College Choice and the Selection of Mechanisms: A Structural Empirical Analysis

José-Raimundo Carvalho
CAEN, Universidade Federal do Ceará

Thierry Magnac
Toulouse School of Economics

Qizhou Xiong
Toulouse School of Economics

December 17, 2015

Abstract

We use rich microeconomic data on performance and choices of students at college entry exams to analyze the impact on student welfare of using different allocation mechanisms between university majors. The allocation procedure in place allows us to model precisely preferences and expectations of students using the set-up of a congestion game. The counterfactuals we consider balance costs arising from congestion and costs arising from using simple exams to rank students. A deferred acceptance type of mechanism or the inversion of the timing of choices and exams are shown to increase welfare. Redistribution among students or schools is sizeable.

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

JEL codes: C57, D47, I21

1 Introduction[1]

The *matching* literature deals with the allocation of goods or relationships between many parties in the absence of a price mechanism (see Roth and Sotomayor, 1992, Roth, 2008). There are many examples ranging from kidney exchange and marriage to school choice. The analysis of school choice as a many-to-one match has been very popular recently in the theoretical and empirical literature (for instance Abdulkadiroğlu, Che and Yasuda (2012), Agarwal and Somaini (2014), Azevedo and Leshno (2014), Budish and Cantillon (2012), Agarwal (2013), Calsamiglia, Fu and Guell (2014), He (2014) and many others) and has had practical value for policy implemented in primary or high schools in various countries.

This literature deals with the question of which allocation is best and which mechanisms lead to it. What is best is defined by using Pareto optimality or a concept of *fairness* or *stability* that translates the elimination of *justified envy (i.e.* no pair of (major, student) might be made better off by changing partners). The absence of disguise by students of their true preferences or *strategy proofness* is also prominent in the debate between scholars (surveyed for instance by Sönmez and Ünver, 2011). Among mechanisms, the well known Gale Shapley student optimal stable mechanism is strategy proof and optimal from the students' perspective among stable mechanisms if school preferences are simple and strict over any pair of students (*e.g.* Abdulkadiroğlu and Sonmez, 2003). Nonetheless, if preferences are not strict as in a system of priority used in primary or high schools, random tie breaking affects the ex-ante optimality of the mechanism (Abdulkadiroğlu, Che and Yasuda, 2012, Buddish and Cantillon, 2012).

If we look at prospective college students however, colleges generally have strict preferences over them so that the coarseness of preferences do not matter. The important distinction between allocation procedures in practice seems to be the degree of their centralization. Some are centralized at the level of local authorities or countries as for instance universities in China (Chen and

[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 talks and interactions with Yinghua He, Philipp Heller and 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, Barcelona, Manchester, Bern and IAAE'15 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.

Kesten, 2014) or Turkey (Balinski and Sonmez, 1999). Usually, students write a general exam, composed of proofs in different fields (maths, language etc) whose results determine students' ranking. Other allocation procedures are decentralized as in the US where colleges compete for students in dimensions which are not just academic. First, centralization avoids the costs of congestion since colleges do not have to deal with all student files. It also streamlines the competition between colleges. In a decentralized setting, the management of offers and acceptance of offers is uncertain and takes time (Che and Koh, 2014), is strategic (Chade, Lewis and Smith, 2014) and the use of waiting lists might lead to unstable mechanisms. Yet, centralization assumes that college preferences are adequately translated by the information revealed at a general exam (Hafalir, Hakimov, Kübler and Kurino, 2014) even though admittedly colleges are free to weight results of multiple subexams in each field of study. In a decentralized system like in the US, many other elements than the SAT score are evaluated and the selection is multidimensional (Che and Koh, 2013). There is thus a tradeoff for colleges between the costs of congestion related to the selection of students and the costs of adopting "proxy" preferences instead of a more precise and adapted ranking.

This is the tradeoff that we want to investigate empirically in this paper. Our observational "experiment" uses full observation of performances (grades at two exam stages and a pre-exam grade as well) at one specific formal *entry exam* at one of the Federal Universities in Brazil in 2004. A mechanism, decentralized at the level of the University and called *vestibular*, is used: during their last high school year, students choose a single major[2] before taking a 2-stage exam at the end of high school. The first stage is common to all majors and selects students for the second stage. This latter stage is more specialized and takes place around a month later than the first stage. The rationale seems to be about (1) avoiding the costs of congestion and (2) eliciting a precise measure of school preferences. The first stage eliminates most candidates through a cost-minimizing multiple question exam while the second stage digs much deeper into eliciting the ability of students in specific fields of study. This is broadly akin to the Japanese experience in which a first stage centralized exam is followed by a second stage exam decentralized at the level of each university *on the same day* which effectively avoids congestion (see Hafalir et al., 2014). Nonetheless, the mechanism in place restricting choices of students to a single major is

[2]We use the terms "major" and "school" interchangeably.

generically not *stable* and students are likely to play *strategically* by balancing preferences and success probabilities.

What we do in this paper aims at evaluating the effects on student allocations and their welfare of adopting other more theoretically sound mechanisms than the existing *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 and model in more detail than in the current literature, the strategies employed by students.

Our first contribution is to adopt an empirical strategy that uses performance data to build up estimates of success probabilities and estimate a model of school choice when students play strategically (Arcidiacono, 2005, Epple, Romano and Sieg, 2006). School choice indeed depends on (1) expected probabilities of success and (2) preferences for colleges including future wages. The advantage of our data lies in the rich information on performance at the two-stage exams before entry (or failure) as well as an initial measure of ability obtained a year before the exams are taken. As usual in the literature, 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 these thresholds.

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 a "congestion" game in which choices of other students affect success probabilities of any student. We explicit the information that students are supposed to have and use the best response functions of the students for estimating parameters. We assume that expectations of success probabilities are perfect that is, they are obtained by infinitely repeating the game with the same players. This is because the strategizing ability seems simpler in our set up than in most cases and many agents like parents or teachers are ready to help out students to form expectations (see Manski, 1993, for a critical appraisal of such assumptions). We show that, conditional on information sets, success probabilities can be obtained by resampling in our single observed sample and by using the Nash equilibrium conditions that deliver random thresholds.

We also provide a proof of 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 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 main economic contribution is to analyze the effects on allocation and welfare of students and schools by contrasting three different counterfactual mechanisms. Our estimated microeconomic model allows us not only to study average welfare effects but also detailed 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 Gale-Shapley deferred acceptance mechanism. Students are allowed to submit a list of two choices instead of a single one 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 utilitarian 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 Matching students to schools has a long history and a brief survey of the recent literature is given in Sonmez and Ünver (2011) in which differences between school choice, college admission and student placement are rigorously defined. Prominently, Gale Shapley mechanisms satisfy both properties of stability and strategy-proofness. 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, is evaluated positively by this major in comparison with other students in the same situation. The seminal analysis of college admissions by Balinski and Sönmez (1999) was theoretical albeit oriented towards the analysis of a specific mechanism. They studied the

optimality of student placement 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 weights that they give to different fields: grades in maths can presumably be given more weight by math colleges.

Most of the empirical litterature is concerned by primary or high school choices. 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) estimate demands in the newly introduced deferrence acceptance mechanism. They are able to compare this allocation in terms of welfare and distributive effects to the previous decentralized allocation. This research generally questions the relative standing of the Gale-Shapley and the Boston mechanisms (Abdulkadiroglu, Pathak, Roth and Sönmez, 2006). Others analyze the Boston mechanism as He (2014) who uses school allocation data from Beijing, finds sizeable strategic moves as well and evaluates the cost of strategizing for sophisticated and naive agents. The question of the importance of truncated lists of preferences used in practice in deferred acceptance mechanisms is high on the agenda in recent research (Calsamiglia et al, 2014, and Fack, Grenet and He, 2015).

School and college choices differ in various dimensions. Demand for colleges is generally much larger than supply. Success probabilities depend on exogenous individual characteristics and not only on priorities. Many agents can help college applicants to assess these probabilities. Demands for colleges 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. Fu (2014) estimates demand and supply equations when students have heterogenous abilities and preferences and when college applications are costly and uncertain.

Decentralized models are studied by Chade, Lewis and Smith (2014), Che and Koh (2013) and Hafalir et al. (2014) among others. Congestion is reduced by either making students pay an application cost or by making them choose only one college. In Chade et al (2014), school preferences are noisy signals of students' abilities and college strategizing can lead to inefficient sorting of students. In Che and Koh (2013), uncertainty of student preferences make school play strategically and this leads to inefficient and unfair assignments. In Hafalir et al. (2014), low and

high ability students are shown to have different preferences over centralized and decentralized mechanisms and a small literature about centralization is surveyed there

To our knowledge, there is no comparative survey of college admission procedures in different countries. There exist empirical papers about the "parallel" mechanism used in China (Chen and Kesten, 2014 or Zhu, 2014) or descriptive anlyses in Turkey (Dogan and Yuret, 2012) or in Egypt (Selim and Salem, 2009). Abizada and Chen (2011) analyze the eligibility restrictions to college access that gives a way of reducing costs of evaluation of students by colleges. A descriptive analysis of the mechanism which is now used in Brazil is provided by Gontijo (2008) and Aygun and Bo (2014).

The paper is organized in the following way. Section 2 describes the Vestibular system as it was defined in 2004, the modeling assumptions of the game and the Nash equilibrium conditions. 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. A Supplementary Appendix, available upon request, gathers the details of our many procedures.

2 Description of the game and modeling

We start by describing how Universidade Federal do Ceara (UFC from here on out) 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 to dispute. As already mentioned, the exam consists in two stages. The access to the second stage is conditionned on the grade obtained at the first stage and students are selected within the population of those who have chosen a given major. The number of students who are accepted to the second stage is a multiple (almost always 4) of the number of final available seats. By ranking students in each major, this defines a first stage grade threshold specific to the major above which students pass to the second stage. Similarly, major specific second stage thresholds determine who passes the exam and is accepted

in the major.[3]

The first subsection defines notation, 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 in which information on future grades is imperfect. Agents are assumed to be informed about observed characteristics of competing students and they know the distributions of unobserved heterogeneity in grades 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 on the uniqueness of pure strategy best responses of this game and the necessary conditions for a pure strategy Nash equilibrium.

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 analyze these two medical schools in the empirical application and since the extension of the theoretical model to any number of majors is easy.[4] 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 measuring students' ability in different subjects is organized one year before *Vestibular* exams begin. It is known

[3]Section S.1 in the Supplementary Appendix gives further details on the mechanism and the exams.

[4]It is much less easy in the empirical application since students' revealed preferences about longer lists of majors are not observed.

as $ENEM$ and is 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 an external option and we do not use the information that we have about other majors in this University.

- **Stage 2 – 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 characteristics of students, Z, as:

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

in which u_1 are random 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 slots 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 school S 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.

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 next year's Vestibular, finding a program outside of the Vestibular system, studying abroad or working.

- **Stage 3 – 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 a linear aggregator of $ENEM$, m_1 and m_2 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 after completion 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 it 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:

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

As decribed above, observed participants are those who get a positive utility level in at least one of the two schools of interest so that $\max(V_S, V_F) > 0$. 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. We test this assumption in the empirical section.

2.2 Expectations and 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 of unknown grade shocks (u_1, u_2). 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 crucially depend 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%.

We distinguish one applicant, indexed by 0, from all other applicants, $i = 1, ., n$, 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). Applicant 0 faces the n other applicants and we shall construct her best response to other players' choices, $\{D_i\}_{i=1,.,n} \equiv D_{(n)}$.

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. We also assume that student 0 observes the characteristics that affect grades of all other students, $Z_{(n)} = \{Z_i\}_{i=1,.,n}$ and that the distribution function, F_u, of u_i for all $i = 0, 1, ., n$ is the same across students as well as functional forms and parameters of grade equations and this information is common knowledge. The information set of student 0 at the initial stage is thus composed of $W_0 = (X_0, Z_0, \varepsilon_0)$ and $Z_{(n)}$.[6]

Student 0 chooses her major $(D_0 \in \{S, F\})$ as a function of expected success probabilities, P_0^S and P_0^F, according to equation (1). Because of continuously distributed unobserved preference

[6]We could have assumed that the information set of the agents is the distribution of the variables, Z_i. As the number of students is large, the empirical and true distribution function are close. We develop the framework that mimics the best our most convenient empirical procedure.

shocks, student 0 plays a pure strategy almost surely. This is her best response to the aggregate behavior of other students on which success probabilities depend. In this sense this is an aggregative game (Jensen, 2010) and we will make use of this characteristic.

