Supplementary Appendix available upon request

S.1 Data appendix

Matching students with university schools 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 schools, 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 schools 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 school. This threshold is calculated based on the following rule: Let N be the number of available places in a specific school previously shown in the Edital. Let r be defined as the ratio of the number of students choosing the school and the number of available seats in the school. 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 school. 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 schools

In total there are 58 schools that students may consider at Universidade Federal do Ceará. We grouped these schools into broad groups according to the type of second-stage exams that students take to access these schools. 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 school 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 school to enter as can be seen in Table S.ii which reports summary statistics in each school 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 schools. school fields are ranked according to the median grade among those who passed the final exam in that school 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 school 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 the two medical 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 (4) for each of the previous n_S set of simulators. We then replace the integration with respect to the thresholds as in equation (12) and the integration in equation (11) by summing over the set of n_S simulators. We experimented with different numbers of simulations to make sure that simulation error is negligible. This allows to compute simulated success probabilities for each student at both stages of the exam and in both schools.

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 2.3.3 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 2.3.3, 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 (7):

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

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 2.3.3, 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 (5). 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 schools, 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
	Min	0.022	0.021	0.050
δ^S	Mean	0.060	0.057	0.066
	Max	0.122	0.248	0.196
	Min	0.015	0.016	0.365
δ^{SF}	Mean	0.385	0.412	0.386
	Max	0.816	0.852	0.559
	Min	0.062	0.027	0.245
δ^F	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%	-0.00029	0.00116	-0.00026	0.00105	-0.00030	0.00119
50%-60%	0.00001	0.00744	0.00032	0.00697	-0.00007	0.00756
60%-70%	0.00674	0.01655	0.00715	0.01594	0.00664	0.01671
70%-80%	0.03122	0.02770	0.03143	0.02727	0.03117	0.02781
80%-82%	0.04070	0.02813	0.04074	0.02806	0.04069	0.02815
82%-84%	0.05491	0.02896	0.05478	0.02917	0.05494	0.02891
84%-86%	0.07304	0.03128	0.07302	0.03129	0.07305	0.03128
86%-88%	0.06124	0.03374	0.06117	0.03381	0.06126	0.03372
88%-90%	0.07932	0.03072	0.07917	0.03106	0.07936	0.03063
90%-92%	0.09239	0.03272	0.09230	0.03293	0.09241	0.03267
92%-94%	0.08806	0.04041	0.08779	0.04087	0.08812	0.04029
94%-96%	0.11009	0.03125	0.11000	0.03153	0.11011	0.03118
96%-98%	0.11178	0.03456	0.11163	0.03498	0.11181	0.03446
98%-100%	0.08939	0.04669	0.08917	0.04707	0.08945	0.04660
$\mathbf{E}(\Delta U_i)$						
	0.01966		0.01975		0.01964	
$\mathbf{s.d.}(\Delta U_i)$						
	0.03785		0.03775		0.03787	
$\mathbf{Pr}(\Delta U_i > 0)$						
	0.4363		0.4363		0.4363	

