College Choice Allocation Mechanisms: 
Structural Estimates and Counterfactuals

José-Raimundo Carvalho  
CAEN, Universidade Federal do Cearà

Thierry Magnac  
Toulouse School of Economics

Qizhou Xiong  
Toulouse School of Economics

February 14, 2014

Abstract

In this paper, we evaluate a simple mechanism that was commonly used in one university in Brazil to select students at entry and allocate them into various majors. Students first choose a single major and then take exams that either select them in the chosen major or select them out. The matching literature analyzing the student placement issue, points out that this decentralized mechanism is not stable and is not strategy-proof. This means that some pairs of major & students can be made better off and that students tend to disguise their preferences using such a mechanism. We build up a model of performance and school choices in which expectations are carefully specified and we estimate it using cross-section data about medical school choices and entry exams at Universidade Federal do Ceará in northeast Brazil in 2004. Using estimation results, we evaluate changes in selection and students’ expected utilities when other mechanisms are implemented. Results highlight the importance of strategic motives and redistributive effects of changes in the allocation mechanisms.

Keywords: Education, two-sided matching, school allocation mechanism, policy evaluation
1 Introduction

University access in Brazil is a very competitive process and even fiercer if one restricts the analysis to public universities, on average the best institutions: More than two millions of students competed to access one of the 331,105 seats in 2006. In some majors, Medicine or Law for instance, the ratio of applications to available seats can be as high as 20 or more (INEP, 2008). Fierce competition is by no means an exclusivity of the process of entrance into Brazilian universities as many developed and developing countries are in a similar situation (Manski and Wise, 1983). What makes Brazil specific is the formality of the selection process. In contrast to countries such as the United States where the predominant selection system uses multiple criteria (for instance, Arcidiacono, 2005), selection using only objective performance under the form of grades at exams is pervasive in Brazil. More than 88% of available seats are allocated through a vestibular as is called the sequence of exams taken by applicants to university degrees (INEP, 2008).

In this paper, we use comprehensive data on the choices of majors by students and the grades that they obtain at the vestibular of the Universidade Federal do Ceará (UFC thereafter) in Northeast Brazil in 2004 and we concentrate on the specifics of this case. The main characteristics of this vestibular is that it is decentralized at the level of each major. Students choose a single undergraduate major before the exams and compete only against those students who made the same choice. Another interesting characteristics is that the exam consists in two stages. The first stage is common to all majors and is comprised of many multiple-choice testing sub-exams, each one evaluating knowledge in a definite subject, i.e. Mathematics, Portuguese etc.. The second stage exams are specific to each major and have a more traditional short-answer or essay format.

The college admission problem has a long history (Roth and Sotomayor, 1990) and a brief survey of the recent literature is given in Roth (2008). The issue at hand is to match students with colleges which are in our case, the schools offering undergraduate majors at the university.

\[1\]This is a much revised version of a previous paper entitled "College Choice and Entry Exams" by two of the coauthors that has been circulated since 2009. Comments by participants at conferences in Brown, Bristol, Atlanta, Northwestern, Shanghai and Rio de Janeiro and seminars at Oxford, CREST, CEMMAP, Cambridge and Amsterdam are gratefully acknowledged. The research leading to these results has received financial support from CNPq (Project 21207) and the European Research Council under the European Community’s Seventh Framework Program FP7/2007-2013 grant agreement N°295298. The usual disclaimer applies.
(medicine, engineering and so on). In the case in which college preferences are simple\(^2\) and consist in attracting students who are the best in each major, it boils down to what is called student placement (Balinski and Sönmez, 1999). The matching mechanisms which are studied are supposed to satisfy certain properties. First, they could be \textit{stable}, or \textit{fair} in the student placement literature, in the sense that there is no pair (student, college) who would like to block the final allocation in order to improve their lot by matching with another partner. Second, mechanisms could be strategy proof i.e. every student has an interest to reveal her true preferences. Stable mechanisms are not unique and some of them are better for the students and others are better for the schools.

Gale Shapley mechanisms satisfy both properties of stability and strategy-proofness (see for instance Abdulkadiroglu and Sonmez, 2003). In its student optimal version, this mechanism consists in deferring acceptance of students in each major until every student who is interested by this major and who has been rejected by any other major (s)he would have preferred, can be evaluated by this major. Comparing the \textit{vestibular} at UFC to the Gale Shapley mechanism is easy since it turns out that the \textit{vestibular} at UFC corresponds to the first step of the GS algorithm. Students are allowed to propose to their first choice only. The \textit{vestibular} is thus neither stable – i.e. there exist pairs of student & school which could be made better off by changing the final allocation – and it is not strategy proof. Some students prefer to disguise their preferences for very demanded disciplines (e.g. medicine) into preferences for less demanded majors (e.g. dentistry) in order to improve their probability of being accepted. As a matter of fact, the \textit{vestibular} seems also restrictive with respect to another mechanism, the so called Boston mechanism, for which acceptance is immediate instead of being deferred although the mechanism continues beyond the first round by further allocating students who were refused by their prefered school in the previous round. Overall, the form under which the vestibular was organized at UFC is difficult to justify. In a nutshell, this mechanism transforms a centralized allocation mechanism, allocating all students to all majors, into a decentralized system in which each major selects its own students from a blocked list of applications.

What we do in this paper is to contribute to the empirical literature on this subject by evaluating the effects on allocation and welfare of using this \textit{vestibular} with respect to other mechanisms\(^3\)

---

\(^2\)Specifically, it eliminates the need to look at preferences over groups of students (i.e. couples for instance)
that would be inspired by the Gale-Shapley and Boston mechanisms. In order to evaluate these
effects, estimating a structural model is key. Our first contribution is to build such an estimable
model of the Bayesian Nash equilibrium of this game. This equilibrium is shown to exist and is
unique under some restrictions. We then propose an original estimation method in which expecta-
tions about success probabilities are obtained by resampling the Nash Bayesian equilibrium. Our
final contribution is to explore the effects on allocation and welfare of three counterfactual mecha-
nisms of college allocation. Microanalysis allows to study average effects as well as redistributive
effects between schools and between students.

We start by constructing a structural model of college or school choices which slightly diverges
from the literature (Arcidiacono, 2006 and Bourdabat and Montmarquette, 2007, for instance).
Our model focusses on choice probabilities of colleges that depend on (1) expected probabilities of
success and (2) reduced-form preferences for colleges including future wages after college. Precise
information on future college success and wages are lacking in our data. The advantage of our
data lays in the rich information on performances at the two-stage exams before entry (or failure).
We observe grades for all students taking the exam at the two stages of this exam as well as
an initial measure of ability obtained a year before the exam is taken. In consequence, we can
carefully model the probability of success of accessing each school by using assumptions about
students’ expectations of their grades. In this paper, we adopt the assumption that expectations
are perfect (see Manski, 1992, for a critical evaluation of this assumption) and thus that players are
sophisticated. The thresholds above which students are accepted into the majors are the results
of a Bayesian Nash equilibrium of this game in which beliefs are given by what is observed in the
data or could be simulated from them.

We estimate this model of performances & preferences using data made available to us by
UFC that we restrict, for simplicity, to the choice process into two majors in medicine, the most
competitive major field. We study non parametric identification of the model although we use
semiparametric models for the grade equations and parametric models for preferences in the empir-
ical application. The estimation procedure in different steps is original. From grade equations, we
can derive success probabilities in the two schools at any thresholds defining success at each exam
stage. Given the conditions of a Bayesian Nash equilibrium, we can then derive the distribution
of thresholds from any distribution of errors and thus derive the expected success probabilities in
the game by resampling the Bayesian Nash equilibrium. This thus allows the estimation of the choice model as a function of these expected success probabilities.

We then explore normative consequences of these results. We focus on three counterfactual experiments that we perform by recomputing the Nash equilibrium under each counterfactual mechanism. In the first experiment, we cut the number of seats available after the first exam to access the second stage. This tends to reduce organization costs for schools at the risk of losing good students. Second, the counterfactual experiment that seems the most worthy of attention is to give students more choices as in a deferred acceptance mechanism. Students are allowed to state two choices instead of one and in consequence, the result should be stable and students play non strategically. We show that indeed, enlarging the choice set has a positive aggregate effect in terms of utilitarian social welfare but has also distributive effects. The strategic effects are shown to be important. Timing change is the third counterfactual experiment that we perform using that the exam is in two stages. We allow students to choose their majors after passing the first-stage exam instead of choosing before this exam. As expected it has strong redistributive effects between schools and between students.

The paper builds upon various strands in the literature and in particular student placement. In a theoretical work but oriented towards the analysis of a specific mechanism, Balinski and Sönmez (1999) study the optimality of the placement of students in Turkish universities although the selection there concerns all students & colleges throughout the country. Students first write exams in various disciplines and scores are constructed by each college. Colleges choose the weight that they give to different fields: grades in maths can presumably be given more weight by math colleges. They show that this mechanism is suboptimal with respect to the Gale & Shapley mechanism.

Abdulkadiroğlu, Pathak, Roth and Sonmez (2006) exhibited strong empirical evidence of both sophisticated and unsophisticated strategic moves by the parents. Using this evidence, they argued that the Boston School Committee should change to the student proposing deferred acceptance algorithm. Their work was one of the main deciding factors which pushed the Boston School Committee to actually change mechanisms in July 2005. Abdulkadiroğlu, Pathak and Roth (2009) study the mechanisms used in the New York high school system and focus on the trade off between efficiency, strategy-proofness and stability. There are other papers concerned by the analysis of
school choice such as in Lai, Sadoulet and de Janvry (2009). Other analyze the Boston mechanism as He (2012) who uses high school allocation data from Beijing and finds sizeable strategic moves as well. More recent research questions questions the relative standing of the Gale-Shapley and the Boston mechanisms (see Abdulkadiroğlu, Che and Yasuda, 2008 in a school choice problem, Budish and Cantillon, 2010 in a multi-unit assignment problem).

The paper is organized in the following way. Section 2 describes the Vestibular system, the modeling assumptions of the game and in particular expectations, and the conditions of a Bayesian Nash equilibrium. Section 3 presents the econometric model of grade equations and college choices and discuss their non parametric identification. Section 4 provides a descriptive analysis of the mechanism in place and the results of the estimation of grade equations and preference shifters. Section 5 details teh results of the three counterfactual experiments. Section 6 will conclude the paper.

2 Description of the game and modeling

We start by describing how Universidade Federal do Ceara (UFC) selected students in 2004 and we formalize the timing and choices that students make. In a nutshell, students first choose one and only one major\(^3\) to dispute. As already mentioned, the exam consists in two stages. The access to the second stage is conditional on the grade obtained at the first stage and students are selected out of those who have chosen a given major. In a chosen major, are accepted all students above a rank at the first stage exam which makes the number of students who write the second stage exam a multiple (usually 4 sometimes 3) of the number of final available seats. These ranks (one for each major) defines a first stage grade threshold. Similarly, second stage thresholds determine who passes the exam and enters the University. Appendix A gives further details on the mechanism and the exams. Note the assymetry between students who state their preferences at the beginning of the game while preferences of schools as described by grades are only gradually revealed over time.

The first subsection defines notations, formalizes the timing of the events for the students and the primitives of the decision problem. We consider a parsimonious theoretical set-up building up

\(^3\)We use the terms "major", "school" or "program" interchangeably.
from models of college choices. Students are supposed to be heterogenous in their performance at
the two exams and students have preferences over different majors which can be monetary or non-
monetary. Monetary rewards or costs include expected earnings that a degree in a specific major
raises in the labor market.

Choices of students are the result of a game between them and the majors, and in which
information is incomplete. Agents will be assumed to be partially informed about the types of
competing students although they are sophisticated in the sense that they know a lot about fellow
students and the distribution of unobserved heterogeneity in the population. The construction of
this set-up in terms of information sets and expectations is presented in the second subsection.
We then derive the Bayesian Nash equilibrium of this game.

2.1 Timing for the Decision maker

We begin with setting up the notations in which we omit the individual index for readability.
Random variable $D$ takes realizations, $d$, a specific major. For simplicity, we restrict the number
of majors to two whose names are $S$ (for the medical school of Sobral) and $F$ (for the medical
school of Fortaleza) since we will use these two medical schools in the empirical application. The
extension to any number of majors is trivial albeit costly in terms of notations. The outside
option is denoted $d = \emptyset$. Observed student characteristics which affect preferences (respectively
performance or grades) are denoted $X$ (respectively $Z$). The sets of variables, $X$ and $Z$, are
overlapping albeit distinct so as to enable parameter identification (see below).

We describe the Vestibular system with a simple sequence of five stages. At each stage, students
obtain information about grades or make decisions.

- **Stage 0 – Pre Vestibular exam**: A standardized national exam is organized one year
before Vestibular exams begin. It is known as ENEM\(^4\) – a broad-range evaluation that we
shall consider as a measure of students’ ability. The result of this exam is denoted $ENEM$.

Beside proxying for ability, it is also used by the University when computing the passing
thresholds in the Vestibular exams.

\(^4\)ENEM is a non-mandatory Brazilian national exam, which evaluates high school education in Brazil. Un-
til 2008, the exam consisted in two tests: a 63 multiple-choice test on different subjects (Portuguese, History,
Geography, Math, Physics, Chemistry and Biology) and a composition test.
• **Stage 1 – Choice of Major:** Before sitting for the *Vestibular* exams, students apply for one major out of all the available options, $d \in \{\emptyset, S, F\}$. The outside option $d = \emptyset$ implies that one renounces the opportunity to get into the two majors under consideration and either chooses another major, another university or any other alternative. After that stage, students are allocated to two sub-samples which are observed in our empirical application, the first one composed of students choosing $S$ and the second one of students choosing $F$.

• **Stage 1 – First Exam:** All students having chosen majors $S$ or $F$, take the first *Vestibular* exam (identical across majors) and obtain grades. Denote the first exam grade $m_1$, and write it as a function of ability and characteristics of students as:

$$m_1 = m_1(Z, u_1; \beta_1)$$

in which $u_1$ are unexpected individual circumstances that affect results at this exam.

After this first exam, students are ranked according to a weighted combination of grades *ENEM* and $m_1$. Those weights are common knowledge *ex-ante*. The thresholds of acceptance to the second-stage exam are given by the rule that the number of available seats is equal to 4 times the number of final seats offered by the major. The number of final seats is known before the majors are chosen. For instance, the number of final seats in *Sobral* is 40 and thus the number of acceptable students after the first exam is 160.

We write the selection rule after the first exam as:

$$m_1 \geq T_1^{D=d}(ENEM) \text{ for } d \in \{S, F\},$$

in which $T_1^D$ is determined by the number of candidates and positions available in the major.

Students who do not pass the first exam get their outside option $D = \emptyset$, with utility, $V_\emptyset$, which is the best among all possible alternatives, for instance, investing another year preparing for the next year’s *Vestibular*, finding a program outside of the *Vestibular* system, studying abroad or working.

• **Stage 2 – Second Exam:** Students who pass the first exam take the second stage exam (identical across majors) and get a second stage grade, denoted $m_2$:

$$m_2 = m_2(Z, u_2; \beta_2)$$
where $u_2$ is an error term whose interpretation is similar to $u_1$’s and $u_2$ is possibly correlated with $u_1$. These students are ranked again according to a known weighted combination of ENEM, $m_1$ and $m_2$, and students are accepted in the order of their ranks until completion of the positions available for each major. Similarly, we write another selection rule:

$$m_2 \geq T_2^{D=d}( \text{ENEM}, m_1 ) \text{ for } d \in \{S, F\}$$

as a function of a second threshold. Students who fail the second stage exam get the same outside utility as students who fail the first stage exam.

- **College entry:** Finally, students who pass the second stage exam get into the majors and enjoy utility $V_D$, which is determined by their preferences and the income prospects of the major.

There could be additional decision nodes to take into account if preferences were evolving over time. For instance, students could leave the game after choosing majors $S$ or $F$ and before taking exams or after passing the first exam. Passing the first stage exam could give students a way to signal their ability to potential employers or other universities and this would modify the value of the outside option after the first stage. Similar arguments could apply to the second stage exam as well.

We do not have any information on students who quit before the exams since our sample consist only of those who take exams. As for quitting before or after the second stage, it seems hard to model those exits and we have abstracted from these issues by selecting medical schools as our two majors of interest. Only 2 students out of more than 700 who pass the first stage exam quit between the two stages.

This makes the model static and the determination of choices is easy. Define the expected probability of success in major $D$ as:

$$P^D = \Pr(m_1(Z, u_1; \beta_1) \geq T_1^D(ENEM), m_2(Z, u_2; \beta_2) \geq T_2^D(ENEM, m_1)),$$

in which we delay until next section the precise definition of the probability measure that we use since this depends on the definition of information sets and expectations. The expected value of major $D$ is given by:

$$\text{EV}_D = P^D V_D + (1 - P^D) V_o.$$
We can normalize $V_\alpha = 0$ and therefore choices are obtained by maximizing expected utility as:

\[
D = S \text{ if } P^S V^S > P^F V^F,
D = F \text{ if } P^S V^S \leq P^F V^F.
\] (1)

We shall specify in the econometric section, preferences as functions $V^S(X, \varepsilon; \zeta)$ and $V^F(X, \varepsilon; \zeta)$ in which $X$ are observed characteristics, $\varepsilon$ is an unobservable preference random term and $\zeta$ are preference parameters. It is enough at this stage to define choices as $D(X, \varepsilon, \zeta, P^S, P^D)$. For simplicity, we shall assume in the following that $\varepsilon$ and $u = (u_1, u_2)$ are independent. This is a testable assumption that will be evaluated in the empirical section.

### 2.2 Expectations and Bayesian Nash equilibrium

Denote $\beta$ (respectively $\zeta$) the collection of parameters entering grade equations (resp. preferences). The list of those parameters will be made more precise when specifying preferences and analyzing identification. We assume that those parameters are common knowledge among students. Denote also $T = (T^S_1, T^S_2, T^F_1, T^F_2)$ the thresholds that determine the passing of exams (stages are indexed by 1 and 2) in each school (superscripts $S$ and $F$). These thresholds are in general random unknowns at the initial stage since they depend on variables that are random unknowns at the initial stage.\(^5\)

Namely, thresholds affect outcomes in two ways. First, realized thresholds, $t^d_j$, command the entry of students into the schools. Second and as a consequence, student expectations of their success probabilities depend on thresholds and those affect directly their school choices. By assuming that expectations are perfect, expectations of thresholds should match the distribution of their realized values. This is this relationship that we construct now.

