DISCUSSION PAPER SERIES

IZA DP No. 8550

College Choice Allocation Mechanisms: Structural Estimates and Counterfactuals

José-Raimundo Carvalho
Thierry Magnac
Qizhou Xiong

October 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

College Choice Allocation Mechanisms: Structural Estimates and Counterfactuals

José-Raimundo Carvalho
CAEN, Universidade Federal do Ceará

Thierry Magnac
Toulouse School of Economics
and IZA

Qizhou Xiong
Toulouse School of Economics

Discussion Paper No. 8550
October 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8550
October 2014

ABSTRACT

College Choice Allocation Mechanisms: Structural Estimates and Counterfactuals

We evaluate a simple allocation mechanism of students to majors at college entry that was commonly used in universities in Brazil in the 1990s and 2000s. Students first chose a single major and then took exams that select them in or out of the chosen major. The literature analyzing student placement, 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 reporting choices between two medical schools and grade performances at the entry exams. Given those estimates, 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 of the allocation mechanisms.

JEL Classification: C57, D47, I21

Keywords: education, two-sided matching, school allocation mechanism, policy evaluation

Corresponding author:

Thierry Magnac
Université de Toulouse 1 Capitole
21 Allée de Brienne
31000 Toulouse
France
E-mail: thierry.magnac@tse-fr.eu

1 Introduction[1]

Matching students with university majors in Brazil is a very competitive process and in particular in public federal universities which are mostly 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 the exclusivity of Brazilian universities. What made Brazil specific in the years 2000s was the formality of the selection process at the level of each university. 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). Moreover, in contrast to countries such as Turkey (Balinski and Sonmez, 1999), the organization of selection was decentralized at the level of universities until 2010.

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 further 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 characteristic is that the exam consists in two stages. The first stage is common to all majors and consists of many multiple-choice tests evaluating knowledge in a definite subject, e.g. mathematics, Portuguese etc.. The second stage exams are specific to each major and have a more traditional short-answer or essay format.

The issue at hand is to match students with colleges which are in our case, the schools offering undergraduate majors at the university (medicine, engineering and so on). Matching students to

[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. Useful discussions with Yinghua He, Jean-Marc Robin and comments by participants at conferences in Brown, Bristol, Atlanta, Northwestern, Shanghai and Rio de Janeiro and seminars at Oxford, CREST, CEMMAP, Cambridge, Amsterdam and Barcelona are gratefully acknowledged. This research 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.

schools has a long history (Roth and Sotomayor, 1992) and a brief survey of the recent literature is given in Roth (2008). 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). Such matching or allocation mechanisms are caracterized by a few theoretical properties. First, they could be *stable*, or *fair* in the student placement literature, in the sense that there is no pair (student, major) 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. revealing their true preferences is a weakly dominant strategy for every student. For instance, 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 in comparison with other students.

In a nutshell, the form under which the vestibular was organized at UFC in 2004 is difficult to justify. This mechanism transforms a centralized allocation of all students to all majors into a decentralized system in which each major selects its own students from a blocked list of applications. The mechanism is thus neither stable – there exist pairs of student & school which could be made better off by changing the final allocation – nor strategy-proof. Students prefer to disguise their preferences for very demanded majors (e.g. medicine) into preferences for less demanded ones (e.g. dentistry) in order to improve their probability of being accepted.

What we do in this paper is to contribute to the empirical literature on this subject by evaluating the effects on student allocations and their welfare of adopting other more theoretically sound mechanisms than the current *vestibular*. In the absence of experiments (Calsamiglia, Haeringer and Klijn, 2010) or quasi-experiments (Pathak and Sonmez, 2013), estimating a structural model is key for our empirical strategy. Furthermore, even if the mechanism does not induce students to reveal true preferences, we are able to take advantage of its specific format and the data we have to model in more detail than in the current literature, the strategies employed by students.

Our first contribution is to build such a structural model of major choices that is derived from the literature on school choices (Arcidiacono, 2005, Epple, Romano and Sieg, 2006 or Bourdabat

[2]Specifically, it eliminates the need to look at preferences over groups of students (e.g. couples)

and Montmarquette, 2007). School choices depend on (1) expected probabilities of success and (2) preferences for colleges including future wages after college. Precise information on success during college years and wages later on is lacking in our data and we use reduced form functions of past educational history and ability. The advantage of our data lies in the rich information on performance at the two-stage exams before entry (or failure). We observe grades for all students taking the exams as well as an initial measure of ability obtained a year before the exams are taken. Entry of students into specific majors is summarized by major-specific thresholds on grades at the two successive exams. Students enter a major if their exam grades are above thresholds determined in the sample.

Our second original contribution is a detailed derivation of the expected success probabilities in entering a major given the observed distribution of grades. Students in our sample play an imperfect information game in which a Bayesian Nash equilibrium exists under general conditions and that we show to be unique in a restricted homogenous setting. In the empirical model, we posit the specific type of information that students have (see Manski, 1993, for a critical appraisal of such assumptions) and we assume that expectations of success probabilities are perfect that is, they are obtained by infinitely repeating the game with the same players. We show that, conditional on information sets, success probabilities can be obtained by resampling in our single observed sample and by using the conditions of the Nash Bayesian equilibrium that deliver random thresholds.

We also provide conditions for non parametric identification of the different objects of interest appearing in grade equations, success probabilities and preferences by using either control functions or/and exclusion restrictions. We estimate grade and preference parameters 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 group of majors. For simplicity also, we use a sequence of semi-parametric regressions and a parametric discrete choice model that depends on the simulated expectations of the success probabilities derived from the procedure summarized above.

Our final contribution is to analyze the effects on allocation and welfare of students and schools using three different counterfactual mechanisms. Microanalysis allows us not only to study average welfare effects but also detail redistributive effects between schools and between students due to changes in the allocation mechanisms. In the first experiment, we restrict 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. We show that this risk is small. Second, the most worthy of attention counterfactual experiment gives students more choices as in a deferred acceptance mechanism. Students are allowed to submit a list of two choices instead of a single choice and in consequence, the result should be stable and strategy-proof. We show that indeed, enlarging the choice set has a positive aggregate effect in terms of utilitarist social welfare but has also distributive effects. This allows us to show that strategic effects in the original mechanism are sizeable. A timing change of choices and exams is our third counterfactual experiment. We allow students to choose their majors after passing the first-stage exam instead of having them choose before this exam. As expected it has strong redistributive effects between schools and between students.

A brief review of the literature The paper builds upon various strands in the literature and in particular student placement. In a theoretical work albeit oriented towards the analysis of a specific mechanism, Balinski and Sönmez (1999) study the optimality of the placement of students in Turkish universities in which selection and competition among students are nationwide. 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.

The theoretical literature on allocation mechanisms and their respective advantages under various conditions is rapidly growing as quickly as the number of institutions which adopt such allocation mechanisms (e.g. Pathak and Sonmez, 2013, Abdulkadiroglu, Che and Yasuda, 2012). As for empirical papers, Abdulkadiroğlu, Pathak, Roth and Sonmez (2006) exhibit strong empirical evidence of the strategic moves by parents in the allocation mechanism used in Boston primary schools. They argued that the Boston School Committee should change the Boston mechanism in place into 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 and Abdulkadiroğlu, Agarwal and Pathak (2013) estimates demands in the newly introduced mechanism which is a deferrence acceptance mechanism. They are able to compare this allocation

in terms of welfare and distributive effects to the previous decentralized allocation.

Demands for schools are estimated in Hastings, Kane and Staiger (2009) to study how the enhancement of choice sets might have unintended consequences for minority students as well as by Agarwal (2013) in which medical schools and medical residents preferences are estimated using a double sided school choice model. There are other papers analyzing school choice such as in Lai, Sadoulet and de Janvry (2009). Others 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 the relative standing of the Gale-Shapley and the Boston mechanisms (see Abdulkadiroğlu, Che and Yasuda, 2012 in a school choice problem, Budish and Cantillon, 2012 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 the conditions of a Bayesian Nash equilibrium. It also explains how expectations can be derived from this structure. Section 3 presents the econometric model of grade equations and college choices, discusses their non parametric identification and explains the estimation procedure. Section 4 provides a descriptive analysis of the mechanism in place and the results of the estimation of grade and preference equations and preference shifters. Section 5 details the results of the three counterfactual experiments. Section 6 concludes the paper.

2 Description of the game and modeling

We start by describing how Universidade Federal do Ceara (UFC) in Northeastern Brazil 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 conditionned by the grade obtained at the first stage and students are selected within the population of those who have chosen a given major. Are accepted to the second stage 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

[3]We use the terms "major", "school" or "program" interchangeably.

University. Appendix A gives further details on the mechanism and the exams.

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 from models of college choice. 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 among 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 conditions for existence and uniqueness of a Bayesian Nash equilibrium of this game in a restricted setting.

2.1 Timing for the Decision maker

We omit the individual index for readability. A random variable, say D, describes school choice and takes realizations, d, a specific major. For simplicity, we restrict the number of majors to two whose names are S (later denoting 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 and since the extension of the model to any number of majors is trivial. The outside option is denoted $d = \varnothing$. 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 identification (see below).

We describe the *Vestibular* system by 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 measures

[4]ENEM is a non-mandatory Brazilian national exam, which evaluates high school education in Brazil. Until 2008, the exam consisted in two tests: a 63 multiple-choice test on different subjects (Portuguese, History,

students' ability. The result of this exam is also used by the University when computing the passing thresholds at the *Vestibular* exams.

- **Stage 1 – Choice of Major:** Students apply for one major among the available options, $d \in \{\varnothing, S, F\}$. The outside option $d = \varnothing$ 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. We do not observe those who choose the outside option.

- **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, Z, 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. This threshold depends on $ENEM$, in other words are individual specific, because students are ranked according to a weighted sum of m_1 and $ENEM$ but we make this dependence implicit in the following.

Geography, Math, Physics, Chemistry and Biology) and writing an essay.

Students who do not pass the first exam get their outside option $D = \varnothing$, with utility, $V_\varnothing$, 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 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. As before, we write the selection rule as:

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

as a function of a second threshold. Again this threshold depends on previous grades since this a grade linearly aggregating $ENEM$, m_1 and m_2 which is used to rank students. 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, say V_D, which is determined by their preferences and expected earnings of this major.

There could be additional decision nodes to take into account when preferences are 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.

Nonetheless, we do not have any information on students who quit before the exams since our sample consists 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 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:

$$\mathbb{E}V_D = P^D V_D + (1 - P^D)V_\varnothing .$$

We can normalize $V_\varnothing = 0$ and therefore choices are obtained by maximizing expected utility as:

$$\begin{aligned} D = S \quad &\text{if } P^S V^S > P^F V^F, \\ D = F \quad &\text{if } P^S V^S \leq P^F V^F. \end{aligned} \tag{1}$$

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, ε is an unobservable preference random term and ζ are preference parameters. It is enough at this stage to define choices as $D(X, \varepsilon, \zeta, P^S, P^F)$. For simplicity, we shall assume in the following that preference shocks, ε and performance shocks, $u = (u_1, u_2)$ are independent. This is a testable assumption that will be evaluated in the empirical section.

2.2 Expectations and Bayesian Nash equilibrium

Denote β (respectively ζ) 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 $T = (T_1^S, T_2^S, T_1^F, T_2^F)$ 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]

[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.

Namely, thresholds affect outcomes in two ways. First, realized thresholds, t_j^d, 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. We assume that expectations of thresholds are perfect in the sense that they should match the distribution of their realized values across any possible sampling scheme. 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, 1993). 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 $n+1 >> n_S + n_F$ 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 one applicant, indexed by 0, from all other applicants to both schools $i = 1, ., n$ and we analyze her decision making. We can proceed this way 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 ε_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, $(X_i, Z_i, \varepsilon_i, u_i)$ $i = 1, ., n$. We assume that characteristics $(X_i, Z_i, \varepsilon_i)$ for all $i = 1, ., n$ 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]

[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 for further research.

[7] Those assumptions are among the strongest that we could make and assume that agents are highly sophisticated.

After this initial stage, student 0 chooses her major ($D_0 \in \{S, F\}$) as a function of their expectations of success P_0^S and P_0^F 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, $i = 1, ., n$, the latter by u_0. Integrating out both risks allows us to derive success probabilities and form what will be the expectations of success of student 0.

2.2.2 Success probabilities

Denote $W_{(n)}^S$ (respectively $W_{(n)}^F$) the characteristics of the sub-sample of students $i = 1, ., n$ applying to Sobral (respectively Fortaleza) observed by student 0. By construction $W_{(n)} = W_{(n)}^S \cup W_{(n)}^F$ and $W_{(n)}^S \cap W_{(n)}^F = \varnothing$. Similarly, we denote $U_{(n)}^S$ and $U_{(n)}^F$ 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

$$\mathbf{1}\{m_1(Z_0, u_0, \beta) \geq T_1^S(W_{(n)}^S, U_{(n)}^S), m_2(Z_0, u_0, \beta) \geq T_2^S(W_{(n)}^S, U_{(n)}^S)\}$$

in which $T_1^d(.)$ and $T_2^d(.)$ 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 $W_{(n)}^d$ and the realization of their grade shocks is equal to $U_{(n)}^d$. 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

We could assume that agents' information set of the agents is reduced to performance shifters and choices $W_{(n)} = (Z_i, D_i)_{i=1,.n}$ and this would not modify the estimation stage. The counterfactuals would however be more difficult to construct. We leave for further work the developments of less informative frameworks.

determined by $\mathbf{1}\{m_1(Z_0,u_0,\beta) \geq T_1^F(W_{(n)}^F, U_{(n)}^F), m_2(Z_0,u_0,\beta) \geq T_2^F(W_{(n)}^F, U_{(n)}^F)\}$. To distinguish those thresholds from the ones defined in the complete sample of $i = 1,.,n$ AND $i = 0$ we denote them as:

$$\tilde{T}_{1,0}^d = T_1^d(W_{(n)}^d, U_{(n)}^d), \tilde{T}_{2,0}^d = T_2^d(W_{(n)}^d, U_{(n)}^d) \text{ for } d = S, F.$$

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

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:[8]

$$\begin{aligned} P_0^d(Z_0,\beta) &= E_{U_{(n)},u_0}\left[\mathbf{1}\{m_1(Z_0,u_0,\beta) \geq \tilde{T}_{1,0}^d, m_2(Z_0,u_0,\beta) \geq \tilde{T}_{2,0}^d\} \mid Z_0, W_{(n)}\right], \\ &= E_{U_{(n)}}\left[p^d(Z_0,\beta,\tilde{T}_{1,0}^d,\tilde{T}_{2,0}^d) \mid Z_0, W_{(n)}\right], \end{aligned} \tag{2}$$

in which the following function concerns the individual shock u_0 only:

$$p^d(Z_0,\beta,\tilde{T}_{1,0}^d,\tilde{T}_{2,0}^d) = E_{u_0}\left[\mathbf{1}\{m_1(Z_0,u_0,\beta) \geq \tilde{T}_{1,0}^d, m_2(Z_0,u_0,\beta) \geq \tilde{T}_{2,0}^d\} \mid Z_0, \tilde{T}_{1,0}^d, \tilde{T}_{2,0}^d\right]. \tag{3}$$

These are the success probabilities that can be computed from observing a single sample, W_n when $\tilde{T}_{1,0}^d, \tilde{T}_{2,0}^d$ $d = S, F$ 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

$$\begin{cases} P_0^S = P^S(Z_0, W_{(n)}, \beta) = E\left[p^S(Z_0,\beta,\tilde{T}_{1,0}^S,\tilde{T}_{2,0}^S) \mid Z_0, W_{(n)}\right], \\ P_0^F = P^F(Z_0, W_{(n)}, \beta) = E\left[p^F(Z_0,\beta,\tilde{T}_{1,0}^F,\tilde{T}_{2,0}^F) \mid Z_0, W_{(n)}\right]. \end{cases} \tag{4}$$

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, ε_0) and they depend on $W_{(n)}$ only through $\tilde{T}_{j,0}^D$ that are computed below.