After choosing one school, the two-stage exams are taken sequentially and students are selected in or out of each school by computing thresholds as functions of observed grades. 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 set $U_{(n)}$ 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 are the rational expectations of success of student 0.

2.2.2 Success probabilities and best responses

Denote $Z^S_{(n)}$ (respectively $Z^F_{(n)}$) the set of characteristics of the sub-sample of students $i = 1, ., n$ applying to Sobral (respectively Fortaleza) considered by student 0 when deriving her best response. By construction $Z_{(n)} = (Z^S_{(n)}, Z^F_{(n)})$. Similarly, we denote $U^S_{(n)}$ and $U^F_{(n)}$ the corresponding components 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

$$\mathbf{1}\{m_1(Z_0, u_0, \beta) \geq T^S_1(Z^S_{(n)}, U^S_{(n)}), m_2(Z_0, u_0, \beta) \geq T^S_2(Z^S_{(n)}, U^S_{(n)})\}$$

in which $T^d_1(.)$ and $T^d_2(.)$ are the values of the thresholds at the two-stage exams for a school $d \in \{S, F\}$ when the characteristics of applicants to this school are described by $Z^d_{(n)}$ and their grade shocks are equal to $U^d_{(n)}$. Notice that when evaluating this event, student 0 is considering only the sample of other students than herself. Because of continuously distributed grades, we can also neglect ties.

The formal construction of these thresholds is explained below after having determined choices but the intuition is clear for instance for the second-stage threshold. The best n_S ranked students after the final exam are accepted by Sobral and the threshold of the final exam is equal to the grade obtained by the worst-ranked accepted student. Respectively, at Fortaleza the success is determined by $\mathbf{1}\{m_1(Z_0, u_0, \beta) \geq T^F_1(Z^F_{(n)}, U^F_{(n)}), m_2(Z_0, u_0, \beta) \geq T^F_2(Z^F_{(n)}, U^F_{(n)})\}$. To distinguish

those counterfactual thresholds from the ones defined in the complete sample of $i = 1, ., n$ AND $i = 0$ we denote them as:

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

These thresholds are indexed by 0 since they refer to the thought experiment that student 0 performs when constructing her expectations as a function of characteristics and strategies 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 $U^d_{(n)}$ (remember that student 0 observes $Z_{(n)}$ only and conditions on $D_{(n)}$) and with respect to the individual source of risk, u_0:[7]

$$\begin{aligned} P^d_0(Z_0, \beta) &= E_{U^d_{(n)}, u_0}\left[\mathbf{1}\{m_1(Z_0, u_0, \beta) \geq \tilde{T}^d_{1,0}, m_2(Z_0, u_0, \beta) \geq \tilde{T}^d_{2,0}\} \mid Z_0, Z^d_{(n)}\right], \\ &= E_{U^d_{(n)}}\left[p^d(Z_0, \beta, \tilde{T}^d_{1,0}, \tilde{T}^d_{2,0}) \mid Z_0, Z^d_{(n)}\right], \end{aligned} \tag{2}$$

in which the following function results from integrating out the individual shock, u_0, only:

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

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

$$\begin{cases} P^S_0 = P^S(Z_0, Z^S_{(n)}, \beta) = E\left[p^S(Z_0, \beta, \tilde{T}^S_{1,0}, \tilde{T}^S_{2,0}) \mid Z_0, Z^S_{(n)}\right], \\ P^F_0 = P^F(Z_0, Z^F_{(n)}, \beta) = E\left[p^F(Z_0, \beta, \tilde{T}^F_{1,0}, \tilde{T}^F_{2,0}) \mid Z_0, Z^F_{(n)}\right]. \end{cases} \tag{4}$$

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 $Z^d_{(n)}$ only through $\tilde{T}^d_{j,0}$ that are computed below.

Denote $D_0(X_0, \varepsilon_0, \zeta, P^S_0, P^F_0) \in \{S, F\}$ the best response 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:

$$Z^S_{(n)} = \{i \in \{1, ., n\}; D_i(X_i, \varepsilon_i, \zeta, P^S_i, P^F_i) = S\}.$$

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

It is thus clear that the application mapping $Z_{(n)}$ into $Z^S_{(n)}$ or $Z^F_{(n)}$ is measurable although it remains to be shown that the application mapping $Z_{(n)}$ 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 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:

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

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.[8]

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 $Z_{(n)}$.

[8] There is a minor complication stemming from the fact that applicants could be in too small a number for one of the schools. In this case the threshold is defined in a trivial way as 0. The average success probability of 5% in our data means that the probability of this event is negligible.

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

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

As choices of other students $\mathbf{1}\{D_i = S\}$ are observed in the sample, 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.

Equations (1) and (6) are necessary conditions for a Nash equilibrium. A sketch of proof of the existence of a Nash equilibrium is spelt out in Appendix A and builds upon tools develop for potential games with weak strategic substitutes (Dubey, Haimanko and Zapechelnyuk, 2006).

3 The Econometric Model : Two stage grades and student preferences

We begin with specifying the two stage grade equations and reviewing 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

$$m_1 = Z\beta_1 + u_1, \tag{7}$$

$$m_2 = Z\beta_2 + u_2, \tag{8}$$

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.

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 postpone 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:

$$\begin{aligned} p^d(Z, \beta, t_1^d, t_2^d) &= Pr\{u_1 \geq m_1^d - Z\beta_1, u_2^* \geq m_2^d - \frac{2}{3}Z\beta_1 - Z\beta_2 - \frac{2}{3}u_1 - g(u_1)\}, \\ &= \int_{m_1^d - Z\beta_1}^{\infty} f_{u_1}(x)(Pr\{u_2^* \geq m_2^d - \frac{2}{3}Z\beta_1 - Z\beta_2 - \frac{2}{3}u_1 - g(u_1)\})dx, \\ &= \int_{m_1^d - Z\beta_1}^{\infty} f_{u_1}(x)[1 - F_{u_2^*}(m_2^d - \frac{2}{3}Z\beta_1 - Z\beta_2 - \frac{2}{3}x - g(x))]dx, \end{aligned} \tag{9}$$

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 preferences for majors and their assessment of the admission or success probabilities. As detailed in the previous section, we assume that students are sophisticated and can compute expected utility of the majors and choose whichever gives them 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 denoted $\delta^S = Pr\{V^S > 0,\, V^F \leq 0\}$. In this regime, 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. In both regions therefore, students reveal their true preferences and do not act strategically. Finally, 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. Those students may disguise their true preferences and act strategically. In this region, we can rewrite the decision model by taking the logarithm of equation (1):

$$\begin{cases} D = S & \text{if } \log(P^S) + \log(V^S) \geq \log(P^F) + \log(V^F), \\ D = F & \text{if } \log(P^S) + \log(V^S) < \log(P^F) + \log(V^F) \end{cases} \tag{10}$$

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[9]

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:

$$\Pr(D = S \mid \Delta(Z), X) = \delta^S(X) + \delta^{SF}(X) F(\Delta(Z) + X\gamma \mid X). \tag{11}$$

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.

[9] Agarwal and Somaini (2014) developed independently after us a proof of identification of preferences that relies also on the exogenous variation of expected probabilities in a more general setting with $n \geq 2$ schools. The proof we present considers in addition the existence of an outside option which creates regions in which expected probabilities have no influence. We also develop more fully the support conditions that are necessary in some regions of the preference space.

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,

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

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:

$$F(0 \mid X) = \frac{1}{2}. \tag{12}$$

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 (12) 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.

Our data exhibit limited variation and this is why we adopt a parametric assumption for $F(. \mid X)$. What non parametric identification arguments above prove is that this parametric assumption is a testable assumption at least in the support in which $\Delta(Z) + X\gamma$ varies.

3.3 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 (9) 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 compute the distribution function of $\tilde{T}_0$ at an arbitrary level of precision using the equilibrium

conditions (6) by simulation of $U_{(n)}$.[10] 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:

$$\hat{P}_{0,C}^d = \frac{1}{C}\sum_{c=1}^{C} p^d(Z_0, \hat{\beta}_n, \tilde{t}_{1,c}^d, \tilde{t}_{2,c}^d). \tag{13}$$

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

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

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

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

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 A Brief Description of Empirical Results

Our empirical analysis uses a sub-sample of applicants to Universidade Federal do Ceará. As the allocation mechanism determines choices *ex-ante*, we can indeed restrict the sample to two medical schools without modifying the argument developed in the economic model. All other majors are summarized by the outside option. In the rest of the analysis, we shall consider only students who take exams in two schools that are part of Medicine, the most competitive group of majors in the University.[11] We restrict the analysis because of the format of the vestibular. As students compete for one major only, it seems impossible to identify rank-ordered lists of schools as is typical in

[10] 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).

[11] See Sections S.1 and S.2 in the Supplementary Appendix which justify these arguments and complement the empirical analysis presented here.

school choice datasets collected in primary schools (for instance, He, 2014). Furthermore, the content of second stage exams is different if majors are far apart in terms of fields and this would introduce an additional dimension of heterogeneity.

We also focus on medical schools because of their attractiveness for the best students. Almost no students desist between the two stage exams if they pass the first stage. Being accepted in those schools is extremely valuable and the care and attention of students, parents and teachers are certainly at the highest for those two schools. They are respectively located in Sobral, the second most populated city in the state of Ceará and Fortaleza the capital city. The first school is small and offers 40 positions only while 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.

The list of variables and descriptive statistics in the pool of applicants to these two schools appear in Table 1. 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. 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.

Because our aim is to focus in the text on counterfactuals, the estimates of our empirical analysis are shown and analyzed in Section S.2 in the Supplementary Appendix. We report and comment there the estimates of grade equations, the predictions of success probabilities and the estimates of preference parameters. Nonetheless, it is in order to discuss briefly in the text our most important modelling choices and our main results.

As described in Table 1, explanatory variables are those that 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,[12] age, gender, educational history, repetitions, parents' education and the undertaking of a preparatory course. Our guidance for selecting variables is that a better fit of grade equations leads to a better prediction of success probabilities in the further steps of our empirical strategy.

Second, as developed in the identification section, Section 3.2.2, one exclusion restriction at least is needed to identify preferences. We chose to exclude from preference shifters all the variables 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 and undertaking a preparatory course). This is even more likely since we condition on ability m_0 measured after the variables describing educational history and which is assumed to be a sufficient statistics for what happened in the past. This dynamic exclusion restriction is akin to those which are assumed in panel data. 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.

Third, the second stage exam has a different format (writing essays) than the first stage multiple choice exam and the second stage grade equation has a much lower R^2. An interesting economic interpretation[13] is that the first stage exam is staged in order to skim out the weaker students and this multiple question exam is quite predictable (large R^2). In the second stage, the examiners can be selective in many more dimensions and try to select students on unobserved traits which are predictive of future behavior (success in the field of studies, drop out, etc) and that the econometrician cannot observe. This vindicates the double stage nature of the exam.

Fourth, Table 2 reports descriptive results on predicted probabilities of success. 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.

[12] 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.

[13] We thank Philipp Heller for suggesting this argument.

For instance, the population of students selected in the second stage exam for the school in Sobral is not the same as the population selected in the second stage exam for Fortaleza. 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.

Finally, students heavily favor Fortaleza over Sobral in their preferences and this confirms that Fortaleza is the most popular medical school in the state. 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%).[14]

5 Evaluation of the Impact of Changes of Mechanisms

We now investigate the impact of various changes of the existing mechanism.

The first counterfactual experiment that we implement is to cut slots proposed at the second-stage exam by 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 would list two ordered choices instead of one only. This means that even if students fail the first stage qualification in one of the two schools they may still get into the second stage exam for the other major. This implies that the average skill level of passing students increases although the difference between the two majors is attenuated.

Finally, since having two stages in the exam allows to cut costs and achieve a more in-depth selection at the second-stage, another experiment consists in changing the timing of choice-making. We allow students to choose their final major after taking the first-exam and learning their grades. This is likeky to generate more opportunistic behavior.

Before entering the details of these new mechanisms, the identification of utilities from estimated preferences and success probabilities is key in these evaluations. We show that expected

[14] Full details and comments of our empirical analysis appear in Section S.2.

utilities are underidentified and suggest how plausible bounds for counterfactual estimates can be constructed. We also explain how to compute counterfactual estimates conditional on observed choices.