#### 2.2.1 Timing of the game and stochastic events

In those models, assumptions about expectations are key because solutions of the model depends crucially on information sets (see Manski, 1992). The timing of information revelation in the game is supposed to be as follows. Before majors are chosen, the number of seats in each school, $n_S$ and $n_F$ are announced and the number of participants, say $n + 1$, is observed. We assume that

\(^5\)We adopt the term random unknowns to signal that the distribution function of those unknowns are common knowledge. Measurability issues are dealt with below.
\( n + 1 \gg n_S + n_F \) because the exam is highly selective. In our data, the average rate of success is 5%. Participants are those who get a positive utility level in applying to one of the two schools of interest\(^6\).

We distinguish somewhat artificially one applicant, indexed arbitrarily by 0, from all other applicants to both schools \( i = 1, \ldots, n \) and we analyze her decision making. This is because we are considering an i.i.d. setting and the model is assumed symmetric between agents (although they differ ex-ante in their observed characteristics and ex-post in their unobserved shocks).

Student 0 observes her characteristics \((Z_0, X_0)\) affecting grades and preferences and the random shocks affecting her preferences \(\varepsilon_0\). Random shocks affecting her grades, \(u_0 = (u_{0,1}, u_{0,2})\) at the two-stage exam later on, remain unobserved but their distribution function, \(F_{u_0}\), is common knowledge as well as the functional forms of grade equations. This observation scheme is also true for characteristics of all other students, \((X_i, Z_i, \varepsilon_i, u_i)\) \(i = 1, \ldots, n\). We assume that characteristics \((X_i, Z_i, \varepsilon_i)\) for all \(i = 1, \ldots, n\) are common knowledge among students as well as the distribution of \(u_i\). The information set of student 0 at the initial stage is thus composed of \(W_0 = (X_0, Z_0, \varepsilon_0)\) and \(W_{(n)} = (X_i, Z_i, \varepsilon_i)_{i=1,..n}\).\(^7\)

After this initial stage, student 0 chooses her major \((D_0 \in \{S, F\})\) as a function of their expectations of success \(P_S^0\) and \(P_D^0\) and other students do as well (say \(D_{(n)}\)) according to equation (1). Later on, the two-stage exams are taken sequentially and students are selected in or out of each school. A realization of the thresholds as a function of observed grades is then computed.

There are two types of risks that student 0 has to face. First, the risks due to random shocks affecting other students’ grades, second the risk induced by her own random shock affecting her grades. The former is described by the random vector \(U_{(n)}\) whose elements are \(u_i, i = 1, \ldots, n\), the latter by \(u_0\). From this, we can derive success probabilities and form what will be the expectations of success of student 0.

\(^6\)We could also study the case in which students do not know the number of competitors when they apply. As we have no element in the data to help us deriving a distribution for this counting variable, we prefer to leave this point aside.

\(^7\)Those assumptions are among the strongest that we could make and they provide a polar case of high sophistication. We leave for further work the developments of less informative frameworks. They seem to be significantly more difficult to estimate.
2.2.2 Success probabilities

Denote $W^S_{(n)}$ (respectively $W^F_{(n)}$) the characteristics of the sub-sample of students $i = 1, \ldots, n$ applying to Sobral (respectively Fortaleza) observed by student 0. By construction $W_{(n)} = W^S_{(n)} \cup W^F_{(n)}$ and $W^S_{(n)} \cap W^F_{(n)} = \emptyset$. Similarly, we denote $U^S_{(n)}$ and $U^F_{(n)}$ the corresponding partition of $U_{(n)}$. We shall see in the next subsection how sub-samples are derived from primitives.

Should Sobral, $S$, be chosen by student 0, her success or failure at Sobral would be determined by the binary condition

$$1\{m_1(Z_0, u_0, \beta) \geq T^S_1(W^S_{(n)}, U^S_{(n)}), m_2(Z_0, u_0, \beta) \geq T^F_2(W^F_{(n)}, U^F_{(n)})\}$$

in which $T^S_1(\cdot)$ and $T^F_2(\cdot)$ are the values of the thresholds at the two-stage exams for a school $d \in \{S, F\}$ when the sample of applicants to this school is described by $W^d_{(n)}$ and the realization of their grade shocks equal to $U^d_{(n)}$. Notice that when evaluating this event, student 0 is considering only the sample of other students than herself. Because of continuously distributed grades, we also neglect ties.

The formal construction of these thresholds is explained below after having determined choices but the intuition is clear for instance for the second-stage threshold. The best $n_S$ ranked students after the final exam are accepted by Sobral and the threshold of the final exam is equal to the grade obtained by the worst-ranked accepted student. Respectively, at Fortaleza the success is determined by $1\{m_1(Z_0, u_0, \beta) \geq T^F_1(W^F_{(n)}, U^F_{(n)}), m_2(Z_0, u_0, \beta) \geq T^F_2(W^F_{(n)}, U^F_{(n)})\}$. To distinguish those thresholds from the ones defined in the complete sample of $i = 1, \ldots, n$ AND $i = 0$ we denote them as:

$$\hat{T}^d_{1,0} = T^d_1(W^d_{(n)}, U^d_{(n)}), \hat{T}^d_{2,0} = T^d_2(W^d_{(n)}, U^d_{(n)}) \text{ for } d = S, F.$$

These thresholds are indexed by 0 since this refers to the construction of expectations of student 0 relative to the sample of other students $i = 1, \ldots, n$.

When student 0 decides upon a school to apply to, she formulates expected probabilities of success by integrating the condition of success with respect to the aggregate source of risk described by $U_{(n)}$ (remember that student 0 observes $W_{(n)}$ only) and with respect to the individual source of risk, $u_0$:

$$P^d_0(Z_0, \beta) = E_{U_{(n)}, u_0}\left[1\{m_1(Z_0, u_0, \beta) \geq \hat{T}^d_{1,0}, m_2(Z_0, u_0, \beta) \geq \hat{T}^d_{2,0}\} \mid Z_0\right],$$

$$= E_{U_{(n)}}\left[p^d(Z_0, \beta, \hat{T}^d_{1,0}, \hat{T}^d_{2,0}) \mid Z_0\right], \quad (2)$$

12
in which the following function concerns the individual shock \( u_0 \) only:

\[
p^d(Z_0, \beta, \tilde{T}^d_{1,0}, \tilde{T}^d_{2,0}) = E \left[ 1\{m_1(Z_0, u_0, \beta) \geq \tilde{T}^d_{1,0}, m_2(Z_0, u_0, \beta) \geq \tilde{T}^d_{2,0} \} \mid Z_0, \tilde{T}^d_{1,0}, \tilde{T}^d_{2,0} \right].
\]  

These are the success probabilities that can be computed from observing a single sample, \( W_n \) when \( \tilde{T}^d_{1,0}, \tilde{T}^d_{2,0} d = S, F \) are equal to their realized values. As the only influence of \( U_{(n)} \) is through these thresholds, those are sufficient statistics and we can rewrite the expected success probabilities as

\[
\left\{ \begin{array}{ll}
P^S_0 = P^S(Z_0, \beta) = E \left[ P^S(Z_0, \beta, \tilde{T}^S_{1,0}, \tilde{T}^S_{2,0}) \mid Z_0 \right], \\
P^F_0 = P^F(Z_0, \beta) = E \left[ P^F(Z_0, \beta, \tilde{T}^F_{1,0}, \tilde{T}^F_{2,0}) \mid Z_0 \right].
\end{array} \right.
\]  

in which risks stemming from the presence of competitors and the individual risk are integrated out. Note that they do not depend on the determinants of the preferences of student 0, \((X_0, \varepsilon_0)\).

Denote \( D_0(Z_0, \varepsilon_0, \zeta, P^S_0, P^F_0) \in \{S, F\} \) the choice of applicant 0 resulting from equation (1). Given that the sample is i.i.d and that 0 is an arbitrary representative element of the sample, \( i = 1, \ldots, n \), we can by substitution construct the samples of applicants to Sobral (say) by using:

\[
W^S_{(n)} = \{ i \in W_n; D_i(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = S \}.
\]

It is thus clear that the application mapping \( W_{(n)} \) into \( W^S_{(n)} \) or \( W^F_{(n)} \) is measurable, therefore that the application mapping \( W_{(n)} \) into thresholds \( \tilde{T}_0 = (\tilde{T}^S_{1,0}, \tilde{T}^S_{2,0}, \tilde{T}^F_{1,0}, \tilde{T}^F_{2,0}) \) is measurable.

### 2.2.3 The Nash equilibrium and the determination of the thresholds

We can now return to the determination of the thresholds \( T \), defined in the complete sample \( i = 0, \ldots, n \) and \( \tilde{T}_0 \) defined in the restricted sample \( i = 1, \ldots, n \).

Starting with \( T \), the equilibrium conditions yield a realization of the thresholds \((t^d_1, t^d_2)_{d \in \{S, F\}}\) for any realizations of \((u_0, u_1, \ldots, u_n)\), are fourfold:

\[
\left\{ \begin{array}{ll}
\sum_{i=0}^{n} [1\{D_i(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = S\}1\{m_1(Z_i, u_i, \beta) \geq t^S_1\}] = 4n_S, \\
\sum_{i=0}^{n} [1\{D_i(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = F\}1\{m_1(Z_i, u_i, \beta) \geq t^F_1\}] = 4n_F, \\
\sum_{i=0}^{n} [1\{D_i(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = S\}1\{m_1(Z_i, u_i, \beta) \geq t^S_1, m_2(Z_i, u_i, \beta) \geq t^S_2\}] = n_S, \\
\sum_{i=0}^{n} [1\{D_i(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = F\}1\{m_1(Z_i, u_i, \beta) \geq t^F_1, m_2(Z_i, u_i, \beta) \geq t^F_2\}] = n_F.
\end{array} \right.
\]  

The first equation translates that given choice \( S \), the number of students admitted after the first-stage exam to the second exam is four times the number of seats available in major \( S \). The
second equation translates the same condition for major \( F \). The third and four equations are the corresponding equilibrium conditions for passing the second-stage exam. The number of students admitted in Sobral is equal to the number of available seats.\(^8\)

As usual with dummy variable equations, this system has many solutions \((t_1^S, t_1^F, t_2^S, t_2^F)\) in an hypercube \( C \) in \( \mathbb{R}^4 \). We retain the solution corresponding to the upper north-west corner i.e. \((\max_C t_1^S, \max_C t_1^F, \max_C t_2^S, \max_C t_2^F)\) and in the absence of ties, this solution is unique. Note that this corresponds to the computation of a finite number of empirical quantiles and in the absence of ties, this is why it yields a unique solution.

Turning to \( \tilde{T}_0 \) we have by the same argument:

\[
\sum_{i=1}^{n} I\{D_i(X_i, \varepsilon_i, \zeta, P_i^S, P_i^F) = S\} I\{m_1(Z_i, u_i, \beta) \geq t_1^S\} = 4n_S,
\]

\[
\sum_{i=1}^{n} I\{D_i(X_i, \varepsilon_i, \zeta, P_i^S, P_i^F) = F\} I\{m_1(Z_i, u_i, \beta) \geq t_1^F\} = 4n_F,
\]

\[
\sum_{i=1}^{n} I\{D_i(X_i, \varepsilon_i, \zeta, P_i^S, P_i^F) = S\} I\{m_1(Z_i, u_i, \beta) \geq t_1^S, m_2(Z_i, u_i, \beta) \geq t_2^S\} = n_S,
\]

\[
\sum_{i=1}^{n} I\{D_i(X_i, \varepsilon_i, \zeta, P_i^S, P_i^F) = F\} I\{m_1(Z_i, u_i, \beta) \geq t_1^F, m_2(Z_i, u_i, \beta) \geq t_2^F\} = n_F.
\]

Notice that the choices of other students \( I\{D_i(X_i, \varepsilon_i, \zeta, P_i^S, P_i^F) = S\} \) are observed in the sample and by student 0 since they depend on observable variables or objects that are common knowledge. Therefore the distribution of \( \tilde{T}_0 \) can be computed using choices and the estimation of grade equations and equations (2) and (3) determine the expectations \( P_0^F \) and \( P_0^S \).

### 2.3 Discussion of the uniqueness of equilibrium

In the current system and the counterfactual experiments below, remains the pending question of the uniqueness of the equilibrium. This equilibrium is defined as a set of choice probabilities and success probabilities that are mutually compatible and compatible with the equilibrium conditions (5).

This property should be proven in each set-up and there is no general result on uniqueness in our setting to our knowledge. Nevertheless, it is possible to prove uniqueness in a simpler context. We assume that the scheme is the current selection scheme with homogenous agents in

\(^8\)There is a minor complication stemming from the fact that applicants could be in too small a number for one of the schools. In this case the threshold is defined in a trivial way by letting the threshold to be 0. The average success probability of 5% in our data means that the probability of this event is negligible.
preferences and performances although we allow here for an arbitrary number of majors. In other words there are no covariates \((X, Z)\) and the model is symmetric between agents because grades
and preferences are affected by i.i.d. shocks.

We define success probabilities, \(\{P^d\}_{d=1,\ldots,D}\), at the first stage exam and \(\{P^d\}_{d=1,\ldots,D}\) at the second stage exam. We assume that in any sample no elements of \(P\) is equal to zero. We pile up these objects into \(D\)-dimensional vectors \(P_1\) and \(P\) and define school choices as \(\{1\{D(\zeta, \varepsilon, P) = d\}\}_{d=1,\ldots,D}\) depending only on overall success \(P\).

School choices are given by the comparison between expected value functions \(\{V^d\}_{d=1,\ldots,D}\) as in equation (1) in which each value function \(V^d\) is independent of the success probability \(P^d\).

Without loss of generality, we can set \(V^d = 0\) when \(V^d < 0\) since we consider a sample in which \(\max_{d=1,\ldots,D} V^d > 0\).

The equilibrium relationships (5) under homogeneity can then be written for any major \(d\):

\[
\sum_{i=0}^{n}\{1\{D(\zeta, \varepsilon, P) = d\}\{m_1(\beta, u_i) \geq t^d_1\}\} \leq 4n_d,
\]

\[
\sum_{i=0}^{n}\{1\{D(\zeta, \varepsilon, P) = d\}\{m_1(\beta, u_i) \geq t^d_1, m_2(\beta, u_i) \geq t^d_2\}\} \leq n_d.
\]

The previous inequalities are equalities when all seats are filled. When they are not, for instance in the second inequality, threshold \(t^d_2\) is equal to zero (and \(t^d_1\) as well if seats after the first stage are not filled).

As thresholds \(t^d_1\) and \(t^d_2\) solve this condition for any realization of \(\{u_i\}_{i=0,\ldots,n}\) we also have by integration over \(u\):

\[
\sum_{i=0}^{n}\{1\{D(\zeta, \varepsilon, P) = d\}\Pr\{m_1(\beta, u_i) \geq t^d_1\}\} = \sum_{i=0}^{n}\{1\{D(\zeta, \varepsilon, P) = d\}\{P^d\}\_1 \leq 4n_d,
\]

\[
\sum_{i=0}^{n}\{1\{D(\zeta, \varepsilon, P) = d\}\Pr\{m_1(\beta, u_i) \geq t^d_1, m_2(\beta, u_i) \geq t^d_2\}\} = \sum_{i=0}^{n}\{1\{D(\zeta, \varepsilon, P) = d\}\{P^d\}\_1 \leq n_d,
\]

or equivalently by defining \(\pi^d(P) = \frac{1}{n}\sum_{i=0}^{n}\{1\{D(\zeta, \varepsilon, P) = d\}\}
\]

\[
\pi^d(P)\{P^d\}_1 \leq \frac{4n_d}{n} \equiv 4\lambda_d,
\]

\[
\pi^d(P)\{P^d\} \leq \frac{n_d}{n} = \lambda_d.
\]

For simplicity, we shall assume that preferences for all majors are sufficiently strong and that
the number of candidates, $n$, is sufficiently large so that inequalities never happen and therefore we obtain the result that:

**Lemma 1** Suppose that the sample under consideration is such that for all $d$, $\min_P \pi^d(P) > 4\lambda_d$. A Bayesian Nash equilibrium necessarily satisfies that seats at the first and second stages for all majors are filled:

$$\begin{cases} 
\pi^d(P)P^d_1 = 4\lambda_d, \\
\pi^d(P)P^d = \lambda_d.
\end{cases}$$

Indices $\{\lambda_d\}_{d=1}^D$ is the fraction of seats in the population attributable to each major. We assume that $\lambda_d \in (0, 1)$, $\sum_{d=1}^D \lambda_d \in (0, 1)$.

By assuming that for all $d$, $\text{Pr}(V^d > 0) > 0$, we have $\pi^d > 0$ for all $P \neq 0$ and $P^d \geq \hat{P}^d \implies \forall d; \pi^d(P) \geq \pi^d(\hat{P})$. Choice probabilities also satisfy adding up:

$$\forall P; \sum_{d=1}^D \pi^d(P^d) = 1.$$

Let $z^d(P) \equiv \pi^d(P)P^d$ and pile up the elements $z^d(P)$ into $z(P)$. The following Lemma shows that equilibrium is unique:

**Lemma 2** Suppose that the sample under consideration is such that for all $d$, $\min_P \pi^d(P) > 4\lambda_d$. Consider any any $(P, \hat{P})$, such that for all $d$, $P^d > 0$, $\hat{P}^d > 0$ and such that $z(P) = z(\hat{P})$. Then $P = \hat{P}$.

**Proof.** The condition $z(P) = z(\hat{P})$ implies that any $d$, $P^d\pi^d = \hat{P}^d\hat{\pi}^d$. We study different cases in which $P \neq \hat{P}$ and show that a contradiction arises.

Consider first that (i) $\hat{P}^d \leq P^d$ for all $d$ and the inequality is strict for at least one $d$. We thus have:

$$P^d\pi^d = \hat{P}^d\hat{\pi}^d \leq P^d\hat{\pi}^d$$

and for one $d$ at least the inequality is strict since for all $d$, $\hat{\pi}^d > 0$. Thus as $P^d > 0$, $\pi^d \leq \hat{\pi}^d$ and one inequality at least is strict. It is a contradiction with $\sum_{d=1}^D \pi^d = 1$. Case (ii) in which $\hat{P}^d \geq P^d$ and one inequality is strict leads to a similar contradiction.