Denote $D_0(X_0, \varepsilon_0, \zeta, P_0^S, P_0^F) \in \{S, F\}$ 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, $i = 1,.,n$, we can by substitution construct the samples of applicants to Sobral (say) by using:

$$W_{(n)}^S = \{i \in \{1,.,n\}; D_i(X_i, \varepsilon_i, \zeta, P_i^S, P_i^F) = S\}.$$

[8] All expectations exist since integrands are measurable and bounded.

It is thus clear that the application mapping $W_{(n)}$ into $W^S_{(n)}$ or $W^F_{(n)}$ is measurable although it remains to be shown 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. That is what we do now.

2.2.3 Bayesian Nash equilibrium and the determination of the thresholds

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

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 $(u_0, u_1, ., u_n)$, are fourfold:

$$\begin{cases} \sum_{i=0}^{n} [\mathbf{1}\{D_i(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = S\}\mathbf{1}\{m_1(Z_i, u_i, \beta) \geq t^S_1] = 4n_S, \\ \sum_{i=0}^{n} [\mathbf{1}\{D_i(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = F\}\mathbf{1}\{m_1(Z_i, u_i, \beta) \geq t^F_1] = 4n_F, \\ \sum_{i=0}^{n} [\mathbf{1}\{D_i(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = S\}\mathbf{1}\{m_1(Z_i, u_i, \beta) \geq t^S_1, m_2(Z_i, u_i, \beta) \geq t^S_2] = n_S, \\ \sum_{i=0}^{n} [\mathbf{1}\{D_i(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = F\}\mathbf{1}\{m_1(Z_i, u_i, \beta) \geq t^F_1, m_2(Z_i, u_i, \beta) \geq t^F_2] = n_F. \end{cases} \tag{5}$$

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. For instance, the number of students admitted in major S is equal to the number of available seats.[9]

As usual with dummy variable equations, this system has many solutions $(t^S_1, t^F_1, t^S_2, t^F_2)$ in an hypercube $\mathcal{C}$ in $\mathbb{R}^4$. We retain the solution corresponding to the upper north-west corner i.e. $(\max_{\mathcal{C}} t^S_1, \max_{\mathcal{C}} t^F_1, \max_{\mathcal{C}} t^S_2, \max_{\mathcal{C}} t^F_2)$ 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 which is a measurable function of Z_0 and $W_{(n)}$.[10]

[9] 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 as 0. The average success probability of 5% in our data means that the probability of this event is negligible.

[10] Note that this aplies to the sophisticated version $W_{(n)} = (X_i, Z_i, \varepsilon_i)$ as well as the more restricted version $W_{(n)} = (Z_i, D_i)$ since equation (5) only depend on the restricted set of variables.

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

$$\begin{cases} \sum_{i=1}^{n}[\mathbf{1}\{D_i(X_i,\varepsilon_i,\zeta,P_i^S,P_i^F)=S\}\mathbf{1}\{m_1(Z_i,u_i,\beta)\geq\tilde{t}_1^S] = 4n_S, \\ \sum_{i=1}^{n}[\mathbf{1}\{D_i(X_i,\varepsilon_i,\zeta,P_i^S,P_i^F)=F\}\mathbf{1}\{m_1(Z_i,u_i,\beta)\geq\tilde{t}_1^F] = 4n_F, \\ \sum_{i=1}^{n}[\mathbf{1}\{D_i(X_i,\varepsilon_i,\zeta,P_i^S,P_i^F)=S\}\mathbf{1}\{m_1(Z_i,u_i,\beta)\geq\tilde{t}_1^S, m_2(Z_i,u_i,\beta)\geq\tilde{t}_2^S] = n_S, \\ \sum_{i=1}^{n}[\mathbf{1}\{D_i(X_i,\varepsilon_i,\zeta,P_i^S,P_i^F)=F\}\mathbf{1}\{m_1(Z_i,u_i,\beta)\geq\tilde{t}_1^F, m_2(Z_i,u_i,\beta)\geq\tilde{t}_2^F] = n_F. \end{cases} \tag{6}$$

Notice that the choices of other students $\mathbf{1}\{D_i(X_i,\varepsilon_i,\zeta,P_i^S,P_i^F)=S\}$ 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

When using the current mechanism or counterfactual experiments below, the question of the uniqueness of the equilibrium is pending. 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. It is easier to prove uniqueness in a simpler context and this is what we do now. We assume that the scheme is the current selection scheme and that agents' preferences and performances are homogenous. 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 allow however for an arbitrary number of majors, K_D.

We define success probabilities, $\{P_1^d\}_{d=1,.,K_D}$, at the first stage exam and $\{P^d\}_{d=1,.,K_D}$ at the second stage exam and we pile up these objects into K_D-dimensional vectors P_1 and P. These probabilities are common among agents. School choices, $D(\zeta,\varepsilon,P)$, are given by the comparison between expected value functions $\{P^dV^d\}_{d=1,.,K_D}$ as in equation (1) in which the value function V^d describes preferences for major d. Without loss of generality, we can set $V^d=0$ when $V^d<0$ since we consider a population in which $\max_{d=1,.,K_D}V^d>0$ so that $V_d\geq 0$.

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

$$\begin{aligned}\sum_{i=0}^{n}[\mathbf{1}\{D(\zeta,\varepsilon_i,P) &= d\}\mathbf{1}\{m_1(\beta,u_i)\geq t_1^d\}]\leq 4n_d,\\ \sum_{i=0}^{n}[\mathbf{1}\{D(\zeta,\varepsilon_i,P) &= d\}\mathbf{1}\{m_1(\beta,u_i)\geq t_1^d, m_2(\beta,u_i)\geq t_2^d\}]\leq n_d.\end{aligned}$$

The previous inequalities are equalities when all seats are filled. When they are not, for instance in the second inequality, threshold t_2^d is set to zero (and t_d^1 as well if seats after the first stage are not filled).

As thresholds t_1^d and t_2^d solve this condition for any realization of $\{u_i\}_{i=0,.,n}$ we also have by integration over u:

$$\begin{aligned}\sum_{i=0}^{n}[\mathbf{1}\{D(\zeta,\varepsilon_i,P) &= d\}\Pr\{m_1(\beta,u_i)\geq t_1^d\}]=\sum_{i=0}^{n}[\mathbf{1}\{D(\zeta,\varepsilon_i,P)=d\}P_1^d\leq 4n_d,\\ \sum_{i=0}^{n}[\mathbf{1}\{D(\zeta,\varepsilon_i,P) &= d\}\Pr\{m_1(\beta,u_i)\geq t_1^d, m_2(\beta,u_i)\geq t_2^d\}]=\sum_{i=0}^{n}[\mathbf{1}\{D(\zeta,\varepsilon_i,P)=d\}P^d\leq n_d,\end{aligned}$$

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

$$\begin{aligned}\pi^d(P)P_1^d &\leq \frac{4n_d}{n}\equiv 4\lambda_d,\\ \pi^d(P)P^d &\leq \frac{n_d}{n}=\lambda_d.\end{aligned}$$

in which parameters $\{\lambda_d\}_{d=1,.,K_D}$ is the fraction of seats in the sample attributable to each major. We assume that $\lambda_d>0$ and that $\sum_{d=1}^{K_D}\lambda_d$ is much lower than 1.

By construction, choice probabilities, $\pi^d(P)$, satisfy adding up:

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

and for all P and $\tilde{P}$ such that $P^d\geq\tilde{P}^d$ for all d, we have that $\pi^d(P)\geq\pi^d(\tilde{P})$.

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 are always equalities as described by:

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:

$$\begin{cases} \pi^d(P)P_1^d = 4\lambda_d, \\ \pi^d(P)P^d = \lambda_d. \end{cases}$$

The suffficient condition is true when assuming that there is a sufficient mass in the sample having $V_d > 0$ and for $d' \neq d$, $V_{d'} = 0$.

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 the equilibrium is unique when success probabilities are positive:

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* $(P, \tilde{P})$, *such that for all* d, $P^d > 0$, $\tilde{P}^d > 0$ *and such that* $z(P) = z(\tilde{P})$. *Then* $P = \tilde{P}$.

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

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

$$P^d\pi^d = \tilde{P}^d\tilde{\pi}^d \geq P^d\tilde{\pi}^d$$

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

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

$$d \in I, P^d\pi^d = \tilde{P}^d\tilde{\pi}^d \Longrightarrow \pi^d = \frac{\tilde{P}^d}{P^d}\tilde{\pi}^d < \tilde{\pi}^d,$$

since $P^d > 0$ and $\tilde{\pi}^d > 0$. It implies that:

$$\sum_{d\in I} \pi^d < \sum_{d\in I} \tilde{\pi}^d. \tag{7}$$

Yet, by definition:

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

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 $d \in I^c$, $\tilde{P}^d \geq P^d$, $\max_{d\in I^c}(\tilde{P}^d V^d) \geq \max_{d\in I^c}(P^d V^d)$, we have:

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

a contradiction with inequality (7).

Therefore, $P = \tilde{P}$. ■

The equilibrium is thus unique and values are obtained as a function of the thresholds:

$$P^d = \Pr(m_1 > T_1^d, m_2 > T_2^d). \tag{8}$$

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 with specifying the two stage grade equations and with detailing 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. $m_1 < T_1^d$ and in the absence of any restriction, the distribution of m_2 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, $E(u_2^* \mid u_1, Z) = 0$.

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 $E(u_2^* \mid m_1 \geq T_1^d, Z) = 0$. This would identify parameters and the control function $g(.)$. 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 each other and of variables Z and simply use the estimated empirical distributions of u_1 and u_2 to recover success probabilities. More elaborate ways available in the literature could be used but we stick in this paper with this simple procedure.

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 of the final admission in each major as described by two inequalities:

$$m_1 + 120 * ENEM/63 \geq \tau_1^d,$$
$$0.4 * (m_1 + 120 * ENEM/63) + 0.6 * m_2 \geq \tau_2^d.$$

Thresholds (τ_1^d, τ_2^d) are derived from those linear combinations of initial grades and first and second stage grades fixed by the University. The individual specific thresholds T_1^d and T_2^d used in

the theoretical section above are derived from those expressions. We pospone the discussion on how we took into account that thresholds are measured with error in the sample and argue here conditional on values, τ_1^d and τ_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:

$$\begin{aligned} u_1 &\geq \tau_1^d - 120 * ENEM/63 - Z\beta_1, \\ u_2^* &\geq \frac{\tau_2^d}{0.6} - \frac{2}{3}(Z\beta_1 + u_1 + 120 * ENEM/63) - Z\beta_2 - g(u_1). \end{aligned}$$

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:

$$\begin{aligned} p^d(Z, \beta, t_1^d, t_2^d) &= Pr\{u_1 \geq m_1^d - Z\beta_1, u_2^* \geq m_2^d - \frac{2}{3}Z\beta_1 - Z\beta_2 - \frac{2}{3}u_1 - g(u_1)\}, \\ &= \int_{m_1^d - Z\beta_1}^{\infty} f_{u_1}(x)(Pr\{u_2^* \geq m_2^d - \frac{2}{3}Z\beta_1 - Z\beta_2 - \frac{2}{3}u_1 - g(u_1)\})dx, \\ &= \int_{m_1^d - Z\beta_1}^{\infty} f_{u_1}(x)[1 - F_{u_2^*}(m_2^d - \frac{2}{3}Z\beta_1 - Z\beta_2 - \frac{2}{3}x - g(x))]dx, \end{aligned} \tag{11}$$

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

$$\begin{cases} m_1^d = & \tau_1^d - 120 * ENEM/63, \\ m_2^d = & \frac{\tau_2^d}{0.6} - \frac{2}{3}(120 * ENEM/63). \end{cases}$$

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 necessarily 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:

$$\begin{cases} D = S \quad \text{if } \log(P^S) + \log(V^S) \geq \log(P^F) + \log(V^F), \\ D = F \quad \text{if } \log(P^S) + \log(V^S) < \log(P^F) + \log(V^F) \end{cases} \tag{12}$$

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 result that both coefficients are 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 ε is an unobserved idiosyncratic preference term. We assume that the distribution of ε 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:

$$\begin{aligned} \Pr(D = S \mid P^S, P^F, X) &= \Pr\{V^S > 0,\, V^F \leq 0 \mid X\} + \\ &\quad \Pr\{V^S > 0,\, V^F > 0 \mid X\}.\Pr\{\log(P^S) + \log(V^S) \geq \log(P^F) + \log(V^F)\} \\ &= \delta^S(X) + \delta^{SF}(X) F(\log(P^S) - \log(P^F) + X\gamma \mid X). \end{aligned}$$

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 (due to the underlying variation in Z)

of the covariate:

$$\Delta(Z) \stackrel{def}{=} \log(P^S) - \log(P^F),$$

from preference shifters, X. For various reasons that will appear more clearly in the following, Δ acts as a price excluded by assumption from utility. As developed in the previous section, Δ is unobserved by the econometrician, yet is a function of observed covariates Z. 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 Xs. 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 the structural objects

$$\{\delta^S(X), \delta^{SF}(X), \gamma, F(. \mid X)\}$$

are deduced from the reduced form choice probabilities $\Pr(D = S \mid \Delta(Z), X)$ using:

$$\Pr(D = S \mid \Delta(Z), X) = \delta^S(X) + \delta^{SF}(X) F(\Delta(Z) + X\gamma \mid X). \tag{13}$$

We first show how to identify functions δs then turn to parameter γ and the distribution function $F(. \mid X)$. Specifically, those who attribute a negative value to one of the schools always choose the other school, no matter how success probabilities change. On the other hand, those whose utilities are both positive are sensitive to the variation in success probabilities. By making success 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 indeed identify δ^S using:

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

A similar approach can be applied to δ^{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:

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

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

$$F(0 \mid X) = \frac{1}{2}. \tag{14}$$

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 Δ whatever X is includes the value 0. Then as developed in Manski (1988), partial identification occurs under the median restriction (14) written above. Parameter γ is identified using the median restriction and $F(\mid X)$ 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 $F(\mid X)$. What non parametric identification arguments above have proven is that this parametric assumption is a testable assumption at least in the support in which $\Delta(Z) + X\gamma$ varies.

3.3 Empirical strategy

We first estimate the parameters of the grade equations and denote them $\hat{\beta}_n$. This in turn allows us to compute the expectation of the success probabilities conditional on thresholds $\tau_j^d, j = 1, 2, d = S, F$ as in equation (11) using the estimated distribution functions for errors in the grade equations.

