s_1, s_2)\} = 1$, $\text{Prob}\{\tilde{P}^1_{-1} = ((s_2, s_1), P_s)\} = 0$ for all $P_s \in \mathcal{P}_s$, and $\text{Prob}\{\tilde{P}^1_{-1} = ((s_1, s_2), P_s)\} \in (0, 1)$. That is, i_1 believes that with probability 0 i_2's preferences are (s_2, s_1), and that each possible priority ranking has positive probability. Thus, student i_1's prior does not have full support. Note that $r_1((s_1, s_2)) \neq (s_1, s_2)$ is not an equilibrium strategy. Indeed, suppose $P_2 = (s_2, s_1)$ (which implies that $r_2((s_2, s_1)) = (s_2, s_1)$) and that $P_s = (i_1, i_2)$. In this case, the unique best response of i_1 is to submit truthfully. Therefore, the unique OBNE is the truth-telling OBNE but student i_1's prior does not have full support.

Theorem 1 suggests when priors do not have full support, the set of OBNE is not singleton, the revelation game has many equilibria because some students with non-full support priors may not submit truthfully.

4.3. **Submission Strategies.** This subsection discusses non-truth-telling strategies that we may expect to occur. Importantly, we introduce self-selection, a strategic behavior by which a student does not top rank her most preferred school.

The analysis from the previous subsection articulates that self-selection is possible at equilibrium only when a student's prior does not have full support. Conceptually, self-selection is not restricted to be just an equilibrium strategy, it can also be a consequence of strategic mistakes. While we cannot directly observe students' priors, we can however identify, using the final exam scores, those cases where self-selection is compatible with equilibrium strategy. Roughly, consider a student who self-selects because she assigns zero probability to rank high enough in schools' priority to be admitted to her top choice. However, once the uncertainty is resolved it turns out that her position in the priority ranking allows her to be admitted to her most preferred school. In this case, self-selection is not an equilibrium strategy but a consequence of strategic mistakes. The direct

implication of our criterion is that when all self-selected students play equilibrium strategies, the outcome is stable and thus, it coincides with the matching under complete information. Therefore, uncertainty about priorities may only hurt students who self-select due to strategic mistakes.

To link our theory to data and identify students' equilibrium strategies, we introduce the following assumption.

Assumption 1 (**Common Top Choice**). Fix a school $\bar{s}$ and suppose that for every student her prior puts positive probability only on those profiles where all students consider $\bar{s}$ as the best school (and this is known by all the students).[9]

This assumption implies that at every OBNE, and for each student, the probability of being assigned to school $\bar{s}$ depends only on the marginal distribution of the student's prior over the set of schools' priority. Therefore, we can focus on the uncertainty about schools' priorities in the data without loss of generality.

Two main reasons stand out why we focus on the top choice. First of all, the top choice, or the most preferred choice, has profound impact on the future. Submitting the top choice truthfully makes sure a student can get it whenever possible. This is no longer the case if the student instead self-selects as a consequence of a strategic mistake. She can never be assigned to her top choice, which she could, had she submitted the top choice truthfully. Second, as we will see in the empirical part, our data allows us to cleanly identify the top choice for a subset of student without imposing further assumptions.

The next two corollaries describes the decision of each student regarding the top choice $\bar{s}$ at equilibrium. For each student i, let $\tilde{P}^i_s$ denote the marginal distribution of $\tilde{P}^i$ over the set $\mathcal{P}_s$, and $P_s(i)$ the position of student i at P_s.

Corollary 1 (**Submitting Top Choice Truthfully**). *Consider a student i with prior $\tilde{P}^i$ satisfying the following condition:*

$$\sum_{P_s : P_s(i) \leq q_{\bar{s}}} Pr\{\tilde{P}^i_s = P_s\} > 0.$$

Then, at every OBNE, and for every $P_i \in \mathcal{P}_i$ such that $Pr\{\tilde{P}^i_i = P_i\} > 0$

$$\pi_1(P_i) = \pi_1(r_i(P_i)) = \bar{s},$$

where, as before, $\tilde{P}^i_i$ is the marginal distribution of $\tilde{P}^i$ over $\mathcal{P}_i$.

[9]Formally, for every $i \in I$, $\tilde{P}^i$ is such that: $Pr\{\tilde{P}^i = P\} = 0$ if there exists $l \in I$ such that $\pi_1(P_l) \neq \bar{s}$ with $P = (P_l, P_{-l})$.

Corollary 1 suggests that under Assumption 1, those students with priors such that the probability of being in a position on schools' priority lower than or equal to $q_{\bar{s}}$ is positive, should submit truthfully their most preferred school at every OBNE. The proof is by contradiction. Suppose a student i who does not top rank $\bar{s}$ at an OBNE, and consider a schools' priority which has positive probability according to $\tilde{P}^i$, and where i is placed at one of the first $q_{\bar{s}}$ positions. Then, for any other students' strategies, $\bar{s}$ still has free seats when student i has to choose her school. Clearly, by not top ranking $\bar{s}$, student i will get a less preferred school than by revealing her top school truthfully. Thus, i is not playing a best response to the strategy of other students at the considered OBNE.

When a student's prior is such that she has zero probability to be assigned to $\bar{s}$, then both top ranking and not top ranking $\bar{s}$ may emerge at equilibrium.[10]

Corollary 2 (**Skipping Top Choice Strategically**)**.** *Consider student i who does not top rank $\bar{s}$ at an OBNE. Then, student i's prior is such that the probability of be assigned to $\bar{s}$ is zero.*

Now we are ready to introduce the definition of self-selection.

Definition 2 (**Self-selection**)**.** A student *self-selects* if she does not top rank her most preferred school.

Self-selection is the phenomenon in which some students self-select. Our definition restricts to the top choice of each student. Therefore, self-selection is not only related to priors about schools' priorities but also (and crucially) to the decision of not top ranking the most preferred school. Our definition does not restrict to only equilibrium behavior. It includes both equilibrium strategy and strategic mistakes.

We expect self-selection to arise in the data. This is because, as mentioned in the introduction, the central clearinghouse COMIPEMS explicitly suggests students to consider their future priority when submitting preference. Then, some students who assign zero probability to get a score high enough to be admitted to their true first option may indeed skip that option. Moreover, the message of COMIPEMS may induce some students who believe they have a very low probability to be assigned to their most preferred school, to adjust downward their priors.

5. Description of Data

Our dataset covers the assignment of public high school options to students in Mexico City and its metropolitan area in 2010. It consists of two parts: administrative data on the assignment,

[10]Note Assumption 1 does not exclude the case where students put zero probability on the realization of some priority order.

and survey responses from students. The first part includes basic information about student's age, gender, home location, rank ordered list of preferences, average grade in secondary school, the COMIPEMS exam score, and the assignment outcome. The survey responses shed light on students' family backgrounds, studying habits, opinions on the education quality of their secondary schools, and future aspiration (see Appendix A for details on data construction). In addition, this data is merged with the official secondary school quality index from ENLACE.[11]

5.1. **Options Characteristics.** There are in total 536 options offered to students. Some schools, especially those with professional orientation, host multiple options. Schools report each option's capacity to COMIPEMS before the start of the match, however capacity can be negotiated in situation where there are multiple students with equal score competing for the last seat of an option. We take the total capacity in the main round as the total number of students assigned, which is 230, 074. Options' capacity range from 16 to 3,976. There are 14 high schools operated by UNAM, which are on the top list of capacities, offering on average 2,446 seats.

5.2. **Students characteristics.**

5.2.1. *Match Information.* In 2010, 315, 848 students submitted their application to COMIPEMS. Students submitted on average 9.7 options, and only 3.3% of the students used the full 20 options. This indicates that the Mexico City high school match is not imperiled by the controlled school choice problem, 20 options are sufficient for most students (see Section 6.2 for more information on the length of students' submitted preferences). Students' submitted first choice is on average 11.1 km away from home, and the nearest UNAM high school is about 11.2 km away. To have a more intuitive picture about the distance, we further check that 13.9% of the students have at least one UNAM high schools within walking distance (a radius of 3 km), and 57.2% of students have at least one UNAM high school within a radius of 10 km, which takes approximately half an hour by public transport.