Therefore its is sufficient to consider case (ii): for all $d \in I$, $\hat{P}^d < P^d$ and for all $d \in I^c$, the complement of $I$, $\hat{P}^d \geq P^d$ in which $I$ and $I^c$ are not empty. We have:

$$d \in I, P^d\pi^d = \hat{P}^d\hat{\pi}^d \implies \pi^d = \frac{\hat{P}^d}{P^d}\hat{\pi}^d \leq \hat{\pi}^d,$$
since $P^d > 0$ and $\tilde{\pi}^d > 0$. It implies that:

$$\sum_{d \in I} \pi^d < \sum_{d \in I} \tilde{\pi}^d. \quad (7)$$

Yet, by definition:

$$\sum_{d \in I} \pi^d = \Pr(\max_{d \in I} (P^d V^d) \geq \max_{d \in J} (P^d V^d)), \quad \sum_{d \in I} \tilde{\pi}^d = \Pr(\max_{d \in I} (\tilde{P}^d V^d) \geq \max_{d \in J} (\tilde{P}^d V^d)).$$

As for all $d \in I$, $\tilde{P}^d < P^d$, $\max_{d \in I}(\tilde{P}^d V^d) \leq \max_{d \in I}(P^d V^d)$ since the value functions $V^d$ are non-negative, and as for all $d \in J$, $\tilde{P}^d \geq P^d$, $\max_{d \in J}(\tilde{P}^d V^d) \geq \max_{d \in J}(P^d V^d)$, we have:

$$\Pr(\max_{d \in I} (P^d V^d) \geq \max_{d \in J} (P^d V^d)) \geq \Pr(\max_{d \in I} (\tilde{P}^d V^d) \geq \max_{d \in J} (\tilde{P}^d V^d)) \implies \sum_{d \in I} \pi^d \geq \sum_{d \in I} \tilde{\pi}^d,$$

a contradiction with inequality (7).

Therefore, $P = \tilde{P}$. □

The equilibrium values are obtained as a function of the thresholds:

$$P^d = \Pr(m_1 > t_1(d), m_2 > t_2(d)). \quad (8)$$

Using the fact that first stage and second stage probabilities are fixed at a certain known ratio $R = 4$, we have:

$$\frac{\Pr(m_1 > t_1(d), m_2 > t_2(d))}{\Pr(m_1 > t_1(d))} = R$$

which determines $t_1(d)$ as the unique solution of:

$$\Pr(m_1 > t_1(d)) = \frac{P^d}{R}.$$ 

The second threshold $t_2(d)$ is then obtained by solving equation (8).

### 3 The Econometric Model: Two stage grades and student preferences

We begin by specifying the two stage grade equations and detail sufficient identifying restrictions. We explain how success probabilities in equation (3) can be derived from such specifications. We then turn to the identification of random preferences and state exclusion restrictions that allow us to recover student preferences for schools.
3.1 Grade equations

As described in the previous Section, only students who pass the first stage exam can write the second stage exam. Therefore in our data, the second stage grades, $m_2$, are censored when first stage grades, $m_1$, are not large enough i.e. $m_1 < t^d_1(ENEM)$ in which initial grades are denoted $ENEM$. In the absence of any restriction, the distribution of $m_2$ given $m_1 < t^d_1(ENEM)$ is not identified.

3.1.1 A control function approach

To proceed we shall write that $(m_1, m_2)$ are functions of covariates

\[ m_1 = Z\beta_1 + u_1, \tag{9} \]
\[ m_2 = Z\beta_2 + u_2, \tag{10} \]

The first stage grade equation is a standard linear model and estimation would proceed under the restriction that $E(u_1 | Z) = 0$. This could be made as flexible and non parametric as we wish. In the second stage grade equation we use a control function approach to describe the influence of the unobservable factor derived from the first grade equation. We assume that:

\[ u_2 = g(u_1) + u^*_2 \]

in which $u^*_2$ is mean independent of $u_1$, $E(u^*_2 | u_1, Z) = 0$.

By doing this, we are now also able to control the selection bias since $u_2$ is supposed to be mean independent of $u_1$ and therefore $E(u_2 | m_1 \geq t_1(ENEM), Z) = 0$. This would identify parameters and the control function $g(\cdot)$. Nonetheless, our goal is not only to estimate these parameters but also to estimate the joint distribution of $(u_1, u_2)$. This is why in the following we assume that $u_1$ and $u_2$ are independent of $Z$ and simply use the estimated empirical distributions of $u_1$ and $u_2$ to recover success probabilities as in the developments below.

More elaborate ways available in the literature could be used but we stick in this paper with this simple procedure. Some of these assumptions are testable and we will assess the robustness of our results to the removal of (some of) these assumptions.
3.1.2 Simulated success probabilities

To predict success probabilities, two important elements are needed: the joint distribution of random terms $u_1$ and $u_2$ and the admission thresholds for the first and second stage grades. We already stated assumptions under which we can recover the former. The latter are derived from the definition using final admission in each major is described by two inequalities:

$$m_1 - 120 \times ENEM/63 \geq t_1^d$$
$$0.4 \times m_1 + 0.6 \times m_2 \geq t_2^d$$

We calculate the major dependent thresholds $t_1^d$ and $t_2^d$ using the grades of students admitted in each major $d$. We postpone the discussion on how we took into account that observed thresholds are measured with error since they are single realizations of random variables. We thus argue here given some arbitrary values, $t_1^d$ and $t_2^d$.

We first transcribe the inequalities above as functions of unobserved heterogeneity terms $u_1$ and $u_2$. For every student, passing the two exams means that the two random terms in the grade equations should be large enough as described by:

$$u_1 \geq t_1^d - 120 \times ENEM/63 - Z\beta_1,$$
$$u_2^* \geq \frac{t_2^d}{0.6} - \frac{2}{3}(Z\beta_1 + u_1 + 120 \times ENEM/63) - Z\beta_2 - g(u_1).$$

Notice that the second inequality depends on first stage grade shocks, $u_1$, because of the correlation between grades. Therefore the success probability in a major $d$ as defined by equation (3) can be expressed as:

$$p^d(Z, \beta, t_1^d, t_2^d) = \Pr\{u_1 \geq m_1^d - Z\beta_1, u_2^* \geq m_2^d - \frac{2}{3}Z\beta_1 - Z\beta_2 - \frac{2}{3}u_1 - g(u_1)\},$$

$$= \int_{m_1^d - Z\beta_1}^{\infty} f_{u_1}(x)(\Pr\{u_2^* \geq m_2^d - \frac{2}{3}Z\beta_1 - Z\beta_2 - \frac{2}{3}u_1 - g(u_1)\})dx,$$

$$= \int_{m_1^d - Z\beta_1}^{\infty} f_{u_1}(x)[1 - F_{u_2^*}(m_2^d - \frac{2}{3}Z\beta_1 - Z\beta_2 - \frac{2}{3}x - g(x))]dx,$$  \hspace{1cm} (11)

in which $m_1^d$ and $m_2^d$ are functions of thresholds:

$$\begin{align*}
m_1^d &= t_1^d - 120 \times ENEM/63, \\
m_2^d &= \frac{t_2^d}{0.6} - \frac{2}{3}(120 \times ENEM/63).
\end{align*}$$
3.2 Identification of Preferences

3.2.1 The decision model

Students make decisions based on their evaluation of the majors and their assessment of the admission or success probabilities. As detailed in the previous section, we assume that students are sophisticated individuals who can form expected utility of the majors and choose whichever gives the largest expected utility as described in equation (1). There are two issues of concern. The first one regards sample selection since only students interested by at least one school are present in the sample so that we condition on the event that $V^S > 0$ or $V^F > 0$. The second issue concerns individuals for whom one school only provides positive utility. This restricts their choice to this school only, the second school being dominated by the outside option. Figure 1 exhibits all different cases. The measure of the north-west quadrant is the probability measure that $V^S > 0$ and $V^F \leq 0$ and is denoted $\delta^S = Pr\{V^S > 0, V^F \leq 0\}$. In this case, school $S$ is necessarily chosen. Similarly, for the south east quadrant $\delta^F = Pr\{V^S \leq 0, V^F > 0\}$ and school $F$ is chosen. The south west quadrant is composed by individuals who are excluded from the sample and its probability measure is not identified.

The north east quadrant which has measure $\delta^{SF} = Pr\{V^S > 0, V^F > 0\}$ is the most interesting since choices can change if success probabilities $P^S$ and $P^F$ change. In this region, we can write the decision model by taking logarithms of the above set of equations:

$$\begin{cases} D = S & \text{if } \log(P^S) + \log(V^S) \geq \log(P^F) + \log(V^F), \\ D = F & \text{if } \log(P^S) + \log(V^S) < \log(P^F) + \log(V^F) \end{cases}$$

(12)

In this set of equations, the two variables $\log(P^S)$ and $\log(P^F)$ are function of covariates and can be estimated as seen in the previous subsection. The fact that both coefficients equal to one provides the usual scale restriction in binary models (and a testable assumption). Nonetheless, the levels of log-utilities is not identified, only their differences are so that we specify:

$$\log(V^S) - \log(V^F) = X\gamma + \varepsilon,$$

in which $X$ contains all variables that affect school utilities and $\varepsilon$ is an unobserved idiosyncratic preference term. We assume that the distribution of $\varepsilon$ in the population defined by $V^S > 0, V^F > 0$ is a function $F(. \mid X)$. We are now in a position to write the choice probability regarding the first
school as:

\[
\Pr(D = S | P^S, P^F, X) = \Pr(V^S > 0, V^F \leq 0 | X) + \\
\Pr(V^S > 0, V^F > 0 | X). \Pr(\log(P^S) + \log(V^S) \geq \log(P^F) + \log(V^F))
\]

\[
= \delta^S(X) + \delta^{SF}(X)F(\log(P^S) - \log(P^F) + X\gamma | X).
\]

We now study the identification of these different objects.

### 3.2.2 Identification analysis

As is well known in binary models since Manski (1988) and Matzkin (1993), the identification of these different objects relies on the independent variation of the covariate $\Delta$ (due to the underlying variation in $Z$):

\[
\Delta(Z) \overset{\text{def}}{=} \log(P^S) - \log(P^F),
\]

from preference shifters, $X$. For various reasons that will appear more clearly in the following, $\Delta$ acts as a price excluded by assumption from utility. As developed in the previous section, $\Delta$ is unobserved by the econometrician, yet is a function of observed covariates $Z$ because information is symmetric across agents and econometricians. Except in very specific circumstances, the effects of price and preference shifters cannot be identified from choice probabilities absent an exclusion restriction of at least one $Z$ from the $X$s. This leads to adopting the following high level assumption:

**Assumption Full Variation (FV):** The support of the conditional distribution of $\Delta(Z)$ conditional on $X$ is the full real line.

We can now proceed to analyze the identification issue whereby we try to identify the structural objects \{\(\delta^S(X), \delta^{SF}(X), \gamma, F(. | X)\)\} from the reduced form choice probabilities $\Pr(D = S | \Delta(Z), X)$ using:

\[
\Pr(D = S | \Delta(Z), X) = \delta^S(X) + \delta^{SF}(X)F(\Delta(Z) + X\gamma | X)
\]

(13)

The idea of identification is to exploit that some changes in $Z$ affects $\Delta(Z)$ without affecting preference shifters, $X$. Specifically for those who have negative utility for one of the schools, no matter how the success probabilities change, they will always decide for the same school. On
the other hand, those whose utilities are both positive are sensitive to the variation in success probabilities. By making these probabilities go to 0 or 1, we can then identify the probabilities of each of the 3 regions in Figure 1. Formally, this is made possible by Assumption FV. We can identify $\delta^S$ using the following:

$$\delta^S(X) = \lim_{\Delta(Z) \to -\infty} \Pr(D = S | \Delta(Z), X) = \inf_{\Delta} \Pr(D = S | \Delta, X).$$

A similar approach can be applied to $\delta^{SF}$ which is identified by,

$$\delta^S(X) + \delta^{SF}(X) = \lim_{\Delta(Z) \to -\infty} Pr(d = 1 | \Delta(Z), X) = \sup_{\Delta} Pr(d = 1 | \Delta, X).$$

We can thus form the expression that:

$$\frac{\Pr(D = S | \Delta(Z), X) - \delta^S(X)}{\delta^{SF}(X)} = F(\Delta(Z) + X\gamma | X).$$

Using standard arguments (Matzkin, 1994), this identifies $\gamma$ and $F(\cdot | X)$ under location restrictions such as the following median restriction:

$$F(0 | X) = \frac{1}{2}. \quad (14)$$

A final remark regards weakening Assumption FV since the support of the conditional distribution of $\Delta(Z)$ conditional on $X$ might not be the full real line. Assume for simplicity though that the support of $\Delta$ whatever $X$ is includes the value 0. Then as developed in Manski (1988), partial identification occurs under the median restriction (14) written above. Parameter $\gamma$ is identified using the median restriction and $F(\cdot | X)$ is identified in the restricted support in which $\Delta(Z) + X\gamma$ varies.

In our data, full variation is not observed and we will adopt a parametric assumption for $F(\cdot | X)$. What non parametric identification arguments above prove is that this parametric assumption is a testable assumption at least in the support in which $\Delta(Z) + X\gamma$ varies.

### 3.3 Estimation

We first estimate the parameters of the grade equation and denote them $\hat{\beta}_n$. This in turn allows us to compute the expectation of the success probabilities conditional on any thresholds $t^d_j, j = 1, 2, d = S, F$ as in equation (11).
Second, in order to compute unconditional success probabilities as in equation (2), we can also compute the distribution function of $\tilde{T}_0$ at an arbitrary level of precision using (6) by simulation of $U(n)$. For any simulation $c = 1, \ldots, C$, let us draw in the distribution of $U(n)$ a size $n$ sample $S_c$. We then derive realizations of $\tilde{T}_0$, say $\tilde{t}_c$ in $C$ samples of size $n$ by fixing choices $\{D_i(Z_i, \varepsilon_i, \zeta, P_i^S, P_i^F) = S\}$, characteristics $X_i$ and solving equation (6). Equation (2) can then be computed by integration as:

$$\hat{P}_{0,C}^d = \frac{1}{C} \sum_{c=1}^{C} p^d(Z_0, \hat{\beta}_n, \tilde{t}_1, \tilde{t}_2).$$

(15)

We can then estimate the preference parameters $\zeta = (\delta, \gamma)$ through the following conditional maximum likelihood approach:

$$\hat{\zeta}_n = \arg \max_{\zeta} l(\zeta | \hat{P}_{0,C}^S, \hat{P}_{0,C}^F).$$

This is a conditional likelihood function since $\hat{P}_{0,C}^S, \hat{P}_{0,C}^F$ depend on the first-step estimate, $\hat{\beta}_n$. Standard results show that when $n \to \infty$:

$$\hat{\zeta}_n \xrightarrow{P} \zeta.$$

The covariance matrix of those estimates can be obtained by the usual methods taking care of the conditional likelihood step or by bootstrap methods.

4 Empirical analysis: Grade and Choice Equations

4.1 Descriptive analysis

The complete original database comprises 41377 students who took the Vestibular exam in 2004. There are several groups of variables in the database that are useful for this study:

- Grades at different exams – the initial national high school evaluation exam (ENEM), the first and second stage of the Vestibular system.

9By construction, $\tilde{T}_0$ depends on observation 0 although this dependence should disappear when $n$ is large. For simplicity, we compute those thresholds in the empirical application using equation (5) instead of equation (6). We tested as a robustness check that this approximation was correct.
• Basic demographic variables – gender, age by discrete values (16, 17.5, 21 and 25), education levels of father and mother, and how much the student benefits or contributes to his/her family earnings.

• Education history – public or private primary or high school as described by discrete values indicating the fraction of time spent in private schools.

• Number of repetitions and undertaking of a preparatory course

• Program selection variables specifying the choice of major from broad major specific groups to university specific groups.

In total there are 58 majors that the students may consider at Universidade Federal do Cearà. We grouped these majors into broad groups according to the type of second-stage exams that students take to access these majors (see Data Appendix A). Table 1 reports the number of student applications, available positions and the rate of success at stages 1 and 2 in each of those major fields. These fields are quite different not only in terms of organization and in terms of contents but also regarding the ratio of the number of applicants to the number of positions. At one extreme lie Physics and Chemistry in which the number of applications is low and the final pass rates reasonably high (20%). At a lesser degree this is also true for Accountancy, Agrosciences and Engineering. At the other extreme, lie Law, Medicine, Other humanities and Pharmacy, Dentist and Other in which the final pass rate is as low as 5 or 6% that is one out of 16 students passes the exam. There are other differentiations in terms of quality.

Medicine is one of the most difficult major to enter as can also be seen in Table 2 which reports summary statistics in each major field and relative to the grades obtained at the first stage of the college exam.\(^{10}\) We report statistics on the distribution of the first stage grades in three samples:\(^{11}\) the complete sample, the sample of students who passed the first stage and the sample of students who passed the second stage and thus are accepted in the majors. Major fields are ranked according to the median grade among those who passed the final exam in that major field.

\(^{10}\)We do not report the second stage grades as they consist in grades in specific fields that are not necessarily comparable across fields.

\(^{11}\)We report for the complete sample the 10th percentile instead of the minimum in order to have a less noisy view of whom are the applicants. There are also a few zeros in the distribution of the initial grades.
These statistics are very informative. The minima tend to be ordered as the median of students who pass (column 6) from 70 to 90 in column 1. The first columns also reveal that some groupings might be somewhat artificial. The whole distribution is for example scattered out in mathematics from a minimum of 70 to a maximum of 222 while in medicine the range is 189 to 224. Other details are worth mentioning. The minimum grade in medicine to pass to the second stage is close to the maximum that was obtained by a successful student in Other fields and somewhat less than in Agrosciences. Medicine and Law are ranked the highest, as a matter of fact by a large amount of difference with other major fields. For instance, in Table 2, the first stage grade among those who passed in Medicine (resp. Law) has a median of 206 (resp. 189) while the next two are Pharmacy, Dentist and Other (175) and Engineering (171) and the minimum is for Agrosciences at 142.