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 the equilibrium conditions (6) by simulation of $U_{(n)}$.[11] For any simulation $c = 1, ., C$, 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 $\mathbf{1}\{D_i(Z_i, \varepsilon_i, \zeta, P_i^S, P_i^F) = S\}$, characteristics X_i and solving the equilibrium conditions (6). Equation (2) can then be computed by integration as:

$$\hat{P}_{0,C}^d = \frac{1}{C}\sum_{c=1}^{C} p^d(Z_0, \hat{\beta}_n, \tilde{t}_{1,c}^d, \tilde{t}_{2,c}^d). \tag{15}$$

We can then estimate the preference parameters $\zeta = (\delta, \gamma)$ using a conditional maximum

[11] By construction, $\tilde{T}_0$ depends on observation 0 although this dependence should matter less and less when n is large. For simplicity, we compute those thresholds in the empirical application using equation (5) instead of equation (6).

likelihood approach:

$$\hat{\zeta}_n = arg\max_{\zeta} \quad l(\zeta|\hat{P}^S_{0,C}, \hat{P}^F_{0,C}).$$

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

$$\hat{\zeta}_n \xrightarrow[n\to\infty]{P} \zeta.$$

We used bootstrap to obtain the covariance matrix of those estimates by replicating the complete estimation procedure as a mixture of non parametric (grade equations) and parametric bootstrap (choice equations).

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 the various exams – the initial national high school evaluation exam (ENEM), the first and second stage of the Vestibular system as well as the number of repetitions of the entry exams.
- Basic demographic variables – gender, age by discrete values (16, 17.5, 21 and 25) and the education levels of father and mother.
- Education history – public or private primary or high school as described by discrete values indicating the fraction of time spent in private schools and undertaking of a preparatory course
- Choices of majors

In total there are 58 majors that students may consider at Universidade Federal do Ceará. 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.

Medicine is one of the most difficult major to enter as can be seen in Table 2 which reports summary statistics in each major field and the grades obtained at the first stage of the college exam.[12] We report statistics on the distribution of the first stage grades in three samples:[13] 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. These statistics are very informative. Distributions remain similar across groups. Minima (column1) tend to be ordered as the median of students who pass (column 6). The first columns also reveal that some groupings might be 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. Medicine and Law are ranked the highest and the difference with other major fields is large. 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. 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 Sample selection

For computational simplicity, the empirical analysis uses a sub-sample of applicants to this University. As the allocation mechanism is decentralized, we can simply restrict the sample without

[12] We do not report the second stage grades as they consist in grades in specific fields that are not necessarily comparable across majors.

[13] 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.

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 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 in the state capital, Fortaleza, is much larger since it offers 150 seats. As shown in the empirical analysis below, this assymmetry 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 209 while it remains around 200 for the two others. 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 is equal to 2867 of which 542 (resp. 2315) apply to Sobral (resp. 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 in the pool of Fortaleza two admitted students only 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.

4.1.2 Performance and preference shifters

Explanatory variables are those which affect exam performance or school preferences. For grade equations, all potential explanatory variables are included: a proxy for ability which is the initial grade obtained at the national exam,[14] age, gender, educational history, repetitions, parents'

[14]When missing (in 5% of cases), we imputed for ability the predicted value of the initial grade $ENEM$ obtained by using all exogenous variables and we denote the result as m_0 to distinguish it from $ENEM$ which is used when computing the passing grades. The administrative rule is to impute 0 when $ENEM$ is missing.

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 the forward looking value 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. private/public sector history, undertaking a preparatory course). This is even more likely since ability is measured after what we call educational history. This exclusion restriction allows us to distinguish performance shifters, Z, from preferences shifters and to identify preferences using results derived in Section 3.2.2. As a consequence, preferences are specified as a function of 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 individual characteristics and performances in each pool of applicants to the two medical schools. Looking at the admission rates, one can see that Sobral admitted $40/527 = 7.6\%$ and Fortaleza $150/2340 = 6.4\%$ and this makes Fortaleza more competitive. Comparing the mean and median of 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 functions, both in Fortaleza and Sobral, are unimodal 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 regarding 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 Estimates of grade equations

4.2.1 First stage exam

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 use splines in this variable although other non-parametric methods such as Robinson (1988) could be used. 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.

Estimates show that 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. Among other 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. Taking a preparatory course and repeating the entry exam have positive and significant effects on grades by presumably increasing capacities and experience of applicants. In the second specification, we tested for the joint exclusion of parents' educations and it 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 yields a large R^2 at around 0.72, and this does not vary much across different specifications. This promises good prediction power of the model and makes the simulation of success probabilities more credible.

4.2.2 Second stage exam

In the second stage grade equation, we again sought for flexibility with respect to two variables – the initial stage grade m_0 and the residual from the first stage grade equation $\hat{u}_1$ as it controls for dependence between stages. 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 enough in terms of predictive power. Results are reported in Table 6. First of all, there exists a strong positive correlation between u_1 and u_2, 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 the first stage 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. Yet, our attempts in previous work to construct a more sophisticated model including endogenous effort did not lead to the confirmation of this model and this is why we decided to use the current simpler model. As for 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 differences are not significant. The second stage exam has a different format (writing essays) than the first stage multiple choice exam and the format could explain gender differences. The lower R^2 at the second stage might also be a consequence of the exam format.

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 although the global F-test does not reject homoskedasticity at a 5% level (p-value of 3.4%). For the second grade equation, the test rejects homoskedasticity at the 5% level 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 heteroskedasticity remains of a limited magnitude. We checked that heteroskedasticity does not generate large differences in the prediction of success probabilities.

4.2.3 Success probabilities

Success probabilities are simulated using the empirical distributions of $\hat{u}_1$ and $\hat{u}_2$ and of the thresholds. We run $n_S = 2000$ sets of n simulations by drawing into the estimated empirical distribution of errors, $\hat{u}_1$ and $\hat{u}_2$. We then compute thresholds by solving equation (5) for each of the previous n_S set of simulators. We then replace the integration with respect to the thresholds as in equation (15) and the integration in equation (11) by summing over the set of n_S simulators. 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 probabilities are close to what is observed and as expected roughly 4 times lower than the first-stage 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 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. The result is shown in Table 8. Some critical quantiles at the top are provided for more detail. The two most important range of percentiles are indeed the 70/75th and 93/95th percentiles since the admission rate at the first exam is slightly less than 30% 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 affected more 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 those

odds ratios at all percentiles.

4.3 Estimates of school preferences

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 assumed to be 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 (that is $V^S > 0, V^F > 0$) as:

$$\delta^{SF}(X) = \frac{1}{1 + \exp(X\delta^{SF})}.$$

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

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

in which $\Phi(.)$ is the zero mean unit normal distribution[15] and the success probabilities P^d are to be replaced by their simulated predictions using grade equations (column 1 of Table 5 and column 2 of Table 6) as developed in the previous Section 4.2.3. In the first part of Table 9, we report the estimated preference coefficients and in the second part we present more readable summary statistics of the estimated probabilities of each region, $\delta^{SF}(X)$. There are three different specifications included in this table. The key difference is how explanatory variables enter the specification of δ^S and δ^{SF}. We chose to use two main variables, ability m_0 and Living in Fortaleza as the main drivers of these probabilities and the three columns of Table 9 include one or both of these variables.

The results are very stable across specifications. As far as δ parameters are concerned, ability significantly affects the probability of the region of jointly positive values, (S, F) (and as a consequence of adding up, also the preference for F alone). Living in Fortaleza decreases preferences for Sobral alone (δ^S) or jointly with Fortaleza (δ^{SF}). The second part of Table 9 shows that the average probability of preferring Sobral alone (resp. Fortaleza alone) to the outside option is

[15] As the range of the log probability difference is not the whole real line as in Section 3.2.2, the scale of the error is not identified and its variance is thus normalized to one.

around 0.06 (respetively 0.55). These frequencies stay almost invariant across specifications. This shows that students heavily favor Fortaleza over Sobral and this confirms that Fortaleza is the most popular medicine school in the state of Cearà. The ratio of those probabilities is 10 which is approximately the ratio between the populations of the two cities albeit much larger than the ratio of final seats in the two schools (150/40). Nonetheless, there is a substantial fraction of students whose utilities for both schools are positive (more than 40%)

We now turn to parameters γ that affect preferences of students who prefer both schools to the 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 colinearity between parents' educations.

Finally, we tested the maintained hypothesis that performance shocks and preference shocks are independent by introducing the residual $\hat{u}_1$ in this preference equation. The hypothesis cannot be rejected at the 10% level (the p-value is equal to 0.184).

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 instead of four times, the number of final seats. 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 increases although the difference between the two majors is attenuated.

Furthermore, there are two stages in the exam because this allows to cut costs and achieve a

more in-depth selection at the second-stage. 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 learning their grades. This would generate more opportunistic behavior.

Before entering the details of these new mechanisms, we first analyze the identification of utilities from estimated preferences and success probabilities 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:

$$\begin{aligned} U_i &= \mathbf{1}\{V_i^S \geq 0, V_i^F < 0\}\mathbf{1}\{\text{Success in } S\}V_i^S \\ &+ \mathbf{1}\{V_i^F \geq 0, V_i^S < 0\}\mathbf{1}\{\text{Success in } F\}V_i^F \\ &+ \mathbf{1}\{V_i^F \geq 0, V_i^S \geq 0\}\left[\mathbf{1}\{D_i = S\}\mathbf{1}\{\text{Success in } S\}V_i^S + \mathbf{1}\{D_i = F\}\mathbf{1}\{\text{Success in } F\}V_i^F\right] \end{aligned}$$

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

$$\begin{aligned} E\left(U_i \mid V_i^S, V_i^F\right) &= \mathbf{1}\{V_i^S \geq 0, V_i^F < 0\}P_i^S V_i^S + \mathbf{1}\{V_i^F \geq 0, V_i^S < 0\}P_i^F V_i^F \\ &+ \mathbf{1}\{V_i^F \geq 0, V_i^S \geq 0\}\left[\mathbf{1}\{D_i = S\}P_i^S V_i^S + \mathbf{1}\{D_i = F\}P_i^F V_i^F\right] \\ &= P_i^S V_i^S\left(\mathbf{1}\{V_i^S \geq 0, V_i^F < 0\} + \mathbf{1}\{V_i^F \geq 0, V_i^S \geq 0\}\mathbf{1}\{D_i = S\}\right) \\ &\quad + P_i^F V_i^F\left(\mathbf{1}\{V_i^F \geq 0, V_i^S < 0\} + \mathbf{1}\{V_i^F \geq 0, V_i^S \geq 0\}\mathbf{1}\{D_i = F\}\right). \end{aligned}$$

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 $\left|V_i^F\right|$ as the scale factor to set:

$$\begin{aligned} V_i^F &= 1 \text{ if } V_i^F > 0, \\ V_i^F &= -1 \text{ if } V_i^F < 0. \end{aligned}$$

Under this normalization:

$$\begin{aligned} E\left(U_i \mid V_i^S, V_i^F\right) &= P_i^S \left(V_i^S \mathbf{1}\{V_i^S \geq 0, V_i^F < 0\} + \frac{V_i^S}{V_i^F} V_i^F \mathbf{1}\{V_i^F \geq 0, V_i^S \geq 0\} \mathbf{1}\{D_i = S\} \right) \\ &\quad + P_i^F V_i^F \left(\mathbf{1}\{V_i^F \geq 0, V_i^S < 0\} + \mathbf{1}\{V_i^F \geq 0, V_i^S \geq 0\} \mathbf{1}\{D_i = F\} \right), \\ &= P_i^S \left(V_i^S \mathbf{1}\{V_i^S \geq 0, V_i^F < 0\} + \frac{V_i^S}{V_i^F} \mathbf{1}\{V_i^F \geq 0, V_i^S \geq 0\} \mathbf{1}\{D_i = S\} \right) \\ &\quad + P_i^F \left(\mathbf{1}\{V_i^F \geq 0, V_i^S < 0\} + \mathbf{1}\{V_i^F \geq 0, V_i^S \geq 0\} \mathbf{1}\{D_i = F\} \right), \end{aligned}$$

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 3.2.2).

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

$$\begin{aligned} E\left(V_i^S \mid V_i^S \geq 0, V_i^F < 0\right) &< E\left(V_i^S \mid V_i^S \geq 0, V_i^F \geq 0\right) = E\left(\frac{V_i^S}{V_i^F} \mid V_i^S \geq 0, V_i^F \geq 0\right) \\ &< \exp(X_i\gamma) E(\exp(\varepsilon_i) \mid V_i^S \geq 0, V_i^F \geq 0) \\ &< \exp(X_i\gamma + .5), \end{aligned}$$

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

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

where $\mu_0 > 0$ captures the positive dependence between V_i^S and V_i^F. This is coherent with the previous equation since :

$$\begin{cases} V_i^S = \exp(X_i\gamma + \varepsilon_i) & \text{if } V_i^F = 1, \\ V_i^S = \exp(X_i\gamma + \varepsilon_i - \mu_0) & \text{if } V_i^F = -1. \end{cases}$$

We will thus evaluate $E\left(U_i \mid V_i^S, V_i^F\right)$ using bounds on $\mu = \exp(-\mu_0)$ that we make vary between 0 (the lower bound for V_i^S) and 1 (the case in which V^S and V^F are uncorrelated).

5.2 Computing equilibria

In every counterfactual experiment, we use the simulation procedure whereby we draw 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 $\bar{P}_i^S$ 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 the pool of potential students and therefore we consider the same sample $i = 0, ., n$. In our experiments, alternative mechanisms act only on success probabilities and not on preferences. It assumes however that these experiments do not modify the predetermined behavior of the students like taking a prep course or the ex-post equilibrium in college and in the labor market.

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

$$\begin{cases} \sum_{i=1}^{n}[\mathbf{1}\{\bar{D}_i(\bar{P}_i^S, \bar{P}_i^F) = S\}\mathbf{1}\{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^S\}] = \bar{n}_S, \\ \sum_{i=1}^{n}[\mathbf{1}\{\bar{D}_i(\bar{P}_i^S, \bar{P}_i^F) = F\}\mathbf{1}\{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^F\}] = \bar{n}_F, \\ \sum_{i=1}^{n}[\mathbf{1}\{\bar{D}_i(\bar{P}_i^S, \bar{P}_i^F) = S\}\mathbf{1}\{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^S, m_2(X_i, \beta, u_i) \geq \tilde{t}_2^S\}] = n_S, \\ \sum_{i=1}^{n}[\mathbf{1}\{\bar{D}_i(\bar{P}_i^S, \bar{P}_i^F) = F\}\mathbf{1}\{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^F, m_2(X_i, \beta, u_i) \geq \tilde{t}_2^F\}] = n_F, \end{cases} \tag{16}$$

have a distribution function that leads to the counterparts of equation (15):[16]

$$\bar{P}_0^d = \mathbb{E}(\mathbf{1}\{m_1(X_0, \beta, u_0) \geq \tilde{t}_1^d, m_2(X_0, \beta, u_0) \geq \tilde{t}_2^d\}) \tag{17}$$

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 vectors $\varepsilon_{(n),c}$ in their distributions conditional to 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 vectors $U_{(n),c}$ and fix them for the rest of the procedure.

[16] Changing the timing of choices requires to acknowledge that there are no choices to make before the first-stage. The first two equations in (16) do not depend on $\bar{D}_i$ and P_i^S, P_i^F are the conditional expectations after the second-stage. Those adaptations do not modify the main principles.

- Set the initial $P_0^{S,0}, P_0^{F,0}$ values at their simulated values $\hat{P}_{0,C}^{d}$ computed from equation (15) replacing ζ by $\hat{\zeta}_n$ and using $U_{(n),c}$ in the observed experiment using equation (6). This implicitly means that choices D_i in the current mechanism are set to their observed values.