[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

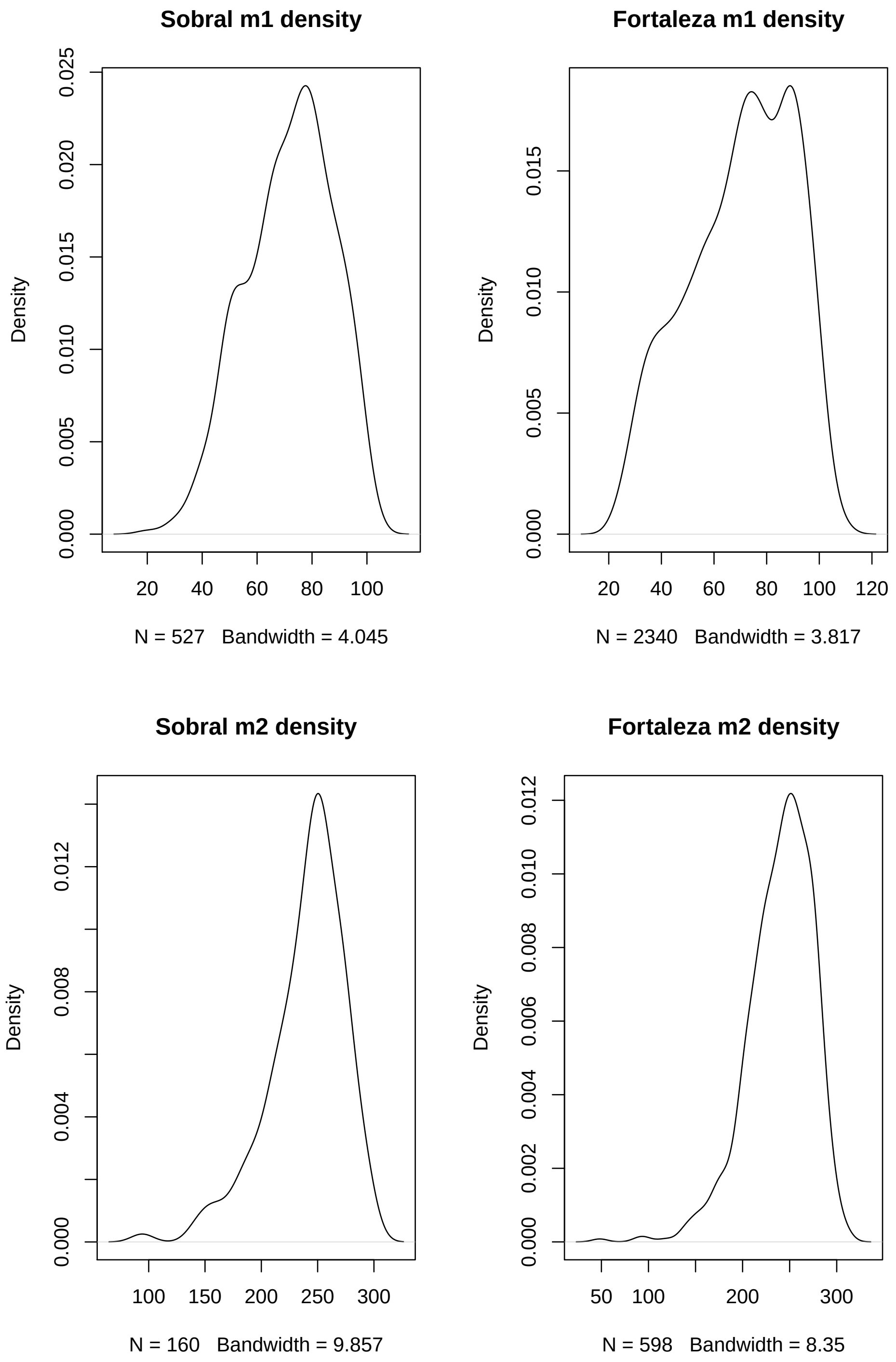