4.1.1 Restricting the sample

For computational simplicity, the empirical analysis will be performed using a sub-sample of applicants to this college entry exam. The mechanism consisting in having only one choice allows us to simply restrict the sample without modifying the argument developed in the economic model. All other majors are summarized by the outside option. In the rest of the analysis, we shall consider only individuals who take exams in two majors that are part of Medicine, the most competitive major field as shown above. There are three majors in this group corresponding to three different locations in the state of Ceará: Barbalha, Sobral and Fortaleza. The first two majors are small and offer 40 positions only while the last one, Fortaleza, is much larger since it offers 160 seats. As shown in the empirical analysis below, this asymmetry turns out to be key for evincing strategic effects.

Table 3 repeats the analysis performed in Table 2 at the disaggregated level of those majors. Fortaleza is the most competitive one since the median of the first-stage grade of those who passed is equal to 208.57 while for the two others, it remains around 200. Nevertheless, the pass rate as shown in Table 3 relating the number of applicants and the number of positions is about the same in Sobral and Fortaleza (7%) while it is slightly lower in Barbalha (5%). At the same time, Barbalha receives applications from the weakest students as shown by the median grades in the sample of all applicants to this major. This is why we restrict the sample further to the two
medical schools Fortaleza and Sobral.

The list of variables and descriptive statistics in the pool of applicants to these two schools appear in Table 4. The number of applicants taking the first exam are in total 2867 and are decomposed into respectively 542 (Sobral) and 2325 (Fortaleza). The number of seats after the first-stage is four times the number of final seats and is thus respectively equal to 160 for the small major and 600 for Fortaleza. Note also that only two applicants in the pool of Fortaleza applicants and none in Sobral fail to go to the second-stage. The utility of taking the second stage exam after the revelation of information after the second-stage is (almost always) positive whatever the probability of success is.

Explanatory variables are those which affect exam performance and educational preference. For grade equations, all potential explanatory variables are included: a proxy for ability which is the initial grade obtained at the national exam,\textsuperscript{12} age, gender, educational history, repetitions, parents’ education and the undertaking of a preparatory course. Our guidance for selecting variables is that a better fit of grade equations leads to a better prediction of success probabilities in the further steps of our empirical strategy.

Regarding the specification of preferences for the majors, we exclude from them any variable related to past educational history. Indeed, preferences are related to forward looking utility of the majors (e.g. wages) which, conditional on the proxy for ability, is unlikely to depend on the precise educational history of the student (e.g. undertaking a preparatory course). This is even more likely since the proxy for ability is measured after the summaries of educational history. As a consequence, the following variables are retained in the specification of preferences: ability, gender, age, education levels of father and mother, and the number of repetitions of the entry exam. The inclusion of gender, age and education of parents is standard in this literature. The number of repetitions reveals either the determination of a student through her strong preference for the majors or the lack of good outside options.

Table 4 reports descriptive statistics regarding the binary decision between medical schools. Looking at the admission rates, one can see that Sobral admitted $\frac{40}{527} = 7.6\%$ and Fortaleza $\frac{150}{2340} = 6.4\%$ and this makes Fortaleza more competitive. Comparing the mean and median of

\textsuperscript{12}When missing, we imputed for ability a predicted value of this initial grade obtained by using all exogenous variables (see Data Appendix) and we denote teh result as $m_0$. Note that $ENEM$ is used when computing the passing grades for which the administrative rule is to impute 0 when missing.
initial and first stage grades, Sobral has better applications than Fortaleza. As to the second stage grades, although both schools have the same mean, selected candidates to Sobral have slightly higher median than those applying to Fortaleza. In conclusion, Fortaleza is more popular among students who apply to a medical school although it is not clear whether this popularity comes from preferences or is the result of strategic behavior of students. Our model is an attempt to disentangle those effects.

Figure 2 reports the estimated density of grades distinguishing Sobral and Fortaleza applicants. The first stage grade density function in Sobral has a regular unimodal shape while Fortaleza has a somewhat irregular modal shape and a fat tail on the left. The second stage grade density function, both in Fortaleza and Sobral behaves as an unimodal distribution, and the Sobral density function has a fatter tail on the left-hand side. The truncation at the first stage plays an important role in removing the fat tails of both densities on the left-hand side.

There are also other interesting differences among applicants to the two schools, such as in gender, age, private high school and preparatory course. There are more female applicants to Fortaleza than to Sobral. Sobral candidates are older on average and repeat more exams than Fortaleza candidates do and these two variables are highly correlated. The average time spent in private high school is higher in Sobral and it is more likely for a Sobral candidate to have taken a preparatory course.

### 4.2 Results of grade equations

We report in Table 5 the results of linear regressions of the first grade equation using three different specifications. We pay special attention to the flexibility of this equation as a function of the ability proxy \( m_0 \), which is the observed ranking of each student with respect to his or her fellow students and the best proxy for the success probability at the exams. We could estimate the control functions using non-parametric methods such as the one proposed in Robinson (1988) for instance or using splines. A thorough specification search made us adopt a 2-term spline specification, which is reported in the first column of Table 5. This specification is used later to predict success probabilities in both schools.

Among explanatory variables, age has a significant negative coefficient in all specifications and this indicates that older students who might have taken one gap year or more are relatively less
successful in the first stage exam. Meanwhile, taking a preparatory course and repeating the entry exam has positive and significant effects on grades, which means that a better preparation or attitude for the exams plays a role in the performance at the exam. As said, this specification includes two spline terms constructed from the initial stage grade $m_0$. More talented students tend to have better grades in exams, since $m_0$ has significant positive effects on the first stage grades although this dependence is slightly non-linear as represented in Figure 3. In the second specification, we tested for the inclusion of parents’ education among explanatory variables. None are significant and joint exclusion is not rejected by a F-test. In the third specification, we restrict the term in $m_0$ to be linear. It shows that results related to other coefficients are stable and robust. The set of explanatory variables we choose give us high $R^2$ around 0.72, and does not vary much across different specifications. This promises good prediction power of the model and makes the simulation of success probabilities credible for the later stage.

Turning to the second stage grade equation, we now have to take into account its dependence on the residual of the first stage grade equation $\hat{u}_1$. We again searched for flexibility in the grade equation with respect to two variables – the initial stage grade $m_0$ and the residual from the first stage grade equation $\hat{u}_1$. Using both non-parametric and spline methods, we found that a two term spline in the initial stage grade $m_0$ and a linear term in $\hat{u}_1$ were giving enough flexibility in this second stage grade equation. Results are reported in Table 6. First of all, there exists a strong positive correlation between $m_2$ and $u_1$, which indicates that unobservable factors on top of the ability proxy affect both equations. All other things being equal, students are more likely to perform well in the second exam if they perform well in the first exam. This may due to some unobservable effort difference or emotional resilience difference between students. The clear significance of $m_1$ residual signals that effort for studying might have been exerted by students during the year separating the initial stage exam revealing $m_0$ and the proper entry exam that we analyze. Our attempts in previous work did not lead to the confirmation of the prediction of this model that effort should be positively related to preferences. As to the other demographic variables, they affect similarly the second stage grade as the first stage grade except for gender. Results suggest that females perform significantly better than males in the second stage exam, while in the first stage grade gender difference are not significant.

Regarding robustness checks, another concern is heteroskedasticity. We perform Breusch-
Pagan tests to see whether there is substantial heteroskedasticity in the grade equations. For the first grade equation, gender is negatively correlated with squared residuals, but the F-test suggests insignificance at 5% level. Thus we do not reject the homoskedasticity assumption. For the second grade equation, the test rejects homoskedasticity and shows that age, private high school and repetition are significant in explaining squared residuals. This is consistent with the common sense that better high school education and more experience makes your performance steadier. However, in the rest of the paper, we adopt the homoskedasticity assumption since we checked that heteroskedasticity does not generate large differences in the prediction of success probabilities.

Finally, success probabilities are simulated using the empirical distribution of $\hat{u}_1$ and $\hat{u}_2$. We run Monte Carlo 1000 times by drawing into the distribution of errors, compute thresholds by solving equation (5) and simulate the integration (with respect to thresholds) of the bivariate integral as in equation (11) by using equation (15) and one set of simulations only. We experimented with different numbers of simulations to make sure that simulation error is negligible. This allows to compute simulated success probabilities for each student at both stages of the exam and in both schools. Table 7 reports descriptive results on these simulated probabilities. The first stage success probability means and medians are around 20-30% in both schools. This is close to what is observed in the sample but not exactly identical since these probabilities are partly counterfactual. For instance, the population of students selected in the second stage exam for Sobral school is not the same as the population selected in the second stage exam for the Fortaleza school. The second stage success probability is much lower and should be expected to be roughly 4 times lower than the previous ones since the number of students passing the first stage is four times the number of students finally admitted.

We also break down the simulated probability to see the difference between the students choosing Fortaleza and choosing Sobral in the original data. In order to see how student choices depend on their actual success probabilities, we compute the odds ratio of success probabilities at both stages. We rank the population with respect to their first stage grades and construct the grid of odd ratios at all percentiles for both stages, which is shown in Table 8. Some critical quantiles at the top are provided to give us some more insights. The two most important range of percentiles are the 70/75th and 93/95th percentiles since the admission rate at the first exam is about 28%
to 29% and the admission rate at the second exam is around 5/7%. Odds ratios are generally larger than 1 and odds ratios are the largest at the middle percentiles for both stages of the exam. It suggests that students who are not at the top of the rankings are making decisions that are more affected by success probabilities than by preferences and might play more strategically. For top students, odd ratios are closer to 1 because preferences matter more for those whose success probabilities are large and strategic effects are less important.

Figure 4 shows a picture of the odds ratio at all percentiles.

4.3 Choice equations: Estimation Results

We build our estimation procedure on the identification results developed in Section 3.2.2 although we adopt two parametric assumptions. First, the distribution of random preferences is a normal distribution when both schools yield positive utility to students. Second, the probabilities that only one school has positive utility are described by logistic functions which depend on a smaller set of covariates. Following the notation of Section 3.2.2, we write the probability measure of the regions in Figure 1, for instance the north-east quadrant as:

\[ SF(X) = \frac{1}{1 + \exp(\delta_{SF}^{0} + \delta_{m0}^{SF}m_0)}. \]

The choice probability is thus derived from equation (13):

\[
\Pr(D = S | \Delta(Z), X) = \delta^S(X) + \delta^{SF}(X)\Phi(\log(P^S) - \log(P^F) + X\gamma)
\]

in which \( \Phi(.) \) is the zero mean unit normal distribution and \( \log(P^d) \) shall be replaced by their simulated predictions using grade equations. In the first part of Table 9, we present the basic summary statistics of the estimated proportions for different specifications, and in the second part we report the estimated preference coefficients. There are three different specifications included in this table. The key difference is how many explanatory variables enter the specification of \( \delta^S \) and \( \delta^{SF} \).

We can see from the results presented in Table 9 that students living in Fortaleza are more likely to prefer Fortaleza medical school. In the first part of Table 9, we also present the estimated probabilities of each subpopulations. The mean probability of only willing to choose Sobral is around 0.06 in all specifications, and the mean probability of only willing to choose Fortaleza
is from 0.48 to 0.55. Despite small differences the proportions stay relatively invariant across specifications. This shows that students heavily favor Fortaleza over Sobral and this confirms Fortaleza as being the most reputed medicine school in the state of Ceará. The probability of only willing to choose Fortaleza is 10 times larger than the probability of only willing to choose Sobral which is approximately the ratio between the populations of the two cities much more than the ratio of final seats in the two schools (150/40). Nonetheless, there a substantial fraction of students whose utilities for both schools are positive (more than 40%).

We now turn to parameters $\gamma$ that affect preferences of students who prefer both schools to any outside option in the north-east quadrant of Figure 1. The variables, "Living in Fortaleza", Age, Gender (female) and ability, $m_0$, have a negative impact on the preference for Sobral, the smaller school. In contrast, the number of repetitions have a positive impact on choosing the medical school in Sobral. A well educated father affects positively preferences for the bigger school in Fortaleza while mother’s education does not have any significant influence on preferences. This is probably because of the collinearity of parents’ education that only leaves significant one of the two variables.

5 Evaluation of the Impact of Changes of Mechanisms

We now turn to the normative implications of our results and we investigate the impact of various changes of the existing mechanism.

The first counterfactual experiment that we implement is to cut seats at the second-stage exam and offering twice the number of final seats instead of four times. The University would incur lower costs in exchange with a possibly degraded selection if good students perform poorly at the first-stage exam. Second, we experiment with enlarging the choice set of students before taking exams. They now can list two ordered choices at most. This means that even if students fail the first stage qualification in one of the two schools they may still get the other major. This implies that the average skill level of passing students increase although the difference between the two majors is attenuated. Finally, there are two stages in the exam because this allows to cut costs of the second-stage exams and achieves a more in-depth selection. Another natural change to experiment is therefore to change the timing of choice-making and allow students to choose their final major after taking the first-exam and knowing their grades. This would generate more
opportunistic behavior.

Before entering the details of these new mechanisms, we first analyze the identification of utilities that are key in these evaluations. We show that expected utilities are underidentified and suggest how we can construct plausible bounds for the counterfactual estimates. Second, we explain how we compute counterfactual estimates conditional on observed choices.

5.1 Identifying Counterfactual Expected Utilities

Let the ex-post utility level be given by:

\[
U_i = \mathbb{1}\{V_i^S \geq 0, V_i^F < 0\}\mathbb{1}\{\text{Success in } S\}V_i^S + \mathbb{1}\{V_i^F \geq 0, V_i^S < 0\}\mathbb{1}\{\text{Success in } F\}V_i^F + \mathbb{1}\{V_i^F \geq 0, V_i^S \geq 0\}\left[\mathbb{1}\{D_i = S\}\mathbb{1}\{\text{Success in } S\}V_i^S + \mathbb{1}\{D_i = F\}\mathbb{1}\{\text{Success in } F\}V_i^F\right]
\]

and thus by taking expectations with respect to grades denoting \(P_i^S, P_i^F\) such expectations:

\[
E(U | V_i^S, V_i^F) = \mathbb{1}\{V_i^S \geq 0, V_i^F < 0\}P_i^S V_i^S + \mathbb{1}\{V_i^F \geq 0, V_i^S < 0\}P_i^F V_i^F + \mathbb{1}\{V_i^F \geq 0, V_i^S \geq 0\}\left[\mathbb{1}\{D_i = S\}P_i^S V_i^S + \mathbb{1}\{D_i = F\}P_i^F V_i^F\right]
\]

As this expected utility can always be rescaled by a scale factor (the location parameter is fixed by the outside option), we will choose the absolute value \(|V_i^F|\) as the scale factor to set:

\[
V_i^F = 1 \text{ if } V_i^F > 0, \quad V_i^F = -1 \text{ if } V_i^F < 0.
\]

Under this normalization:

\[
E(U | V_i^S, V_i^F) = P_i^S \left(V_i^S \mathbb{1}\{V_i^S \geq 0, V_i^F < 0\} + \frac{V_i^S}{V_i^F} V_i^F \mathbb{1}\{V_i^F \geq 0, V_i^S \geq 0\} \mathbb{1}\{D_i = S\}\right) + P_i^F \left(V_i^F \mathbb{1}\{V_i^F \geq 0, V_i^S < 0\} + \frac{V_i^F}{V_i^S} V_i^S \mathbb{1}\{V_i^F \geq 0, V_i^S \geq 0\} \mathbb{1}\{D_i = F\}\right).
\]
the only unknown is \( V_i^S \) when \( V_i^S \geq 0, V_i^F < 0 \) since \( \frac{V_i^S}{V_i^F} \) when \( V_i^F \geq 0, V_i^S \geq 0 \) is identified (see Section ??).

Various assumptions are possible. If there is some positive correlation between \( V_i^F \) and \( V_i^S \), we would expect that

\[
E \left( V_i^S \mid V_i^S \geq 0, V_i^F < 0 \right) < E \left( V_i^S \mid V_i^S \geq 0, V_i^F \geq 0 \right) = E \left( \frac{V_i^S}{V_i^F} \mid V_i^S \geq 0, V_i^F \geq 0 \right) < \exp(X_i \gamma) E(\exp(\varepsilon_i) \mid V_i^S \geq 0, V_i^F \geq 0) < \exp(X_i \gamma + .5),
\]

the last expression being obtained under the normality assumption. This is why we assume that when \( V_i^S > 0 \):

\[
\log V_i^S = \frac{\mu_0}{2} V_i^F + (\log \left( \frac{V_i^S}{V_i^F} \right) - \frac{\mu_0}{2}) |V_i^F| = \frac{\mu_0}{2} V_i^F + (X_i \gamma + \varepsilon_i - \frac{\mu_0}{2}) |V_i^F|
\]

where \( \mu_0 > 0 \) captures the common factor between \( V_i^S \) and \( V_i^F \). This is coherent with the previous equation since:

\[
\left\{
\begin{array}{ll}
V_i^S = \exp(X_i \gamma + \varepsilon_i) & \text{if } V_i^F = 1, \\
V_i^S = \exp(X_i \gamma + \varepsilon_i - \mu_0) & \text{if } V_i^F = -1.
\end{array}
\right.
\]

We will thus evaluate \( E \left( U_i \mid V_i^S, V_i^F \right) \) using the bounds on \( \mu = \exp(-\mu_0) \) that varies between 0 and 1.

5.2 Computational Framework

In every counterfactual experiment, we use the following framework. We proceed by simulation and draw the unknown random terms conditional on observed choices. This insures that observed choices are compatible with simulated choices in the observed data. In each simulation, let \( \bar{D}_i \) be the counterfactual choices of the students that depend on counterfactual expectations \( P_i^S \) and \( \bar{P}_i^F \). Denote \( \bar{n}_S \) and \( \bar{n}_F \) the new number of seats in the cutting-seat counterfactual. In other cases \( \bar{n}_S = 4n_S \) and \( \bar{n}_F = 4n_F \) as in the original system.