5.1 Identifying Counterfactual Expected Utilities

Taking expectations with respect to grades using success probabilities P_i^S, P_i^F of ex-post utility levels, U_i, leads to:

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

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). This partial identification issue comes from the fact that ordinal preferences only are recovered in the case in which only one of the value function is positive and when both value functions are positive, relative cardinal utilities only can be identified.

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 normality of ε_i. 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. We assume 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 expectations require that the counterfactual random thresholds, $\tilde{T}_0$, as defined as the solution $(\tilde{t}_1^S, \tilde{t}_2^S, \tilde{t}_1^F, \tilde{t}_2^F)$ to the counterfactual counterpart of equation (6):

$$\begin{cases} \sum_{i=1}^{n}[\mathbf{1}\{\bar{D}_i(\bar{P}_i^S, \bar{P}_i^F) = S\}\mathbf{1}\{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^S\}] = \bar{n}_S, \\ \sum_{i=1}^{n}[\mathbf{1}\{\bar{D}_i(\bar{P}_i^S, \bar{P}_i^F) = F\}\mathbf{1}\{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^F\}] = \bar{n}_F, \\ \sum_{i=1}^{n}[\mathbf{1}\{\bar{D}_i(\bar{P}_i^S, \bar{P}_i^F) = S\}\mathbf{1}\{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^S, m_2(X_i, \beta, u_i) \geq \tilde{t}_2^S\}] = n_S, \\ \sum_{i=1}^{n}[\mathbf{1}\{\bar{D}_i(\bar{P}_i^S, \bar{P}_i^F) = F\}\mathbf{1}\{m_1(X_i, \beta, u_i) \geq \tilde{t}_1^F, m_2(X_i, \beta, u_i) \geq \tilde{t}_2^F\}] = n_F, \end{cases} \tag{14}$$

have a distribution function that leads to the counterparts of equation (13):[15]

$$\bar{P}_0^d = \mathbb{E}(\mathbf{1}\{m_1(X_0, \beta, u_0) \geq \tilde{t}_1^d, m_2(X_0, \beta, u_0) \geq \tilde{t}_2^d\}) \tag{15}$$

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

1. Initialization:

   - Draw C random vectors $\varepsilon_{(n),c}$ in their distributions conditional to observed choices, D_i, (see Supplementary Appendix S.3.1.1 for details). This implicitly means that choices D_i in the current mechanism whne computed using preference parameter estimates $\hat{\zeta}_n$ and these random shocks are set to their observed values. Fix those $\varepsilon_{(n),c}$ for the rest of the procedure. We use $C = 3000$.
   - Draw C random vectors $U_{(n),c}$ and fix them for the rest of the procedure.
   - Set the initial $P_0^{S,0}, P_0^{F,0}$ values at their simulated values $\hat{P}_{0,C}^d$ derived from equation (13) and using $U_{(n),c}$ in the observed experiment using equation (6).

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

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

[15] Changing the timing of choices requires to acknowledge that there are no choices to make before the first-stage. The first two equations in (14) 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.

(b) Compute a sequence of $\tilde{t}_c$ for $c = 1, ., C$ using $U_{(n),c}$ and equations (14).

(c) Derive $\hat{P}_{0,C}^{d,k+1}$ from equation (15).

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. We develop in Appendix A a simplified model with no covariates in which we show that a Nash equilibrium is obtained in a finite number of steps. The general case with covariates is more involved and we leave this for future research.

5.3 Cutting seats at the second stage exam

We start with the easiest policy change that reduces admission rates to the second stage. As said, the existing *Vestibular* system allows the number of students who take the second exam to be four times the number of final seats. In the experiment, the number of available 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 2. We explore the possible consequences of this policy and investigate two main issues – who among students benefit from this policy change and whether schools lose good students.

Some discussion about the expected effects are in order. Cutting seats in the second exam reduces schools' administrative costs although this 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 first stage best achievers 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 and the computation of expected utility follow the procedures described in Section 5.1 and Section 5.2.

5.3.1 Changes in thresholds

In Table 3 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. In the cutting seat experiment, the counterfactual first stage thresholds are much higher than in the original experiment since fewer students are admitted after the first stage exam. In contrast, the thresholds of the second stage exam are sligthly 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, estimated 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) would get a higher chance of admission and the worse students would have a lower chance. This is why we evaluate changes in success probabilities in relation to an index of students' abilities. As our ability index, we use the expected final grade which is, as already said, a combination of the initial, first and second stage grades. We also choose to focus 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 2 for Sobral. 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.[16] The dispersion of these changes is due to the heterogeneity of observed characteristics across students (conditional on expected grade).

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 1 that second stage grades have a much larger variance than first-stage

[16] The corresponding Figure for Fortaleza appears in Section S.3 of the Supplementary Appendix (Figure S.iv).

grades. The risk of failing is lower when fewer students participate in the second stage exam. In contrast, for students who are between the median and first stage threshold in terms of expected final grades, this is the converse. They are much less often admitted after the first stage exam and even if the second stage exam has less competition, it is the former negative effect that dominates overall. In particular, students who are around the first stage threshold in the current system suffer the most simply because they are more likely to be selected out at the first stage.

5.3.3 Changes in students' utilities and the impact on schools

Table 4 presents summaries of changes in students' expected utility. We construct groups according to various quantiles of the distribution of the expected final grade. The closer to the top of the distribution, the smaller the groups are (two percent of the population only). As defined in Section 5.1, we set the unknown weight in utilities at $\mu = 0.8$.

Consistently with changes in success probabilities, top students have significant utility improvements although this is also true for lower ranked students (above the 80% quantile). Nonetheless, focussing on means of expected utility hide very large dispersions in the 80-90% quantiles. This is best seen in the distribution of changes in utility (Supplementary Appendix S.1, Figure S.v) in which students in the 8th and 9th deciles are the ones whose changes in utility is the most dispersed. Furthermore, students just 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 less than the ones who opted for Sobral. Overall, these results about this counterfactual experiment bring out a significant total utilitarian welfare change. Yet, there are strong distributional effects and top students are better off and less able students are worse off.[17]

The impact on schools of cutting seats seems favourable 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

[17] We also performed a robustness analysis by using different values for the weight μ (see Section 5.1). Results are shown in Table S.viii in the Supplementary Appendix. 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 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.

Nonetheless, 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. Therefore, there are trade-offs between costs of organizing exams and the importance that schools would attach to the more informative second stage exam. The limitations of the data do not allow us to say more about these issues.

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 is the preferred major. Since our sample of interest only comprises students who positively value at least one of the majors, we have $d_1 \in \{S, F\}$. Yet, students can now provide a second choice and $d_2 \in \{\varnothing, S, F\}/\{d_1\}$ in which $d_2 = \varnothing$ is the null option chosen by students who do not value tehs econd major. 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 the list (S, F). If after the first-exam, she is above the threshold for school S, her second choice does not matter.[18] 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.[19] She fails altogether when her grades are lower than both thresholds.

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

[18] In particular, we discard the possibility of choosing a second ranked school after the second stage exam.

[19] See also the third experiment in which students choose according to the information they have on their performance at the first stage for a variation around these constraints.

- $m_1 \geq t_1^S$: 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 (see Supplementary Appendix S.3.2). 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

The new thresholds for this counterfactual experiment are also shown in Table 3. For the first stage, the threshold of Sobral is now slightly larger than the original one while the threshold of Fortaleza remains roughly unchanged. This is an indication that Sobral is admitting better students while the effect on Fortaleza is negligible. Some of the students who were failing Fortaleza before can now compete for Sobral and get admitted after the first stage. Furthermore, some of the students who were choosing Sobral for strategic reasons in the original mechanism can now at no risk choose Fortaleza first and Sobral second. Deferred acceptance mechanisms lessen strategic motives and make choices more truthful (Abdulkadiroglu and Sonmez, 2003). In the original system, students tended to choose Sobral as a "safety school" even when they truly preferred Fortaleza since success probabilities were higher at the former school. Giving students two choices attenuates the "safety school" effect although it does not eliminate it completely because of the two-stage nature of the exam. 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 S.vii.

Large standard errors for thresholds at the second stage exam make differences with the original ones insignificant. Yet, 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 as in the first stage.

5.4.3 Changes in success probabilities

Figure 3 reports changes in success probabilities for Sobral (see Figure S.vi for Fortaleza). Unlike the previous counterfactual experiment, the changes in Sobral and Fortaleza are now quite different. In Fortaleza, the change in success probabilities is negligible as thresholds are constant and the reallocation of choices from Sobral to Fortaleza not strong enough. In contrast, a larger fraction of students below the first admission threshold and above median seem to have a lower success probability in Sobral in the counterfactual experiment. This is because better students who fail Fortaleza switch to Sobral to compete with them and lower ranked students are evicted since first-stage thresholds are now higher in Sobral. In other words, getting Sobral if failing Fortaleza is acting as an insurance device and students just above the first stage threshold benefits from the existence of this insurance. Last, note that the change in success probabilities is small in this counterfactual compared with the previous cutting seat one since it affects the allocation of students only through the mechanism.

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 – 60% – will be (strictly) better off as shown in Table 5. 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. Deferred acceptance restricts less the possibilities of very top students since they can keep an outside option. In particular, students who preferred Sobral initially, benefit much more than those who preferred Fortaleza initially seemingly because the pressure of competition at the top in Sobral is lower since it loses its safety school status. In contrast, since Sobral has a lower threshold at the first stage exam, students

who prefer Sobral and are ranked around the first stage threshold suffer from more competition from evicted students from Fortaleza. However for those who prefered Fortaleza in the original system, 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 larger than the change in success probabilities.

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 students ranked around the first stage threshold in the original system and 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 Sobral 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 students when they choose majors **after** learning their first stage exam grade and not any longer before. Schools continue to admit students to the second stage exam according to the ranking given by the same combination of $ENEM$ and m_1.

The new selection procedure is a serial dictatorship mechanism. In the case of a single exam, it is shown in Balinski and Sonmez (1999) to be optimal. It 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 $d' \neq d$ or to opting out 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^F remain the same while this new mechanism affects the proba-

bilities of success $P^S_{m_1} = Pr\{m_2 > t^S_2|m_1\}$ and $P^F_{m_1} = Pr\{m_2 > t^F_2|m_1\}$ which are now conditional on the first-stage grade m_1. To define choices, suppose that $t^S_1 > t^F_1$ which means in practice that Sobral seats are filled in faster than Fortaleza's. A student can face three cases:

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

This algorithm is easily adapted to the case in which $t^S_1 < t^F_1$ prevails.

5.5.1 Changes in thresholds

The new thresholds in this counterfactual experiment are shown in Table 3. Sobral has now a slightly higher threshold at the first stage and a slightly higher threshold at the second stage exam while this is true only at the second stage for Fortaleza. The school in Fortaleza is overall more popular (see Table S.vii) and even more than 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 less likely to be filled earlier than Fortaleza's in spite of the one to four ratio (160/600). For instance, if more than 80% of the top 750 students prefer Fortaleza to Sobral, the 600 seats at Fortaleza would be filled in after those 750 students reveal their choices while Sobral would still have 10 seats to fill in. Note that such an argument can be very unstable and depend very much on revealed preferences and the first stage randomness in selecting the set of students who can go to the second stage.

5.5.2 Changes in success probabilities

Changes in success probabilities as shown in Figure 4, are becoming more disperse with clear changes in the mean probabilities. This is likely due to the fact that success probabilities now depend more on the first stage than before so that students performing well at the first stage increase their overall success probabilities while those performing worse have now lower success probabilities. As expected, the ex-post dispersion increases proportionally with the level of these

initial probabilities because of the increasing importance of the first-stage grade. These conclusions are true for Fortaleza (see Section S.3) as well.

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 mechanically larger than in the original system. Indeed, the probability of an increase in expected utility is quite large. This mechanism is mainly attractive for the top students as shown in Table 6. 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 were choosing Fortaleza in the original system would benefit less than those who prefered Sobral and this seems to be due to the difference in the sizes of the school through the mechanism exposed when we were analyzing the impact on thresholds. Sobral seats are filled less quickly than Fortaleza's.

Overall, this counterfactual seems more friendly to top students.

6 Conclusion

In this paper, we use data from entry exams and an allocation mechanism to college majors to provide an evaluation of other allocation mechanisms. 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 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 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 be 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 this new system.

Nonetheless, the previous mechanism allowed schools to tailor their selection procedures to the information they had about prerequisites for their courses and any predictors of success or drop out of the students they selected. This fine tuning is lost in the new centralized procedure which abstracts away the question of acquiring information that determine school preferences (Coles, Kushnir and Niederle, 2013).