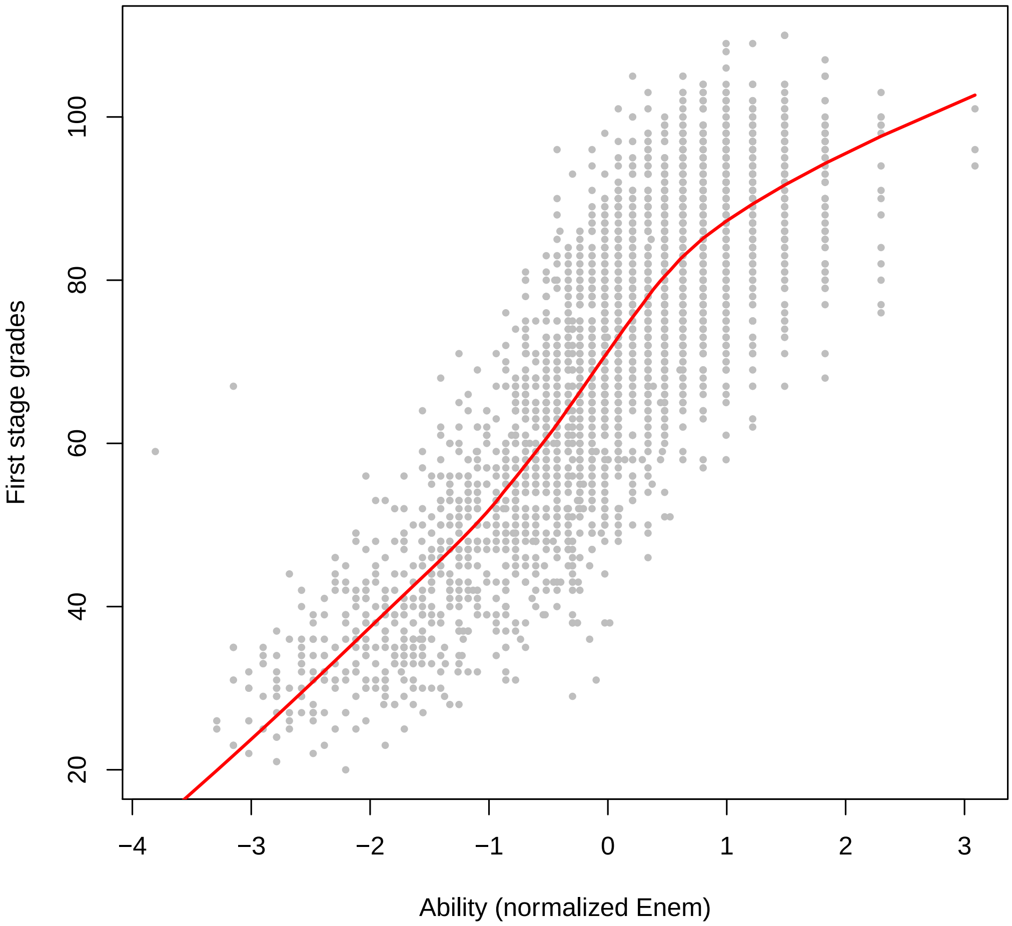
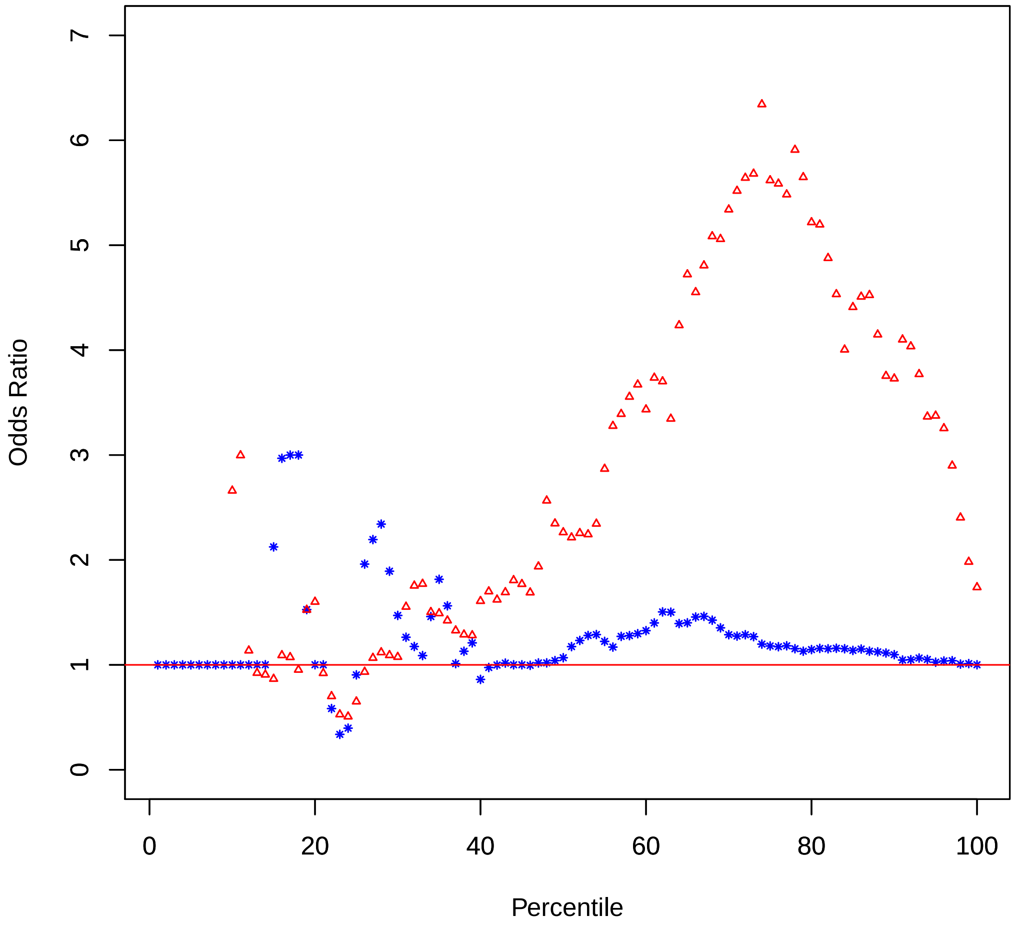