2. At step k, denote $P_i^{S,k}, P_i^{F,k}$ the expected success probabilities

   (a) Compute counterfactual choices $D_i(Z_i, \varepsilon_{i,c}, \hat{\zeta}_n, P_i^{S,k}, P_i^{F,k})$.

   (b) Compute a sequence of $\tilde{t}_c$ for $c = 1, ., C$ replacing ζ 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 view of the organization cost of exams, it is tempting 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. In the experiment, the number of final positions is kept unchanged but half as many students are allowed to take the second exam. In other words the admission rate after the first stage exam is divided 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 seats in the second exam reduces schools' administrative costs although this 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 Section 5.1 and we compute expected utility as in Section 5.2.

5.3.1 Changes in thresholds

In Table 10 we present estimates of the new threshold distributions at both stage exams in the three counterfactual experiments and in particular in the cutting seat experiment. Standard errors are computed by using the bootstrap simulations that were generated to compute standard errors of grade and preference parameter estimates and thus 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 in the original system because there is now less competition in the second stage exam when half as many 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 how this counterfactual brings benefit to schools and students, we study in turn changes in success probabilities and changes in students' utilities.

5.3.2 Changes in success probabilities

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. This is why we evaluate changes in success probabilities in relation to an index of students' abilities and we use the expected final grade (a combination of the initial, first and second stage grades) 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 and Figure 6 for Fortaleza. In those Figures three vertical lines are drawn at the median of expected final grade and at the quantiles associated to the first and second-stage thresholds on average *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 a much larger variance than first-stage grades. The chance is lower when fewer students participate in 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, they will be admitted to the second stage exam with less competition and this entails a higher success probability. The chance that they perform not that well in the first exam is however 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 in the second stage exam due to the system change.

5.3.3 Changes in students' utilities and the impact on schools

Table 11 presents summaries of changes in students' expected utility in which students are ranked in percentile groups according to their expected final grade. As defined in Section 5.1, we set the unknown weight in utilities at $\mu = 0.8$. Consistently with changes in success probabilities, only the very top students – above the 94% quantile of ability – have significant utility improvements. Nonetheless, students above the 88% quantile also 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 school 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 individual changes in expected utility in Figure 7.

We also performed a robustness analysis by using different values for the weight μ (see Section

5.1). Results are shown in Table 12. When μ is at the lower bound – 0 – utility changes are slightly smaller. When μ is at the upper bound – 1 – utility changes become slightly larger. Overall, differences are very limited and our previous results are quantatively robust to the value of μ.

The impact of cutting seats on schools seems to be positive 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. This benefit comes in addition to 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.

5.4 Enlarging the choice set

In this experiment, students can submit an enlarged list of two majors if they wish. 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 comprises students who positively value at least one of the majors so that $d_1 \neq \varnothing$) and $d_2 \in \{\varnothing, S, F\}$. We thus still allow students to provide a single choice if $d_2 = \varnothing$. This mechanism belongs in the deferred-acceptance family with the additional twist that we keep the sequence of two exams as it is. The allocation of students after the first exam needs however to be adapted and this is the design that we now explain.

5.4.1 Design of the experiment

To fix ideas, consider first a student who (1) has $V_S > 0$ and $V_F > 0$ (2) chooses (S, F). If after the first-exam, she is above the threshold for school S, her second choice does not matter. It is only if she is NOT accepted to the second stage exam in school S that she could compete for the second stage exam in school F. She fails when her grades are lower than both thresholds.[17]

Consider first that at equilibrium $t_1^S > t_1^F$. After the first stage exam, there are three possible

[17] There are alternative experiments that could be explored as well such as the one in which students are allocated to majors after the second-stage exam.

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 and takes the outside option.

While if $t_1^S > t_1^F$ (the probability of a tie being equal to zero),

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

This sequence is easily adapted to students choosing the list (F, S). Moreover, for students submitting a list $(d_1, \varnothing)$, the sequence of actions is the same as in the original mechanism. Students are selected into the second-stage exam for school d_1 if their grade is above d_1 first stage threshold.

Furthermore, given any choice among the four lists, $\{(S, F), (F, S), (S, \varnothing), (F, \varnothing)\}$ we can construct counterfactual success probabilities in each major P^S and P^F by adapting the algorithm we used before. For any value of success probabilities, we can then compute the optimal choice between $\{(S, F), (F, S), (S, \varnothing), (F, \varnothing)\}$. Details about how we get counterfactual thresholds and choices follow the lines of what was developed in Section 5.2.

5.4.2 Changes in thresholds

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. A few 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 for strategic reasons in the original mechanism can now at no risk choose Sobral first and Fortaleza second. Deferred acceptance mechanisms lessen strategic motives and make choices more truthful (Abdulkadiroglu and Sonmez, 2003). 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 by preference estimates in Table 9.

Thresholds at the second stage exam are slightly less than the original ones although large standard errors point out that those differences are unlikely to be significant (since both threshold estimates are correlated positively). Moreover, even if 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.

5.4.3 Changes in success probabilities

Figure 8 for Sobral and Figure 9 for Fortaleza report changes in success probabilities. Unlike the previous counterfactual experiment, the changes in Sobral and Fortaleza are now quite different. In Fortaleza, almost all students whose ability is above the first stage admission quantile have now higher success probabilities and the ones who benefit the most are the very top students. This might be due to the fact that Fortaleza is not any longer a safety school for some top students. In contrast, a larger portion of students below the first admission threshold and above median have a lower success probability in Sobral in the counterfactual experiment. This is because good students who fail Fortaleza switch to Sobral to compete with them and first-stage thresholds are now higher in Sobral and medium ranked studenst are evicted. 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.

5.4.4 Changes in expected utilities and the impact on schools

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 because of the same reasons as for success probabilities. Since Sobral has a lower threshold at the first stage exam, students who prefer Sobral are now bearing more competition from top evicted students from Fortaleza rankings and those who choose Sobral and fail have no second chance. Therefore expected utility increases for those who opt for Sobral are

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. The majority of students are better off except the medium ranked students who prefer the smallest school. From the perspective of the schools, Sobral should be more favourable to this mechanism since it can now attract higher ranked students. Fortaleza's thresholds remain the same although the composition of their recruitment might have changed since it lost its safety school status. This seems however to moderately affect top students.

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 the last counterfactual experiment, we try to evaluate the impact on the allocation and expected utility of students when they choose majors **after** learning their first stage exam grade and not any longer before. 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 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_1^d. 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 $P^S_{m_1} = Pr\{m_2 > t_2^S | m_1\}$ and $P^F_{m_1} = Pr\{m_2 > t_2^F | m_1\}$ which are now conditional to the first-stage grade m_1. To define choices, suppose that $t_1^S > t_1^F$. A student can face three

cases:

- $m_1 > t_1^S$: the choice set is complete and consists in $\{S, F\}$. Majors are chosen by comparing $P_{m_1}^S V^S$ and $P_{m_1}^F V^F$ (since either $V^S > 0$ or $V^F > 0$).
- $m_1 < t_1^S$ and $m_1 \geq t_1^F$: 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.
- $m_1 < t_1^F$: the only choice left is the outside option.

This algorithm is easily adapted to the case in which $t_1^S < t_1^F$ prevails.

5.5.1 Changes in thresholds

Thresholds in 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 Table 9) but this does not compensate the difference in offered seats. By making students choose in the order of first stage grades, positions in Sobral at the second-stage exam 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 smaller school (Sobral) relatively to revealed preferences.

5.5.2 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, students have lower success probabilities and the impact is the largest for top students. In contrast, the success probability in Fortaleza 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. The school in Fortaleza on average loses many elite students since it recruits at a lower level in terms of thresholds. This seems to be the most assymmetric experiment between the schools, Sobral gaining a lot.

5.5.3 Changes in expected utilities and the impact on schools

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. Indeed, the probability of an increase in expected utility is equal to 1. This mechanism is mainly attractive for the top students as shown in Table 14 where students above the 70% quantile are gaining significantly more than students below this quantile. In a nutshell, top students in the first stage are better protected from the competition of lower ranked students.

There are clear differences in utility changes among the top students conditional on their preferences for the schools. On average, students who prefer Fortaleza would benefit more than those who prefer Sobral. This is then consequence of the fact that Sobral seats fill much more quickly than Fortaleza's. Given the differential success probabilities across schools, Fortaleza is now easier to get than Sobral. This is confirmed by Figure 13 in which individual utility changes are plotted.

Overall, this counterfactual seems more friendly to top students and to the small school. Sobral whould rather have this mechanism because it would be able to enroll much 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 to college majors in medicine to provide an evaluation of the mechanism in place. We first estimate a model of major choices as well as performance to derive the parameters governing success probabilities and preferences. Expectations of sophisticated students are obtained by sampling into the Bayesian Nash equilibrium conditions. Using those estimates, we can compute in a second step the impact of three counterfactual experiments on success probabilities and expected utility of the students. This shows at what benefits and costs the current mechanism could be changed, not only in terms of aggregate utilitarian welfare but also in terms of potentially strong redistributive effects between schools and between students.

These cost and benefit analyses show that the choice of an allocation mechanism has sizeable

consequences for both schools and students. The mechanism in place is neither fair nor strategic although it might be rationalized by the fact that some majors and/or groups of students would lose if it were changed. The political economy of such a choice of an allocation mechanism remains to be documented and analyzed and it would interesting to develop the analysis of the ex-ante game between schools and/or students that leads to the adoption of such or such mechanism. As a matter of fact, federal universities in Brazil have adopted since 2010, under the pressure of the Federal government, a national allocation mechanism and some of us are in the process of collecting data to evaluate the new system.

On the modeling side, much remains to be done. Specifically, the modelling assumptions about expectations are strong and weakening them is high on the agenda. Identification however is bound to be weak since there is nothing in the data that might indicate whether agents are sophisticated, well or badly informed or even naïve (Pathak and Sonmez, 2013, He, 2012). The analysis shall thus proceed as an analysis of robustness that could lead to partial identification of the costs and benefits we have been describing above. It is also true that the question of why so many students are taking this exam although they have no chances to succeed remains pending. They could be overly optimistic and this relates to assumptions about expectations but they could also use the exam as a training device for the following year or for other exams of a similar type. However, this type of behaviour seems to be easier to accomodate in the current framework.

References

Abdulkadiroğlu, A., Agarwal, N., & Pathak, P. A., 2014, "The Welfare Effects of Congestion in Uncoordinated Assignment: Evidence from the NYC HS Match", working paper.

Abdulkadiroğlu, A., Y., K., Che and Y. Yasuda, 2012, "Expanding Choice in School Choice", Working paper.

Abdulkadiroğlu, A., P.A. Pathak and A. Roth, 2009, "Strategy-proofness versus Efficiency in Matching with Indifferences; Redesigning the NYC High School Match", *American Economic Review*, Vol. 99, No. 5, pp. 1954-1978.

Abdulkadiroğlu, A., Pathak, P., Roth, A. E., & Sonmez, T. (2006), "Changing the Boston school choice mechanism", WP 11965, National Bureau of Economic Research.

Abdulkadiroğlu, A. and T., Sonmez, 2003, "School Choice: A Mechanism Design Approach", *American Economic Review*, Vol. 93, No. 3, pp. 729-747

Agarwal, N., 2013, "An Empirical Model of the Medical Match," Unpublished working paper, MIT.

Arcidiacono, P., 2005, "Affirmative Action in Higher Education: How Do Admission and Financial Aid Rules Affect Future Earnings?", *Econometrica*, Vol. 73, No. 5, pp. 1477-1524.

Balinski M., and T., Sönmez, 1999, "A Tale of Two Mechanisms: Student Placement", *Journal of Economic Theory* 84, 73-94.

Bourdabat B. and Montmarquette C., 2007, "Choice of Fields of Study of Canadian University Graduates: The Role of Gender and their Parents' Education", IZA Discussion Paper No.2552.

Budish, E. and E. Cantillon, 2012, "The Multi-unit Assignment Problem: Theory and Evidence from Course Allocation at Harvard", *American Economic Review*, 102(5):2237-2271

Calsamiglia, C., Haeringer, G., & Klijn, F., 2010, "Constrained school choice: An experimental study", *The American Economic Review*, 1860-1874.

Epple, D., R. Romano and H. Sieg, 2006, "Admission, Tuition, and Financial Aid Policies in the Market for Higher Education", *Econometrica*, Vol. 74, No. 4, pp. 885-928

Hastings, J., T. J. Kane, and D. O. Staiger, 2009, "Heterogenous Preferences and the Efficacy of Public School Choice," Working paper, Yale University.

He, Y., 2012, "Gaming the School Choice Mechanism in Beijing", unpublished manuscript, Toulouse School of Economics.

Lai, F., E., Sadoulet and A., de Janvry, 2009, "The Adverse Effect of Parents' School Selection Errors on Academic Achievement: Evidence from the Beijing Open Enrollment Program", *Economics of Education Review*, 28:485-496.

Instituto Nacional de Estudos e Pesquisas (INEP), 2008, "Sinopses estatísticas da educação superior", available at http://www.inep.gov.br/superior/censosuperior/sinopse/.

Olive, A. C., 2002, "Histórico da educação superior no Brasil", in: Soares, M. S. A. (coord.). *Educação superior no Brasil.* Brasília, p. 31-42.

Manski, C. F., 1988, "Identification of binary response models", *Journal of the American Statistical Association*, 83(403), 729-738.

Manski C., 1993, "Adolescent Econometricians: How Do Youths Infer the Returns to Schooling?"in *Studies of Supply and Demand in Higher Education*, edited by Charles T.Clotfelter and Michael Rothschild. Chicago: University of Chicago Press.

Matzkin, R. L., 1993, "Nonparametric identification and estimation of polychotomous choice models", *Journal of Econometrics*, 58(1), 137-168.

Pathak P.A., and T., Sonmez, 2013, "Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism," *American Economic Review*, 98(4), 1636–1652.

Robinson, P. M., 1988, "Root-N-consistent semiparametric regression", *Econometrica,* 56:931-954.

Roth, A.E, 2008, "Deferred acceptance algorithms: history, theory, practice, and open questions", *International Journal of Game Theory*, 36:537–569

Roth, A. E., & Sotomayor, M. A. O., 1992, *Two-sided matching: A study in game-theoretic modeling and analysis* (No. 18). Cambridge University Press.

A Data appendix

A.1 Description

The Vestibular, an entrance exam whereby different universities develop their own format of testing students restricted by some federal constraints, has its roots in 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 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 its 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.

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) and they are set according to the requirements of each major. 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.