A total of 230,074 eligible students were assigned in the main round, of whom 85,019 (37%) were assigned to their first option, 78,649 (77.7%) were assigned to one of their top 5 options, and 28,529 (12.4%) students remained unassigned. In terms of distance, students were assigned to a school on average 8.2 km away from home.

5.2.2. *Survey Information.* The survey questionnaire is handed out to each student together with the registration form, and student is required to return them together. Survey responses are not

[11]ENLACE stands for National Assessment of Academic Achievement in Schools, a standardized test for primary and secondary schools in Mexico.

compulsory, but are strongly recommended by COMIPEMS. The response rate for the variables we are interested in is high: 81.2% of students reported their family income, 78.1% responded to the type of university wish to attend, 81.1% to parent education level, and 81.2% to the student's expected education level.

The application manual advises students explicitly that their responses to the survey has no impact on their assignment outcome. Therefore, we have no compelling reason to think that the survey information does not reflect the real decision environment faced by the students.

6. Evidence of self-selection

In this section, we first identify a sub-sample of students who agree that their most preferred school is one of the high schools affiliated to UNAM. Then based on this identification, we find evidence of self-selection under the strategy-proof serial dictatorship mechanism.

6.1. **Selected Sample.**

Although UNAM high schools are generally viewed as the best, individual preferences may still differ from the norm. We leverage the survey information. In particular, we select a sample of students using the following question:

"*Which type of university would you like to attend after high school?*"

Students are asked to pick one type from a list of choices including: UNAM, IPN, private universities, technology universities or colleges, and other type of universities. If a student's answer is UNAM, then we consider her top choice is one of the UNAM high schools. A total of 134, 706 students (42.6 % of all registered applicants) responded that they would like to attend UNAM, making them the sample of students whose top choice is one of the UNAM high schools.

The reasons why we use the question about preference for universities to deduce preference for high schools are the following. First, given UNAM is the most competitive and recognized public university in Mexico, UNAM high schools provide the best education quality. Second, UNAM high schools offer easiest access to UNAM. These two points are supported by survey responses.[12] For the second point, a quick look at the UNAM admission statistics in the year 2009-2010, a year before students in our data were making decisions, shows that 24,445 out of 27,987 students from UNAM high schools were admitted (87%) and only 17,813 out of 121,950 from other type of high schools (14.6%) were admitted (DGP, 2010). When we trace the 2010 cohort in our data to university application (see Appendix D Table D.3), students from UNAM high schools take the

[12]Since 2013, COMIPEMS asks students if quality and easy access to universities are their main concerns among others when choosing their top choice. In 2014, for example, 91% students consider quality being one of the main concerns when select top choice, and 66% of the students declare easy access to universities as one of the main reasons.

majority share in UNAM university.[13] One main cause behind this situation is that students from UNAM high schools have priority to get a seat in UNAM over other students. [14]

Two caveats are worth mentioning about our identification. First, our approach is cautious, and potentially neglecting students who prefer other type of universities but nevertheless prefer UNAM high schools as well. Second, our strategy refrains from the potential heterogeneous preferences students may have within the group of UNAM high schools, and the only criterion we impose for truth-telling is just to top rank one of them. Moreover, given that all UNAM high schools have the same priority to be admitted to UNAM at the university level, there is no clear rule to distinguish between them. Thus, we adopt a coarse criterion, under which, we may overlook strategic behaviors within the set of UNAM high schools.[15]

Our selected sample resembles the full sample (see Appendix B for more details on summary statistics). For instance, in terms of average grade, for students in the selected sample it is slightly higher by 0.11 point on a scale of maximum 10. Similarly, they perform slightly better in the final exam (2.5 points on a scale of maximum 128). In addition, they live 1 km closer to the nearest UNAM high school, 2% more students come from households with higher income, and 4% more students have parents with high school diploma. Both selected and full sample have more girls than boys, and there is 4% more girls in our selected sample compared to the full sample.

6.2. **Evidence.** Out of 134,706 students who prefer UNAM high schools the most, 30,308 (22%) do not submit any of the UNAM high schools as their first choice, while 104,398 (78%) do so. Thus, one fifth of the students do not submit their true first option, even though the mechanism in place is strategy-proof. This is our first piece of evidence of *self-selection*.

This evidence is not undermined by the maximal number of 20 options allowed on the preference list. Indeed, students in the selected sample submit on average 10 options, and only around 3.8% submit a list with 20 schools. Moreover, self-selected students submit on average 9 options, and 2.5% use the full list, a even smaller percentage than all students from the selected sample. Therefore it is reasonable to think students' submitted preferences are unconstrained.

[13]A student can also apply to UNAM after attending a private high school. However, the admission rate of these students has little difference from those who attended a non-UNAM public high school, and significantly lower than those from a UNAM high school. Thus, attending a UNAM high school maximizes the probability of being admitted to UNAM (see Appendix D Table D.4).

[14]The other group of university-affiliated elite schools are operated by IPN. However, students who go to IPN high schools do not have priority to IPN University. Thus, the same arguments as with UNAM may not hold for the case of the high schools operated by IPN.

[15]As we mention in Section 3, UNAM high schools require a minimum average grade in secondary school of 7 in order to be eligible. Thus, one might be concerned that some students, even when their true first choice is a UNAM high school, do not top rank it because of this eligibility constraint. One of the robustness check in Appendix C deals with this concern, and shows that the results are still valid after considering this requisite.

Next, we compare the characteristics of the self-selected students with truth-telling students. After removing missing observations, the sample reduces to 106,623 students, with 21% of self-selected students. Panel A of Table 1 shows that self-selected students have a mean average grade of 7.9 from a scale of 10, 0.4 point lower than truth-telling students. The final score in the COMIPEMS exam is also lower among the self-selected students, by about 9 points out of a scale of 128. Regarding the geographic location, self-selected students live about 1.6 times further from the nearest UNAM high school.

Panel B summarizes variables related to socio-economic backgrounds. Family income is divided into three levels: low, middle and high income. Low income students account for 47% of the self-selected, up by about 12% compared to truth-telling students. Another important variable often received attention in empirical analysis is parent's education. Following the literature, we take into account mother's education level, and only use father's education level when the former is not available. The data shows that students whose parents have an education level lower than or equal to primary education, account for about 30% of the self-selected population, higher than the share among those non-self-selected ones by about 12%.

TABLE 1. Self-selection vs truth-telling: key characteristics

	Self-selection		Truth-telling	
Panel A:	Mean	Std.Dev.	Mean	Std.Dev.
Average grade	7.90	0.86	8.30	0.84
Exam score	60.89	18.00	69.86	19.23
Distance to nearest UNAM HS	14.28	9.42	8.73	7.56
Distance to submitted 1st choice	9.56	9.39	11.15	8.44
Panel B:	Freq	Col %	Freq	Col %
Family income				
- Low income	10,392	46.58	29,039	34.44
- Middle income	10,905	48.88	46,272	54.88
- High income	1,012	4.54	9,003	10.68
Parent education				
- $\leq$ Primary	6,713	30.09	15,808	18.75
- Secondary	12,149	54.46	43,301	51.36
- $\geq$ HS	3,447	15.45	25,205	29.89
N	22,309		84,314	

7. What influences self-selection?

The descriptive statistics hint to the correlation between average grade, distance and socio-economic backgrounds with self-selection, this section further explores what influences self-selection using a probit model.

7.1. **Probit model.** We estimate the following equation in Table 2:

$$\Pr\{y_i = 1|\mathbf{x}_i\} = \Phi(\mathbf{x}_i, \beta), \tag{1}$$

where y_i is a binary variable indicating whether student i self-selects (with $y_i = 1$ meaning self-select, i.e. not top rank a UNAM high school), Φ is the cumulative distribution function of normal distribution, and $\mathbf{x}_i$ is a vector of observable characteristics.