Figure S.ii: The relation between ability and first stage grades



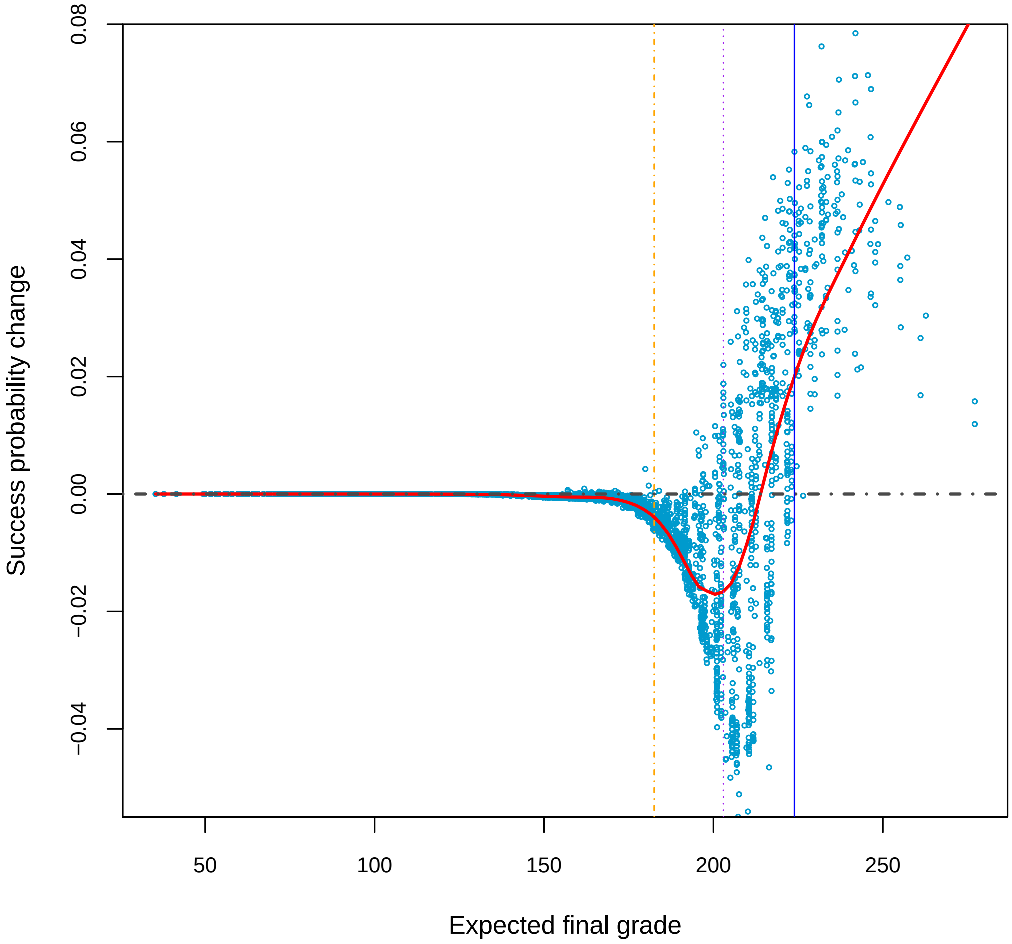
[1] The round grey points are the scatter plots of first stage grade on ability (normalized Enem); [2] The curve is the LOWESS curve of first stage grade on ability (normalized Enem).