A more specific 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 success probabilities $P_i^S(Z_i,\beta)$ and $P_i^F(Z_i,\beta)$, the decision is determined by:

$$\begin{aligned} D_0(X_i,\varepsilon_i,\zeta,P_i^S,P_i^F) &= S \Longleftrightarrow \ln(V^S(X_i,\varepsilon_i,\zeta)/V^F(X_i,\varepsilon_i,\zeta)) + \ln(P_i^S/P_i^F) \geq 0, \\ D_0(X_i,\varepsilon_i,\zeta,P_i^S,P_i^F) &= F \Longleftrightarrow \ln(V^S(X_i,\varepsilon_i,\zeta)/V^F(X_i,\varepsilon_i,\zeta)) + \ln(P_i^S/P_i^F) < 0. \end{aligned}$$

B.1.2 Simulations of $\varepsilon_{(i)}$ conditional on choices

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

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

in which $\tilde{\varepsilon}_{i,c}^{(1)} \sim U[0,1]$. Then, if $\varepsilon_{i,c}^{(1)} < \delta^S(X_i)$ the observed choice is necessarily $D_i = S$. 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:

$$X_i\gamma + \varepsilon_i^{(2)} + \ln(P_i^S/P_i^F) > 0$$

as derived from equation (12). This is easily done by drawing in a truncated normal distribution. Draw $\tilde{\varepsilon}_{i,c}^{(2)}$ into a $U[0,1]$ and write:

$$\varepsilon_{i,c}^{(2)} = \Phi^{-1}(\Phi(-\ln(P_i^S/P_i^F) - X_i\gamma) + (1 - \Phi(-\ln(P_i^S/P_i^F) - X_i\gamma))\tilde{\varepsilon}_{i,c}^{(2)}),$$

or equivalently:

$$\varepsilon_{i,c}^{(2)} = -\Phi^{-1}(\Phi(\ln(P_i^S/P_i^F) + X_i\gamma)(1 - \tilde{\varepsilon}_{i,c}^{(2)})).$$

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

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

$$\varepsilon_{i,c}^{(2)} = \Phi^{-1}(\Phi(-\ln(P_i^S/P_i^F) - X_i\gamma)(1 - \tilde{\varepsilon}_{i,c}^{(2)})), \tilde{\varepsilon}_{i,c}^{(2)} \sim U[0,1].$$

B.2 The counterfactual experiment 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), $(S,\varnothing)$, $(F,\varnothing)$ 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 $(d, \varnothing)$, we have that:

$$U^d = V^d \Pr\{m_1 > t_1^d, m_2 > t_2^d\}$$

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

$$U^{d_1 d_2} = V^{d_1} \Pr\{m_1 > t_1^{d_1}, m_2 > t_2^{d_1}\} + V^{d_2} \Pr\{m_1 \in [t_1^{d_1}, t_1^{d_2}), m_2 > t_2^{d_2}\}$$

in which $\Pr\{m_1 \in [t_1^{d_1}, t_1^{d_2}) = 0$ if $t_1^{d_2} < t_1^{d_1}$. The choice model can now be described by four success probabilities:

$$\begin{cases} P^d = \Pr\{m_1 > t_1^d, m_2 > t_2^d\}, d = S, F \\ P^{d_1 d_2} = \Pr\{m_1 \in [t_1^{d_1}, t_1^{d_2}), m_2 > t_2^{d_2}\}, (d_1, d_2) \in \{(S, F), (F, S)\}, \end{cases}$$

which are functions of thresholds t_1^d, t_2^d. Those thresholds remain sufficient statistics in order to derive these success probabilities.

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

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

Source: Vestibular cross section data in 2004.

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)

Subgroup	10th percentile	Min	Min	Median			Maximum		
	All	Firststage	Pass	All	First stage	Pass	All	First stage	Pass
Agrosciences	71,1	91,2	100,1	106,9	128,1	141,6	192,6	192,6	192,6
Other	66,1	102,1	104,8	102,0	136,7	143,3	187,5	187,5	187,5
Physics & Chemistry	76,8	33,0	50,0	115,2	128,9	144,6	210,2	210,2	210,2
Humanities	67,9	96,3	99,2	104,2	133,6	147,1	203,3	203,3	203,3
Social Sciences	68,9	101,0	102,0	109,4	138,6	147,9	214,3	214,3	214,3
Accountancy	80,5	120,5	122,9	120,3	139,9	151,5	200,7	200,7	198,6
Economics	71,8	113,3	121,1	110,9	133,8	152,3	209,2	209,2	209,2
Administration	68,6	108,5	121,0	108,7	140,9	154,2	212,3	212,3	212,3
Mathematics	75,8	70,3	73,0	122,1	151,7	158,9	222,1	222,1	222,1
Engineering	84,3	130,2	137,6	133,7	156,3	170,8	210,5	210,5	210,5
Pharmacy, Dentist & Other	73,8	142,0	143,8	123,0	160,2	175,1	208,1	208,1	208,1
Law	77,4	165,5	168,0	139,5	179,4	189,5	215,2	215,2	215,2
Medicine	89,6	182,0	186,9	169,0	200,2	206,4	224,3	224,3	224,3

Source: Vestibular cross section data in 2004.

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)

Major	10th percentile	Min	Min	Median			Maximum			Observations
	All	Firststage	Pass	All	First stage	Pass	All	First stage	Pass	
Barbalha	66,19	182,05	186,86	152,62	191,67	199,62	214,29	214,29	214,29	739
Sobral	121,57	185,05	186,86	171,76	196,52	200,76	214,38	214,38	214,19	542
Fortaleza	93,05	193,67	193,86	172,95	202,57	208,57	224,29	224,29	224,29	2325

Source: Vestibular cross section data in 2004.

Table 4: Descriptive statistics in the two medical majors

Sobral: 40 positions

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

Fortaleza: 150 positions

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

Source: Vestibular cross section data in 2004.

Table 5: First stage exam grade equation

	Specification 1	Specification 2	Specification 3
(Intercept)	27.28	26.59	78.00
	(3.59)***	(3.66)***	(2.23)***
Female	0.54	0.47	0.44
	(0.40)	(0.40)	(0.40)
Age	-0.86	-0.86	-0.87
	(0.11)***	(0.11)***	(0.11)***
Special high school	-6.54	-6.46	-6.65
	(1.73)***	(1.74)***	(1.75)***
Private high school	2.67	1.99	2.14
	(0.56)***	(0.67)***	(0.65)***
Preparatory course	1.67	1.51	1.51
	(0.48)***	(0.50)***	(0.50)***
Repetitions	2.83	2.86	2.87
	(0.35)***	(0.37)***	(0.37)***
Ability(m_0)			12.96
			(0.65)***
Spline(1)(m_0 Residual)	48.18	48.72	
	(4.03)***	(4.00)***	
Spline(2)(m_0 Residual)	89.17	89.20	
	(4.54)***	(4.49)***	
Living in Fortaleza	3.72	3.69	3.60
	(0.66)***	(0.67)***	(0.67)***
Living in Fortaleza*Ability	2.02	1.98	1.93
	(0.68)***	(0.66)***	(0.66)***
Mother's education		0.11	0.10
		(0.31)	(0.31)
Father's education		0.33	0.33
		(0.29)	(0.29)
R^2	0.7196	0.7199	0.7198

[1] Living in Fortaleza is a dummy which indicates whether the student is currently living in Fortaleza.

[2] Standard errors are between brackets and * (resp. ** and ***) denotes significance at a 10 (resp 5 and 1) percent level.

[3] The coefficients and their standard errors are computed by bootstrapping the procedure 499 times using the empirical distribution of residuals.

Table 6: Second stage exam grade equation

	Specification 1	Specification 2
(Intercept)	232.65	171.69
	(13.72)***	(20.08)***
Female	7.36	7.16
	(2.27)***	(2.28)***
Age	-3.90	-3.96
	(0.75)***	(0.74)***
Special high school	-11.48	-12.68
	(21.76)	(20.25)
Private high school	8.82	9.11
	(4.15)***	(4.27)***
Preparatory course	9.15	8.95
	(3.38)***	(3.44)***
Repetitions	13.91	14.14
	(2.21)***	(2.25)***
u_1 (m_1 residual)	2.51	
	(0.18)***	
Spline(1)(m_1 residual)		68.09
		(28.38)***
Spline(2)(m_1 residual)		153.07
		(11.47)***
Ability (m_0)	35.23	35.05
	(3.52)***	(2.63)***
R^2	0.2284	0.2286

[1] Standard errors are computed by bootstrapping 499 times using both grade equations and the empirical distributions of residuals.
[2] Standard errors are between brackets and starred signs are defined as in Table 5.

Table 7: Simulated success probabilities

	Sobral		Fortaleza	
	Stage 1	Final Success	Stage 1	Final Success
Min.	0.000	0.000	0.000	0.000
25%	0.001	0.001	0.000	0.000
Median	0.088	0.011	0.012	0.004
Mean	0.314	0.076	0.203	0.062
75%	0.676	0.103	0.360	0.071
Max.	1.000	0.934	1.000	0.920

[1] Success probabilities are constructed using 1000 Monte Carlo simulations.

Table 8: Odds ratio of success probabilities

Percentile	First stage	Second stage
10	1.00	2.66
20	1.00	1.60
30	1.47	1.08
40	0.86	1.61
50	1.07	2.26
60	1.33	3.43
70	1.29	5.34
75	1.18	5.62
80	1.15	5.22
85	1.14	4.41
90	1.10	3.73
95	1.03	3.37
100	1.00	1.74

[1] The first column reports the odds ratio of success probabilities at the first stage between subsamples of those who choose Sobral and choose Fortaleza $\frac{p1sob|d_i=s}{p1fort|d_i=s} / \frac{p1sob|d_i=f}{p1fort|d_i=f}$.

[2] The second column reports the odds ratio of final success probability at the second stage between subsamples of those who choose Sobral and choose Fortaleza $\frac{psob|d_i=s}{pfort|d_i=s} / \frac{psob|d_i=f}{pfort|d_i=f}$.

[3] Percentiles in rows are computed using first stage exam grades.

Table 9: Estimated preferences for Sobral's medical school

Parameters

	Specification 1	Specification 2	Specification 3
δ_0^S	-2.782	-1.132	-1.167
	(0.303)***	(0.309)***	(0.277)***
$\delta_{m_0}^S$	0.261	0.166	
	(0.189)*	(0.146)*	
$\delta_{LivinginFortaleza}^S$		-1.815	-1.586
		(0.522)***	(0.283)***
δ_0^{SF}	-0.453	0.521	0.484
	(0.271)*	(0.312)**	(0.296)**
$\delta_{m_0}^{SF}$	0.979	1.062	
	(0.198)***	(0.179)***	
$\delta_{LivinginFortaleza}^{SF}$		-1.314	-1.225
		(0.326)***	(0.393)***
Intercept	0.075	0.334	0.0482
	(0.707)	(0.387)	(0.393)
Ability (m_0)	-1.079	-0.977	-0.020
	(0.261)***	(0.247)***	(0.095)
Living in Fortaleza		-0.248	-0.558
		(0.301).	(0.314)**
Female	-0.325	-0.240	-0.373
	(0.139)***	(0.152)***	(0.186)***
Age	-0.038	-0.045	-0.048
	(0.039)	(0.027)**	(0.026)**
Repetitions	0.688	0.851	0.911
	(0.144)***	(0.141)***	(0.210)***
Father's education	-0.278	-0.257	-0.341
	(0.111)***	(0.119)***	(0.154)***
Mother's education	0.084	0.046	0.216
	(0.106)	(0.114)	(0.145)

Proportions

		Specification 1	Specification 2	Specification 3
δ^S	Min	0.022	0.021	0.050
	Mean	0.060	0.057	0.066
	Max	0.122	0.248	0.196
δ^{SF}	Min	0.015	0.016	0.365
	Mean	0.385	0.412	0.386
	Max	0.816	0.852	0.559
δ^F	Min	0.062	0.027	0.245
	Mean	0.555	0.531	0.548
	Max	0.963	0.962	0.585

[1] The second part of the table reports summaries of the probabilities of being in one of the three regions of Figure 1.
[2] The coefficients and their standard errors are computed by bootstrapping 499 times the whole procedure (including grade equations).
[3] Standard errors are between brackets and starred signs are defined as in Table 5.

Table 10: Thresholds in the original and counterfactual experiments

School			Sobral	Fortaleza
Stage 1	Original system	Mean Thresholds	183.13	190.10
		Standard Errors	(0.850)	(0.397)
	Cutting seats	Mean Thresholds	196.75	200.71
		Standard Errors	(0.962)	(0.510)
	Two-Choices	Mean Thresholds	187.39	190.16
		Standard error	(0.552)	(0.434)
	Timing-Change	Mean Thresholds	205.59	186.89
		Standard error	(0.508)	(0.390)

School			Sobral	Fortaleza
Stage 2	Original system	Mean Thresholds	239.03	244.63
		Standard Errors	(2.907)	(1.426)
	Cutting seats	Mean Thresholds	234.25	237.43
		Standard Errors	(3.124)	(1.610)
	Two-Choices	Mean Thresholds	235.21	241.25
		Standard error	(2.562)	(1.296)
	Timing-Change	Mean Thresholds	259.77	235.49
		Standard error	(2.315)	(1.319)

[1] The coefficients and their standard errors are computed by using the 499 bootstrapped estimates of preference and grade parameters and applying the procedure in the text.

[2] The cutting seats counterfactual has a few cases in which the computation developed in Section 5.2 does not converge after many repetitions, and we have excluded those bootstrap values that do not converge after 500 iterations.

Table 11: Cutting seats: Expected utility changes

Expected	ALL		D=Sobral		D=Fortaleza	
Final Grade	mean	s.d.	mean	s.d.	mean	s.d.
0% -50%	-0.00033	0.00059	-0.00048	0.00083	-0.00030	0.00052
50%-60%	-0.00383	0.00208	-0.00448	0.00254	-0.00360	0.00185
60%-70%	-0.00751	0.00633	-0.00872	0.00617	-0.00716	0.00635
70%-80%	-0.00694	0.01290	-0.01089	0.00926	-0.00623	0.01334
80%-82%	-0.00851	0.01289	-0.00754	0.00575	-0.00885	0.01460
82%-84%	-0.00075	0.01430	0.00073	0.00584	-0.00107	0.01558
84%-86%	0.01203	0.01002	-0.00200	0.01422	0.01433	0.00700
86%-88%	0.00229	0.01111	-0.00033	0.01141	0.00302	0.01105
88%-90%	0.01226	0.00860	0.00803	0.00319	0.01291	0.00899
90%-92%	0.01657	0.01203	0.00157	0.01103	0.01828	0.01102
92%-94%	0.01695	0.01015	0.00671	0.01245	0.01942	0.00779
94%-96%	0.02622	0.00550	0.01191	0.00157	0.02739	0.00376
96%-98%	0.03126	0.00648	0.01385	0.00293	0.03311	0.00305
98%-100%	0.02890	0.01028	0.01041	0.00230	0.03398	0.00343
$\mathbf{E}(\Delta U_i)$	0.00080		-0.00218		0.00149	
$\mathbf{s.d.}(\Delta U_i)$	0.01107		0.00710		0.01170	
$\mathbf{Pr}(\Delta U_i > 0)$	0.3907		0.2934		0.4133	

[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.

[3] $\mathbf{E}(\Delta U_i)$ (resp. $\mathbf{s.d.}(\Delta U_i)$) is the sample average (resp. standard deviation) of the total utilitarian welfare change.