The first important thing to note is that the population of reference does not change in the counterfactual experiments. Only those whose utilities are such that \( V^S > 0 \) or \( V^F > 0 \) remain in

\[\text{footnote: Changing the timing of choices, requires to acknowledge that there are no choices to make before the first-stage. The first two equations do not depend on } \bar{D}_i \text{ and } P_i^S, P_i^F \text{ are the conditional expectations after the second-stage. Those adaptations do not modify the main principles.}\]
the pool of potential students and therefore we consider the same sample \( i = 0, \ldots, n \). In our model, alternative mechanisms act only on success probabilities and not on utilities.

Moreover, consistency of choices and perfect expectations require that the counterfactual random thresholds, \( \tilde{T}_0 \), as defined as the solution \( (\tilde{t}_1^S, \tilde{t}_2^S, \tilde{t}_1^F, \tilde{t}_2^F) \) to their counterfactual counterpart to equation (6):

\[
\begin{align*}
\sum_{i=1}^{n} \{D_i(\tilde{P}_i^S, \tilde{P}_i^F) = S\} & \{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^S\} = \bar{n}_S, \\
\sum_{i=1}^{n} \{D_i(\tilde{P}_i^S, \tilde{P}_i^F) = F\} & \{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^F\} = \bar{n}_F, \\
\sum_{i=1}^{n} \{D_i(\tilde{P}_i^S, \tilde{P}_i^F) = S\} & \{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^S, m_2(X_i, \beta, u_i) \geq \tilde{t}_2^S\} = n_S, \\
\sum_{i=1}^{n} \{D_i(\tilde{P}_i^S, \tilde{P}_i^F) = F\} & \{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^F, m_2(X_i, \beta, u_i) \geq \tilde{t}_2^F\} = n_F.
\end{align*}
\] (16)

have a distribution function that leads to the counterpart to equation (15):

\[
\tilde{P}_0^d = \mathbb{E}(\{m_1(X_0, \beta, u_0) \geq \tilde{t}_1^d, m_2(X_0, \beta, u_0) \geq \tilde{t}_2^d\})
\] (17)

We thus propose to iterate the following algorithm (we explain it for observation 0 and extend it naturally to any index \( i \)):

1. Initialization:

   - Draw \( C \) random vector \( \varepsilon_{(n),c} \) in their distributions conditional to the observed choices, \( D_i, \) (see Appendix B.1.2 for details). Fix those \( \varepsilon_{(n),c} \) for the rest of the procedure.
   - Draw \( C \) random vector \( U_{(n),c} \) and fix them for the rest of the procedure.
   - Set the initial \( P_{0,0}^{S,k}, P_{0,0}^{F,k} \) values at their simulated values \( \hat{P}_0^d, \) computed from equation (15) replacing \( \zeta \) by \( \hat{\zeta}_n \) and using \( U_{(n),c} \) in the observed experiment that is through equations (6). This implicitly means that choices \( D_i \) are set to their observed values.

2. At step \( k \), denote \( P_{i,k}^{S,k}, P_{i,k}^{F,k} \) the expected success probabilities

   (a) Compute choices \( D_i(Z_i, \varepsilon_{i,c}, \hat{\zeta}_n, P_{i,k}^{S,k}, P_{i,k}^{F,k}) \).
   (b) Compute a sequence of \( \tilde{t}_c \) for \( c = 1, \ldots, C \) replacing \( \zeta \) by \( \hat{\zeta}_n \) and using \( U_{(n),c} \) and equations (16).
   (c) Derive \( \hat{P}_{0,C}^{d,k+1} \) from equation (17).
3. Repeat the previous step until a measure of distance $d(P^{(k+1)}, P^{(k)})$ is small enough.

If this algorithm converges then this is the fixed point we are looking for.

5.3 Cutting seats at the second stage exam

We start with the easiest interesting policy change that assigns a different admission rate after the first stage. In consideration of the organization cost of exams, it is logical to reduce the admission rate after the first stage. As said, the existing Vestibular system usually allows the number of students who take the second exam to be four times the number of available seats. We keep the number of final positions unchanged but allow only half of the students to take the second exam. In other words, the admission rate after the first stage exam is reduced by a factor of 2. We explore the possible consequences of this policy and investigate two main issues – which type of students will benefit from this policy change and are schools losing good students?

Some discussion about the expected effects are in order. Cutting the seats in the second exam reduces the schools’ administrative costs in terms of correcting exams although it also comes with the risk of losing talented students. Students may not be always consistent in their exam performance and even the most talented students may have a strong negative shock in the first exam. Those students would be eliminated too early without being given a second chance. Nonetheless, it could also be that cutting seats protect the best achievers at the first stage from competition and thus from the risk of losing ranks at the second stage exam. The net result is unclear theoretically and this is why an empirical analysis is worthy of attention.

The simulation of the counterfactual follows the procedure described in Sections 5.1 and we compute expected utility as in Section 5.2.

5.3.1 Results

In Table 10 we present estimates of the new thresholds distribution of all three counterfactual experiments distinguishing the two stage exams. Standard errors are computed from 1000 Monte Carlo simulations although they do not take into account parameter uncertainty. In the cutting seat experiment, the counterfactual first stage thresholds are much higher and this is expected since fewer students are admitted after the first stage exam. In contrast, the thresholds of the second stage exam are lower than the original system because there is now less competition in
the second stage exam when fewer students are admitted. In both first and second stage exams, thresholds in Sobral are more volatile than the ones in Fortaleza because Sobral is a much smaller school.

To evaluate whether the counterfactual brings benefit to schools and students, we first need to investigate changes in the expected success probabilities. We then turn to the evaluation of changes in students’ utilities.

**Changes in success probabilities** We evaluate those changes in relation to an index of students’ ability that proxies schools’ preferences. Namely, schools would find that the admittance procedure has improved if abler students (in expectation) get a higher chance of admission and the worse students have a lower chance. In the analysis that follows, we use the expected final grade combination as our ability index. We also choose to concentrate on the top 50% of students because the lower 50% of the sample have almost no chance of getting admitted whether the original or counterfactual mechanisms are used.

We represent changes in success probabilities in Figure 5 for Sobral (respectively Figure 6 for Fortaleza). In those Figures three vertical lines are drawn at various expected final grades positions standing for the median and the associated first and second-stage thresholds *in the original system*. Changes in probabilities are very similar in the two schools with a slightly larger success probability improvement for Fortaleza.

The very top students who are above the second stage admission quantile, have better chances in the counterfactual system since they now face less competition in the second stage exam. We have seen from Table 4 that second stage grades have much larger variance than first-stage grades. The chance of losing the final position is low when fewer students participate the second stage exam. For students who are between the median and second stage admission expected final grade, the situation is worse. If they happen to perform well in the first exam, then they will be admitted and enter the second stage exam with less competition and that leads to a higher success probability. However the chance that they perform not so well in the first exam is much higher since fewer students are admitted and it is this negative effect that dominates overall. Finally, for students between first stage and second stage admission thresholds, they tend to have a higher chance of success at the first stage and thus benefit from less competition in the second stage. It is the students who are around the first admission thresholds who suffer the most simply because they
are more likely to be the students who lose the chance of participating the second stage exam due to the system change.

In conclusion, schools seem to benefit from cutting seats since the most able students now have a higher chance of admission since they are protected from the competition of less able students at the second stage that is more uncertain. This benefit comes in addition to the cutting the costs of organizing and correcting the second-stage exam proofs. Note that the policy in place is enacted at the level of the University and not the medical schools under consideration and it may well be that these conclusions are reversed when analyzing the entry into other majors. It might also be that the schools have additional information about the correlation of second-stage exams and future success in undergraduate studies and favor more second-stage exams that what we posit here. Again, the weighing of first-stage and second-stage exam grades in the final grade is decided at the level of the University and medical schools might not be able to affect these weighing schemes.

Changes in students’ utilities Table 11 shows the changes in students’ expected utility ranked in percentile groups according to their ability as described by their expected final grade. As defined in Section 5.1, we set the unknown weight in utilities, $\mu = 0.8$. Consistently with changes in success probabilities, only the very top students – above the 94% ability quantile – have significant utility improvements. Students above the 88% ability quantile are more likely to have a positive change in utility. Students above the median tend to have lower expected utility in the counterfactual system and this is also consistent with what we obtained for success probabilities. If we divide the sample by the original major choice, an indication of their preference, students who chose Fortaleza tend to benefit more than the ones who opted for Sobral. Overall, these results about this counterfactual experiment bring out no significant total utilitarian welfare change. Yet, there are strong distributional effects and top students are better off and less able students are worse off. We can visualize changes in expected utility in Figure 7.

We also performed a bound analysis by using different values for the weight $\mu$. Results are shown in Table 12. When $\mu$ is at the lower bound – 0 – utility changes are slightly smaller. When $\mu$ is at the upper bound – 1 – utility changes become slightly larger. Overall, differences are very limited and our previous results are robust to the value of $\mu$.  

37
5.4 Enlarging the choice set

In this experiment, students interested in at least one major, can nevertheless choose a list of two majors. This would correspond to what happens in a deferred-acceptance mechanism with the additional twist that we keep the sequence of two exams as it is. We chose to allocate students in majors after the first exam in a design that we now explain.

First, a choice list contains two elements $d_1$ and $d_2$ in which $d_1 \in \{S, F\}$ is the preferred major since our sample of interest refers to students who positively value at least one of the majors so that $d_1 \neq \emptyset$, and $d_2 \in \{\emptyset, S, F\}$. We thus still allow student to provide a single choice if $d_2 = \emptyset$. In those cases, the sequence of choices do not change for students. They take the first exam and pass it if their grade is above some threshold. The pool of students to which they are compared however change since students who have positive utilities in both majors now provide a list of two choices.

To fix ideas, consider a student who has $V_S > 0$ and $V_F > 0$ and chooses $(S, F)$. If after the first-exam, she gets in the second-exam for school $S$, she will not compete with students taking the second exam in school $F$. It is only if she is not accepted to the second stage exam in school $S$ that she can compete for the second stage exam seat in school $F$. She fails only when her grades are lower than both thresholds. Suppose first that at the equilibrium $t_1^S > t_1^F$. After the first stage exam, there are then three possible outcomes for the student:

- $m_1 \geq t_1^S$: she takes the second exam of major $S$,
- $m_1 < t_1^S$ and $m_1 \geq t_1^F$: she takes the second stage exam of major $F$,
- $m_1 < t_1^F$: she fails the exam this year, and takes the outside option.

While if $t_1^S > t_1^F$,

- $m_1 \geq t_1^S$: she takes the second exam of major $S$;
- $m_1 < t_1^S$: she fails the exam this year, and takes the outside option.

This sequence is easily adapted to students choosing the list $(F, S)$. Therefore given any choice among the four lists, \{$(S, F), (F, S), (S, \emptyset), (F, \emptyset)$\} we can construct counterfactual success probabilities in each major $P^S$ and $P^D$ by adapting the algorithm we used before. Reciprocally, for any value of success probabilities, we can define their optimal choices between
\{(S, F), (F, S), (S, \emptyset), (F, \emptyset)\}. Details about how we get counterfactual thresholds follow the lines of what was developed in Section 5.2.

An alternative that could be explored as well is to allocate students in majors after the second-stage exam.

5.4.1 Results

Thresholds for this counterfactual experiment are also shown in Table 10. For the first stage, the threshold of Sobral is now larger than the original one while the threshold of Fortaleza remains roughly unchanged. This is an indication that Sobral is admitting better students without hurting Fortaleza. Some top students who were failing Fortaleza before can now compete for Sobral and get admitted after the first stage. Furthermore, some students who were choosing Fortaleza before for strategic reasons can now at no risk choose Sobral first and Fortaleza second. Deferred acceptance mechanisms lessen strategic motives and make choices more truthful. In the original system, students tended to choose Fortaleza as a ”safety school” even when they truly preferred Sobral. Giving students two choices cancels the ”safety school” effect. Yet, thresholds for the school in Fortaleza remains higher than for Sobral at both stages because it attracts more top-ability \(m_0\) students as was shown in the preference estimation in Table 9.

Moreover, the thresholds at the second stage exam are not significantly different from the original ones. The second stage thresholds are determined by the quality of the very top students in either school. This counterfactual experiment moves some of the relatively good students after the first stage exam from Fortaleza to Sobral. Sobral however still attract less able students than Fortaleza in the second stage competition.

Changes in success probabilities Figure 8 for Sobral and Figure 9 for Fortaleza report changes in success probabilities. Unlike the previous counterfactual experiment, the change in Sobral is now quite different from Fortaleza. In Sobral, almost all students whose ability is above the first stage admission quantile has now higher success probabilities and the ones who benefit the most are the very top students. In contrast, a larger portion of students below the first admission threshold and above median have a lower success probability in the counterfactual experiment because top students who fail Fortaleza can switch to Sobral to compete with them. In Fortaleza, almost everyone has a higher success probability because some top students are switching to Sobral
and this makes it easier for top students who prefer Fortaleza.

From the perspective of the schools, both Fortaleza and Sobral are happy to see such changes in the system because they can now attract higher ranked students. Sobral should be more favourable to this mechanism since the school can now recruit many more top students. The last point that should be noted is that the change in success probabilities is small in this counterfactual compared with the previous one when we cut seats.

**Changes in expected utilities** From the perspective of the students, this mechanism is also attractive since a majority of students –88% – will be better off as shown in Table 13. Moreover, top students benefit more from the change than less able students because they are more likely to pass to the second-stage exam even if they happen to fail their preferred school. Students who prefer Fortaleza benefit much more than those who prefer Sobral. Since Sobral has a lower threshold at the first stage exam, those who choose Sobral and fail have no second chance. Therefore the expected utility increase for those who opt for Sobral is purely derived from the change in the success probability and that is why less able students are hurt in the counterfactual. However for those who prefer Fortaleza, expected utility mainly increases because of the second chance they get to compete for Sobral when they fail Fortaleza. The effect on expected utility is thus much larger than the change in success probabilities. Expected utility changes are graphed in Figure 10.

In summary, enlarging the choice set improves the average ability of those who pass the first stage exam in both schools and the small school has a better chance of recruiting top students. The majority of students are better off except the medium ranked students who prefer the smallest school. This confirms theoretical insights that the move to a deferred acceptance mechanism is likely to make both schools and the majority of students better off.

### 5.5 Changing the timing

In this counterfactual experiment, we try to evaluate the impact on the allocation and expected utility of students when they choose majors after the first stage exam and not any longer before. They are thus informed about their first stage grades. As in the original system, schools admit students to the second stage exam according to the ranking given by a combination of ENEM
and $m_1$ and students’ preferences.

The new selection procedure proceeds as follows. Starting from the first-ranked student at the first-stage exam and going down in the distribution of first stage grades in sequence, each student chooses major $S$ or $F$ until the number of admitted students in one of the majors, say $d$, reaches four times the number of final seats in this major. This defines threshold $t^d_1$. The sequence continues going down grades although choice is now restricted to the other major $d' \neq d$ until the number of admitted students in that major reaches four times the number of final seats. The allocation of students to the second-stage exam is then complete. The game continues afterwards as in the current system.

As before, utilities $V^S$ and $V^D$ remain the same while this new mechanism affects the probabilities of success $P^S_{m_1} = Pr\{m_2 > t^S_2|m_1\}$ and $P^F_{m_1} = Pr\{m_2 > t^F_2|m_1\}$ which are now conditional to first-stage grade $m_1$. To define choices, suppose that $t^S_1 > t^F_1$. A student can face three cases:

- $m_1 > t^S_1$: the choice set is complete and consists in $\{S, F\}$. Majors are chosen by comparing $P^S_{m_1} V^S$ and $P^F_{m_1} V^F$ (since either $V^S > 0$ or $V^F > 0$).
- $m_1 < t^S_1$ and $m_1 > t^F_1$: the choice set is restricted to $F$ and the student either opts for the second stage exam in $F$ if $V^F > 0$ or the outside option if not.
- $m_1 > t^F_1$: the only choice left is the outside option.

This can be adapted easily if the other case $t^S_1 < t^F_1$ prevails.

### 5.5.1 Results

The new thresholds for this counterfactual experiment are shown in Table 10. Sobral has now much higher thresholds at both stages thanks to her smaller size. The school in Fortaleza is overall more popular (see 9) but this does not compensate the difference in offered seats. By making students choose in the order of first stage grades, Sobral positions for passing to the second-stage are more likely to be filled earlier than Fortaleza’s because of the one to four ratio (160/600). For instance, if 25% of the top 640 students prefer Sobral to Fortaleza, the 160 seats at Sobral would be filled after those 640 students reveal their choices while Fortaleza will still have 120 seats to fill in. Such a mechanism favours the smallest school in Sobral.
Changes in success probabilities  Changes in success probabilities as shown in Figure 11 and Figure 12, are a straightforward consequence of thresholds changes. In Sobral, all students have lower success probability and top students suffer the most. In Fortaleza, the success probability becomes larger for everyone, especially the top students who are above the final admission thresholds. The school of Sobral has on average better top students at the second stage exam than those in the original system. Although Fortaleza school is likely to lose a few elite students the success probability of top students on average increases more than for medium-ranked students. Fortaleza will still be able to recruit better students on average in this counterfactual.

Changes in expected utilities  As this mechanism introduces an element of flexibility for the students since they can condition their choices on their first stage grades, their expected utility is on average larger than in the original system. This mechanism however is mainly attractive for the top students as shown in Table 14 where only students above the 70% quantile are significantly better off. Moreover, top students are more likely to pass the first stage exam and into their preferred programs.

There is no clear difference in utility increase among the top students with different preferences. On average, the students who prefer Fortaleza would benefit more than those who prefer Sobral. Moreover the utility increase for those who prefer Sobral is more uncertain as we see that a few very top students only gain a little while other gain a lot in Figure 13. This is due the fact that now Sobral has much higher thresholds. For the students who are only interested in Sobral, on one hand the uncertainty is reduced by the mechanism, which increases significantly the expected utility; on the other hand, the higher thresholds have also made it harder for them to get the final success in Sobral.