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



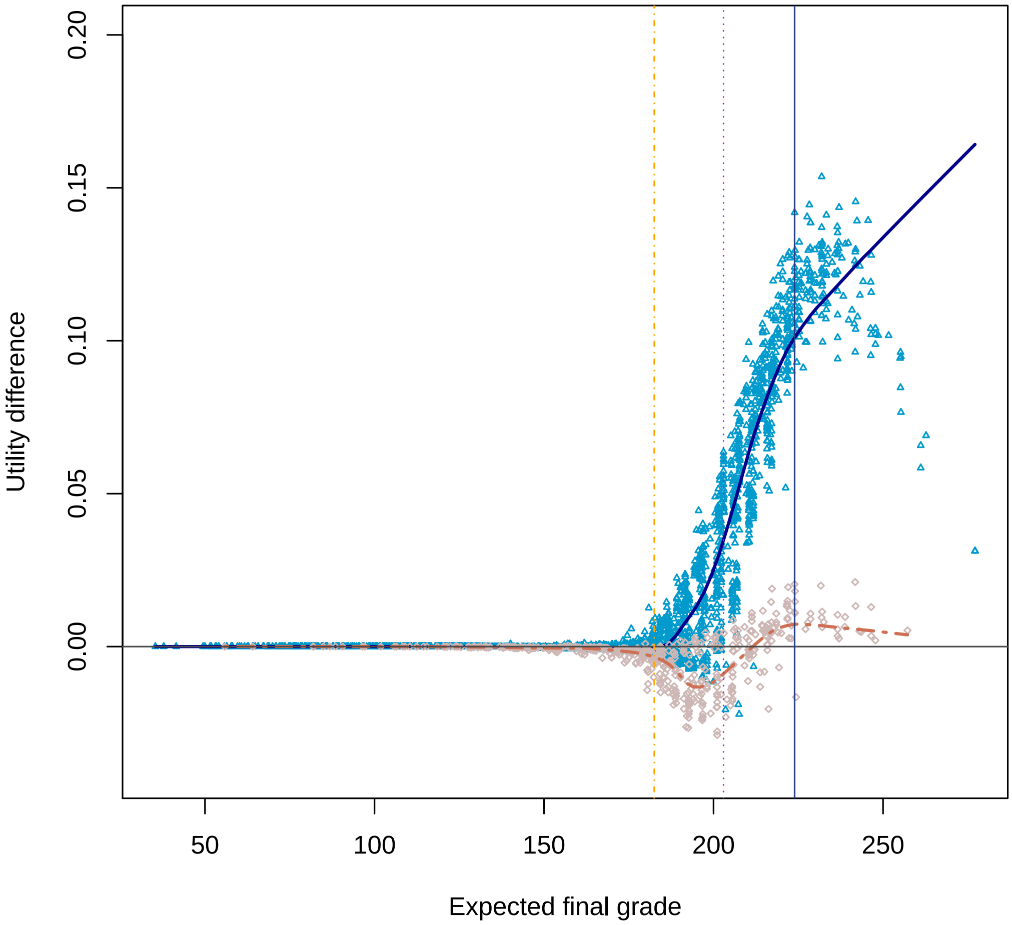
[1] The star points are odds ratio at the first stage ; [2] the triangular points are the odds ratio at the second stage; [3] percentiles are computed using first stage grades.