By restricting congestion through the single choice of a major, schools were also reducing opportunistic behavior that might arise when using a centralized mechanism. The last counterfactual experiment is a good example of this since the fist stage exam allows students to strategize better their choices of schools without appropriate control for abilities of these students.

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, 2014). 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. This 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

Abizada, A.. and S. Chen, 2011, "The College Admission Promblem with Entrance Criterion", unpublished manuscript.

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

Agarwal, N., and P., Somaini, 2014, "Demand Analysis using Strategic Reports: An Application to a School Choice Mechanism", NBER WP 20775.

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.

Aygün, O., & Bo, I. 2014, "College Admission with Multidimensional Privileges: The Brazilian Affirmative Action Case",. mimeo..

Azevedo, E.M., and J.D., Leshno, 2014, "A Supply and Demand Framework for Two-Sided Matching Markets", unpublished manuscript.

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

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

Calsamiglia, C., C., Fu and M.Güell, 2014, "Structural Estimation of a Model of School Choices: the Boston Mechanism vs its alternatives", WP 2014-21, Universidad Autonoma, Barcelona.

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

Chade, H., Lewis, G., & Smith, L. (2014). "Student portfolios and the college admissions problem", *The Review of Economic Studies*, 81(3), 971-1002.

Che, Y.K., and Y. Koh, 2014, "Decentralized College Admissions", unpublished manuscript.

Chen, Y., & Kesten, O., 2014,."Chinese College Admissions and School Choice Reforms: Theory and Experiments", unpublished manuscript..

Coles, P., A. Kushnir and M. Niederle, 2013, "Preference Signaling in Matching Markets", *American Economic Journal: Microeconomics,* 2013, 5(2): 99–134

Dogan, M. K., & Yuret, T. (2013). "Publication Performance and Student Quality of Turkish Economics Departments/Türkiye'de Iktisat Bölümlerinin Yayin Performansi ve Ögrenci Kalitesi". *Sosyoekonomi*, (1), 71.

Dubey, P., O. Haimanko and A., Zapechelnyuk, 2006, "Strategic Complements and Substitutes and Potential Games", *Games and Economic Behavior*, 54:77-94.

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

Fack, G., J., Grenet and Y., He, 2015, "Estimating Preferences in School Choice Mechanisms", unpublished manuscript.

Fu C., 2014, "Equilibrium tuition, applications, admissions, and enrollment in the college market", *Journal of Political Economy*, 122(2), 225-281.

Gontijo, M.F., 2008, "Uma Aplicacão da Teoria dos Jogos ao Mercado do Vestibular Brasileiro", U. São Paulo.

Hafalir, I.E., R., Hakimov, D. Kübler and M. Kurino, 2014, "College Admissions with Entrance Exams: Centralized versus Decentralized, WP 2014-208, WZB, Berlin.

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., 2014, "Gaming the School Choice Mechanism in Beijing", unpublished manuscript, Toulouse School of Economics.

Jensen, M.K., 2010**, "**Aggregative games and best-reply potentials**",** *Economic Theory*, 43:45-66

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.

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.

Selim T., and S., Salem, 2009, "Student Placement in Egyptian Colleges", MPRA Paper No. 17596

Sönmez, T., & Ünver, M. U. (2011). "Matching, allocation, and exchange of discrete resources", *Handbook of Social Economics*, 1, 781-852.

Zhu, M., 2014, "College admissions in China: A mechanism design perspective", *China Economic Review* 30 (2014) 618–631

A Existence of a Nash equilibrium and convergence to an equilibrium

When using the current mechanism or counterfactual experiments, the question of the existence of a Nash equilibrium is pending. This equilibrium is defined as the solution to the best response equations (1) and success probabilities that are mutually compatible and compatible with the equilibrium conditions (5). We rely on the theory of pseudo potential games as developed in Dubey et al (2006)

In this discussion, we sketch the proof in a simpler game restricted to a single stage exam and imposing some weak conditions. The extension to two exams complicates notation but does not affect the intuition. Conditions (5) become:

$$\begin{aligned}\sum_{i=1}^{n}[\mathbf{1}\{D_i = S\}\mathbf{1}\{m_1(Z_i, u_i, \beta) \geq \tilde{t}_1^S] = 4n_S,\\ \sum_{i=1}^{n}[\mathbf{1}\{D_i = F\}\mathbf{1}\{m_1(Z_i, u_i, \beta) \geq \tilde{t}_1^F] = 4n_F.\end{aligned}$$

We will also assume that both schools are overdemanded by students who do not value positively both schools i.e.

$$\begin{aligned}\sum_{i=1}^{n}\mathbf{1}\{V_i^S > 0 \geq V_i^F\} > 4n_S,\\ \sum_{i=1}^{n}\mathbf{1}\{V_i^F > 0 \geq V_i^S] > 4n_F,\end{aligned} \tag{16}$$

so that thresholds in (6) are always defined by equalities.

Setting $\lambda_d(D_{(n)}) = \frac{4n_S}{\sum_{i=1}^{n}\mathbf{1}\{D_i=S\}}$, we can write an explicit definition of the thresholds as the empirical $(1-\lambda_d)-$quantile of the distribution of grades in the sample of applicants to d:

$$T_1^d(Z_{(n)}^d, U_{(n)}^d) = F^{-1}_{\{m_1(Z_i,u_i,\beta),D_i=d\}}(1-\lambda_d).$$

Note that the strategies of other students affect λ_d as well as the quantile so that expected success probabilities can be written as:

$$P_0^d(D_{(n)}) = \mathbb{E}(\mathbf{1}\{m_1(Z_0, u_i, \beta) \geq T_1^d(Z_{(n)}^d, U_{(n)}^d)\}).$$

It is easy to formulate deep assumptions about the distribution function of grades that imply that the success probabilities strictly decrease when adding an additional competitor to the set of applicants to d. Indeed, let order the strategy set $\{S, F\}$ as $S > F$. Extend the order to a partial order in strategies $D_{(n)}$ in the sample by positing that:

$$D_{(n)} > D'_{(n)} \text{ iff } D_i \geq D'_i \text{ and for at least one } i \ D_i > D'_i.$$

If the distribution of grade shocks is unbounded, adding competitors creates congestion and we have that:

$$D_{(n)} > D'_{(n)} \Longrightarrow P_0^S(D_{(n)}) < P_0^S(D'_{(n)}) \text{ and } P_0^F(D_{(n)}) > P_0^F(D'_{(n)}).$$

It is now straighforward to prove that the game satisfies the *dual strong single crossing property.* Suppose indeed that $V_0^S > 0$ and that:

$$P_0^S(D'_{(n)})V_0^S \leq P_0^F(D'_{(n)})V_0^F.$$

This implies

$$P_0^S(D_{(n)})V_0^S < P_0^S(D'_{(n)})V_0^S \leq P_0^F(D'_{(n)})V_0^F < P_0^F(D_{(n)})V_0^F.$$

This is also trivially satisfied when $V_0^S \leq 0$ and $V_0^F > 0$.

As this property of *dual strong single crossing* implies that this is a game of weak strategic substitutes with aggregation (Dubey et al, 2006) , it is a pseudo potential game (Theorem 1, p.81) and it has a Nash equilibrium (Proposition 1, p.84). Furthermore, since the strategy set is finite, there are no best response cycles in the game. "If players start with an arbitrary strategic profile and each player (one at a time) unilaterally deviates to his unique best reply then the process terminates in a Nash equilibrium after finitely many steps" (Remark 1, p.85)

Table 1: 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 2: 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 3: Thresholds of the Counterfactuals

	School		Sobral	Fortaleza
Stage 1	Original system	Mean Thresholds	184.48	189.88
		Standard Errors	(1.257)	(0.401)
	Cutting seats	Mean Thresholds	195.79	201.04
		Standard Errors	(0.996)	(0.506)
	Two-Choices	Mean Thresholds	186.98	190.13
		Standard error	(0.564)	(0.458)
	Timing-Change	Mean Thresholds	183.05	190.11
		Standard error	(0.859)	(0.447)

	School		Sobral	Fortaleza
Stage 2	Original system	Mean Thresholds	235.41	241.44
		Standard Errors	(1.669)	(0.898)
	Cutting seats	Mean Thresholds	233.34	237.77
		Standard Errors	(3.094)	(1.603)
	Two-Choices	Mean Thresholds	235.38	241.19
		Standard error	(2.589)	(1.302)
	Timing-Change	Mean Thresholds	239.07	244.30
		Standard error	(2.722)	(1.408)

[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 4: Cutting seats: Expected utility changes

Expected	ALL		D=Sobral		D=Fortaleza	
Final Grade	mean	s.d.	mean	s.d.	mean	s.d.
0% -50%	-0.00053	0.00094	-0.00068	0.00119	-0.00049	0.00087
50%-60%	-0.00441	0.00371	-0.00508	0.00438	-0.00418	0.00343
60%-70%	-0.00553	0.01176	-0.00430	0.01677	-0.00589	0.00983
70%-80%	0.00437	0.02200	0.00735	0.02674	0.00384	0.02106
80%-82%	0.01908	0.02556	0.03641	0.02165	0.01316	0.02427
82%-84%	0.03067	0.03307	0.06246	0.02099	0.02376	0.03123
84%-86%	0.04921	0.03228	0.09114	0.05772	0.04234	0.01951
86%-88%	0.05236	0.03565	0.09170	0.04433	0.04138	0.02352
88%-90%	0.06702	0.03518	0.12892	0.02309	0.05750	0.02580
90%-92%	0.06480	0.03649	0.13972	0.05993	0.05623	0.02087
92%-94%	0.09246	0.05419	0.17484	0.04767	0.07253	0.03263
94%-96%	0.09604	0.04910	0.24901	0.01415	0.08355	0.02213
96%-98%	0.12623	0.07049	0.31976	0.01292	0.10564	0.03147
98%-100%	0.17586	0.14463	0.43675	0.07173	0.10424	0.03388
$\mathbf{E}(\Delta U_i)$	0.01508		0.02851		0.01195	
$\mathbf{s.d.}(\Delta U_i)$	0.04815		0.08726		0.03216	
$\mathbf{Pr}(\Delta U_i > 0)$	0.3129		0.2915		0.3178	

[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 5: Two choices: Expected utility changes

Expected	ALL		D=Sobral		D=Fortaleza	
Final Grade	mean	s.d.	mean	s.d.	mean	s.d.
0% -50%	0.00014	0.00094	0.00018	0.00124	0.00014	0.00086
50% -60%	0.00295	0.00340	0.00282	0.00384	0.00300	0.00324
60% -70%	0.01020	0.00869	0.01232	0.00908	0.00957	0.00849
70% -80%	0.01913	0.01426	0.03370	0.01450	0.01653	0.01258
80% -82%	0.03901	0.01943	0.06275	0.00742	0.03090	0.01508
82% -84%	0.03246	0.01940	0.06150	0.00856	0.02615	0.01474
84% -86%	0.02987	0.03002	0.09661	0.01325	0.01895	0.01270
86% -88%	0.04578	0.02972	0.08945	0.01194	0.03360	0.02003
88% -90%	0.04533	0.03059	0.10819	0.01345	0.03566	0.01850
90% -92%	0.03725	0.03898	0.13392	0.02348	0.02620	0.02061
92% -94%	0.05836	0.04560	0.14043	0.01358	0.03851	0.02217
94% -96%	0.04575	0.04787	0.19825	0.01551	0.03330	0.01926
96% -98%	0.05615	0.06551	0.24437	0.01413	0.03612	0.02192
98% -100%	0.09306	0.13622	0.34371	0.06213	0.02425	0.01814
$\mathbf{E}(\Delta U_i)$	0.01321		0.03004		0.00928	
$\mathbf{s.d.}(\Delta U_i)$	0.03369		0.06785		0.01571	
$\mathbf{Pr}(\Delta U_i > 0)$	0.5999		0.6309		0.5926	

[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 4.