Column 1 presents a parsimonious estimation on average grade, distance, and income. All coefficients show the expected signs at 1% significance level in line with the descriptive statistics presented in the previous section. Students with high secondary school grades are less likely to self-select, students from low and middle income families have a higher probability of self-selection than those from high income families, and those who live closer to a UNAM high school tend to self-select less than those who live further away.

Columns 2 and 3 show that, after controlling for students' and their family characteristics, average grade, distance and income are still significant. The first group of controls reveals students' characteristics such as age, gender, whether the student works with salary, and hours studied per week. Age shows a positive and significant relation. Gender also affects self-selection. Male students are less likely to do so compared with female students. Work with salary is a dummy variable which accounts for options outside schooling, and students with paid work may be less motivated to continue schooling or go to UNAM high schools, however this variable is not significant when controlling for family characteristics. The controls for family characteristics contain information about parent's education level, whether the student is from single parent family or no parent, whether indigenous language is mother tongue, the number of siblings and persons at home, whether receives need-based fellowship, and parent's occupation. Students of parents without high school diploma are more likely to self-select, with a significance level of 1%.

Column 4 includes two variables to account for secondary school fixed effects. The first one is school quality. School quality is an education quality rating monitored by ENLACE. We see that students coming from a secondary school with better quality are less likely to self-select. The second variable is the percentage of self-selecting students constructed at school level. This variable

TABLE 2. Probit results

Self-select	(1) Baseline	(2) Students' controls	(3) Families' controls	(4) Schools' controls	(5) Interactions and others
Average grade	-0.418***	-0.389***	-0.390***	-0.399***	-0.322***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.026)
Dist to nearest UNAM HS	0.051***	0.052***	0.050***	0.025***	0.020***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Family income (base: high)					
- Low	0.640***	0.578***	0.390***	0.353***	1.098***
	(0.020)	(0.020)	(0.022)	(0.023)	(0.218)
- Middle	0.371***	0.344***	0.233***	0.195***	0.681**
	(0.020)	(0.020)	(0.021)	(0.022)	(0.216)
Student's characteristics controls					
Age		0.091***	0.082***	0.079***	0.080***
		(0.005)	(0.005)	(0.005)	(0.005)
Male		-0.075***	-0.066***	-0.078***	-0.078***
		(0.010)	(0.010)	(0.010)	(0.010)
Work with salary		0.070**	0.045	0.023	0.022
		(0.023)	(0.023)	(0.024)	(0.024)
Family's characteristics controls					
Parent's education (base: $\geq$ HS)					
- Primary and below			0.320***	0.253***	0.254***
			(0.016)	(0.017)	(0.017)
- Secondary			0.199***	0.164***	0.166***
			(0.013)	(0.014)	(0.014)
Mother tongue = indigenous			0.013	0.002	0.001
			(0.029)	(0.029)	(0.030)
Single parent			0.021	0.002	0.003
			(0.012)	(0.012)	(0.012)
No. of siblings			0.020***	0.008	0.008
			(0.004)	(0.004)	(0.004)
No. of persons at home			0.019***	0.019***	0.019***
			(0.003)	(0.003)	(0.003)
Fellowship			-0.004	0.026	0.025
			(0.014)	(0.014)	(0.014)
Secondary schools controls					
School quality				-0.001***	-0.001***
				(0.000)	(0.000)
Pct of self-selection				0.074***	0.075***
				(0.001)	(0.001)
Constant	1.570***	0.128	0.070	0.199	-0.375
	(0.050)	(0.101)	(0.103)	(0.128)	(0.236)
Hours studied/week	No	Yes	Yes	Yes	Yes
Parent occupation	No	No	Yes	Yes	Yes
Income $\times$ Average grade	No	No	No	No	Yes
Income $\times$ distance	No	No	No	No	Yes
Observations	106623	106623	106623	106623	106623
Pseudo R^2	0.13	0.14	0.15	0.21	0.21

Robust standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 3. Average marginal effects

Self-select	(1) Baseline	(2) Students' controls	(3) Families' controls	(4) Schools' controls	(5) Interactions and others
Average grade	-0.1038***	-0.0957***	-0.0948***	-0.0896***	-0.0897***
	(0.0014)	(0.0015)	(0.0015)	(0.0015)	(0.0015)
Dist to nearest UNAM HS	0.0126***	0.0128***	0.0122***	0.0056***	0.0057***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Family income (base: High)					
- Low	0.1450***	0.1304***	0.0898***	0.0759***	0.0777***
	(0.0037)	(0.0039)	(0.0045)	(0.0045)	(0.0046)
- Middle	0.0749***	0.0704***	0.0503***	0.0397***	0.0408***
	(0.0035)	(0.0036)	(0.0041)	(0.0042)	(0.0042)
Parent's education (base: $\geq$ HS)					
- Primary and below			0.0773***	0.0563***	0.0565***
			(0.0039)	(0.0037)	(0.0037)
- Secondary			0.0460***	0.0355***	0.0357***
			(0.0030)	(0.0029)	(0.0029)
School quality				-0.0002***	-0.0002***
				(0.0000)	(0.0000)
Pct of self-selection				0.0167***	0.0167***
				(0.0002)	(0.0002)
Observations	106623	106623	106623	106623	106623

Robust standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

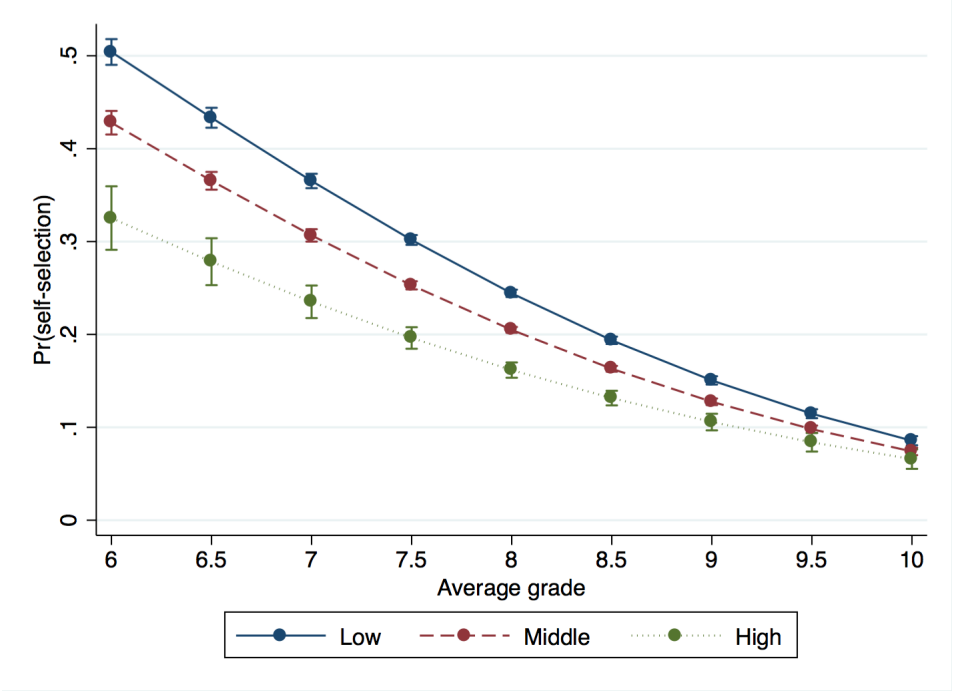
aims at capturing peer effects. The positive relation indicates that self-selection is influenced by the share of students that follow the same strategy in the secondary school.

Column 5 includes further interactions between income and average grade, and income and distance.