[3] $\mathbf{Pr}(\Delta U_i > 0)$ is the frequency of students whose expected utility changes are positive

Table 12: Cutting seats: Robustness

Expected	$\mu = 0.8$		$\mu = 0$		$\mu = 1$	
Final Grade	mean	s.d.	mean	s.d.	mean	s.d.
0% -50%	-0.00033	0.00059	-0.00030	0.00054	-0.00034	0.00061
50%-60%	-0.00383	0.00208	-0.00351	0.00187	-0.00391	0.00214
60%-70%	-0.00751	0.00633	-0.00698	0.00584	-0.00764	0.00646
70%-80%	-0.00694	0.01290	-0.00660	0.01232	-0.00702	0.01304
80%-82%	-0.00851	0.01289	-0.00842	0.01266	-0.00853	0.01295
82%-84%	-0.00075	0.01430	-0.00085	0.01394	-0.00072	0.01439
84%-86%	0.01203	0.01002	0.01185	0.00984	0.01208	0.01007
86%-88%	0.00229	0.01111	0.00209	0.01084	0.00234	0.01118
88%-90%	0.01226	0.00860	0.01186	0.00848	0.01236	0.00863
90%-92%	0.01657	0.01203	0.01630	0.01191	0.01664	0.01206
92%-94%	0.01695	0.01015	0.01651	0.00997	0.01706	0.01019
94%-96%	0.02622	0.00550	0.02579	0.00555	0.02633	0.00549
96%-98%	0.03126	0.00648	0.03077	0.00652	0.03138	0.00647
98%-100%	0.02890	0.01028	0.02852	0.01031	0.02900	0.01027
$\mathbf{E}(\Delta U_i)$						
	0.00080		-0.00218		0.00149	
$\mathbf{s.d.}(\Delta U_i)$						
	0.01107		0.00710		0.01170	
$\mathbf{Pr}(\Delta U_i > 0)$						
	0.3907		0.2934		0.4133	

[1] Results as in Table 11 using different values of μ.
[2] See notes of Table 11

Table 13: Two choices: Expected utility changes

Expected	ALL		D=Sobral		D=Fortaleza	
Final Grade	mean	s.d.	mean	s.d.	mean	s.d.
0%-50%	0.00019	0.00041	-0.00006	0.00015	0.00025	0.00043
50%-60%	0.00177	0.00206	-0.00046	0.00041	0.00254	0.00183
60%-70%	0.00601	0.00642	-0.00040	0.00038	0.00790	0.00613
70%-80%	0.01226	0.01022	-0.00009	0.00050	0.01446	0.00953
80%-82%	0.02100	0.01562	0.00019	0.00012	0.02810	0.01123
82%-84%	0.01962	0.01363	0.00023	0.00015	0.02384	0.01120
84%-86%	0.01433	0.01039	0.00023	0.00013	0.01664	0.00934
86%-88%	0.02760	0.01886	0.00034	0.00012	0.03520	0.01363
88%-90%	0.02968	0.01713	0.00033	0.00010	0.03420	0.01355
90%-92%	0.02497	0.02098	0.00049	0.00025	0.02777	0.02034
92%-94%	0.03179	0.02228	0.00054	0.00021	0.03935	0.01790
94%-96%	0.03469	0.01907	0.00048	0.00022	0.03748	0.01699
96%-98%	0.04127	0.02469	0.00073	0.00026	0.04558	0.02186
98%-100%	0.03756	0.02944	0.00063	0.00030	0.04769	0.02496
$\mathbf{E}(\Delta U_i)$						
	0.00780		- 0.00008		0.00964	
$\mathbf{s.d.}(\Delta U_i)$						
	0.01507		0.00039		0.01619	
$\mathbf{Pr}(\Delta U_i > 0)$						
	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.
[3] Notes: See notes of Table 11.

Table 14: Timing change: Expected utility changes

Expected	ALL		D=Sobral		D=Fortaleza	
Final Grade	mean	s.d.	mean	s.d.	mean	s.d.
0%- 50%	0.00221	0.00383	0.00146	0.00223	0.00237	0.00408
50%-60%	0.01885	0.01194	0.00784	0.00807	0.02231	0.01082
60%-70%	0.04189	0.02198	0.01841	0.01606	0.04957	0.01783
70%-80%	0.09641	0.03976	0.04911	0.04427	0.10524	0.03194
80%-82%	0.13401	0.04577	0.06602	0.04275	0.15440	0.01940
82%-84%	0.14342	0.04150	0.07605	0.04903	0.15914	0.01653
84%-86%	0.17757	0.03936	0.12277	0.07257	0.18806	0.01539
86%-88%	0.19540	0.04594	0.13027	0.06396	0.21277	0.01406
88%-90%	0.22733	0.01694	0.20861	0.02565	0.22913	0.01503
90%-92%	0.27670	0.08313	0.14890	0.09739	0.30017	0.05478
92%-94%	0.28371	0.06732	0.19483	0.08725	0.30548	0.03824
94%-96%	0.34077	0.05782	0.21722	0.13380	0.35337	0.02115
96%-98%	0.42198	0.07927	0.28169	0.17460	0.43916	0.03365
98%-100%	0.54759	0.16372	0.35510	0.24210	0.59972	0.07894
$\mathbf{E}(\Delta U_i)$						
	0.07259		0.03853		0.08053	
$\mathbf{s.d.}(\Delta U_i)$						
	0.12694		0.08966		0.13292	
$\mathbf{Pr}(\Delta U_i > 0)$						
	1		0.99		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.

[3] See notes of Table 11

Figure 1: Choice space

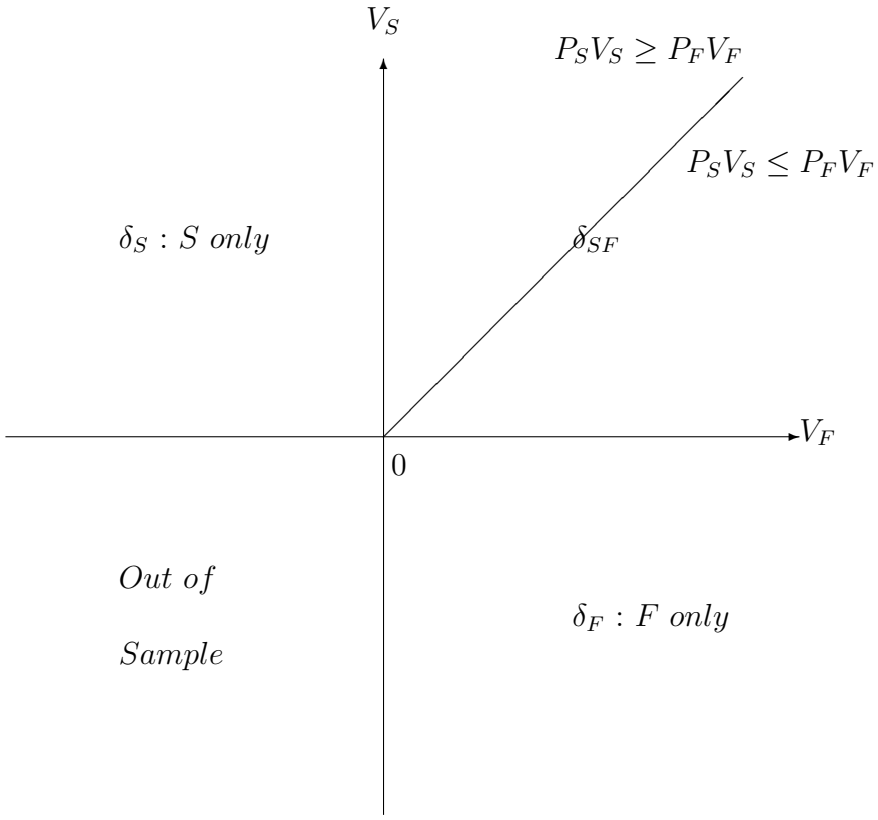


Figure 2: Density plots of the grades

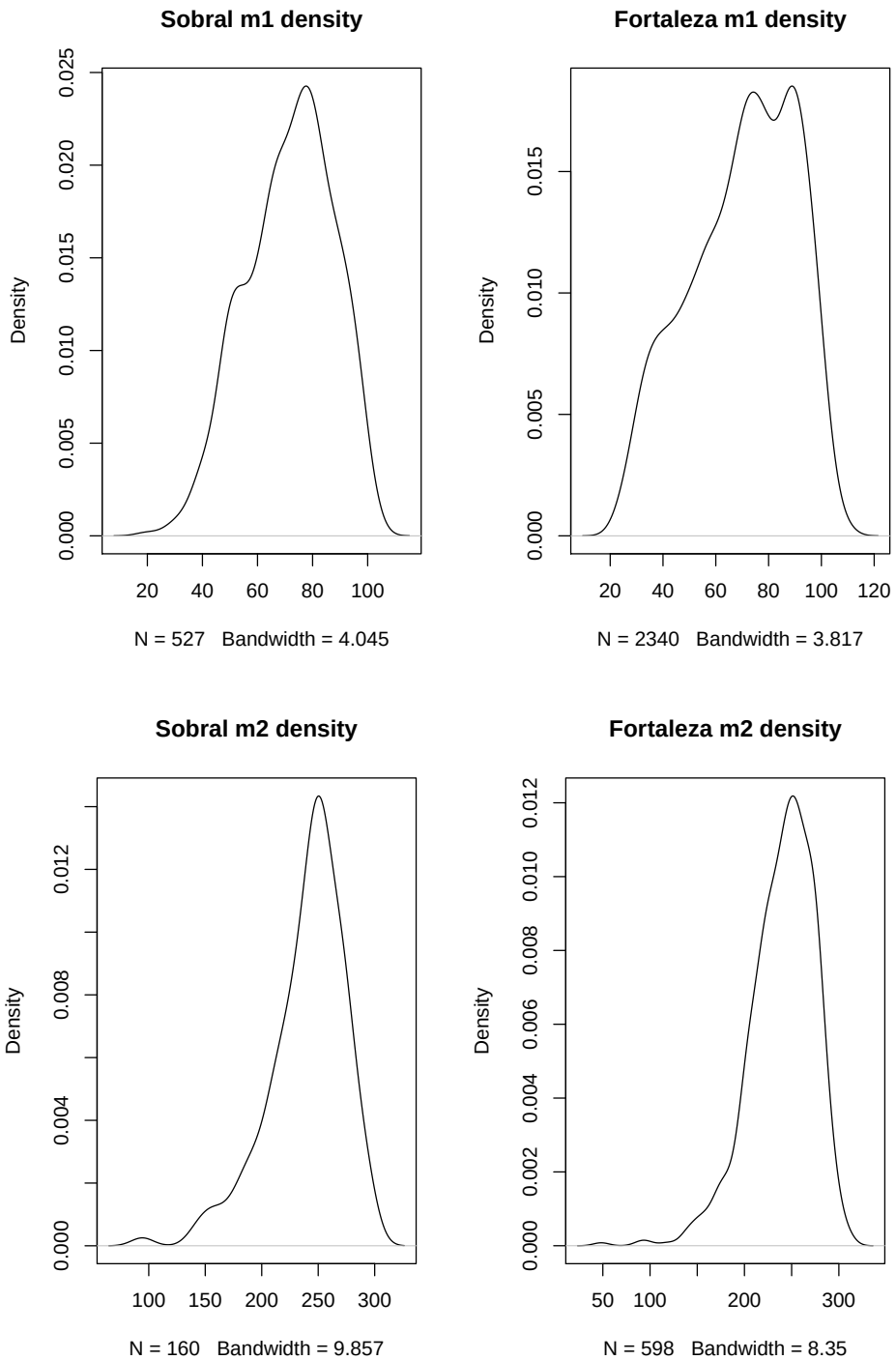
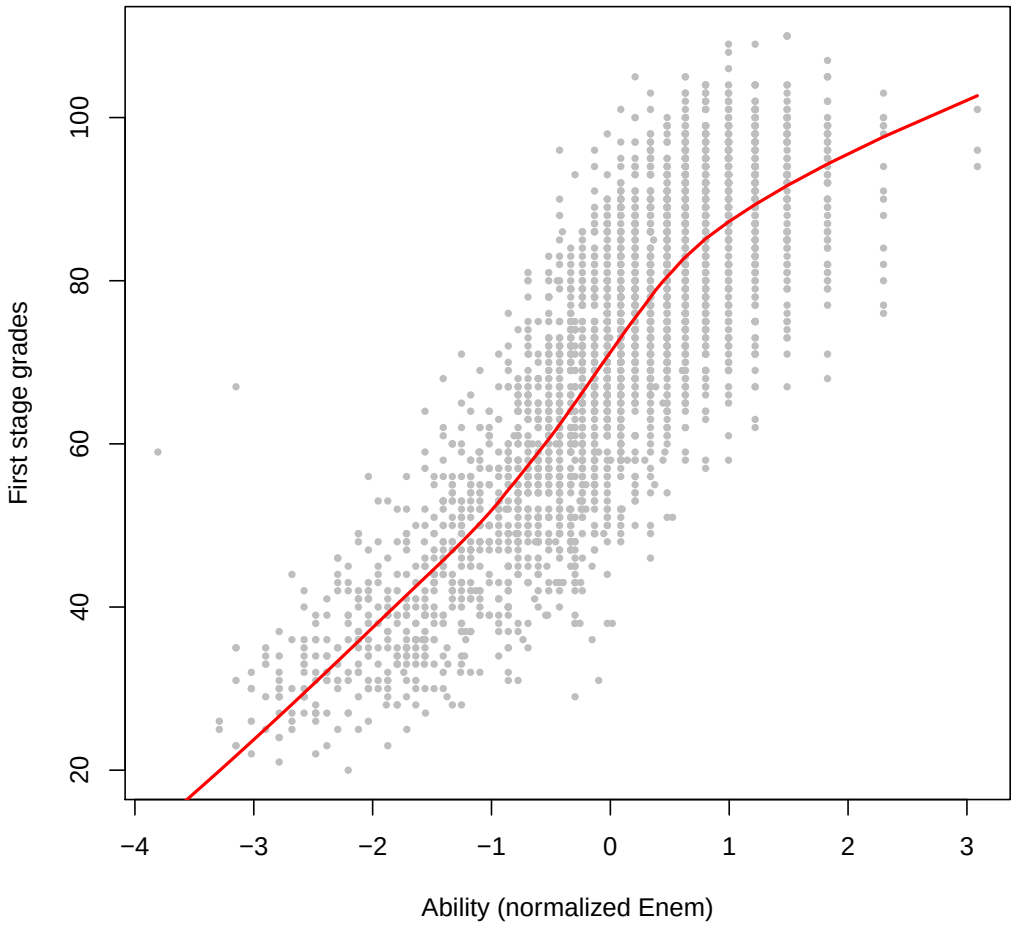
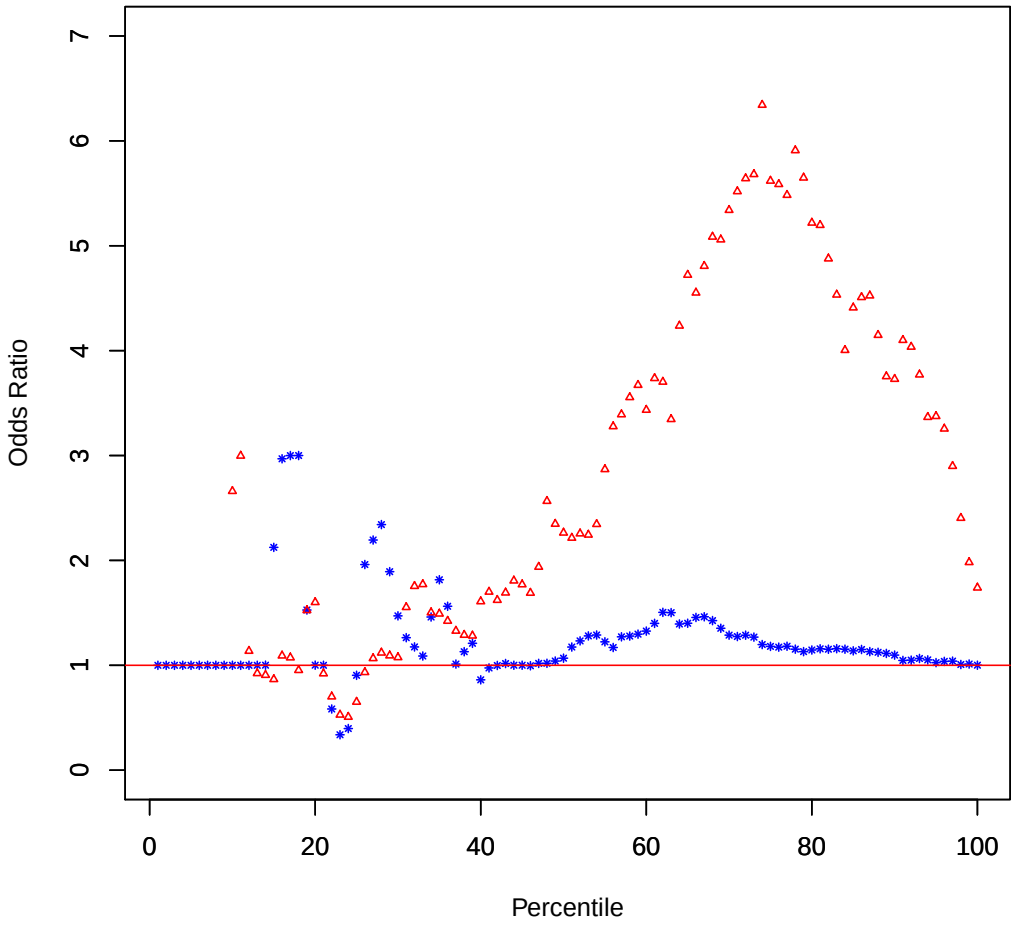