Table 6: Timing change: Expected utility changes

Expected	ALL		D=Sobral		D=Fortaleza	
Final Grade	mean	s.d.	mean	s.d.	mean	s.d.
0% -50%	0.00024	0.00084	0.00033	0.00094	0.00022	0.00081
50% -60%	0.00354	0.00290	0.00364	0.00274	0.00350	0.00296
60% -70%	0.01118	0.00823	0.01320	0.00759	0.01059	0.00834
70% -80%	0.02093	0.01396	0.03611	0.01273	0.01823	0.01237
80% -82%	0.04186	0.01917	0.06311	0.00531	0.03461	0.01659
82% -84%	0.03754	0.02117	0.07019	0.00628	0.03045	0.01585
84% -86%	0.03228	0.02916	0.09724	0.01069	0.02165	0.01251
86% -88%	0.05397	0.03023	0.09905	0.01604	0.04138	0.01910
88% -90%	0.05149	0.03143	0.11720	0.01238	0.04138	0.01845
90% -92%	0.04445	0.04059	0.14048	0.01318	0.03347	0.02484
92% -94%	0.07033	0.04987	0.15939	0.01221	0.04878	0.02534
94% -96%	0.06149	0.05029	0.21510	0.01133	0.04895	0.02460
96% -98%	0.07580	0.06961	0.27027	0.01618	0.05511	0.02830
98% -100%	0.13073	0.14474	0.39325	0.06788	0.05866	0.03198
$\mathbf{E}(\Delta U_i)$	0.01603		0.03326		0.01201	
$\mathbf{s.d.}(\Delta U_i)$	0.03879		0.07621		0.02045	
$\mathbf{Pr}(\Delta U_i > 0)$	0.6582		0.7196		0.6439	

[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 4

Figure 1: Choice space

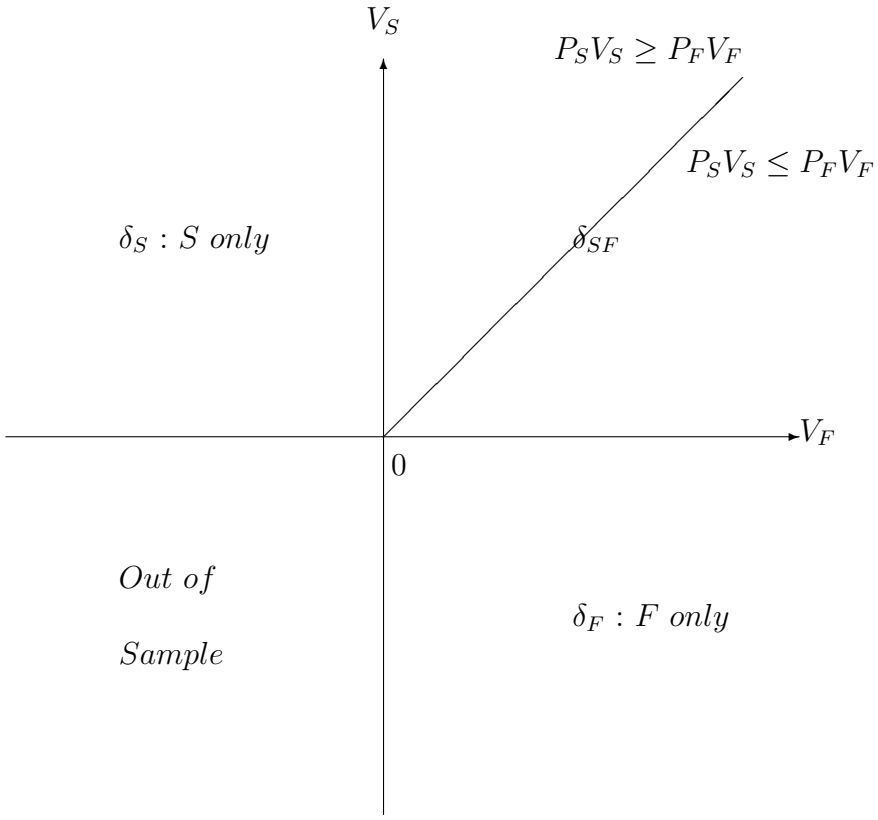
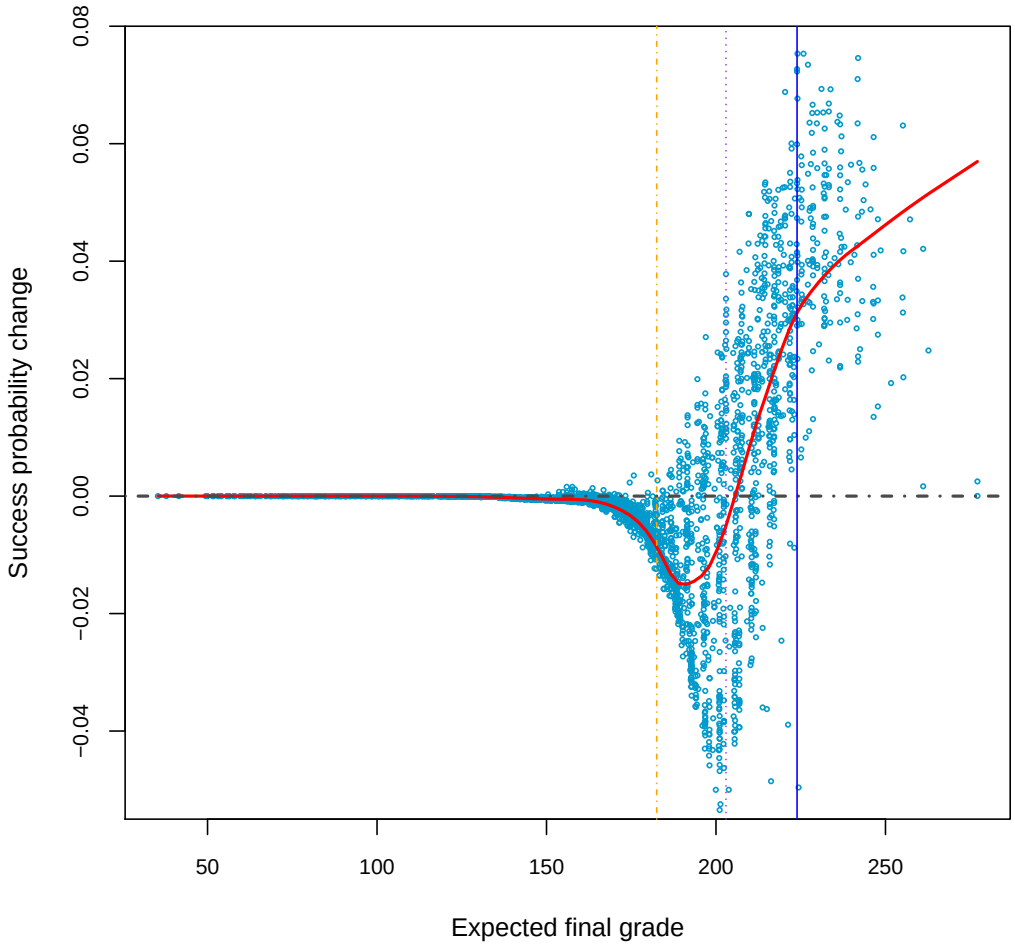
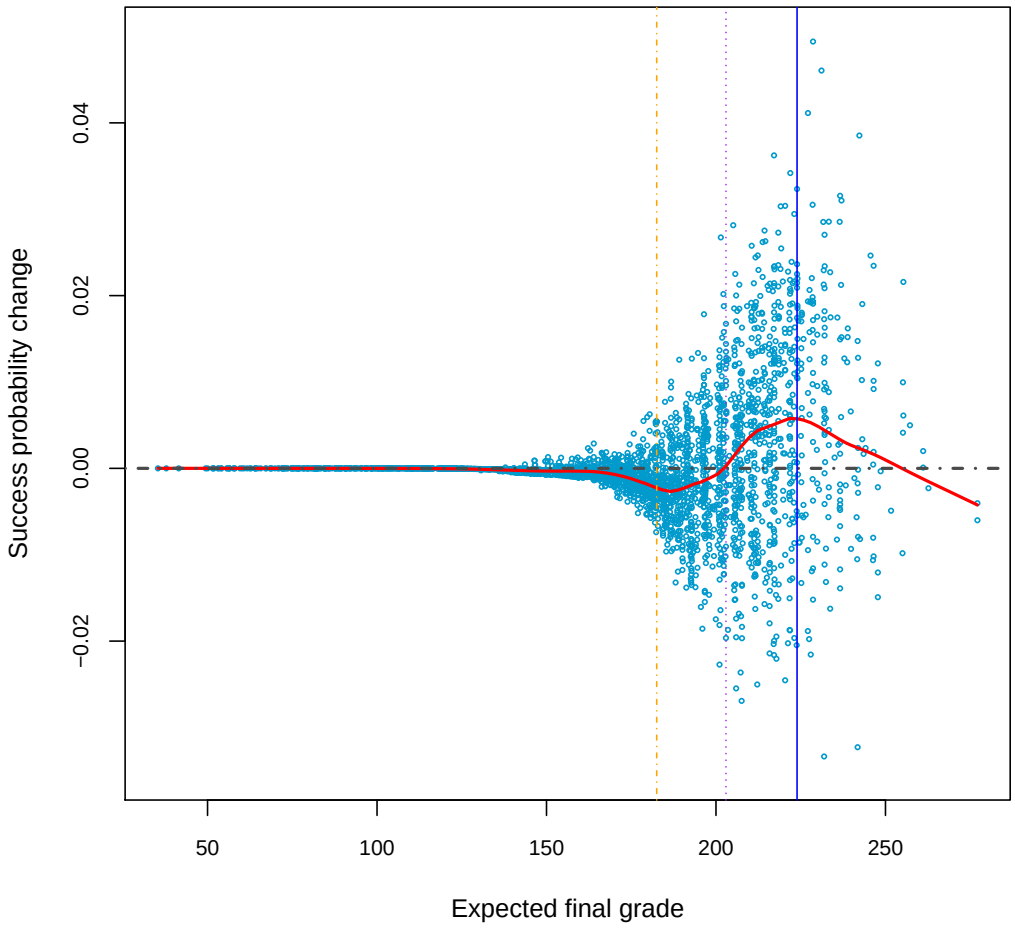


Figure 2: Cutting seats: Changes of success probabilities in Sobral



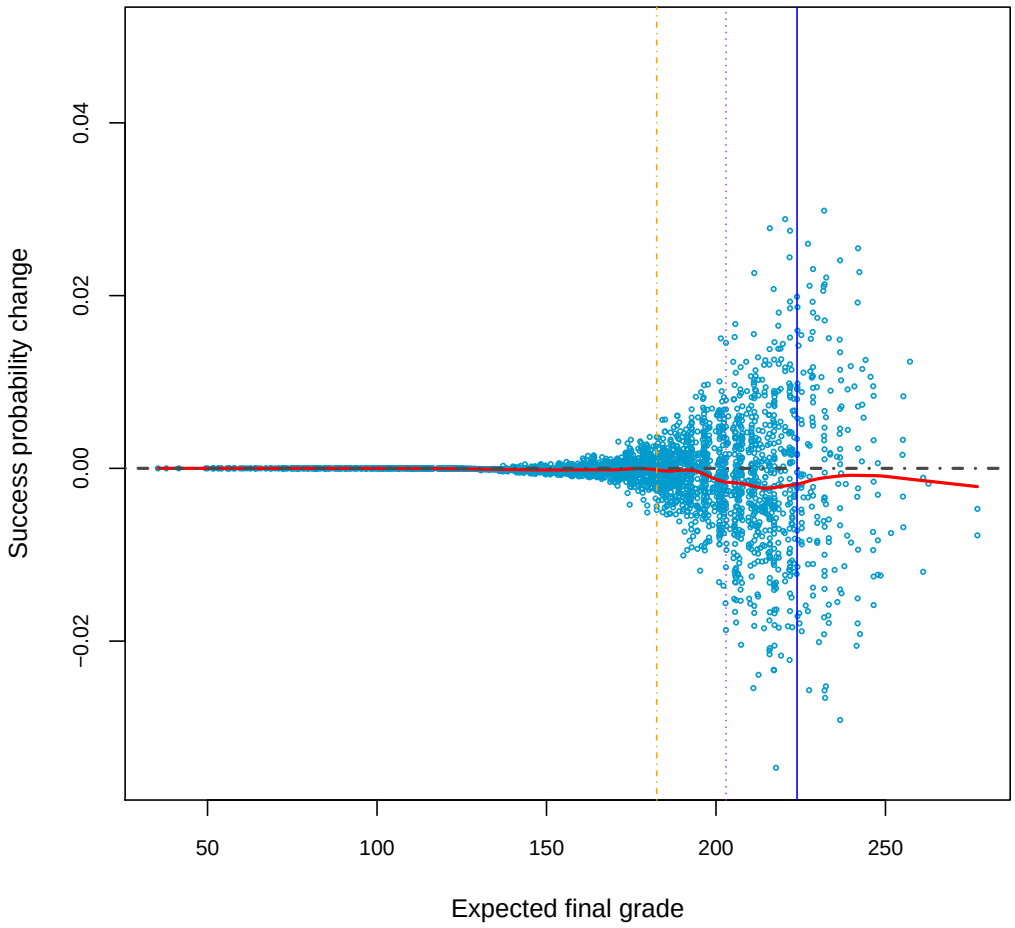
[1] The circles are individual success probability changes vs expected final grades; [2] From left to right, 1) the first vertical line is the median, 2) the second line is the quantile of 1st stage admission – $(1 - \frac{4(nos+nof)}{nobs}) \times 100\%$, and 3) the third line is the quantile of 2nd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



Notes: See notes of Figure 2

Figure 4: Timing change: Success probability changes in Sobral



Notes: See notes of Figure 2

Supplementary Appendix available upon request

S.1 Data appendix

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.