To interpret our results, we compute in Table 3 the average marginal effects of main variables on self-selection. The results confirm our theoretic explanation that through influencing priors, average grade plays an important role on self-selection. Take for example the full specification from Column 5, if a student increases her average grade by 1 point (and fixing other variables), then the probability of self-selection decreases by about 9 percentage points. The average marginal effects of income evaluate the differences in probabilities of self-selection when varying a student's family income level. A student coming from a low income family, is about 7.8 percentage points more likely to self-select with respect to someone from a high income family. Finally, it is worth noting that, although significant, distance has a small influence on the probability of self-selection.

Average grade and family income stand out as two driving variables for self-selection. We further use the results from Column 5 to compute the predicted probability of family income for students with average grade from 6 to 10 by a grid of half point. Figure 2 illustrates *income-typical* behavior

FIGURE 2. Predictive margins of income by average grade with 95% CIs



Note: the predicted margins are calculated for each income group, by substituting the observation's average grade with grade 6, 6.5, 7 and so on. The whiskers are the confidence intervals for the predicted margins.

and *performance-typical* behavior. Overall, students from low income family backgrounds are more likely to self-select. However, as grade improves, the gaps across income groups are narrowed down. In fact, the middle income group behaves almost the same as high income group when students grades belong to the top 10% (higher than 9). This indicates that better performance convinces students to submit top choice truthfully, yet it happens more often for middle and high income students than for their low income counterparts.

To summarize, the evidence presented in this section shows that students are more likely to self-select if they have a poor performance in the secondary school. Moreover, past grades have different effects on self-selection across income groups: given the same grade, those students from low economic backgrounds tend to self-select more often, even if for the same grades families with high economic background and more educated do not. We perform additional robustness checks in Appendix C, and we show that the main variables affecting self-selection remain important and significant.

TABLE 4. Distangling self-selection

	Strategic mistake		Equilibrium strategy	
	Freq	Row %	Freq	Row %
Low income	2,261	16.83	11,170	78.83
Middle income	3,592	26.66	9,883	69.84
High income	500	41.95	692	52.99
N	6,353	22.61	21,745	77.39

8. STRATEGIC MISTAKES AND EQUILIBRIUM STRATEGY

Self-selection is a strategy by which a student does not top rank her most preferred school. In this section we first disentangle those cases where self-selection is an equilibrium strategy from those where it is a strategic mistake. Then, after correcting strategic mistakes we compute the matching that results when all students play equilibrium strategies.

Our theoretical framework shows that self-selection is possible at equilibrium only when students' priors do not have full support. In particular, when there is one high school which is the most preferred by all students, self-selection can arise at equilibrium only if the student assigns zero probability to be in a position on schools' priority lower than or equal to the capacity of the school (Corollary 1).

We say that self-selection is a strategic mistake for a student if she self-selects and finally obtains a score high enough to enter one of the UNAM high schools. Self-selection in this case is not compatible with equilibrium and it may harm the student in terms of her final assignment.

When students self-select, we do not observe what are their most preferred UNAM high schools. Thus, we need to agree on a proper threshold in the definition of mistakes. We take the minimum acceptance threshold among all UNAM high schools in 2010 as benchmark score (74). This method is in line with our empirical identification, treating all UNAM high schools as a single high school and the minimum threshold of all UNAM high schools becomes the acceptance threshold of the new single school.[16] This gives us the following criteria.

Definition 3 (**Equilibrium Strategy vs Strategic Mistakes**)**.** For a self-selected student, if her final exam score is lower than 74, then self-selection is an equilibrium strategy. Otherwise, we say that self-selection is a strategic mistake.

[16]We experimented with other criteria including the mean and maximum of all UNAM high schools' thresholds. When taking the mean score 88 as benchmark, more than 7% of the self-selected students make strategic mistakes. The maximum score is 101, given this criterion, about 1.5% of the students self-select as a consequence of strategic mistakes. This last ratio provides a very conservative lower-bound for scope of strategic mistakes.

Table 4 shows nearly 23% of self-selected students make strategic mistakes, whereas about 77% of the students are playing equilibrium strategies consistent with Corollary 2. This finding implies stability of the matching is under threat since almost 23% of the self-selected students may form a blocking pair with a school.

The share of self-selection due to strategic mistakes increases as we move from low income to high income: 17%, 27%, and 42% among low, middle and high income students, respectively. This can be explained by the fact that students from better socio-economic backgrounds obtain better scores, and as a result self-selection is more likely due to strategic mistakes.[17]

8.1. **Counterfactual matching.** Among the students currently admitted to all UNAM high schools, 23.5% are from low income backgrounds. Comparing to the population who regard UNAM high schools as their most desirable choices, the low income students are under-represent by 15 percentage points. Given this discrepancy, we are intrigued to ask: if there are no strategic mistakes, will the social diversity within UNAM high schools become more representative of the given population?

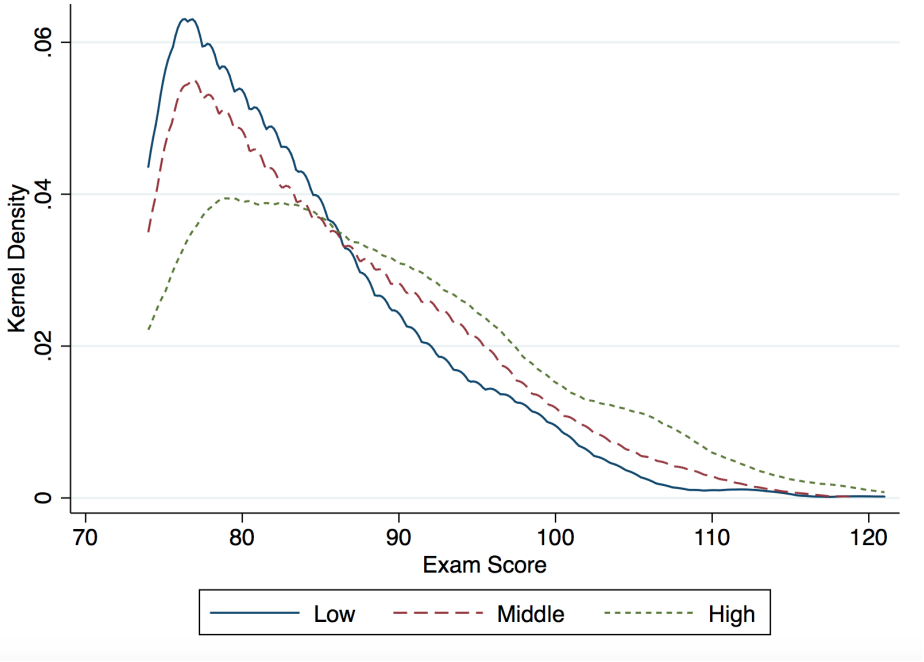
To address this concern of low participation from low income students, we simulate a new matching with no strategic mistakes assuming all students play truthfully (which we call equilibrium matching). As Section 4.3 explained, the outcome under this hypothesis coincides with the one under complete information. Thus, the scenario without mistakes may be also viewed as a situation where students have complete information about schools' priorities.

Why do we expect a change of social mix within UNAM high schools after correcting strategic mistakes? First, within all self-selected students making strategic mistakes, high scored students from low income backgrounds do not perform much differently with respect to those from the other income groups (see Figure 3). Second, low-income students takes a sizable share, representing almost 36% of all students make strategic mistakes (compared to 8% high income).

Table 5 confirms such a change in income distribution between the current and the new equilibrium matching. In the equilibrium matching the participation of students from low income families increases. In particular, by comparing the change in the number of students that are assigned to a UNAM high school in the equilibrium matching for each income group, low income students are most impacted. In the new matching, participation increases by 5.2% and 1.6% for low and middle income groups, respectively. Whereas the number for high income students is reduced by 0.5%. In terms of the social diversity within UNAM high schools, this means that the share of students from

[17] Students from low income group achieve a mean exam score of 61.27 with 17.98 standard deviation, students from middle income group obtain a mean exam score of 68.58 and standard deviation 19.17, and finally students from high income group have a mean exam score of 77.7 with 19.72 standard deviation.