Overall, this counterfactual is more friendly for top students. From the perspective of the schools, Sobral should prefer this mechanism because it will be able to enroll better students. Fortaleza loses its ”safety school” feature and can no longer attract because of strategic reasons those risk averse top students who prefer Sobral.
6 Conclusion

In this paper, we use data from entry exams and an allocation mechanism in a University to provide an evaluation of this mechanism. We first estimate a model of major choices as well as performance to derive the parameters governing success probabilities and preferences. We then estimate three counterfactual experiments and show that the current mechanism could be adapted with benefits to schools and students although with potentially strong redistributive effects.
References


A Data appendix

A.1 Description

The Vestibular, an entrance exam where different universities develop their own format of testing restricted to some federal constraints, has its roots during the creation of the first undergraduate course in Brazil 200 hundred years ago. Only in 1970, with the creation of the National Commission of the Vestibular, the system started to develop a regulatory background in order to rationalize the increasing demand for undergraduate education in the country. The final step that shaped the format of the Vestibular in place in 2004 was taken in 1996 with the approval of the Law of Directives and Basis of the National Education (LDB). The LDB, among other things, set the minimum requirements of the exam and made explicit some constraints regarding the form and content that universities must obey if they choose to select their students through a Vestibular. Also, Olive (2002) asserts that LDB introduced a regular and systematic process of evaluation and credentialing that initiated a new era of meritocracy in Brazilian universities. Even though LDB reinforced regulation and as a consequence brought about many new restrictions, law abiding universities still have in practice a lot of degrees of freedom to adapt their entrance exams to their needs.

Roughly, the Vestibular has the following features:

1. The student chooses the undergraduate degree before the test, and compete only against those students who made the same choice;
2. It is comprised of many sub-exams, each one evaluating knowledge in Mathematics, Physics, Chemistry, Biology, Portuguese, History, Geography and a Foreign Language;
3. The exams are almost exclusively developed with objective (multiple choice) questions;
4. Different undergraduate courses can weight the sub-exams differently in order to reflect their priorities in terms of required knowledge;
5. More than one stage is allowed during the process of testing.
6. Almost all universities developed their own exam, however it is possible to form groups of universities to develop unified exams;
7. After the exams, students are ranked according to their grades and a pre-determined protocol. Places are filled from top to bottom, and if there are remaining free seats, other students
might be recalled.

8. Those who do not exercise their right of initiating the university course in the same year they took the Vestibular cannot make it later on. However, any student can take the entrance exam as many times as they want to.

A.2 The Vestibular at UFC

The Vestibular at UFC shares the same features described above regarding its protocol. However, we give a rather detailed description of some of its feature in order to gain insight when developing and estimating econometrics models. An important first thing to know is the fact that by law all entrance exams in public universities must be preceded by the release of a document called Edital. An Edital is a public document that must contain the whole set of regulations regarding the exam. It must contain, among others, a specific timeline for exams, a detailed list of syllabus for all disciplines required in the exams, the majors offered as well as the available spots in each one, how scores are calculated, how students are ranked, forbidden actions that may cause elimination from the exams, minimum requirements in terms of grades and so on. Accordingly to Brazilian law the Edital is a document that possesses the status of legislation, i.e., any dispute of rights with respect to details of the Vestibular must use the contents of the Edital as a first guiding line in order to settle the dispute.

The first stage, called General Knowledge (GK), is composed of a unique 66 objective questions (multiple choice, with five alternatives A, B, C, D and E) exam whose content is exactly the core high school curricula, i.e., Portuguese (Grammar and Writing), Geography, History, Biology, Chemistry, Mathematics, Physics and Foreign Language. In order to understand the grading system for this first exam note that there are two types of scores: raw and standardized, respectively. The raw score for each subject is given in Table A1 and the standardized scores in Table A2.

Adding up all standardized scores gives the total standardized score $X_s^{GK}$. In order to pass to the following second stage and take the so called Specific Knowledge (SK) exam, the student must obey the following rules:

1. Get a grade in each subject appearing in the GK exam;
2. After being ranked accordingly to his/her overall standardized score $X_s^{GK}$, the student must be placed in a position equal or above the threshold specific to his/her chosen major. This
threshold is calculated based on the following rule: Let \( N \) be the number of available places in a specific major previously shown in the Edital. Let \( r \) be defined as the ratio of the number of students choosing the major and the number of available seats in the major. If \( r < 10 \) then the threshold is \( 3N \), otherwise it is \( 4N \). Note that the threshold is not known by the candidate when choosing majors. This information is disclosed after chosen a major.

The SK exam is comprised of two separated sub-exams (realized in two consecutive days apart only two weeks after the release of first stage exam results). The SK is described below:

The two specific exams are set according to the requirements of each major. Again, this list in known before majors are chosen and is given by the following table:

The standardized scores are calculated according to the following formulas:

The sum of all standardized scores taken in the second stage gives the second stage grade. The sum of all first stage standardized scores and all second stage standardized scores gives the final grade. All students are ranked again and available seats are allocated to the best ranked students.

An issue with the data is that the initial stage grade, \( ENEM \), which we would like to treat as the proxy for ability is not observable for all individuals. An imputation method is needed to complete the observations so that we do not lose any information due to the missing \( ENEM \). We regress non-missing \( ENEM \) onto the basic demographic variables such as age, gender and education history and predict values for missing data. This yields our proxy for ability, \( m_0 \).

B Technical appendix

B.1 Preference model and Simulations conditional on observed choices

B.1.1 Set-up

Recall that we describe three groups of students according to their preferences: those only interested in Sobral, those only interested in Fortaleza and those interested in both. The probability of each of these three groups are denoted as \( \delta_i^S, \delta_i^F, \delta_i^{SF} \) and these probabilities are heterogeneous across students since they depend on \( X_i \). Let \( \varepsilon_i = (\varepsilon_i^{(1)}, \varepsilon_i^{(2)}) \) be such that \( \varepsilon_i^{(1)} \sim U[0,1] \) and \( \varepsilon_i^{(2)} \sim N(0,1) \). The first random term allocates student 0 to one of the three groups i.e. \( \varepsilon_i^{(1)} \leq \delta^S(X_i) \) means that she prefers Sobral only to the outside option and \( \varepsilon_i^{(1)} \geq \delta^S(X_i) + \delta^{SF}(X_i) \) means that she prefers Fortaleza only to the outside option. If \( \varepsilon_i^{(1)} \in (\delta^S, \delta^S + \delta^{SF}) \), both schools
bring positive utility to her. It is only in the latter case that expected success probabilities matter.

Let the function of $X_i$ and the second random term:

$$\ln(V^F(X_i, \varepsilon_i, \zeta)/V^S(X_i, \varepsilon_i, \zeta)) = X_i \gamma + \varepsilon_i^{(2)}$$

be the relative utility in logarithms of Sobral and Fortaleza. using the success probabilities $P^S_i(Z_i, \beta)$ and $P^F_i(Z_i, \beta)$, the decision is determined by:

$$D_0(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = S \iff \ln(V^S(X_i, \varepsilon_i, \zeta)/V^F(X_i, \varepsilon_i, \zeta)) + \ln(P^S_i/P^F_i) \geq 0,$$

$$D_0(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = F \iff \ln(V^S(X_i, \varepsilon_i, \zeta)/V^F(X_i, \varepsilon_i, \zeta)) + \ln(P^S_i/P^F_i) < 0.$$

### B.1.2 Simulations of $\varepsilon(i)$ conditional on choices

We shall simulate $\varepsilon_{i,c}$ in its distribution conditional on the observed choice $D_i = F$ (say). This necessarily means that $\varepsilon_{i,c}^{(1)} \sim U[0,1]$ conditional on $\varepsilon_{i,c}^{(1)} < \delta^S(X_i) + \delta^SF(X_i)$ so that we can write:

$$\varepsilon_{i,c}^{(1)} = (\delta^S(X_i) + \delta^SF(X_i))\varepsilon_{i,c}^{(1)}$$

in which $\varepsilon_{i,c}^{(1)} \sim U[0,1]$. Then, if $\varepsilon_{i,c}^{(1)} < \delta^S(X_i)$ the observed choice is necessarily $D_i = F$. In the other case, if $\varepsilon_{i,c}^{(1)} > \delta^S(X_i)$, we should condition the drawing of $\varepsilon_0^{(2)}$ on the restriction that:

$$X_i \gamma + \varepsilon_i^{(2)} + \ln(P^S_i/P^F_i) > 0$$

as derived from equations (12). This is easily done by drawing in a truncated normal distribution.

Draw $\varepsilon_i^{(2)}$ into a $U[0,1]$ and write:

$$\varepsilon_i^{(2)} = \Phi^{-1}(\Phi(-\ln(P^S_i/P^F_i) - X_i \gamma)) + (1 - \Phi(-\ln(P^S_i/P^F_i) - X_i \gamma))\varepsilon_{i,c}^{(2)}),$$

or equivalently:

$$\varepsilon_i^{(2)} = -\Phi^{-1}(\Phi(\ln(P^S_i/P^F_i) + X_i \gamma))(1 - \varepsilon_{i,c}^{(2)})).$$

Adaptations should be made to this construction when the choice is $D_i = S$. In this case,

$$\varepsilon_{i,c}^{(1)} = \delta^S(X_i) + (1 - \delta^S(X_i))\varepsilon_{i,c}^{(1)}, \varepsilon_{i,c}^{(1)} \sim U[0,1],$$

$$\varepsilon_i^{(2)} = \Phi^{-1}(\Phi(-\ln(P^S_i/P^F_i) - X_i \gamma)) + (1 - \Phi(-\ln(P^S_i/P^F_i) - X_i \gamma))\varepsilon_{i,c}^{(2)}), \varepsilon_{i,c}^{(2)} \sim U[0,1].$$

49
B.2 Counterfactual with lists of two choices

Here we describe how to compute the model of choice between two majors, $S$ and $F$. This allows four possible choices: $(S, F)$, $(F, S)$, $(S, \emptyset)$, $(F, \emptyset)$ and their respective expected values: $U^{SF}$, $U^{FS}$, $U^S$, $U^F$. Those values depend on probabilities of success and on thresholds in the following way.

Starting with the singleton lists $(d, \emptyset)$, we have that:

$$U^d = V^d \Pr\{m_1 > t_{d_1}^d, m_2 > t_{d_2}^d\}$$

as before. For the lists $(d_1, d_2) \in \{(S, F), (F, S)\}$, we use the description of the text to state that:

$$U^{d_1, d_2} = V^{d_1} \Pr\{m_1 > t_{d_1}^{d_1}, m_2 > t_{d_2}^{d_1}\} + V^{d_2} \Pr\{m_1 \in [t_{d_1}^{d_2}, t_{d_2}^{d_2}), m_2 > t_{d_2}^{d_2}\}$$

in which $\Pr\{m_1 \in [t_{d_1}^{d_1}, t_{d_2}^{d_2}) = 0$ if $t_{d_1}^{d_2} < t_{d_2}^{d_1}$. The choice model can now be described by four success probabilities:

$$\begin{align*}
P^d & = \Pr\{m_1 > t_{d_1}^d, m_2 > t_{d_2}^d\}, d = S, F \\
P^{d_1, d_2} & = \Pr\{m_1 \in [t_{d_1}^{d_1}, t_{d_2}^{d_2}), m_2 > t_{d_2}^{d_2}\}, (d_1, d_2) \in \{(S, F), (F, S)\},
\end{align*}$$

which are functions of thresholds $t_{d_1}^d, t_{d_2}^d$. Those thresholds remain sufficient statistics in order to derive these success probabilities.
### Tables and Figures

**Table 1: Number of applications, number of positions and success probabilities**

<table>
<thead>
<tr>
<th>Groups of majors</th>
<th>Applications</th>
<th>% Pass 1st stage</th>
<th>% Pass 2nd stage</th>
<th>Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accountancy</td>
<td>1,374</td>
<td>40%</td>
<td>13%</td>
<td>185</td>
</tr>
<tr>
<td>Administration</td>
<td>2,474</td>
<td>29%</td>
<td>8%</td>
<td>200</td>
</tr>
<tr>
<td>Agrosciences</td>
<td>2,996</td>
<td>41%</td>
<td>13%</td>
<td>390</td>
</tr>
<tr>
<td>Economics</td>
<td>1,516</td>
<td>37%</td>
<td>11%</td>
<td>160</td>
</tr>
<tr>
<td>Engineering</td>
<td>2,648</td>
<td>40%</td>
<td>14%</td>
<td>360</td>
</tr>
<tr>
<td>Humanities</td>
<td>4,897</td>
<td>17%</td>
<td>9%</td>
<td>430</td>
</tr>
<tr>
<td>Law</td>
<td>3,625</td>
<td>20%</td>
<td>5%</td>
<td>180</td>
</tr>
<tr>
<td>Mathematics</td>
<td>2,425</td>
<td>37%</td>
<td>11%</td>
<td>269</td>
</tr>
<tr>
<td>Medicine</td>
<td>4,024</td>
<td>23%</td>
<td>6%</td>
<td>230</td>
</tr>
<tr>
<td>Other</td>
<td>2,778</td>
<td>21%</td>
<td>6%</td>
<td>165</td>
</tr>
<tr>
<td>Pharmacy, Dentist &amp; Other</td>
<td>5,312</td>
<td>24%</td>
<td>6%</td>
<td>320</td>
</tr>
<tr>
<td>Physics &amp; Chemistry</td>
<td>1,734</td>
<td>58%</td>
<td>20%</td>
<td>349</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>5,574</td>
<td>26%</td>
<td>7%</td>
<td>385</td>
</tr>
</tbody>
</table>

Table 2: Summary statistics of first stage grades in the samples of (1) all, (2) pass after first stage (3) definite pass after second stage (The order of subgroups is given by the median of the first stage grades in the pass sample, column 6)

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>10th percentile</th>
<th>Min First stage</th>
<th>Min Pass</th>
<th>Median First stage</th>
<th>Median Pass</th>
<th>Maximum First stage</th>
<th>Maximum Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agrosciences</td>
<td>71.1</td>
<td>91.2</td>
<td>100.1</td>
<td>106.9</td>
<td>128.1</td>
<td>141.6</td>
<td>192.6</td>
</tr>
<tr>
<td>Other</td>
<td>66.1</td>
<td>102.1</td>
<td>104.8</td>
<td>102.0</td>
<td>136.7</td>
<td>143.3</td>
<td>187.5</td>
</tr>
<tr>
<td>Physics &amp; Chemistry</td>
<td>76.8</td>
<td>33.0</td>
<td>50.0</td>
<td>115.2</td>
<td>128.9</td>
<td>144.6</td>
<td>210.2</td>
</tr>
<tr>
<td>Humanities</td>
<td>67.9</td>
<td>96.3</td>
<td>99.2</td>
<td>104.2</td>
<td>133.6</td>
<td>147.1</td>
<td>203.3</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>68.9</td>
<td>101.0</td>
<td>102.0</td>
<td>109.4</td>
<td>138.6</td>
<td>147.9</td>
<td>214.3</td>
</tr>
<tr>
<td>Accountancy</td>
<td>80.5</td>
<td>120.5</td>
<td>122.9</td>
<td>120.3</td>
<td>139.9</td>
<td>151.5</td>
<td>200.7</td>
</tr>
<tr>
<td>Economics</td>
<td>71.8</td>
<td>113.3</td>
<td>121.1</td>
<td>110.9</td>
<td>133.8</td>
<td>152.3</td>
<td>209.2</td>
</tr>
<tr>
<td>Administration</td>
<td>68.6</td>
<td>108.5</td>
<td>121.0</td>
<td>108.7</td>
<td>140.9</td>
<td>154.2</td>
<td>212.3</td>
</tr>
<tr>
<td>Mathematics</td>
<td>75.8</td>
<td>70.3</td>
<td>73.0</td>
<td>122.1</td>
<td>151.7</td>
<td>158.9</td>
<td>222.1</td>
</tr>
<tr>
<td>Engineering</td>
<td>84.3</td>
<td>130.2</td>
<td>137.6</td>
<td>133.7</td>
<td>156.3</td>
<td>170.8</td>
<td>210.5</td>
</tr>
<tr>
<td>Pharmacy, Dentist &amp; Other</td>
<td>73.8</td>
<td>142.0</td>
<td>143.8</td>
<td>123.0</td>
<td>160.2</td>
<td>175.1</td>
<td>208.1</td>
</tr>
<tr>
<td>Law</td>
<td>77.4</td>
<td>165.5</td>
<td>168.0</td>
<td>139.5</td>
<td>179.4</td>
<td>189.5</td>
<td>215.2</td>
</tr>
<tr>
<td>Medicine</td>
<td>89.6</td>
<td>182.0</td>
<td>186.9</td>
<td>169.0</td>
<td>200.2</td>
<td>206.4</td>
<td>224.3</td>
</tr>
</tbody>
</table>

Table 3: Summary statistics of initial grades in the samples of (1) all, (2) pass after first stage, (3) definite pass after second stage (Medicine sample composed by three majors: Barbalha, Sobral and Fortaleza)

<table>
<thead>
<tr>
<th>Major</th>
<th>10th percentile</th>
<th>Min</th>
<th>Median</th>
<th>Maximum</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Pass</td>
<td>All</td>
<td>Pass</td>
<td>All</td>
</tr>
<tr>
<td>Barbalha</td>
<td>66.19</td>
<td>182.05</td>
<td>186.86</td>
<td>152.62</td>
<td>191.67</td>
</tr>
<tr>
<td>Sobral</td>
<td>121.57</td>
<td>185.05</td>
<td>186.86</td>
<td>171.76</td>
<td>200.76</td>
</tr>
<tr>
<td>Fortaleza</td>
<td>93.05</td>
<td>193.67</td>
<td>193.86</td>
<td>172.96</td>
<td>202.76</td>
</tr>
</tbody>
</table>

Table 4: Descriptive statistics in the two medical majors

**Sobral: 40 positions**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade: National Exam (m_0)</td>
<td>50.43</td>
<td>52.00</td>
<td>7.29</td>
<td>18.00</td>
<td>61.00</td>
<td>527</td>
</tr>
<tr>
<td>Grade: First stage</td>
<td>71.67</td>
<td>73.00</td>
<td>15.74</td>
<td>20.00</td>
<td>103.00</td>
<td>527</td>
</tr>
<tr>
<td>Grade: Second stage</td>
<td>240.0</td>
<td>246.5</td>
<td>33.98</td>
<td>94.3</td>
<td>296.6</td>
<td>160</td>
</tr>
<tr>
<td>Female</td>
<td>0.47</td>
<td>0</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
<td>527</td>
</tr>
<tr>
<td>Age</td>
<td>19.58</td>
<td>21.50</td>
<td>2.48</td>
<td>16.00</td>
<td>25.00</td>
<td>527</td>
</tr>
<tr>
<td>Private High School</td>
<td>0.87</td>
<td>1</td>
<td>0.33</td>
<td>0</td>
<td>1</td>
<td>527</td>
</tr>
<tr>
<td>Repetitions</td>
<td>0.99</td>
<td>1</td>
<td>0.88</td>
<td>0</td>
<td>2</td>
<td>527</td>
</tr>
<tr>
<td>Preparatory Course</td>
<td>0.71</td>
<td>1</td>
<td>0.45</td>
<td>0</td>
<td>1</td>
<td>527</td>
</tr>
<tr>
<td>Father’s education</td>
<td>2.09</td>
<td>2</td>
<td>1.03</td>
<td>0</td>
<td>3</td>
<td>527</td>
</tr>
<tr>
<td>Mother’s education</td>
<td>2.21</td>
<td>3</td>
<td>0.98</td>
<td>0</td>
<td>3</td>
<td>527</td>
</tr>
</tbody>
</table>