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

S.1.1 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 explicit constraints regarding the form and

content that universities must obey if they choose to select their students through a Vestibular. 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.

S.1.2 The Vestibular at UFC

The Vestibular at UFC shares the same features described above regarding its protocol. However, we give a detailed description of some of its feature in order to gain insight when developing and estimating econometrics models. First, all entrance exams in public universities must be preceded, by law, by the release of a document called Edital which contains the whole set of regulations regarding the exam: 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 63 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 $4N$.

The second stage 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.

S.1.3 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 Cearà. We grouped these majors into broad groups according to the type of second-stage exams that students take to access these majors. Table S.i 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 S.ii which reports summary statistics in each major field and the grades obtained at the first stage of the college

exam.[S.1] We report statistics on the distribution of the first stage grades in three samples:[S.2] 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.

This is why eventually, we chose to analyze only two medical schools in Sobral and Fortaleza.

S.2 Empirical Analysis: Estimates of Grade and Preference Equations

We present here the results of the estimation of grade equations, success probabilities and preferences.

S.2.1 Descriptive statistics

Table S.iii summarises the distribution of grades in teh two mediacl schools Sobral and Fortaleza in three samples: the complete sample, the sample of students who passed the first stage and the sample of students who pass the final stage. 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 Sobral. In conclusion, Fortaleza is more popular among students who apply to a medical school

[S.1] We do not report the second stage grades as they consist in grades in specific fields that are not necessarily comparable across majors.

[S.2] 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.

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.

There are also other interesting differences among applicants to the two schools regarding gender, age, private high school and preparatory course as appears in Table 1. 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.

Among explanatory variables, the initial grade obtained at the national exam $ENEM$ receives a special treatment. 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.

S.2.2 Estimates of grade equations

S.2.2.1 First stage exam

We report in Table S.iv 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 S.iv. 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 S.ii. 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 abilities and experience of applicants. In the second specification, we tested for the joint exclusion of parents' education 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.

S.2.2.2 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 S.v. 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 failed in the sense that the influence of effort never came out significantly. 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.

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 1% level (p-value of 3.4%). For the second grade equation, the test rejects homoskedasticity at the 1% 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 we checked that heteroskedasticity does not generate large differences in the prediction of success probabilities.

S.2.2.3 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 (13) and the integration in equation (9) 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.

In addition to summaries of predicted probabilities reported in the text in Table 2, we 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 S.vi. 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.

S.2.3 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 (11):

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

in which $\Phi(.)$ is the zero mean unit normal distribution[S.3] and the success probabilities P^d are to be replaced by their simulated predictions using grade equations (column 1 of Table S.iv and column 2 of Table S.v) as developed in the previous Section S.2.2.3. In the first part of Table S.vii, 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 S.vii 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 S.vii shows that the average probability of preferring Sobral alone (resp. Fortaleza alone) to the outside option is around 0.06 (respetively 0.55). These frequencies stay almost invariant across specifications. These results lead to what is commented in the text.

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).

[S.3] 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.

S.3 Complements to the Counterfactual Analysis

S.3.1 Simulated preferences conditional on observed choices

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:

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

S.3.1.1 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 (10). 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,

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

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

S.3.2 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:

$$U^{d_1d_2} = V^{d_1} \Pr\{m_1 > t_1^{d_1}, m_2 > t_2^{d_1}\} + V^{d_2} \Pr\{m_1 \in [t_1^{d_1}, t_1^{d_2}), m_2 > t_2^{d_2}\}$$

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:

$$\begin{cases} P^d = \Pr\{m_1 > t_1^d, m_2 > t_2^d\}, d = S,F \\ P^{d_1d_2} = \Pr\{m_1 \in [t_1^{d_1}, t_1^{d_2}), m_2 > t_2^{d_2}\}, (d_1,d_2) \in \{(S,F),(F,S)\}, \end{cases}$$

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

S.3.3 Additional Tables and Figures

Figure S.i 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.

Figure S.iii shows a picture of those odds ratios at all percentiles. We can visualize individual changes in expected utility in the cutting seat counterfactual in Figure S.v . Figure S.vi (respectively Figure S.viii) report changes in success probabilities for Fortaleza in the two choice experiment (resp. timing change). Changes in expected utility for the two choice experiment (resp. timing change) are graphed in Figure S.vii (resp. Figure S.ix).

Other references:

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.

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

Table S.i: 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 S.ii: 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 S.iii: 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 two majors: Sobral and Fortaleza)

Major	10th percentile	Min	Min	Median			Maximum			Observations
	All	Firststage	Pass	All	First stage	Pass	All	First stage	Pass	
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 S.iv: 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 S.v: 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 S.iv.

Table S.vi: 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 S.vii: 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 S.iv.

Table S.viii: Cutting seats: Robustness

Expected	$\mu = 0.8$		$\mu = 0$		$\mu = 1$	
Final Grade	mean	s.d.	mean	s.d.	mean	s.d.
0% -50%	-0.00053	0.00094	-0.00048	0.00087	-0.00054	0.00096
50%-60%	-0.00441	0.00371	-0.00400	0.00350	-0.00452	0.00376
60%-70%	-0.00553	0.01176	-0.00482	0.01123	-0.00571	0.01190
70%-80%	0.00437	0.02200	0.00479	0.02126	0.00427	0.02218
80%-82%	0.01908	0.02556	0.01923	0.02517	0.01904	0.02566
82%-84%	0.03067	0.03307	0.03041	0.03253	0.03074	0.03321
84%-86%	0.04921	0.03228	0.04880	0.03199	0.04931	0.03235
86%-88%	0.05236	0.03565	0.05201	0.03534	0.05244	0.03572
88%-90%	0.06702	0.03518	0.06634	0.03491	0.06719	0.03525
90%-92%	0.06480	0.03649	0.06434	0.03613	0.06491	0.03658
92%-94%	0.09246	0.05419	0.09165	0.05381	0.09267	0.05429
94%-96%	0.09604	0.04910	0.09526	0.04886	0.09624	0.04916
96%-98%	0.12623	0.07049	0.12532	0.07017	0.12646	0.07057
98%-100%	0.17586	0.14463	0.17519	0.14440	0.17602	0.14469
$\mathbf{E}(\Delta U_i)$						
	0.01508		0.01516		0.01506	
$\mathbf{s.d.}(\Delta U_i)$						
	0.04815		0.04778		0.04824	
$\mathbf{Pr}(\Delta U_i > 0)$						
	0.3129		0.3163		0.3114	