FIGURE 3. Density distribution of the score for students who make strategic mistakes by income



low income families increases now by about 1 percentage point comparing to the current matching. Based on the counterfactual exercise, a change in the timing of preference submission after students learn their scores (as a way to eliminate strategic mistakes due to incomplete information on priorities) will benefit those from low socio-economic backgrounds.

TABLE 5. Social mix

	Current matching	Equilibrium matching	Δ %	Assigned at both matchings	Rejected at new matching	Newly assigned
Low income	6,902	7,259	5.2	6,020	882	1,239
Middle income	17,532	17,804	1.6	15,595	1,937	2,209
High income	4,897	4,873	-0.5	4,515	382	357
Missing obs	4,919	4,314		4,253	666	61
N	34,250	34,250		30,653	3,267	3,867

Notes:
(1) Δ% stands for the percentage change.
(2) Rejected is the number of students who were assigned in the current matching but not in the new matching. As there are 666 rejected students who we do not observe their family income, we assign these students with missing information to each income by assuming the income distribution in the whole population. The ratios are taken from the year book of UNAM high schools (DGP, 2011). This adjustment results the net change at low, middle and high income level to be $248, -194, -104$, with a percentage net change of $3.6, -1.1, -2.1$ respectively. Through different measurements, the access for low income students always improves.

9. Conclusions

Economic theory has played an increasingly important role in designing real life matching markets such as school choice. The standard literature in school choice assumes complete information: students know each others' preferences, and their priorities at schools. However, complete information may not always happen in practice, and yet practitioners have to design the market through trial and error, and often on *ad hoc* basis.

In this paper, we explore uncertainty on students' priorities at schools. It is motivated by the high school match in Mexico City, where students have to submit their preferences before priorities are known. Our theory first characterizes, using an incomplete information framework, truth-telling as the unique equilibrium if students' priors have full support. Importantly, this suggests that even when the mechanism in place is strategy-proof, strategic behaviors may happen at equilibrium, which in turn may create a loss of information preventing us from treating submitted preferences as true preferences.

Then, we confirm with data the existence of one important strategic behavior, self-selection. We found that self-selection concentrates more on the low socio-economic group, that is on those students coming from low income families, and whose parents have low education levels. This raises further concerns when high-achieving students from low socio-economic group also self-select. We attribute the strategy of self-selection to uncertainty exposed to students or parents, rather than intrinsic preferences. Therefore, changing the timing of submission after knowing their priorities, as a way to eliminate strategic mistakes, can improve the access of these students.

By studying in details the Mexico City high school match, we found an alerting phenomenon of self-selection which is so far neglected in designing school choice. Centralized school choice using strategy-proof mechanism is designed to provide students from all socio-economic backgrounds with equal opportunities to attend good schools, contrary to the widely debated school choice programs based on catchment areas. However, the evidence of self-selection makes us to ponder if this goal is fulfilled, and suggests the importance of some details such as timing of submission in school choice design. Our findings suggest a change of the timing from submitting preferences before the exam to after knowing priorities. If this change is indeed difficult, for example, due to logistics reasons, then the clearinghouse has to be very clear in giving their advice. In our current case, COMIPEMS should clearly advise students to think only about their preferences, and not about the future realization of their priorities.

References

Abdulkadiroglu, A., and T. Sönmez (2003): "School choice: A mechanism design approach," *The American Economic Review*, 93(3), 729–747.

Avery, C., C. Hoxby, C. Jackson, K. Burek, G. Pope, and M. Raman (2006): "Cost should be no barrier: An evaluation of the first year of Harvard's financial aid initiative," Discussion paper, National Bureau of Economic Research.

Balinski, M., and T. Sönmez (1999): "A tale of two mechanisms: student placement," *Journal of Economic theory*, 84(1), 73–94.

Budish, E., and E. Cantillon (2012): "The Multi-unit Assignment Problem: Theory and Evidence from Course Allocation at Harvard," *American Economic Review*, 102(5), 2237–71.

Chakraborty, A., A. Citanna, and M. Ostrovsky (2010): "Two-sided matching with interdependent values," *Journal of Economic Theory*, 145(1), 85–105.

Chen, Y., and T. Sönmez (2006): "School choice: an experimental study," *Journal of Economic theory*, 127(1), 202–231.

DGP (2010): "Perfil de Aspirantes y asignados a bachillerato y Licenciatura de la UNAM 2009-2010," *Dirección General de Planeación - UNAM*.

——— (2011): "Perfil de Aspirantes y asignados a bachillerato y Licenciatura de la UNAM 2010-2011," *Dirección General de Planeación - UNAM*.

Dillon, E. W., and J. A. Smith (2013): "The determinants of mismatch between students and colleges," Discussion paper, National Bureau of Economic Research.

Ehlers, L., and J. Massó (2007): "Incomplete information and singleton cores in matching markets," *Journal of Economic Theory*, 136(1), 587–600.

——— (2015): "Matching Markets under (In) complete Information," *Journal of Economic Theory*, forthcoming.

Fack, G., J. Grenet, and Y. He (2015): "Beyond Truth-Telling: Preference Estimation with Centralized School Choice," .

Featherstone, C., and M. Niederle (2013): "Improving on strategy-proof school choice mechanisms: An experimental investigation," Discussion paper, Working Paper, Stanford University.

Gale, D., and L. S. Shapley (1962): "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, 69(1), 9–15.

Hoxby, C., and S. Turner (2015): "What High-Achieving Low-Income Students Know About College," Discussion paper, National Bureau of Economic Research.

Hoxby, C. M., and C. Avery (2012): "The missing" one-offs": The hidden supply of high-achieving, low income students," Discussion paper, National Bureau of Economic Research.

INEE (2011): "Estructura y Dimensión del Sistema Educativo Nacional," *Panorama Educativo de México*.

——— (2012): "Panorama educativo de México," *Indicadores del Sistema Educativo Nacional 2010. Educación básica y media superior*.

Liu, Q., G. J. Mailath, A. Postlewaite, and L. Samuelson (2014): "Stable Matching with Incomplete Information," *Econometrica*, 82(2), 541–587.

Pais, J., and Á. Pintér (2008): "School choice and information: An experimental study on matching mechanisms," *Games and Economic Behavior*, 64(1), 303–328.

Pallais, A. (2013): "Small differences that matter: mistakes in applying to college," Discussion paper, National Bureau of Economic Research.

Roth, A. E. (1989): "Two-sided matching with incomplete information about others' preferences," *Games and Economic Behavior*, 1(2), 191–209.

——— (2008): "What Have We Learned from Market Design?*," *The Economic Journal*, 118(527), 285–310.

Appendix A. Data Construction

A.1. **Distance.** First, we use the coordinates of post codes as students' home location. COMIPEMS collected information on the post codes where students reside. In Mexico city and its surrounding, each post code refers to a neighborhood, known as "colonias", usually consists of a few streets. In total, 3,845 post codes are reported by all students, 3,146 codes can be correctly retrieved their coordinates. The rest of 699 are wrong codes, or codes which do not match the reported neighborhood name. For the affected students (4.3% of total students), we use their secondary school's coordinates as proxy for their home locations, Admission to secondary schools is based on catchment areas, meaning students attend nearby secondary schools. Therefore, the location of secondary school is the best proxy for the location of students' home. The geographic coordinates for secondary schools and high schools are obtained from the Secretary of Public Education.

Finally, we use the Google Distance Matrix Application Programming Interface (API) and Python to compute the walking distance between students' home and high school options.