**Fortaleza: 150 positions**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade: National Exam (m_0)</td>
<td>49.16</td>
<td>52.00</td>
<td>10.03</td>
<td>12.00</td>
<td>63.00</td>
<td>2340</td>
</tr>
<tr>
<td>Grade: First stage</td>
<td>70.06</td>
<td>72.00</td>
<td>20.01</td>
<td>20.01</td>
<td>110.00</td>
<td>2340</td>
</tr>
<tr>
<td>Grade: Second stage</td>
<td>240.0</td>
<td>245.1</td>
<td>34.37</td>
<td>48.3</td>
<td>311.1</td>
<td>600</td>
</tr>
<tr>
<td>Female</td>
<td>0.54</td>
<td>1</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
<td>2340</td>
</tr>
<tr>
<td>Age</td>
<td>19.13</td>
<td>17.50</td>
<td>2.43</td>
<td>16.00</td>
<td>25.00</td>
<td>2340</td>
</tr>
<tr>
<td>Private High School</td>
<td>0.77</td>
<td>1</td>
<td>0.41</td>
<td>0</td>
<td>1</td>
<td>2340</td>
</tr>
<tr>
<td>Repetitions</td>
<td>0.69</td>
<td>1</td>
<td>0.83</td>
<td>0</td>
<td>2</td>
<td>2340</td>
</tr>
<tr>
<td>Preparatory Course</td>
<td>0.59</td>
<td>1</td>
<td>0.49</td>
<td>0</td>
<td>1</td>
<td>2340</td>
</tr>
<tr>
<td>Father’s education</td>
<td>2.13</td>
<td>2</td>
<td>1.00</td>
<td>0</td>
<td>3</td>
<td>2340</td>
</tr>
<tr>
<td>Mother’s education</td>
<td>2.15</td>
<td>2</td>
<td>0.98</td>
<td>0</td>
<td>3</td>
<td>2340</td>
</tr>
</tbody>
</table>

Table 5: First stage exam grade equation

<table>
<thead>
<tr>
<th></th>
<th>Specification 1</th>
<th>Specification 2</th>
<th>Specification 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spline(1)($m_0$)</td>
<td>48.9840</td>
<td>48.2575</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.8893)***</td>
<td>(3.9997)***</td>
<td></td>
</tr>
<tr>
<td>Spline(2)($m_0$ Residual)</td>
<td>89.4969</td>
<td>88.9733</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.7092)***</td>
<td>(4.5442)***</td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>26.1537</td>
<td>27.1143</td>
<td>78.1881</td>
</tr>
<tr>
<td></td>
<td>(3.4519)***</td>
<td>(3.5949)***</td>
<td>(2.1733)***</td>
</tr>
<tr>
<td>Female</td>
<td>0.4875</td>
<td>0.4320</td>
<td>0.4071</td>
</tr>
<tr>
<td></td>
<td>(0.4243)</td>
<td>(0.4068)</td>
<td>(0.4063)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.8326</td>
<td>-0.8747</td>
<td>-0.8886</td>
</tr>
<tr>
<td></td>
<td>(0.1093)***</td>
<td>(0.1127)</td>
<td>(0.1120)***</td>
</tr>
<tr>
<td>Special high school</td>
<td>-6.6897</td>
<td>-6.5658</td>
<td>-6.5589</td>
</tr>
<tr>
<td></td>
<td>(1.6927)***</td>
<td>(1.7256)***</td>
<td>(1.7234)***</td>
</tr>
<tr>
<td>Private high school</td>
<td>2.1812</td>
<td>2.0571</td>
<td>2.1960</td>
</tr>
<tr>
<td></td>
<td>(0.6496)***</td>
<td>(0.6772)**</td>
<td>(0.6662)***</td>
</tr>
<tr>
<td>Preparatory course</td>
<td>1.5672</td>
<td>1.5241</td>
<td>1.5378</td>
</tr>
<tr>
<td></td>
<td>(0.4385)***</td>
<td>(0.4912)**</td>
<td>(0.4910)**</td>
</tr>
<tr>
<td>Repetitions</td>
<td>2.7910</td>
<td>2.8861</td>
<td>2.8967</td>
</tr>
<tr>
<td></td>
<td>(0.3368)***</td>
<td>(0.3516)***</td>
<td>(0.3515)***</td>
</tr>
<tr>
<td>Ability($m_0$)</td>
<td></td>
<td></td>
<td>12.9015</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.6587)***</td>
</tr>
<tr>
<td>Living in Fortaleza</td>
<td>3.7139</td>
<td>3.6994</td>
<td>3.6414</td>
</tr>
<tr>
<td></td>
<td>(0.6545)***</td>
<td>(0.6600)***</td>
<td>(0.6581)***</td>
</tr>
<tr>
<td>Living in Fortaleza*Ability</td>
<td>1.9752</td>
<td>1.9843</td>
<td>1.9373</td>
</tr>
<tr>
<td></td>
<td>(0.7287)***</td>
<td>(0.6819)**</td>
<td>(0.6807)**</td>
</tr>
<tr>
<td>Mother’s education</td>
<td>0.1048</td>
<td>0.0503</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.3102)</td>
<td>(0.3116)</td>
<td></td>
</tr>
<tr>
<td>Father’s education</td>
<td>0.3578</td>
<td>0.4223</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.2924)</td>
<td>(0.2937)</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.7196</td>
<td>0.7199</td>
<td>0.7198</td>
</tr>
</tbody>
</table>

Notes:
1. Living in Fortaleza is a region dummy which indicates whether the student is currently living in Fortaleza.
2. Standard errors are between brackets and * sign marks the significance.
3. Standard errors in the first column are estimated by using bootstrap (200 replications); in other columns standard errors are robust asymptotic SEs.
Table 6: Second stage exam grade equation

<table>
<thead>
<tr>
<th></th>
<th>Specification 1</th>
<th>Specification 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>234.0629</td>
<td>171.8725</td>
</tr>
<tr>
<td></td>
<td>(12.3670)***</td>
<td>(17.7571)***</td>
</tr>
<tr>
<td>Female</td>
<td>7.2225</td>
<td>7.3699</td>
</tr>
<tr>
<td></td>
<td>(2.0446)**</td>
<td>(2.2697)**</td>
</tr>
<tr>
<td>Age</td>
<td>-3.9717</td>
<td>-3.9443</td>
</tr>
<tr>
<td></td>
<td>(0.71450)***</td>
<td>(0.7552)***</td>
</tr>
<tr>
<td>Special high school</td>
<td>-11.8078</td>
<td>-12.4952</td>
</tr>
<tr>
<td></td>
<td>(22.3426)</td>
<td>(21.8549)</td>
</tr>
<tr>
<td>Private high school</td>
<td>8.7807</td>
<td>8.9230</td>
</tr>
<tr>
<td></td>
<td>(4.4395)*</td>
<td>(4.1535)*</td>
</tr>
<tr>
<td>Preparatory course</td>
<td>9.3419</td>
<td>9.1989</td>
</tr>
<tr>
<td></td>
<td>(3.3598)**</td>
<td>(3.3812)**</td>
</tr>
<tr>
<td>Repetitions</td>
<td>14.1166</td>
<td>14.0344</td>
</tr>
<tr>
<td></td>
<td>(2.1688)***</td>
<td>(2.2251)***</td>
</tr>
<tr>
<td>Spline(1)(m₁ residual)</td>
<td>67.7974</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(16.7723)***</td>
<td></td>
</tr>
<tr>
<td>Spline(2)(m₁ residual)</td>
<td>152.3348</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(11.1541)***</td>
<td></td>
</tr>
<tr>
<td>u₁ (m₁ residual)</td>
<td>2.5066</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1416)***</td>
<td></td>
</tr>
<tr>
<td>Ability (m₀)</td>
<td>35.0695</td>
<td>35.1335</td>
</tr>
<tr>
<td></td>
<td>(2.4023)***</td>
<td>(3.5228)***</td>
</tr>
<tr>
<td>R²</td>
<td>0.2284</td>
<td>0.2286</td>
</tr>
</tbody>
</table>

Notes:
1. Standard errors are between brackets and * sign marks the significance.
2. Standard errors in the first column are estimated by using bootstrap (200 replications); in other columns standard errors are robust asymptotic SEs not correcting for the generated regressor issues.

Table 7: Simulated success probabilities

<table>
<thead>
<tr>
<th></th>
<th>Sobral</th>
<th>Fortaleza</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stage 1</td>
<td>Final Success</td>
</tr>
<tr>
<td>Min.</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>25%</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Median</td>
<td>0.088</td>
<td>0.011</td>
</tr>
<tr>
<td>Mean</td>
<td>0.314</td>
<td>0.076</td>
</tr>
<tr>
<td>75%</td>
<td>0.676</td>
<td>0.103</td>
</tr>
<tr>
<td>Max.</td>
<td>1.000</td>
<td>0.934</td>
</tr>
</tbody>
</table>

1. Success probabilities are constructed by 1000 Monte Carlo simulations.
Table 8: Simulated success probability
Odds Ratios

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Odds ratio 1</th>
<th>Odds ratio 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.00</td>
<td>2.66</td>
</tr>
<tr>
<td>20</td>
<td>1.00</td>
<td>1.60</td>
</tr>
<tr>
<td>30</td>
<td>1.47</td>
<td>1.08</td>
</tr>
<tr>
<td>40</td>
<td>0.86</td>
<td>1.61</td>
</tr>
<tr>
<td>50</td>
<td>1.07</td>
<td>2.26</td>
</tr>
<tr>
<td>60</td>
<td>1.33</td>
<td>3.43</td>
</tr>
<tr>
<td>70</td>
<td>1.29</td>
<td>5.34</td>
</tr>
<tr>
<td>75</td>
<td>1.18</td>
<td>5.62</td>
</tr>
<tr>
<td>80</td>
<td>1.15</td>
<td>5.22</td>
</tr>
<tr>
<td>85</td>
<td>1.14</td>
<td>4.41</td>
</tr>
<tr>
<td>90</td>
<td>1.10</td>
<td>3.73</td>
</tr>
<tr>
<td>95</td>
<td>1.03</td>
<td>3.37</td>
</tr>
<tr>
<td>100</td>
<td>1.00</td>
<td>1.74</td>
</tr>
</tbody>
</table>

1 Odds ratio 1 is the odds ratio of success probability at the first stage between subsamples of those who choose Sobral and choose Fortaleza $\frac{p_{sob|d_i=s}}{p_{sob|d_i=f}} / \frac{p_{fort|d_i=s}}{p_{fort|d_i=f}}$.

2 Odds ratio 2 is the odds ratio of final success probability at the second stage between subsamples of those who choose Sobral and choose Fortaleza $\frac{p_{sob|d_i=s}}{p_{sob|d_i=f}} / \frac{p_{fort|d_i=s}}{p_{fort|d_i=f}}$.

3 The percentile is computed with respect to the first stage exam grades.
Table 9: The determinants of preferences towards Sobral’s program

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Specification 1</th>
<th>Specification 2</th>
<th>Specification 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>δ₀</td>
<td>-2.514</td>
<td>-1.535</td>
<td>-1.413</td>
</tr>
<tr>
<td></td>
<td>(0.161)***</td>
<td>(0.500)***</td>
<td>(0.465)***</td>
</tr>
<tr>
<td>δₘ₀</td>
<td>0.114</td>
<td>0.186</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.109)</td>
<td>(0.184)</td>
<td></td>
</tr>
<tr>
<td>δₘₐ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>δₘ₀ₐ</td>
<td>-1.579</td>
<td>-1.537</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.522)***</td>
<td>(0.484)***</td>
<td></td>
</tr>
<tr>
<td>δₘₐₐ</td>
<td>-0.318</td>
<td>1.145</td>
<td>0.825</td>
</tr>
<tr>
<td></td>
<td>(0.260)</td>
<td>(0.348)***</td>
<td>(0.343)***</td>
</tr>
<tr>
<td>δₘₐₐₐ</td>
<td>1.158</td>
<td>0.942</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.189)***</td>
<td>(0.207)***</td>
<td></td>
</tr>
<tr>
<td>δₘₐₐₐₐ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.057</td>
<td>0.628</td>
<td>0.838</td>
</tr>
<tr>
<td></td>
<td>(0.420)</td>
<td>(0.563)</td>
<td>(0.563)</td>
</tr>
<tr>
<td>Ability (ₘ₀)</td>
<td>-1.012</td>
<td>-0.574</td>
<td>-0.346</td>
</tr>
<tr>
<td></td>
<td>(0.217)***</td>
<td>(0.221)***</td>
<td>(0.133)***</td>
</tr>
<tr>
<td>CFF</td>
<td>-0.528</td>
<td>-0.730</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.326)</td>
<td>(0.341)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.305</td>
<td>-0.276</td>
<td>-0.346</td>
</tr>
<tr>
<td></td>
<td>(0.122)***</td>
<td>(0.125)***</td>
<td>(0.133)***</td>
</tr>
<tr>
<td>Age</td>
<td>-0.044</td>
<td>-0.054</td>
<td>-0.071</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.030)</td>
<td>(0.031)***</td>
</tr>
<tr>
<td>Repetitions</td>
<td>0.708</td>
<td>0.708</td>
<td>0.765</td>
</tr>
<tr>
<td></td>
<td>(0.117)***</td>
<td>(0.127)***</td>
<td>(0.144)***</td>
</tr>
<tr>
<td>Father’s education</td>
<td>-0.265</td>
<td>-0.257</td>
<td>-0.280</td>
</tr>
<tr>
<td></td>
<td>(0.115)***</td>
<td>(0.087)***</td>
<td>(0.097)***</td>
</tr>
<tr>
<td>Mother’s education</td>
<td>0.088</td>
<td>0.083</td>
<td>0.155</td>
</tr>
<tr>
<td></td>
<td>(0.111)</td>
<td>(0.089)</td>
<td>(0.097)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Proportions</th>
<th>Specification 1</th>
<th>Specification 2</th>
<th>Specification 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>δₘₐ</td>
<td>Min 0.022</td>
<td>0.021</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>Mean 0.060</td>
<td>0.057</td>
<td>0.066</td>
</tr>
<tr>
<td></td>
<td>Max 0.122</td>
<td>0.248</td>
<td>0.196</td>
</tr>
<tr>
<td>δₘₐₐ</td>
<td>Min 0.015</td>
<td>0.016</td>
<td>0.365</td>
</tr>
<tr>
<td></td>
<td>Mean 0.385</td>
<td>0.412</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td>Max 0.816</td>
<td>0.852</td>
<td>0.559</td>
</tr>
<tr>
<td>δₘₐₐₐ</td>
<td>Min 0.062</td>
<td>0.027</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td>Mean 0.555</td>
<td>0.531</td>
<td>0.548</td>
</tr>
<tr>
<td></td>
<td>Max 0.963</td>
<td>0.962</td>
<td>0.585</td>
</tr>
</tbody>
</table>

1. In the first specification, we do not use the regional dummy CFF, which indicates whether the student is living in Fortaleza. Only the ability (ₘ₀) is used to explain the proportions of the choice set.
2. In the second specification, we use both CFF and ability.
3. The third specification only keeps CFF to explain the proportions.
4. The second part of the table shows the estimated probability of being in one of the three choice sets.
5. The coefficients and standard errors of the first specification are estimated by 200 bootstraps; and the coefficients and standard errors are obtained by one estimation of the observed data for the second specification.
Table 10: Thresholds of the Counterfactuals

<table>
<thead>
<tr>
<th>School</th>
<th>Stage 1</th>
<th>Sobral</th>
<th>Fortaleza</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original system</td>
<td>Mean Thresholds</td>
<td>184.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard Errors</td>
<td>(1.257)</td>
</tr>
<tr>
<td>Cutting seats</td>
<td></td>
<td>Mean Thresholds</td>
<td>195.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard Errors</td>
<td>(0.996)</td>
</tr>
<tr>
<td>Two-Choices</td>
<td></td>
<td>Mean Thresholds</td>
<td>186.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard error</td>
<td>(0.543)</td>
</tr>
<tr>
<td>Timing-Change</td>
<td></td>
<td>Mean Thresholds</td>
<td>205.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard error</td>
<td>(0.519)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>School</th>
<th>Stage 2</th>
<th>Sobral</th>
<th>Fortaleza</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original system</td>
<td>Mean Thresholds</td>
<td>235.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard Errors</td>
<td>(1.669)</td>
</tr>
<tr>
<td>Cutting seats</td>
<td></td>
<td>Mean Thresholds</td>
<td>233.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard Errors</td>
<td>(3.094)</td>
</tr>
<tr>
<td>Two-Choices</td>
<td></td>
<td>Mean Thresholds</td>
<td>236.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard error</td>
<td>(2.589)</td>
</tr>
<tr>
<td>Timing-Change</td>
<td></td>
<td>Mean Thresholds</td>
<td>260.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard error</td>
<td>(2.341)</td>
</tr>
</tbody>
</table>

The mean and standard errors of the thresholds are computed by resimulating the system for 1000 times using the convergent success probability and program choices.
The mean thresholds and standard errors of the original system are computed by 200 bootstraps.
The standard error of the counterfactual systems are conditional on the estimated $\beta$, and thus not the true standard errors.
Table 11: Utility change analysis of cutting seats counterfactual