Figure 3: The relation between ability and first stage grades



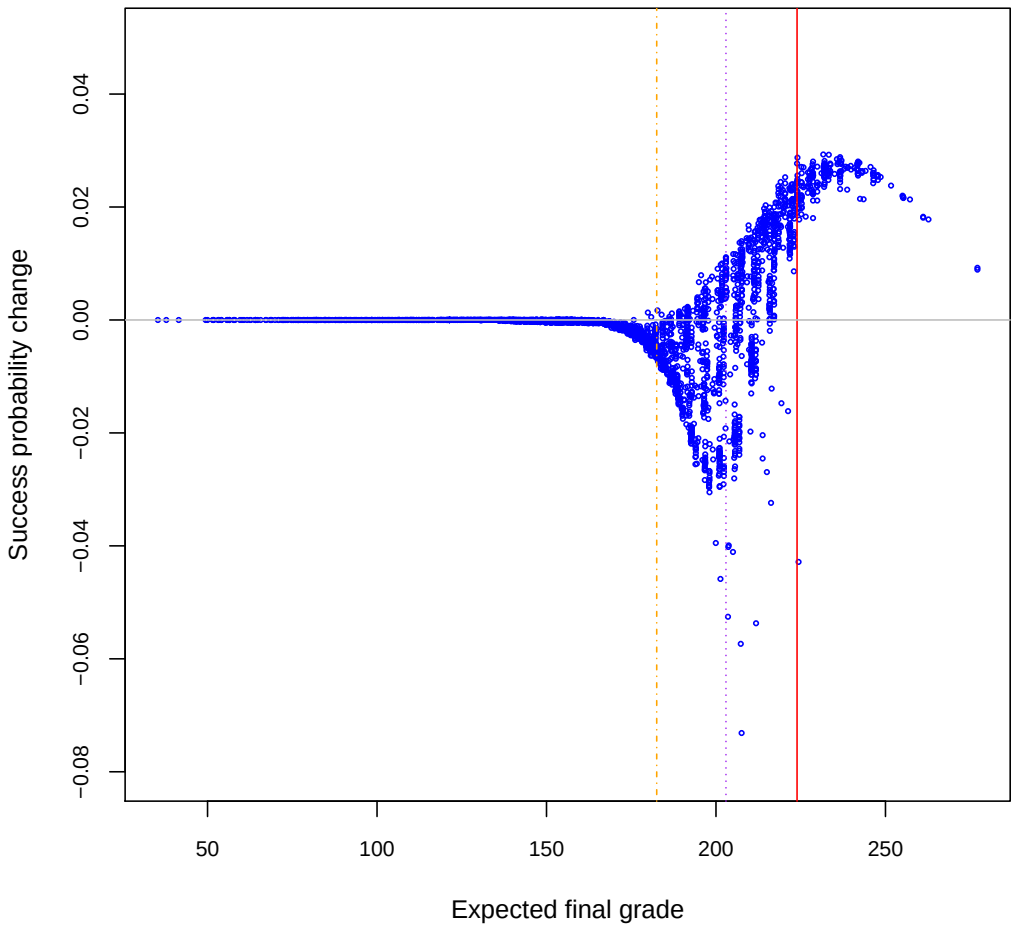
[1] The round grey points are the scatter plots of first stage grade on ability (normalized Enem); [2] The curve is the LOWESS curve of first stage grade on ability (normalized Enem).

Figure 4: The Odds ratio plot of simulated success probabilities



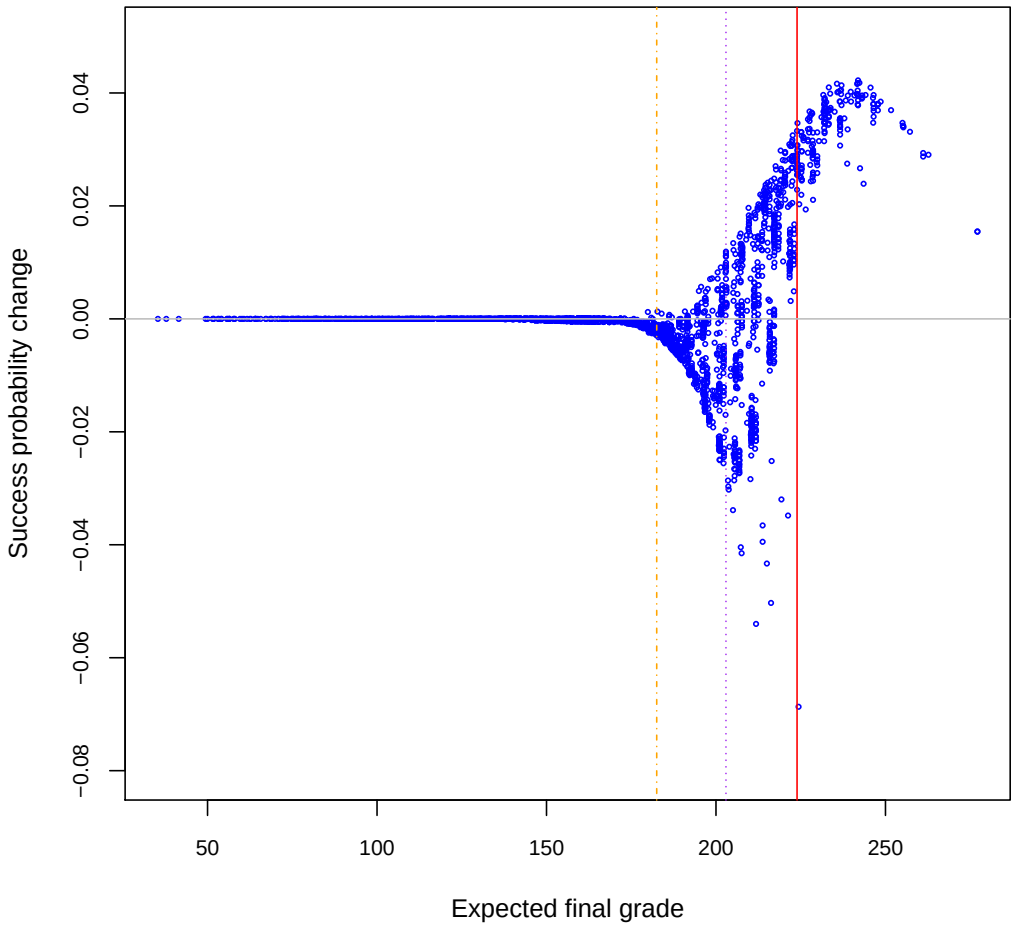
[1] The star points are odds ratio at the first stage ; [2] the triangular points are the odds ratio at the second stage; [3] percentiles are computed using first stage grades.

Figure 5: Cutting seats: Changes of success probabilities in Sobral



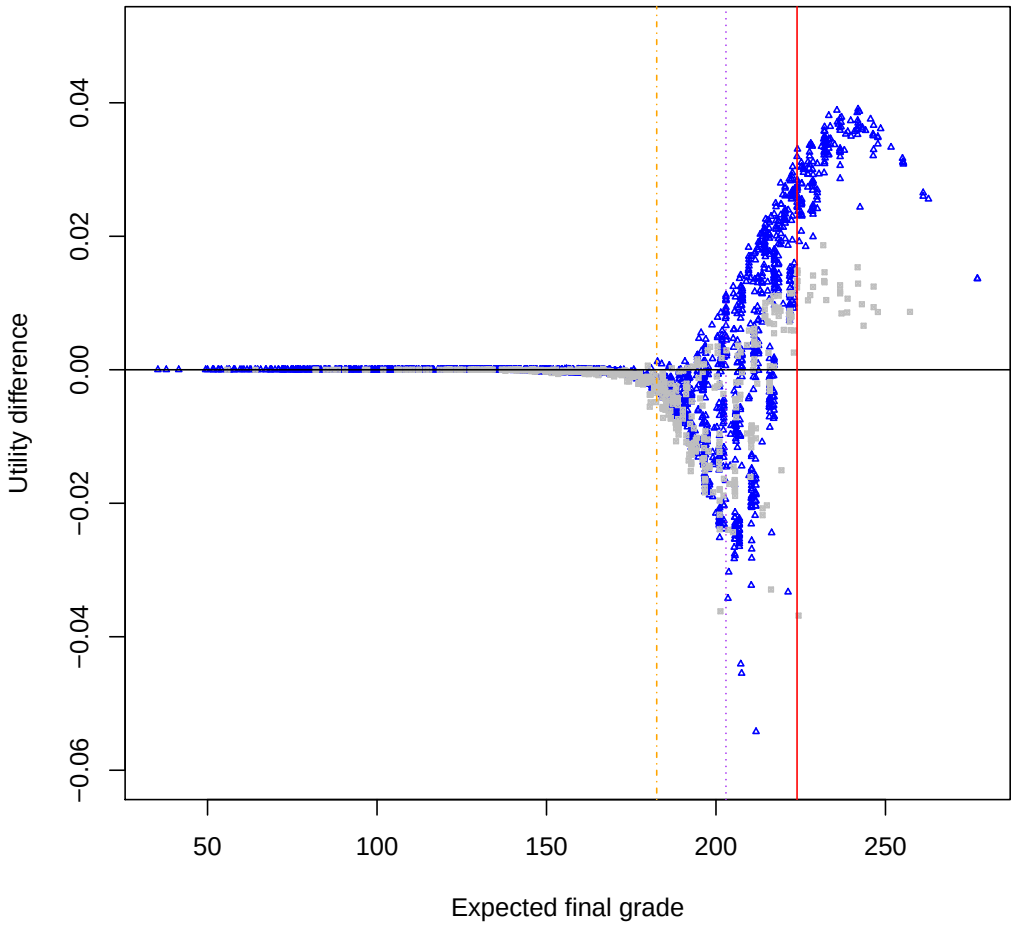
[1] The circles are individual success probability changes vs expected final grades; [2] From left to right, 1) the first vertical line is the median, 2) the second line is the quantile of 1st stage admission – $(1 - \frac{4(nos+nof)}{nobs}) \times 100\%$, and 3) the third line is the quantile of 2nd 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 6: Cutting seats: Changes of success probabilities in Fortaleza