A.2. **Income.** We reclassify the original 15 family monthly income categories into 3 levels: low income (below 232 USD in Mexico City, and below 165 USD in Mexico State), middle income (from 232 to 746 USD in Mexico City, and from 165 to 630 USD in Mexico State), and high income (above 746 USD in Mexico City, and above 630 in Mexico State). In the new classification, low income families have a monthly income which is equal to the bottom 10 % of Mexico City and Mexico State by the standards of the 2010 Council of Social Development Assessment in Mexico City.[18]

[18]Source: http://www.evalua.df.gob.mx/encuestas.php

Appendix B. Students' characteristics

Table B.1. Students' characteristics of full and selected samples

	Full		Selected sample	
Panel A:	Mean	Std.Dev.	Mean	Std.Dev
Number of submitted options	9.87	3.80	10.00	3.86
Age	15.25	1.16	15.24	1.10
Average grade	8.11	0.86	8.22	0.86
Exam score	65.42	19.82	68.01	19.32
Distance to nearest UNAM HS	11.10	8.76	9.90	8.30
Distance to submitted 1st choice	11.06	9.27	10.81	8.64
No. of siblings	1.96	1.32	1.85	1.26
No. of Persons at home	4.87	1.60	4.78	1.55
School quality	520.65	45.06	525.95	47.00
Pct of self-selection	9.67	6.021	8.92	6.16
Panel B:	Freq	Col%	Freq	Col%
Gender				
- Female	103,335	52.00	59,928	56.21
- Male	95,383	48.00	46,695	43.79
Family income				
- Low	82,613	41.57	39,431	36.98
- Middle	101,320	50.99	57,177	53.63
- High	14,785	7.44	10,015	9.39
Parental education				
- $\leq$ Primary	49,863	25.09	22,521	21.12
- Secondary	104,704	52.69	55,450	52.01
- $\geq$ HS	44,151	22.22	28,652	26.87
Work with salary	9,811	4.94	4,415	4.14
Mother tongue = indigenous	6,207	3.12	2,629	2.47
Single Parent	52,029	26.18	28,142	26.39
Fellowship	29,663	14.93	16,678	15.64
N	198,718		106,623	

Note: Column 2 to 3 report descriptive statistics for the full sample after removing missing observations in line with probit regression. Column 4 to 5 are for the selected sample, i.e. students who wish to attend UNAM.

Appendix C. Robustness check

We perform additional checks for robustness of our main empirical findings.

The first concern rises from the fact that UNAM high schools require a minimum average grade of 7 in the secondary school for admission. Thus, some students may self-select with the fear of not being able to fulfill the minimum grade. Column 1 of Table C.1 considers only those students with secondary grade higher than or equal to 7. Results show that in this restricted sample, average grade, distance and family income are still significant.

Table C.1. Robustness check for probit results

Self-select	(1) Min 7	(2) # COMIPEMS exams	(3) Average distance	(4) Teachers' attention	(5) Aspiration
Self-select					
Average grade	-0.239***	-0.320***	-0.324***	-0.321***	-0.299***
	(0.028)	(0.026)	(0.026)	(0.026)	(0.026)
Dist to nearest UNAM HS	0.021***	0.021***		0.020***	0.021***
	(0.002)	(0.002)		(0.002)	(0.002)
Family income (base: High)					
- Low	0.927***	1.102***	1.101***	1.085***	1.106***
	(0.239)	(0.219)	(0.219)	(0.220)	(0.221)
- Middle	0.641**	0.686**	0.679**	0.667**	0.700**
	(0.237)	(0.217)	(0.217)	(0.217)	(0.219)
Parent's education (base: $\geq$ HS)					
- Primary and below	0.276***	0.254***	0.249***	0.251***	0.219***
	(0.018)	(0.017)	(0.017)	(0.017)	(0.017)
- Secondary	0.179***	0.165***	0.163***	0.163***	0.141***
	(0.015)	(0.014)	(0.014)	(0.014)	(0.014)
No. of COMIPEMS exams taken		0.033*			
		(0.014)			
Avg dist to UNAM HS			0.016***		
			(0.002)		
Expected education < postgraduate					0.259***
					(0.010)
Teachers' evaluation	No	No	No	Yes	No
Observations	100035	106121	106623	105229	104198
Pseudo R^2	0.21	0.21	0.21	0.21	0.22

Robust standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Second, some students participated more than once in the COMIPEMS exam, and they may have different application strategies than the first-time applicants. Column 2 includes a variable measuring the number of times a student has taken the COMIPEMS exam, and the more one has taken, the more likely she will self-select.

The third concern relates to our construction of distance using the postal codes of students' home (see more details in Appendix A). The importance of distance may be underestimated as the result of measurement bias. Moreover, our definition of self-selection is agnostic about the exact choice for UNAM high school, however in the main regression we take the distance to the nearest UNAM high school for each student, imposing that the distance to nearest UNAM high school actually reflects the real traveling cost faced by students. For these reasons, Column 3 uses another variable, the average distance to all UNAM high schools, while keeping everything else the same as in the full specification from Column 5 of Table 2. The term average distance to all UNAM high schools now has a smaller impact, but this change does not undermine the significance of average grade nor family income.

Column 4 adds an additional variable taken from the survey where students are asked to rate how frequently their teachers evaluate their studies. If a teacher evaluates students almost all the time, implying the teacher pays a high attention to students. The coefficients are omitted due to insignificance.

The last robustness check introduces the student's expected education level, which measures students aspiration level. A student who wishes to reach a higher level of study can be more motivated to top rank an elite high school, therefore controlling for expected education level could reduce the impact of average grade and family income. Additionally, the chance to be admitted in postgraduate studies in Mexico is higher if the student graduates from UNAM, and the chance to go to UNAM is higher if the student goes to UNAM high schools.

Table C.2 reports the average marginal effects of our robustness checks. As we expected, the impact of average grade and family income declines after adding robustness controls, however it is still important and significant.

TABLE C.2. Robutness check for average marginal effects

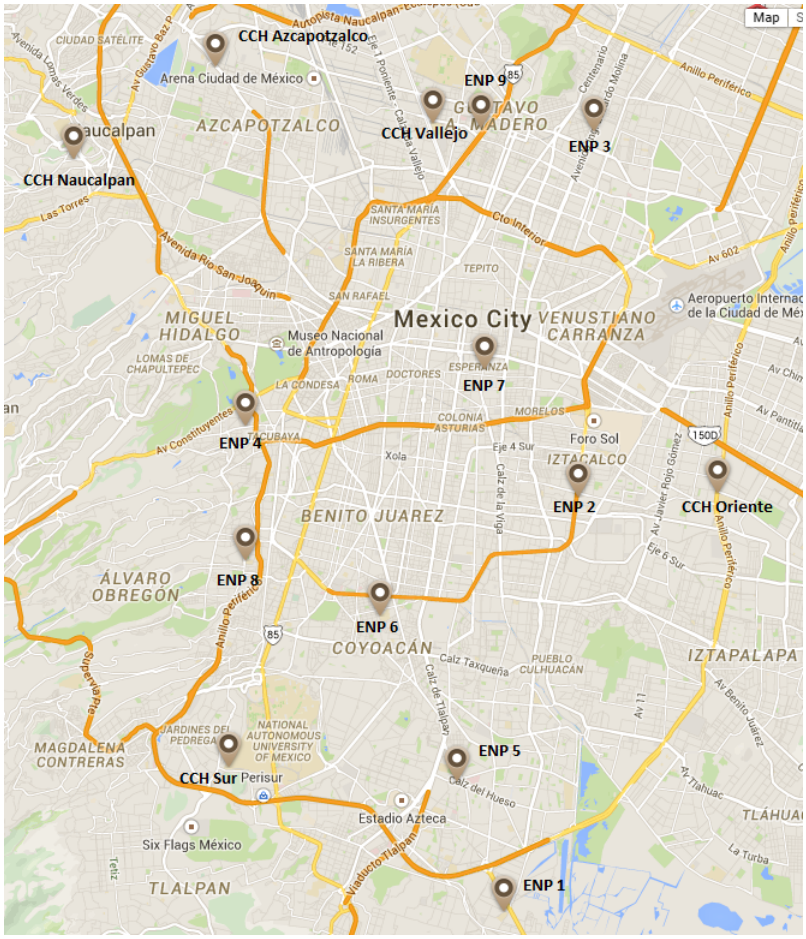
	(1) Min 7	(2) # COMIPEMS exams	(3) Average distance	(4) Teachers' attention	(5) Aspiration
Average grade	-0.065***	-0.089***	-0.090***	-0.089***	-0.085***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Dist to nearest UNAM HS	0.006***	0.006***		0.006***	0.006***
	(0.000)	(0.000)		(0.000)	(0.000)
Family income (base: High)					
- Low	0.071***	0.078***	0.078***	0.077***	0.069***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
- Middle	0.036***	0.041***	0.041***	0.040***	0.035***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Parent's education (base: $\geq$ HS)					
- Primary and below	0.058***	0.057***	0.055***	0.056***	0.048***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
- Secondary	0.036***	0.036***	0.035***	0.035***	0.030***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
No. of COMIPEMS exams taken		0.007*			
		(0.003)			
Avg dist to UNAM HS			0.004***		
			(0.001)		
Expected education < postgraduate					0.057***
					(0.002)
Observations	100035	106121	106623	105229	104198