<table>
<thead>
<tr>
<th>Expected Final Grade</th>
<th>ALL mean</th>
<th>s.d.</th>
<th>D=Sobral mean</th>
<th>s.d.</th>
<th>D=Fortaleza mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0% -50%</td>
<td>-0.00033</td>
<td>0.00059</td>
<td>-0.00048</td>
<td>0.00083</td>
<td>-0.00030</td>
<td>0.00052</td>
</tr>
<tr>
<td>50%-60%</td>
<td>-0.00383</td>
<td>0.00208</td>
<td>-0.00448</td>
<td>0.00254</td>
<td>-0.00360</td>
<td>0.00185</td>
</tr>
<tr>
<td>60%-70%</td>
<td>-0.00751</td>
<td>0.00633</td>
<td>-0.00872</td>
<td>0.00617</td>
<td>-0.00716</td>
<td>0.00635</td>
</tr>
<tr>
<td>70%-80%</td>
<td>-0.00694</td>
<td>0.01290</td>
<td>-0.01089</td>
<td>0.00926</td>
<td>-0.00623</td>
<td>0.01334</td>
</tr>
<tr>
<td>80%-82%</td>
<td>-0.00851</td>
<td>0.01289</td>
<td>-0.00754</td>
<td>0.00575</td>
<td>-0.00885</td>
<td>0.01460</td>
</tr>
<tr>
<td>82%-84%</td>
<td>-0.00075</td>
<td>0.01430</td>
<td>0.00073</td>
<td>0.00584</td>
<td>-0.00107</td>
<td>0.01558</td>
</tr>
<tr>
<td>84%-86%</td>
<td>0.01203</td>
<td>0.01002</td>
<td>-0.00200</td>
<td>0.01422</td>
<td>0.01433</td>
<td>0.00700</td>
</tr>
<tr>
<td>86%-88%</td>
<td>0.00229</td>
<td>0.01111</td>
<td>-0.00033</td>
<td>0.01141</td>
<td>0.00302</td>
<td>0.01105</td>
</tr>
<tr>
<td>88%-90%</td>
<td>0.01226</td>
<td>0.00860</td>
<td>0.00803</td>
<td>0.00319</td>
<td>0.01291</td>
<td>0.00899</td>
</tr>
<tr>
<td>90%-92%</td>
<td>0.01657</td>
<td>0.01203</td>
<td>0.00157</td>
<td>0.01103</td>
<td>0.01828</td>
<td>0.01102</td>
</tr>
<tr>
<td>92%-94%</td>
<td>0.01695</td>
<td>0.01015</td>
<td>0.00671</td>
<td>0.01245</td>
<td>0.01942</td>
<td>0.00779</td>
</tr>
<tr>
<td>94%-96%</td>
<td>0.02622</td>
<td>0.00550</td>
<td>0.01191</td>
<td>0.00157</td>
<td>0.02739</td>
<td>0.00376</td>
</tr>
<tr>
<td>96%-98%</td>
<td>0.03126</td>
<td>0.00648</td>
<td>0.01385</td>
<td>0.00293</td>
<td>0.03311</td>
<td>0.00305</td>
</tr>
<tr>
<td>98%-100%</td>
<td>0.02890</td>
<td>0.01028</td>
<td>0.01041</td>
<td>0.00230</td>
<td>0.03398</td>
<td>0.00343</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
E(\Delta U_i) & | 0.00080 & -0.00218 & 0.00149 \\
\text{s.d.}(\Delta U_i) & | 0.01107 & 0.00710 & 0.01170 \\
\Pr(\Delta U_i > 0) & | 0.3907 & 0.2934 & 0.4133 \\
\end{align*}
\]

1 ALL contains all the students no matter what the original choices are.
2 D=Sobral means the sub-population who choose Sobral in the original system; and D=Fortaleza means the sub-population who choose Fortaleza in the original system.
Table 12: Robustness check in cutting seats counterfactual with different $\mu$

<table>
<thead>
<tr>
<th>Expected Final Grade</th>
<th>$\mu = 0.8$</th>
<th>$\mu = 0$</th>
<th>$\mu = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>s.d.</td>
<td>mean</td>
</tr>
<tr>
<td>0%-50%</td>
<td>-0.00033</td>
<td>0.00059</td>
<td>-0.00030</td>
</tr>
<tr>
<td>50%-60%</td>
<td>-0.00383</td>
<td>0.00208</td>
<td>-0.00351</td>
</tr>
<tr>
<td>60%-70%</td>
<td>-0.00751</td>
<td>0.00633</td>
<td>-0.00698</td>
</tr>
<tr>
<td>70%-80%</td>
<td>-0.00694</td>
<td>0.01290</td>
<td>-0.00660</td>
</tr>
<tr>
<td>80%-82%</td>
<td>-0.00851</td>
<td>0.01289</td>
<td>-0.00842</td>
</tr>
<tr>
<td>82%-84%</td>
<td>-0.00075</td>
<td>0.01430</td>
<td>-0.00085</td>
</tr>
<tr>
<td>84%-86%</td>
<td>0.01203</td>
<td>0.01002</td>
<td>0.01185</td>
</tr>
<tr>
<td>86%-88%</td>
<td>0.00229</td>
<td>0.01111</td>
<td>0.00209</td>
</tr>
<tr>
<td>88%-90%</td>
<td>0.01226</td>
<td>0.00860</td>
<td>0.01186</td>
</tr>
<tr>
<td>90%-92%</td>
<td>0.01657</td>
<td>0.01203</td>
<td>0.01630</td>
</tr>
<tr>
<td>92%-94%</td>
<td>0.01695</td>
<td>0.01015</td>
<td>0.01651</td>
</tr>
<tr>
<td>94%-96%</td>
<td>0.02622</td>
<td>0.00550</td>
<td>0.02579</td>
</tr>
<tr>
<td>96%-98%</td>
<td>0.03126</td>
<td>0.00648</td>
<td>0.03077</td>
</tr>
<tr>
<td>98%-100%</td>
<td>0.02890</td>
<td>0.01028</td>
<td>0.02852</td>
</tr>
</tbody>
</table>

$E(\Delta U_i)$  
$s.d.(\Delta U_i)$  
$Pr(\Delta U_i > 0)$

<table>
<thead>
<tr>
<th></th>
<th>0.00080</th>
<th>-0.00218</th>
<th>0.00149</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01107</td>
<td>0.00710</td>
<td>0.01170</td>
</tr>
<tr>
<td></td>
<td>0.3907</td>
<td>0.2934</td>
<td>0.4133</td>
</tr>
</tbody>
</table>

1 The same sample is considered using different values of $\mu$. 
Table 13: Utility change analysis of two choice counterfactual

<table>
<thead>
<tr>
<th>Final Grade</th>
<th>Expected</th>
<th>ALL</th>
<th>D=Sobral</th>
<th>D=Fortaleza</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>s.d.</td>
<td>mean</td>
<td>s.d.</td>
</tr>
<tr>
<td>0%-50%</td>
<td>0.00019</td>
<td>0.00041</td>
<td>-0.00006</td>
<td>0.00015</td>
</tr>
<tr>
<td>50%-60%</td>
<td>0.00177</td>
<td>0.00206</td>
<td>-0.00046</td>
<td>0.00041</td>
</tr>
<tr>
<td>60%-70%</td>
<td>0.00601</td>
<td>0.00642</td>
<td>-0.00040</td>
<td>0.00038</td>
</tr>
<tr>
<td>70%-80%</td>
<td>0.01226</td>
<td>0.01022</td>
<td>-0.00009</td>
<td>0.00050</td>
</tr>
<tr>
<td>80%-82%</td>
<td>0.02100</td>
<td>0.01562</td>
<td>0.00019</td>
<td>0.00012</td>
</tr>
<tr>
<td>82%-84%</td>
<td>0.01962</td>
<td>0.01363</td>
<td>0.00023</td>
<td>0.00015</td>
</tr>
<tr>
<td>84%-86%</td>
<td>0.01433</td>
<td>0.01039</td>
<td>0.00023</td>
<td>0.00013</td>
</tr>
<tr>
<td>86%-88%</td>
<td>0.02760</td>
<td>0.01886</td>
<td>0.00034</td>
<td>0.00012</td>
</tr>
<tr>
<td>88%-90%</td>
<td>0.02968</td>
<td>0.01713</td>
<td>0.00033</td>
<td>0.00010</td>
</tr>
<tr>
<td>90%-92%</td>
<td>0.02497</td>
<td>0.02098</td>
<td>0.00049</td>
<td>0.00025</td>
</tr>
<tr>
<td>92%-94%</td>
<td>0.03179</td>
<td>0.02228</td>
<td>0.00054</td>
<td>0.00021</td>
</tr>
<tr>
<td>94%-96%</td>
<td>0.03469</td>
<td>0.01907</td>
<td>0.00048</td>
<td>0.00022</td>
</tr>
<tr>
<td>96%-98%</td>
<td>0.04127</td>
<td>0.02469</td>
<td>0.00073</td>
<td>0.00026</td>
</tr>
<tr>
<td>98%-100%</td>
<td>0.03756</td>
<td>0.02944</td>
<td>0.00063</td>
<td>0.00030</td>
</tr>
</tbody>
</table>

E($\Delta U_i$) | 0.00780 | -0.00008 | 0.00964
s.d.($\Delta U_i$) | 0.01507 | 0.0039 | 0.01619
Pr($\Delta U_i > 0$) | 0.8814 | 0.3726 | 1

1. ALL contains all students no matter what the original choices are.
2. D=Sobral means the sub-population of those who choose Sobral in the original system; and D=Fortaleza means the sub-population of those who choose Fortaleza in the original system.
Table 14: Utility change analysis of timing change counterfactual

<table>
<thead>
<tr>
<th>Expected Final Grade</th>
<th>ALL mean</th>
<th>s.d.</th>
<th>D=Sobral mean</th>
<th>s.d.</th>
<th>D=Fortaleza mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%- 50%</td>
<td>0.00221</td>
<td>0.00383</td>
<td>0.00146</td>
<td>0.00223</td>
<td>0.00237</td>
<td>0.00408</td>
</tr>
<tr>
<td>50%-60%</td>
<td>0.01885</td>
<td>0.01194</td>
<td>0.00784</td>
<td>0.00807</td>
<td>0.02231</td>
<td>0.01082</td>
</tr>
<tr>
<td>60%-70%</td>
<td>0.04189</td>
<td>0.02198</td>
<td>0.01841</td>
<td>0.01606</td>
<td>0.04957</td>
<td>0.01783</td>
</tr>
<tr>
<td>70%-80%</td>
<td>0.09641</td>
<td>0.03976</td>
<td>0.04911</td>
<td>0.04427</td>
<td>0.10524</td>
<td>0.03194</td>
</tr>
<tr>
<td>80%-82%</td>
<td>0.13401</td>
<td>0.04577</td>
<td>0.06602</td>
<td>0.04275</td>
<td>0.15440</td>
<td>0.01940</td>
</tr>
<tr>
<td>82%-84%</td>
<td>0.14342</td>
<td>0.04150</td>
<td>0.07605</td>
<td>0.04903</td>
<td>0.15914</td>
<td>0.01653</td>
</tr>
<tr>
<td>84%-86%</td>
<td>0.17757</td>
<td>0.03936</td>
<td>0.12277</td>
<td>0.07257</td>
<td>0.18806</td>
<td>0.01539</td>
</tr>
<tr>
<td>86%-88%</td>
<td>0.19540</td>
<td>0.04594</td>
<td>0.13027</td>
<td>0.06396</td>
<td>0.21277</td>
<td>0.01406</td>
</tr>
<tr>
<td>88%-90%</td>
<td>0.22733</td>
<td>0.01694</td>
<td>0.20861</td>
<td>0.02565</td>
<td>0.22913</td>
<td>0.01503</td>
</tr>
<tr>
<td>90%-92%</td>
<td>0.27670</td>
<td>0.08313</td>
<td>0.14890</td>
<td>0.09739</td>
<td>0.30017</td>
<td>0.05478</td>
</tr>
<tr>
<td>92%-94%</td>
<td>0.28371</td>
<td>0.06732</td>
<td>0.19483</td>
<td>0.08725</td>
<td>0.30548</td>
<td>0.03824</td>
</tr>
<tr>
<td>94%-96%</td>
<td>0.34077</td>
<td>0.05782</td>
<td>0.21722</td>
<td>0.13380</td>
<td>0.35337</td>
<td>0.02115</td>
</tr>
<tr>
<td>96%-98%</td>
<td>0.42198</td>
<td>0.07927</td>
<td>0.28169</td>
<td>0.17460</td>
<td>0.43916</td>
<td>0.03365</td>
</tr>
<tr>
<td>98%-100%</td>
<td>0.54759</td>
<td>0.16372</td>
<td>0.35510</td>
<td>0.24210</td>
<td>0.59972</td>
<td>0.07894</td>
</tr>
</tbody>
</table>

\[
\mathbb{E}(\Delta U_i) = 0.07259, \quad \text{s.d.}(\Delta U_i) = 0.12694, \quad \text{Pr}(\Delta U_i > 0) = 1
\]

1. ALL contains all the students no matter what the original choices are.
2. D=Sobral means the sub-population of those who choose Sobral in the original system; and D=Fortaleza means the sub-population of those who choose Fortaleza in the original system.
Figure 1: Choice space

\[ P_S V_S \geq P_F V_F \]
\[ P_S V_S \leq P_F V_F \]
Figure 2: Density plots of the grades

Sobral m1 density

Fortaleza m1 density

Sobral m2 density

Fortaleza m2 density

N = 527  Bandwidth = 4.045

N = 2340  Bandwidth = 3.817

N = 160  Bandwidth = 9.857

N = 598  Bandwidth = 8.35
Figure 3: The relation between ability and first stage grades

[1] The round grey points are the scatter plots of first stage grade on ability (normalized Enem); [2] The lines are the LOWESS of first stage grade on ability (normalized Enem).
Figure 4: The Odds ratio plot of simulated success probabilities

[1] The star points is the plot of odds ratio at the first stage; [2] the triangular points is the plot of odds ratio at the second stage.
Figure 5: Cutting seats counterfactual: Success probability change in Sobral

[1] The circle dots are the success probability change plot with respect to expected final grade; [2] From left to right, 1) the first vertical line is the median, 2) the second line is the quantile of 1st stage admission $- (1 - \frac{4(n+nof)}{nobs}) \times 100\%$, and 3) the third line is the quantile of 2nd stage admission $- (1 - \frac{3(n+nof)}{nobs}) \times 100\%$. [3] nos is the number of final seats in Sobral, nof is the number of final seats in Fortaleza and nobs is the number of total applicants.
Figure 6: Cutting seats counterfactual: Success probability change in Fortaleza

[1] The circle dots are the success probability change plot with respect to expected final grade; [2] From left to right, 1) the first vertical line is the median, 2) the second line is the quantile of 1st stage admission – (1 – \( \frac{nos+nof}{nobs} \)) \times 100\%, and 3) the third line is the quantile of 2nd stage admission – (1 – \( \frac{nos+nof}{nobs} \)) \times 100\%. [3] nos is the number of final seats in Sobral, nof is the number of final seats in Fortaleza and nobs is the number of total applicants.
Figure 7: Utility change in Cutting seats counterfactual

Utility Comparison by exam performance

[1] the gray star dots are the scatter plot of utility difference onto expected final grade for those choose Sobral in the original system. [2] the blue triangle dots are the scatter plot of utility difference onto expected final grade for those choose Fortaleza; [3] the red line is the 0 level.
Figure 8: Two choice counterfactual: Success probability change in Sobral

[1] The circle dots are the success probability change plot with respect to expected final grade; [2] From left to right, 1) the first vertical line is the median, 2) the second line is the quantile of 1st stage admission \(- (1 - \frac{4(nos+nof)}{nobs}) \times 100\%\), and 3) the third line is the quantile of 2nd stage admission \(- (1 - \frac{nos+nof}{nobs}) \times 100\%\). [3] nos is the number of final seats in Sobral, nof is the number of final seats in Fortaleza and nobs is the number of total applicants.
Figure 9: Two choice counterfactual: Success probability change in Fortaleza

[1] The circle dots are the success probability change plot with respect to expected final grade; [2] From left to right, 1) the first vertical line is the median, 2) the second line is the quantile of 1st stage admission \(-\frac{\text{flos} + \text{nof}}{nobs}\) \times 100\%, and 3) the third line is the quantile of 2nd stage admission \(-\frac{\text{flos} + \text{nof}}{nobs}\) \times 100\%. [3] nos is the number of final seats in Sobral, nof is the number of final seats in Fortaleza and nobs is the number of total applicants.
Figure 10: Utility change in Two choice counterfactual

Utility Comparison by exam performance

[1] the gray star dots are the scatter plot of utility difference onto expected expected final grade for those choose Sobral in the original system. [2] the blue triangle dots are the scatter plot of utility difference onto expected expected final grade for those choose Fortaleza; [3] the red line is the 0 level.
Figure 11: Timing change counterfactual: Success probability change in Sobral

[1] The circle dots are the success probability change plot with respect to expected final grade; [2] From left to right, 1) the first vertical line is the median, 2) the second line is the quantile of 1st stage admission \(- (1 - \frac{4(nos+nof)}{nobs}) \times 100\%\), and 3) the third line is the quantile of 2nd stage admission \(- (1 - \frac{nos+nof}{nobs}) \times 100\%\). [3] nos is the number of final seats in Sobral, nof is the number of final seats in Fortaleza and nobs is the number of total applicants.
Figure 12: Timing change counterfactual: Success probability change in Fortaleza

[1] The circle dots are the success probability change plot with respect to expected final grade; [2] From left to right, 1) the first vertical line is the median, 2) the second line is the quantile of 1st stage admission – \(1 - \frac{n_{\text{nos}} + n_{\text{nof}}}{n_{\text{obs}}}\) × 100%, and 3) the third line is the quantile of 2nd stage admission – \(1 - \frac{n_{\text{nos}} + n_{\text{nof}}}{n_{\text{obs}}}\) × 100%. [3] nos is the number of final seats in Sobral,nof is the number of final seats in Fortaleza and nob is the number of total applicants.
Figure 13: Utility change in Timing change counterfactual

[1] the gray star dots are the scatter plot of utility difference onto expected expected final grade for those choose Sobral in the original system. [2] the blue triangle dots are the scatter plot of utility difference onto expected expected final grade for those choose Fortaleza; [3] the red line is the 0 level.