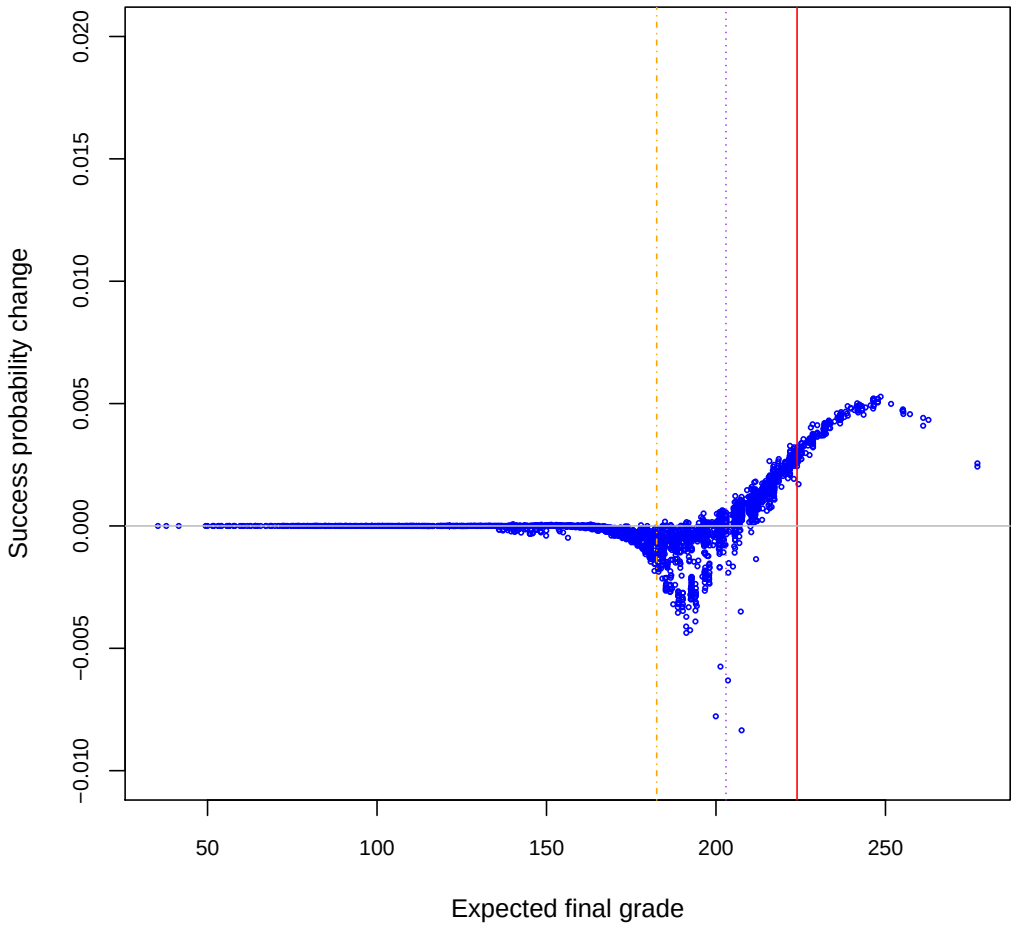
See notes of Figure 5

Figure 7: Cutting seats: Expected utility changes



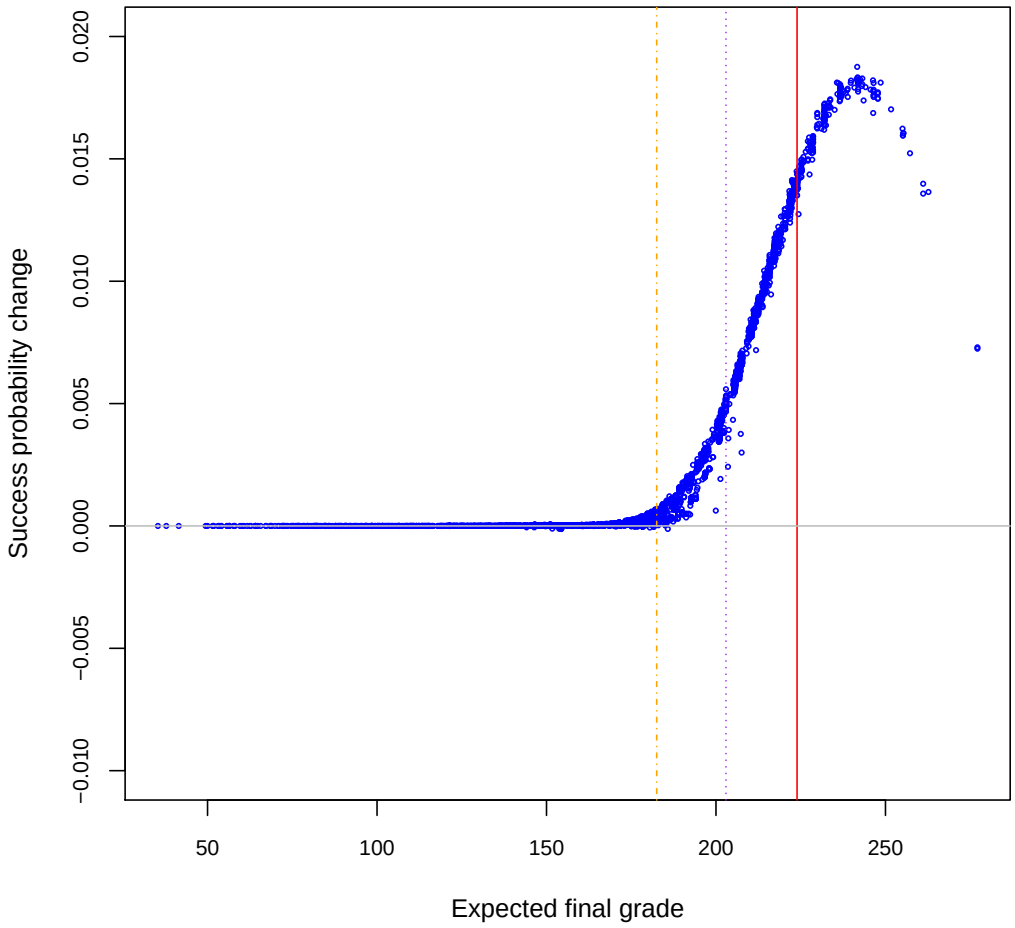
[1] the grey squares (resp. blue triangles) report changes in expected utilities and expected final grades for those who choose Sobral (resp. Fortaleza) in the original system. [2] the red line is the 0 level; [3] the vertical lines are as in Figure 6.

Figure 8: Two choices: Success probability change in Sobral



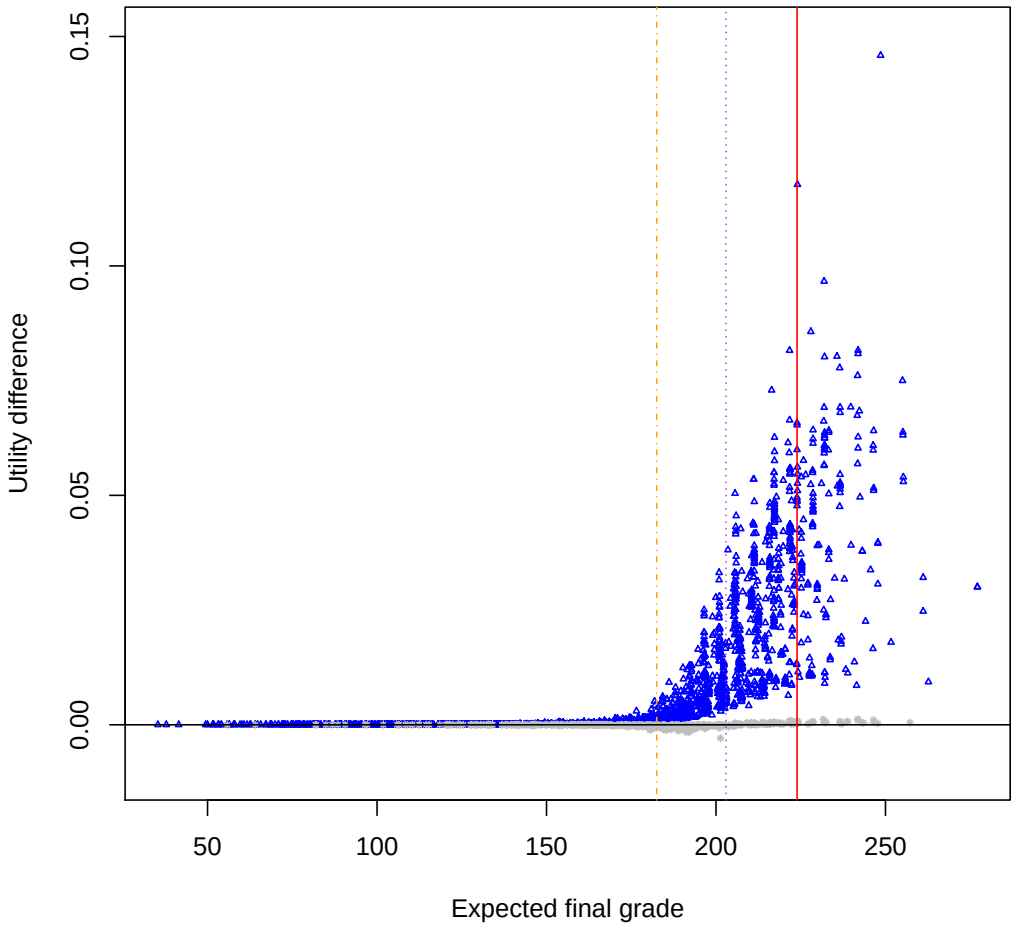
Notes: See notes of Figure 5

Figure 9: Two choices: Success probability change in Fortaleza



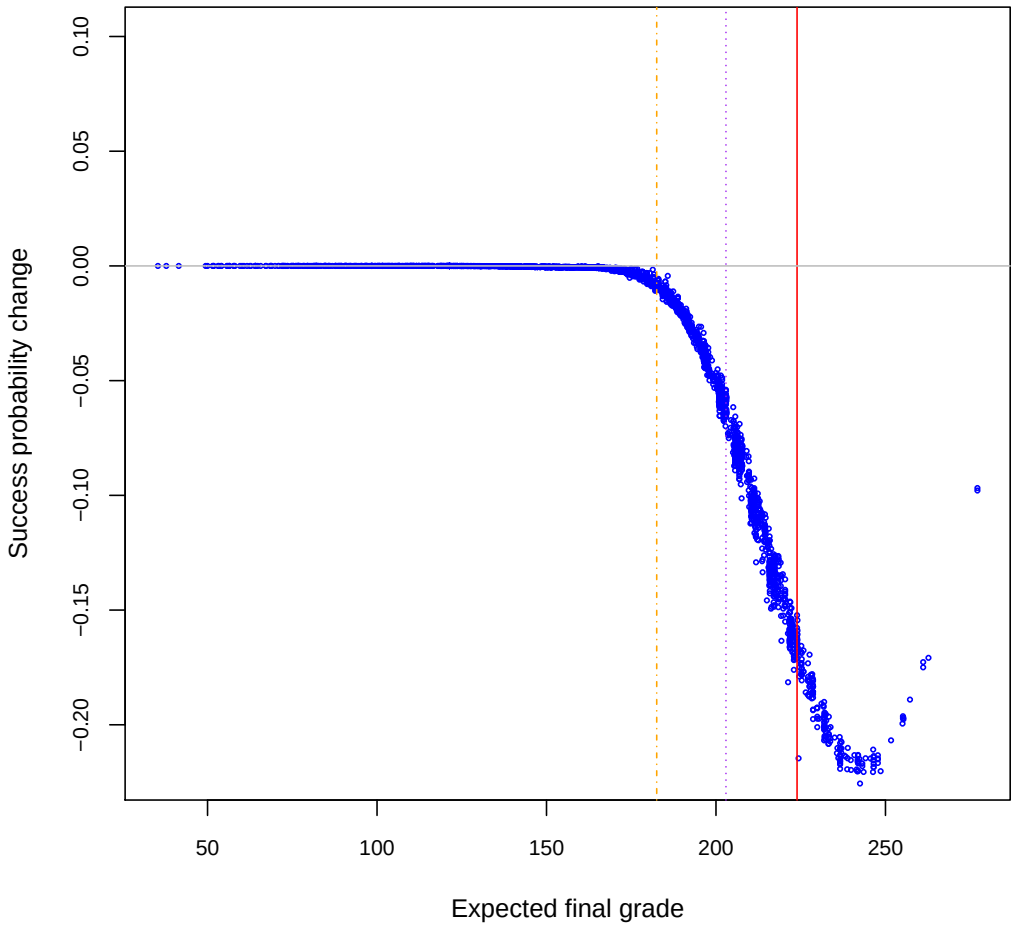
Notes: See notes of Figure 5

Figure 10: Two choices: Expected utility changes



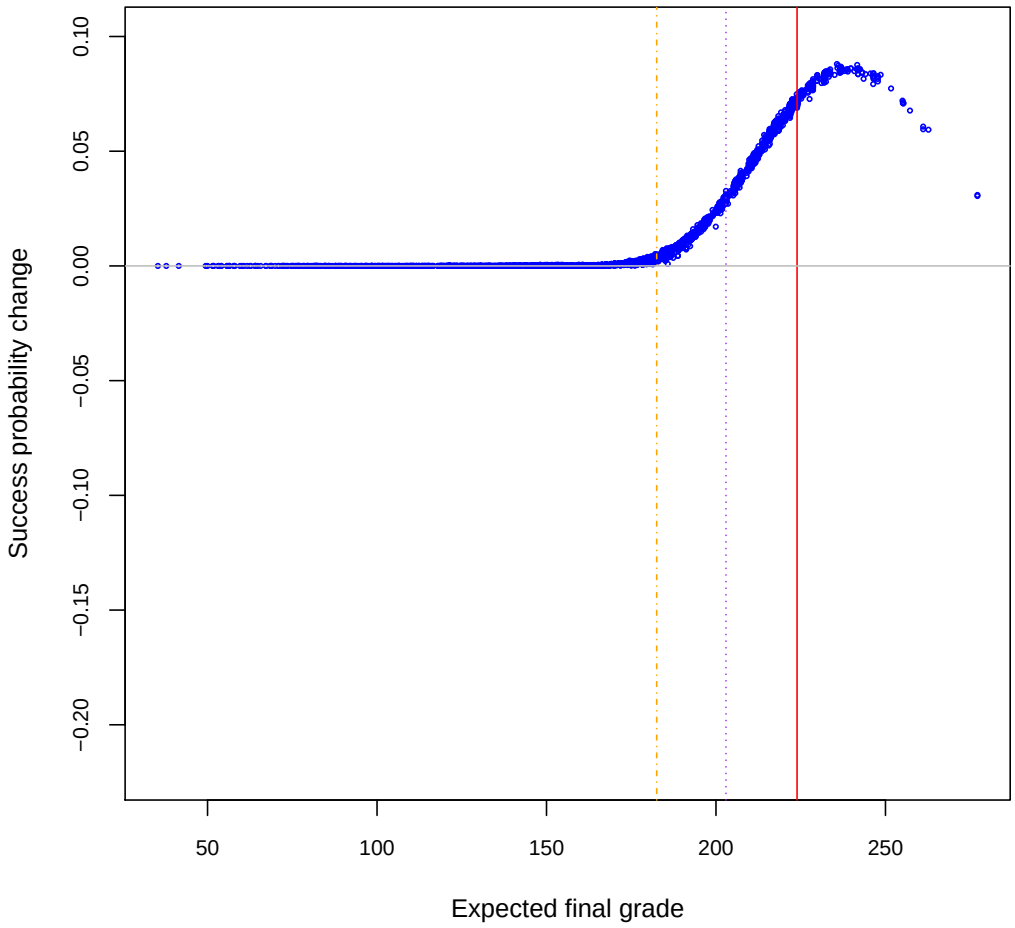
Notes: See notes of Figure 7

Figure 11: Timing change: Success probability changes in Sobral



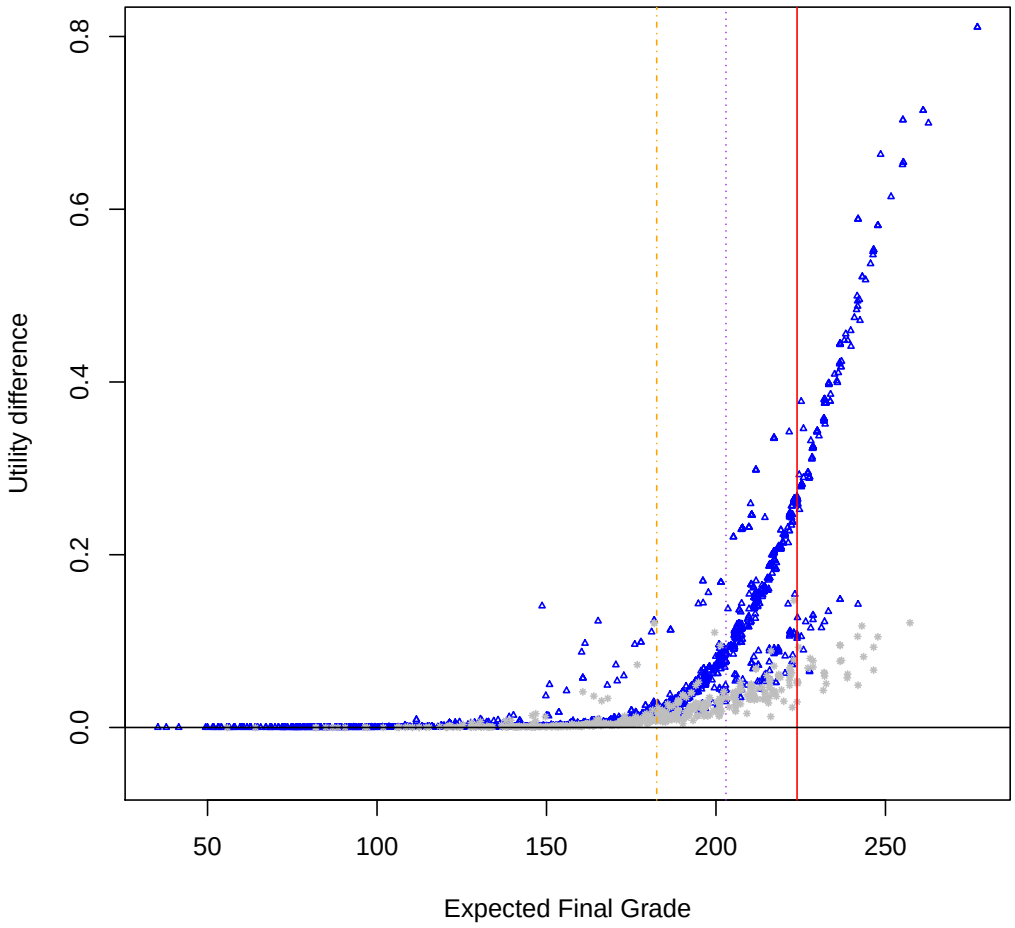
Notes: See notes of Figure 5

Figure 12: Timing change: Success probability changes in Fortaleza



Notes: See notes of Figure 5

Figure 13: Timing change: Expected utility changes



Notes: See notes of Figure 7