Robust standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

THE FOLLOWING APPENDICES ARE FOR ONLINE PUBLICATION

Appendix D. Information on UNAM high schools.

Figure D.1. Location of UNAM high schools



Note: Each dot marks the location of the corresponding UNAM high school. Legends correspond to those presented in Table D.2

TABLE D.1. Cutoffs by organizing institutions

	2010			**2011**		
	Min	Max	Mean	Min	Max	Mean
COLEGIO DE BACHILLERES	31	75	55	31	77	54
CONALEP	31	71	40	31	71	39
SE CONALEP ESTADO DE MEXICO	31	67	34	31	66	33
DIRECCIN GENERAL DEL BACHILLERATO	60	65	63	59	64	62
DGETA	31			31		
DGETI	31	67	47	31	70	45
IPN	75	99	82	77	101	84
UNAM	74	101	88	77	107	91
SE	31	85	46	31	87	44
UAEM	81			83		

Note: Cutoff is defined as the lowest exam score between all admitted students. By definition is between 31 and 128.
Source: COMIPEMS.

TABLE D.2. Capacity and demand by UNAM high school

	2008			2009			2010			2011		
High School	Quota	Demand	Ratio	Quota	Demand	Ratio	Quota	Demand	Ratio	Quota	Demand	Ratio
ENP 1 Gabino Barreda	1,200	7,471	6.2	1,200	7,416	6.2	1,200	7,718	6.4	1,200	7,419	6.2
ENP 2 Erasmo Castellanos Quinto	1,200	16,008	13.3	1,200	16,528	13.8	1,200	16,856	14.0	1,200	17,093	14.2
ENP 3 Justo Sierra	1,300	8,987	6.9	1,300	8,653	6.7	1,300	8,052	6.2	1,300	7,876	6.1
ENP 4 Vidal Castañeda y Nájera	1,500	4,552	3.0	1,500	4,531	3.0	1,500	4,402	2.9	1,500	4,624	3.1
ENP 5 José Vasconcelos	2,667	11,673	4.4	2,667	12,165	4.6	2,667	12,177	4.6	2,667	12,667	4.7
ENP 6 Antonio Caso	1,631	15,476	9.5	1,623	15,745	9.7	1,623	16,721	10.3	1,623	19,637	12.1
ENP 7 Ezequiel A. Chávez	1,812	3,977	2.2	1,820	3,966	2.2	1,820	3,989	2.2	1,820	3,955	2.2
ENP 8 Miguel E. Schulz	1,810	6,223	3.4	1,810	6,516	3.6	1,810	6,507	3.6	1,810	6,294	3.5
ENP 9 Pedro de Alba	1,880	19,984	10.6	1,880	20,440	10.9	1,880	20,756	11.0	1,880	21,939	11.7
CCH Azcapotzalco	3,600	10,738	3.0	3,600	10,422	2.9	3,600	9,861	2.7	3,600	10,258	2.8
CCH Naucalpan	3,600	10,673	3.0	3,600	10,735	3.0	3,600	10,729	3.0	3,600	11,303	3.1
CCH Vallejo	3,600	6,755	1.9	3,600	6,777	1.9	3,600	6,186	1.7	3,600	6,474	1.8
CCH Oriente	3,600	15,617	4.3	3,600	16,089	4.5	3,600	15,814	4.4	3,600	16,293	4.5
CCH Sur	3,600	9,422	2.6	3,600	9,384	2.6	3,600	9,094	2.5	3,600	9,434	2.6

Note: The demand for each high school is computed as the number of students that top rank that high school. Ratio stands for the demand over the capacity of each high school.
Source: UNAM.

TABLE D.3. Applicant Statistics at UNAM by UNAM High Schools

High School		**UNAM Admission (2013-2014)**			**Position in Cutoff Order**						
ENP (Plantel)	**Assigned 2010**	**Applications**	**Assigned**	%	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
1 Gabino Barreda	1,359	1,192	854	71.64	7	7	7	7	7	7	7
2 Erasmo Castellanos Quinto	1,218	1,486	1,086	73.08	2	2	3	2	2	2	3
3 Justo Sierra	1,425	1,374	961	69.94	5	5	5	6	6	6	6
4 Vidal Castañeda y Nájera	1,565	1,363	870	63.83	17	18	17	16	15	15	18
5 José Vasconcelos	2,692	2,490	1,765	70.88	6	6	6	5	5	5	5
6 Antonio Caso	1,756	1,432	1,099	76.75	1	1	1	1	1	1	1
7 Ezequiel A. Chávez	1,903	1,654	1,040	62.88	12	13	13	11	9	8	10
8 Miguel E. Schulz	1,847	1,560	1,098	70.38	11	12	12	10	11	10	9
9 Pedro de Alba	1,886	1,677	1,199	71.50	4	4	4	4	3	3	4
Sub-Total	*15,651*	*14,228*	*9,972*	*70.09*							
CCH (Plantel)											
Azcapotzalco	3,976	3,294	2,440	74.1	27	26	28	26	23	21	23
Naucalpan	3,837	3,187	2,509	78.7	41	40	40	37	30	27	34
Vallejo	3,643	3,262	2,493	76.4	21	21	21	21	17	16	20
Oriente	3,588	3,399	2,752	81.0	16	17	16	15	14	14	17
Sur	3,555	3,215	2,535	78.8	14	14	14	12	12	12	13
Sub-Total	*18,599*	*16,357*	*12,729*	*77.82*							
Total	**34,250**	**30,585**	**22,701**	**74.22**							

Source: UNAM annual reports and COMIPEMS data.

Note:

(1) Assigned 2010 refers to the number of students assigned to each high school in the 2010 match.

(2) UNAM Admission (2013-2014) refers to the number of applications, assigned students and the percentage of assigned students over the number of applications in the academic year 2013-2014 for the UNAM University, for each of the high schools.