[1] Results as in Table 4 using different values of μ.
[2] See notes of Table 4

Figure S.i: Density plots of the grades

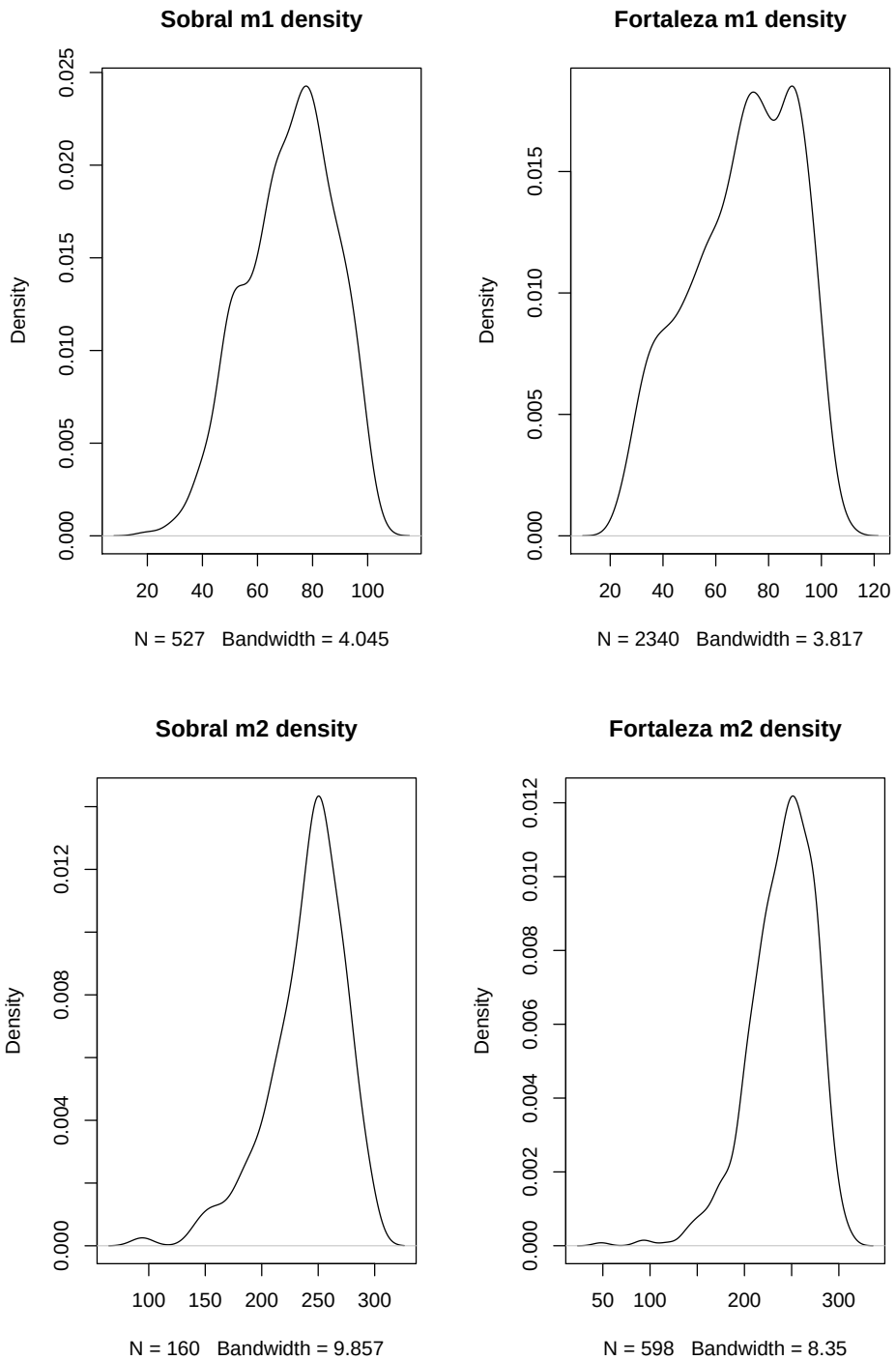
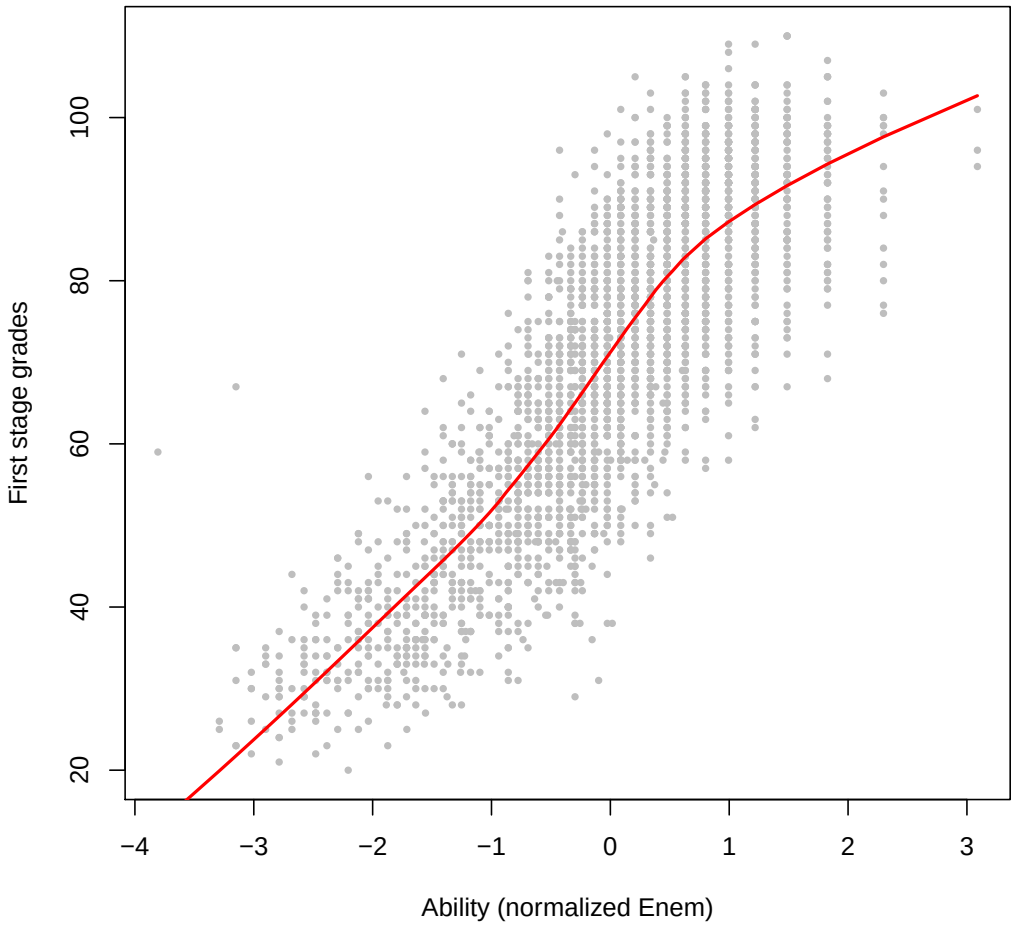
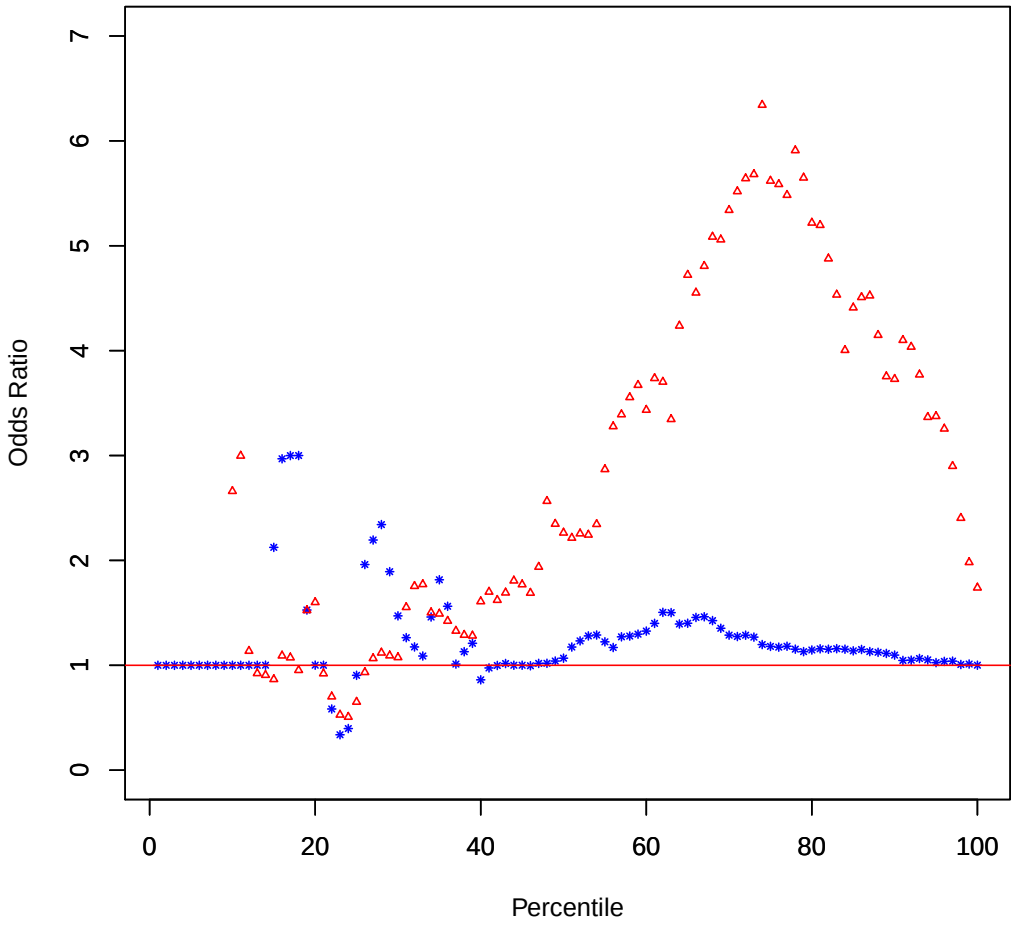


Figure S.ii: The relation between ability and first stage grades



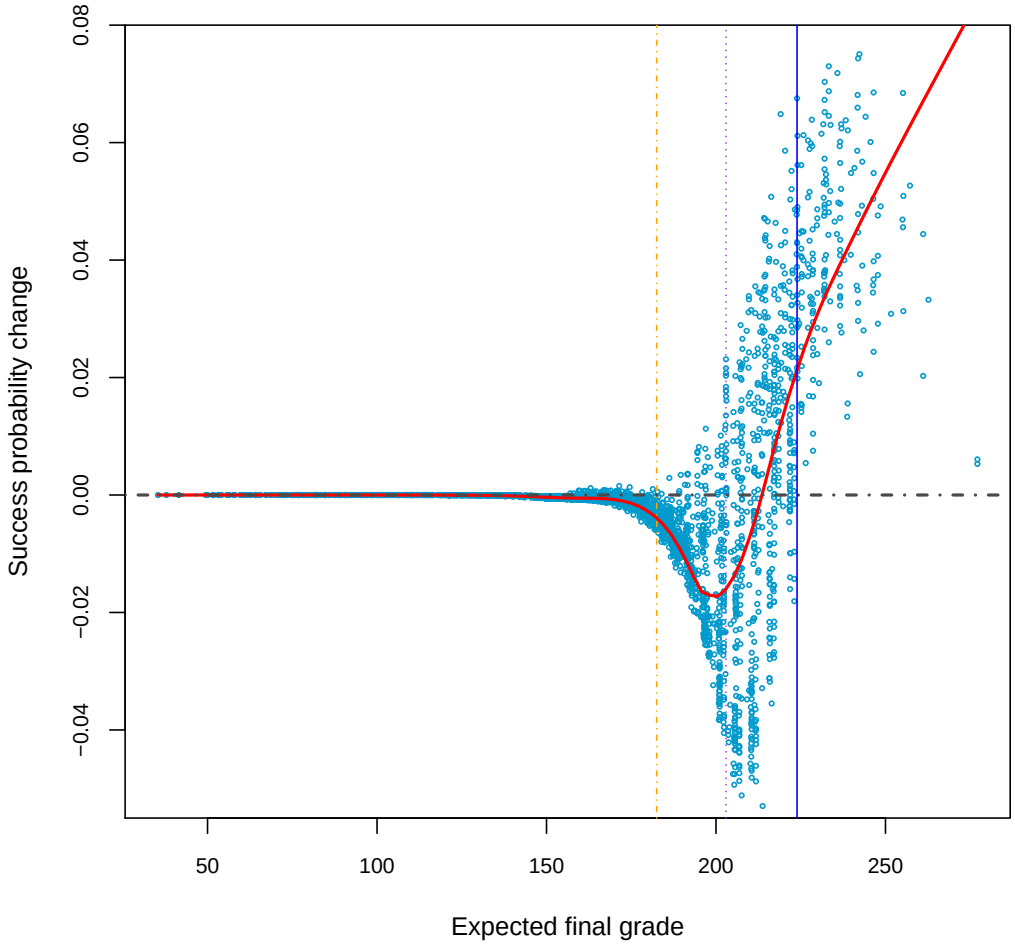
[1] The round grey points are the scatter plots of first stage grade on ability (normalized Enem); [2] The curve is the LOWESS curve of first stage grade on ability (normalized Enem).

Figure S.iii: The Odds ratio plot of simulated success probabilities



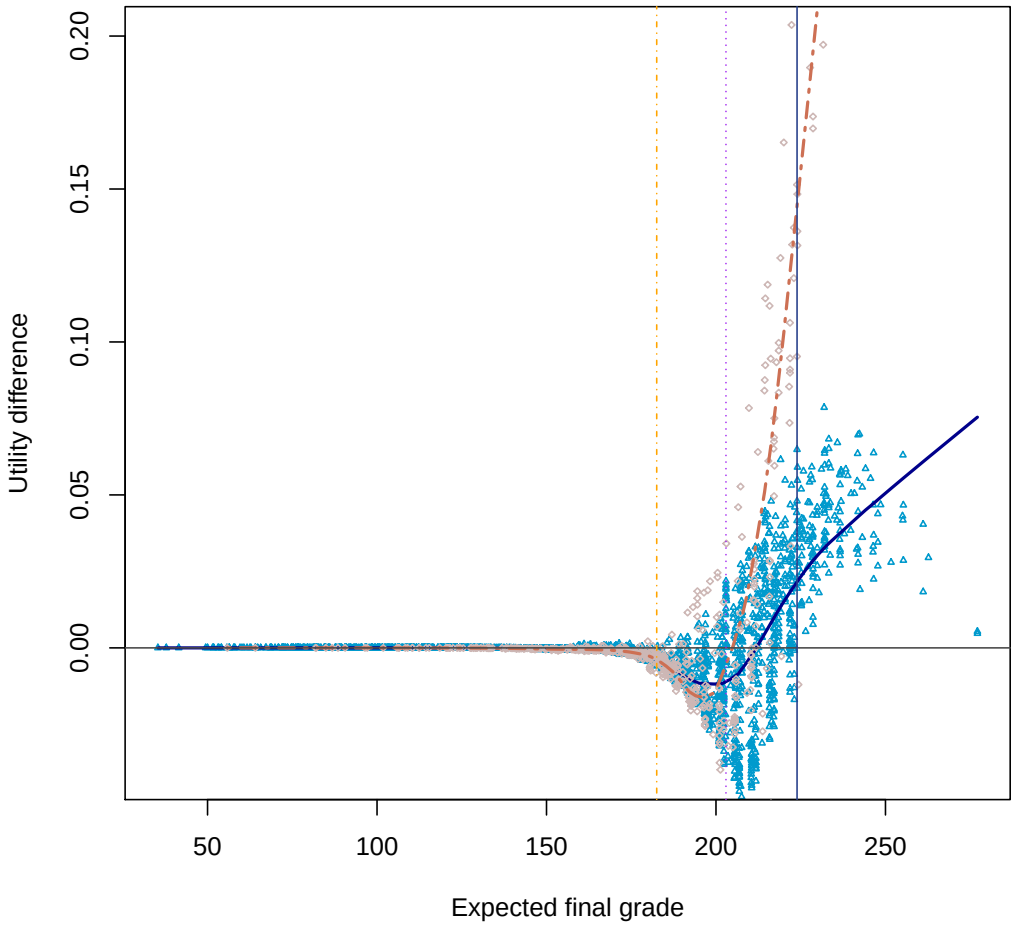
[1] The star points are odds ratio at the first stage ; [2] the triangular points are the odds ratio at the second stage; [3] percentiles are computed using first stage grades.