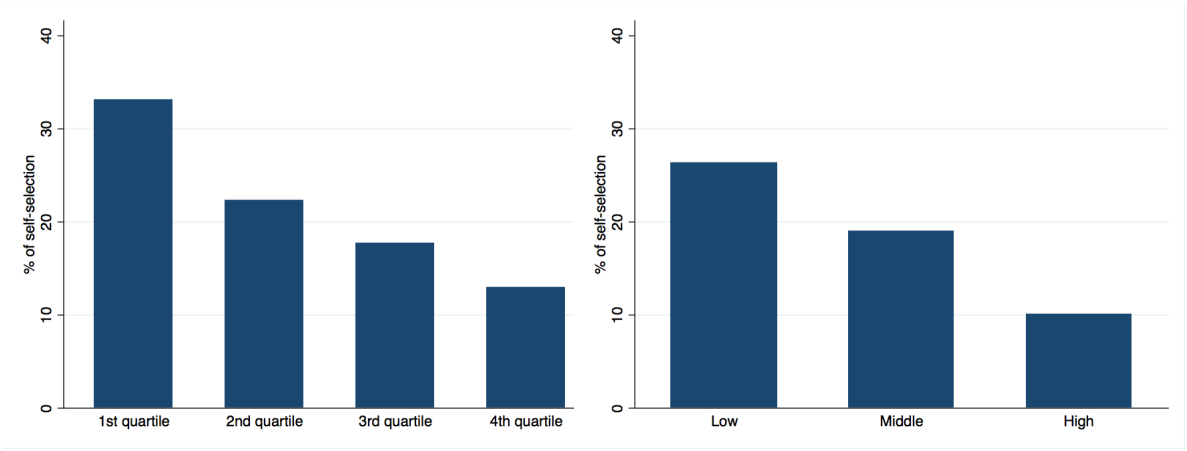
(3) Position in Cutoff Order is defined as follows. For each option we consider the minimum over all the assigned students of the exam score. Then, we order all options using the minimum cutoff, from the highest to the lowest. This information refers to the position of each UNAM high school in this ranking (for example, the high school 603000 is the first high school, with the highest cutoff, in every year during 2008-2012).

TABLE D.4. Admissions to UNAM academic year 2013-2014, by attended high school

	Applications		**Admissions**		**Admission Rate**
Attended high school type		%		%	
UNAM HS	30585	20.4	22701	60.9	74.22
Public non UNAM HS	91150	60.8	10473	28.1	11.49
Private HS	23301	15.5	3400	9.1	14.59
Both Public non UNAM and Private*	4398	2.9	638	1.7	14.51
No information	480	0.3	43	0.1	
Total	149914	100	37255	100	24.85

Source: UNAM Annual report 2013.
*Students who attended some years a public high school and other years a private one.

FIGURE E.1. Self-selection behavior types



(A) Performance-typical behavior (B) Income-typical behavior

APPENDIX E. SELF-SELECTED STUDENTS: MAIN CHARACTERISTICS.

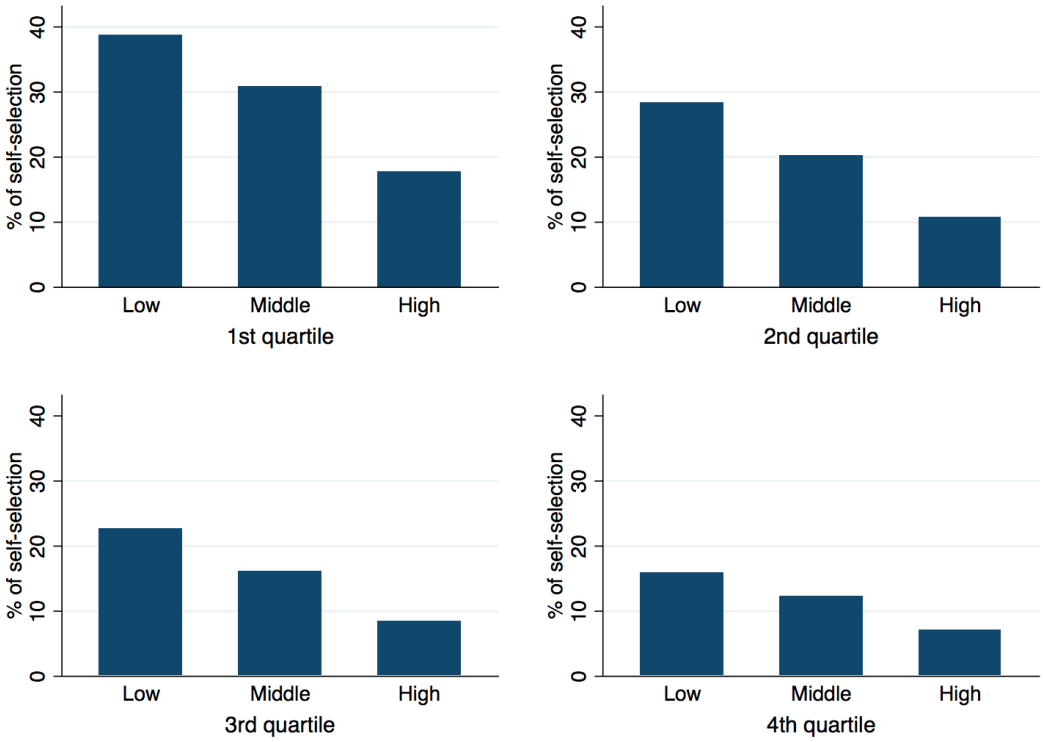
This Appendix takes a closer examination at average grade, income and distance of self-selected students.

E.1. **Average grade and Income.** Average grade measures students' academic performance over the three years of secondary school, on a scale from 0 to 10, with 6 being the minimum pass grade.[19] It serves as a natural benchmark for students to predict their future score in the COMIPEMS exam, and consequently the probability of being admitted to one of the UNAM high schools. A first look into the data confirms this, the Pearson's correlation coefficient between the past average grade and the final COMIPEMS score being 0.36 and significant at 5% level. Figure E.1(a) illustrates indeed that students exhibit this type of *performance-typical* behavior: as average grade increases, the percentage of students that self-select decreases. Consider the bottom quartile of the distribution, 33.1% of the students self-select, this figure drops to 13% when moving to the top quartile students.

We also find another type of application behavior: *income-typical* behavior. As Figure E.1(b) demonstrates, ignoring other influences, 26.4% of students from low income families self-select, 16.3% higher than students from high income families.

[19] The average grade in our data might not be the one students observed at the time of application, since students need to submit their preferences in March, only after the first semester of the last year of secondary school. However, we consider average grade as a good measure for the past grades students have observed. First, to be eligible for the match, students need to present their secondary school certificates which requires an average of 6, thus students still need to get sufficient grades in the last semester, so their grades from the last semester should not be significantly lower than the past. It can be that students were motivated to study harder in the last semester and resulting in higher grade in this semester, but the influence after being averaged out should not be too large neither.

FIGURE E.2. Self-selection by average grade and income



We turn our attention further to the behavior of students by income group at different quartiles of average grade. In particular, we are interested in *high-achieving low-income* students. Following Hoxby and Avery (2012), we define *high-achieving low-income* students as those low income students who achieved the top quartile performance. From Figure E.2 we know that despite having an average grade from top quartile, 15.9% of the low income students self-select, which is similar to the high income students with an average grade from bottom quartile.

E.2. **Distance.** Table E.1 shows first, overall self-selected students face longer traveling distance to nearest UNAM high school than truth-telling ones by about 4.9 km. The gap is much smaller if we only look at students residing in Mexico City, which is not surprising given that almost all high schools operated by UNAM (13 out of 14) are inside Mexico City. However distance alone is not able to explain completely self-selection. If students self-selects because of traveling distance, then we shall expect that they submit a choice closer to home than a nearby UNAM high school, and the data suggests that this is not always the case. If we look at all self-selected students, the distance to their submitted first choice is 4.7 km closer than the nearest UNAM high school. In the State of Mexico, the difference is widened to 7.9 km. If we consider only those students in Mexico

City, we find the opposite: the distance to the submitted first choice is 0.9 km further than the nearest UNAM high school.

TABLE E.1. Self-selection by distance (km)

	All		Mexico City		Mexico State	
Distance to	Self-select	Not	Self-select	Not	Self-select	Not
Submitted 1st choice	9.56	11.14	7.97	7.80	10.29	16.50
	(9.39)	(8.41)	(6.00)	(4.90)	(10.39)	(9.50)
Nearest UNAM HS	14.28	9.42	6.08	5.56	18.19	13.77
	(9.42)	(7.56)	(3.92)	(3.83)	(8.47)	(8.63)

Standard errors are reported in parentheses.