Figure S.iv: Cutting seats: Changes of success probabilities in Fortaleza



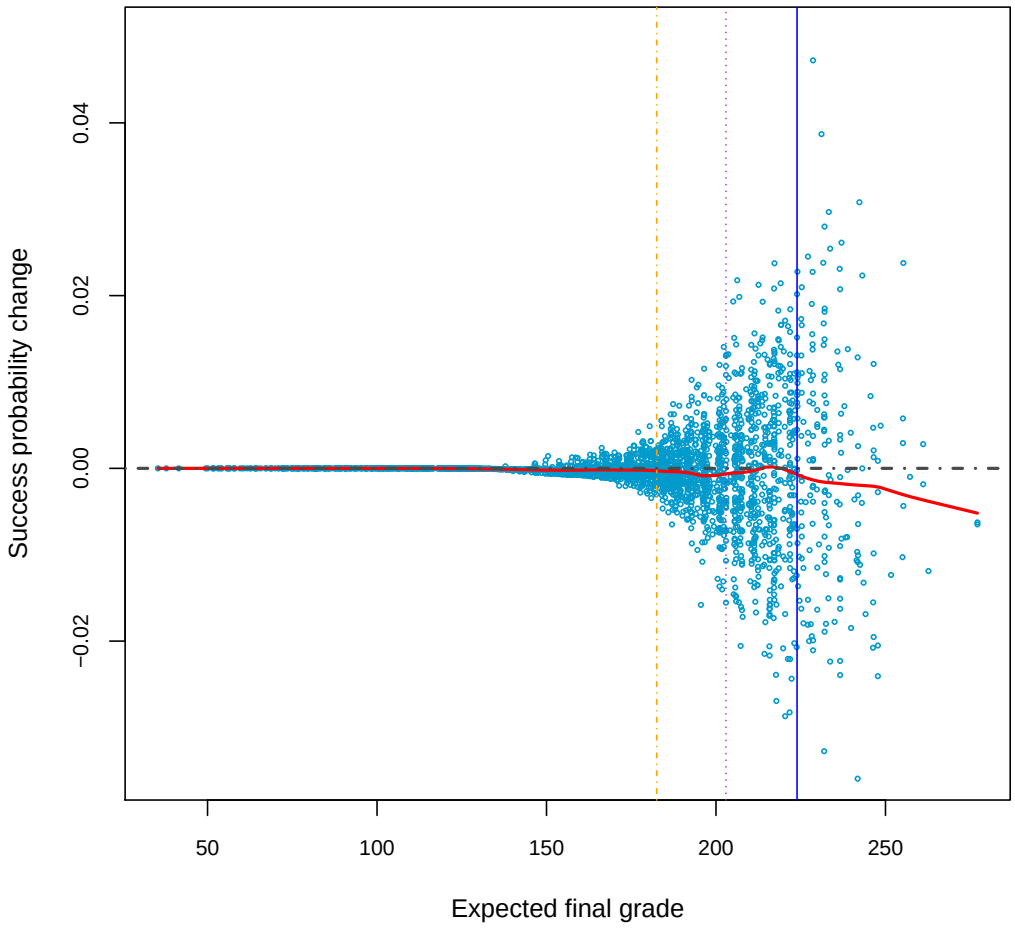
See notes of Figure 2

Figure S.v: Cutting seats: Expected utility changes



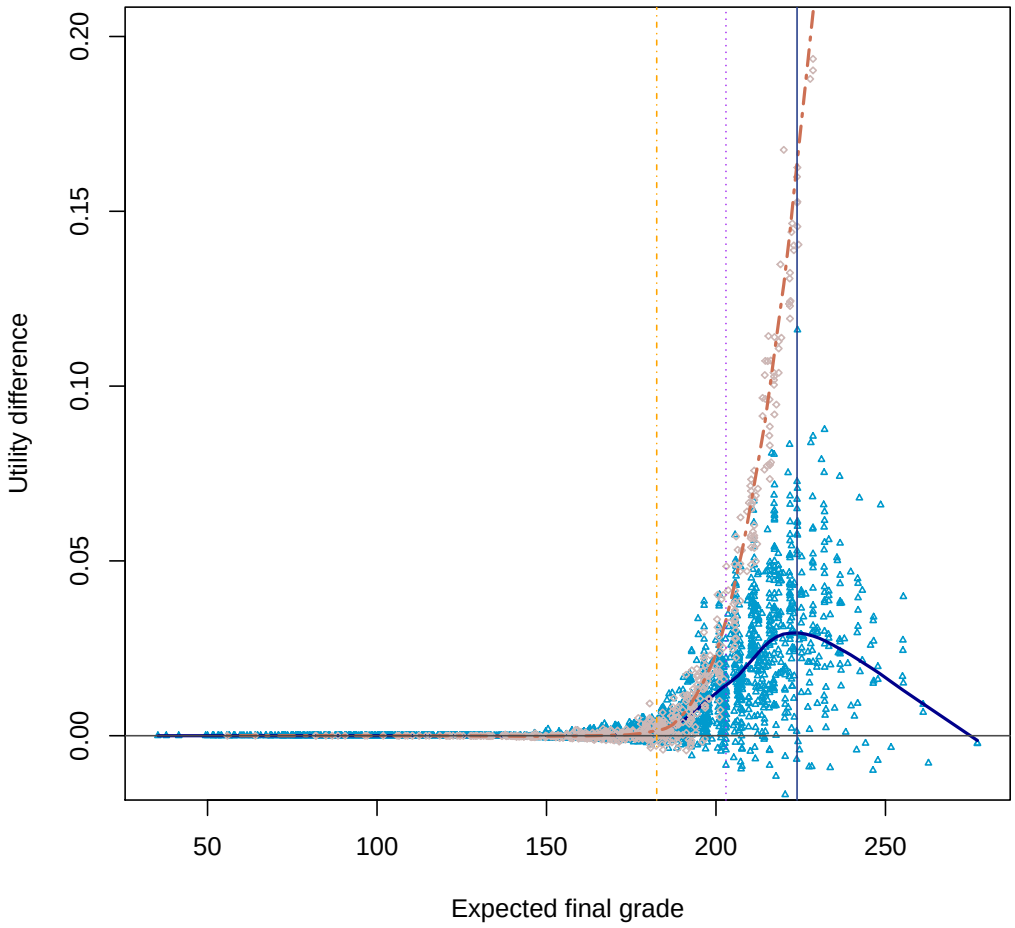
[1] the grey squares (resp. blue triangles) report changes in expected utilities and expected final grades for those who choose Sobral (resp. Fortaleza) in the original system. [2] the red line is the 0 level; [3] the vertical lines are as in Figure S.iv.

Figure S.vi: Two choices: Success probability change in Fortaleza



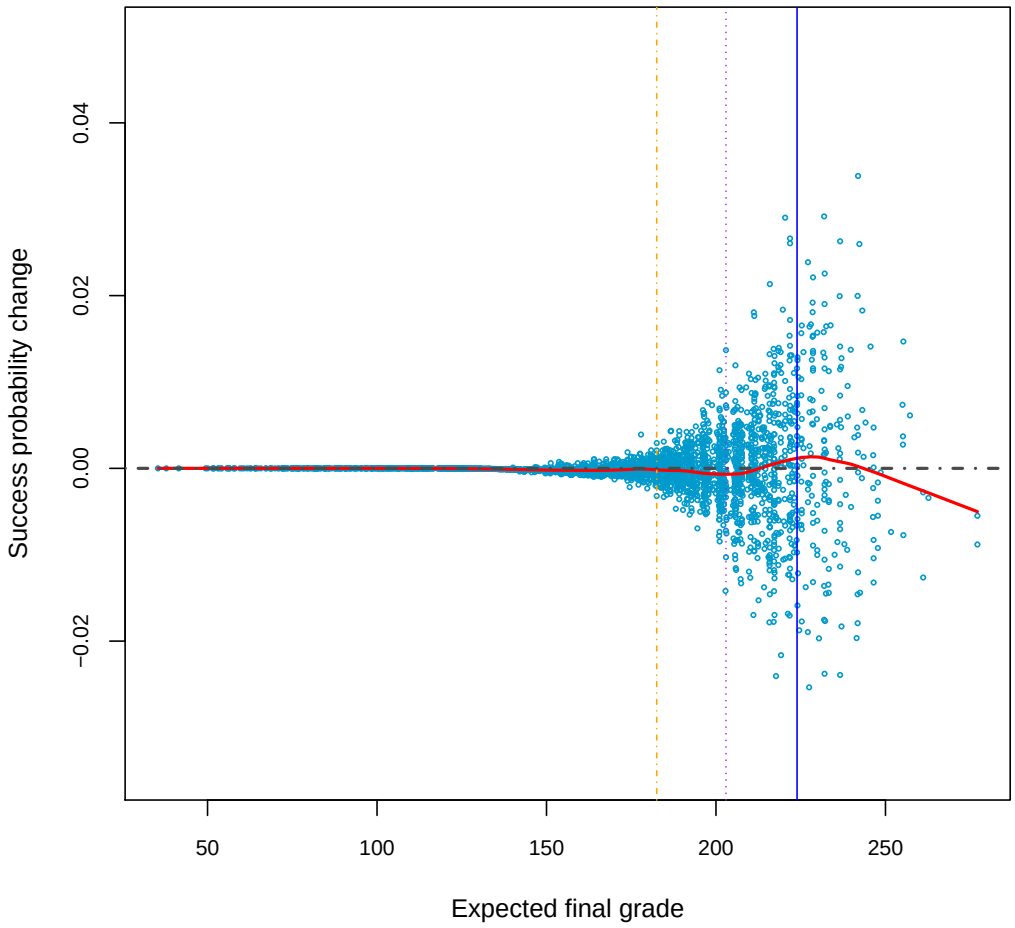
Notes: See notes of Figure 2

Figure S.vii: Two choices: Expected utility changes



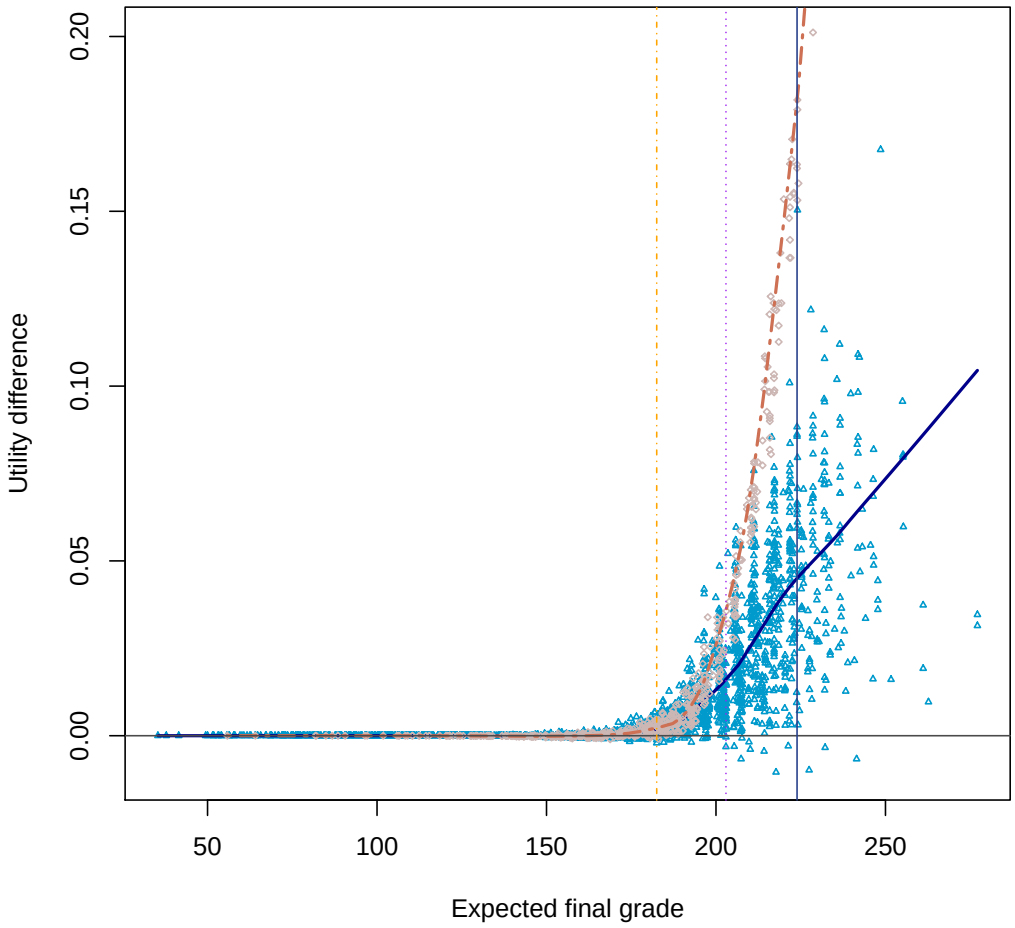
Notes: See notes of Figure S.v

Figure S.viii: Timing change: Success probability changes in Fortaleza



Notes: See notes of Figure 2

Figure S.ix: Timing change: Expected utility changes



Notes: See notes of Figure S.v