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



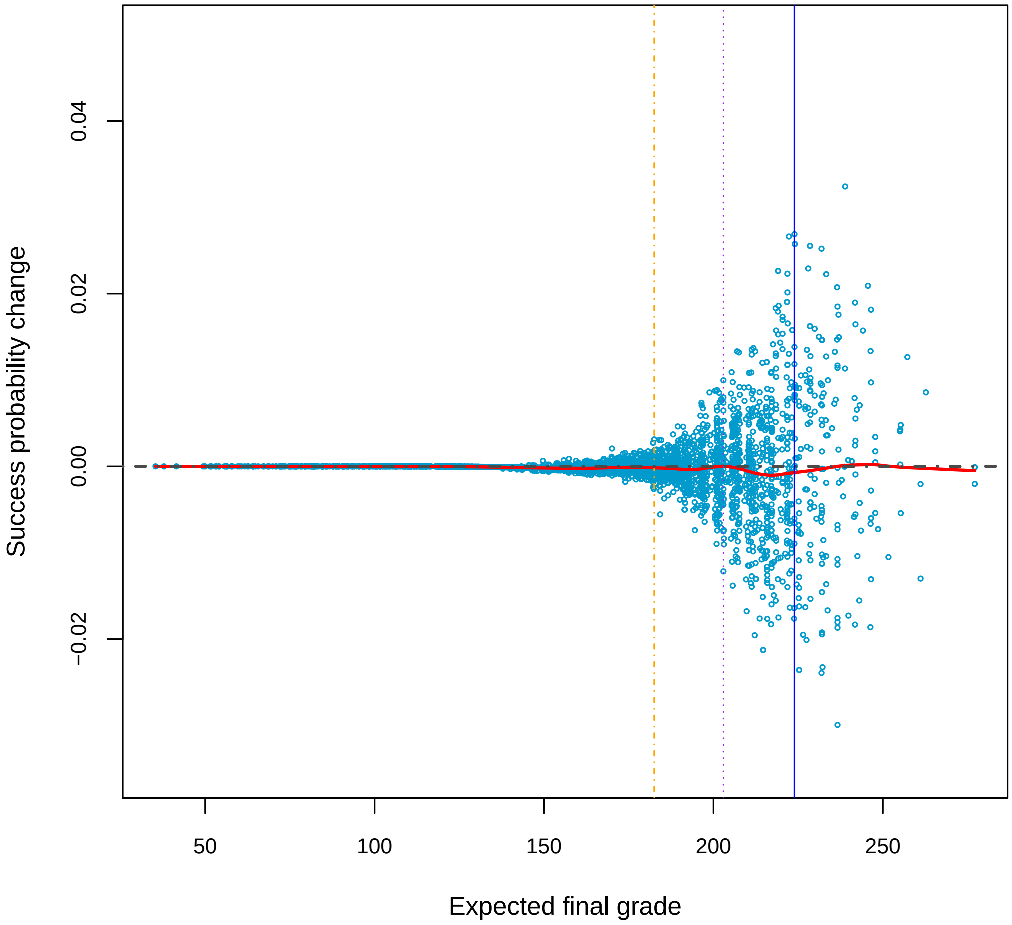
See notes of Figure 2

Figure S.v: Cutting seats: Expected utility changes



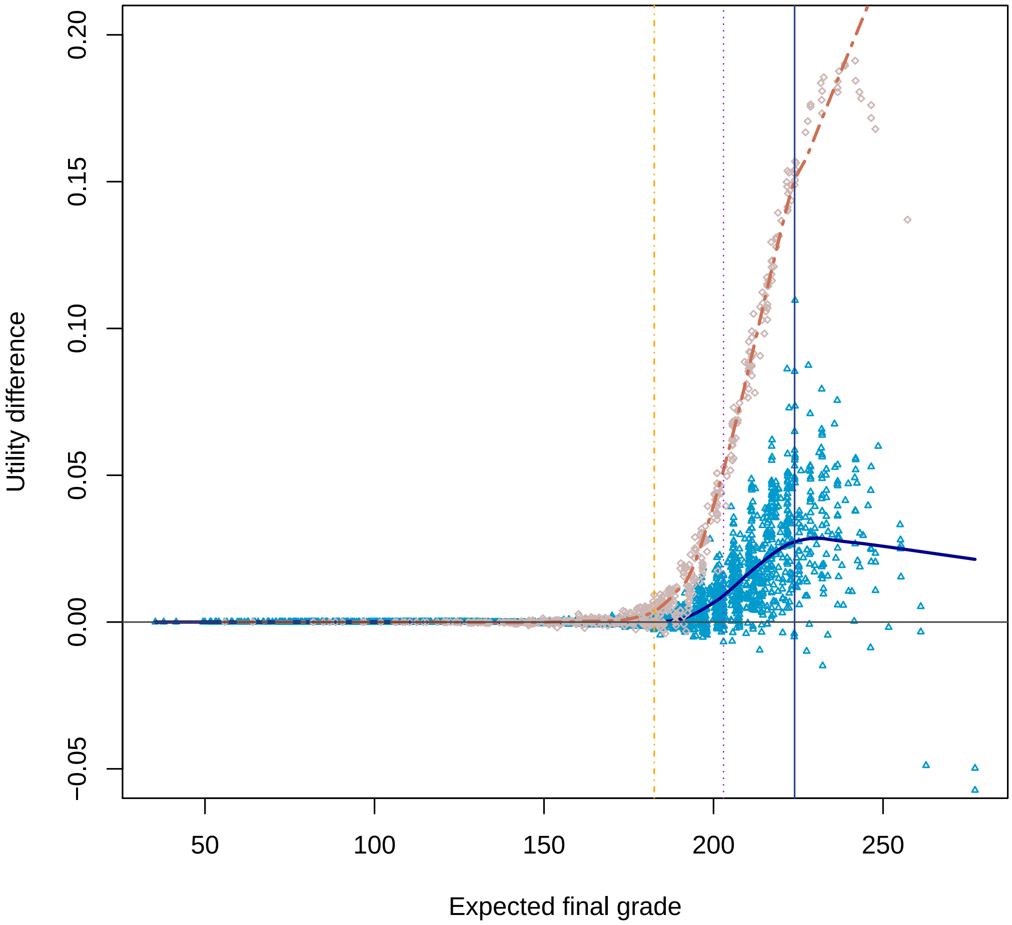
[1] the grey squares (resp. blue triangles) report changes in expected utilities and expected final grades for those who choose Sobral (resp. Fortaleza) in the original system. [2] the red line is the 0 level; [3] the vertical lines are as in Figure S.iv.

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



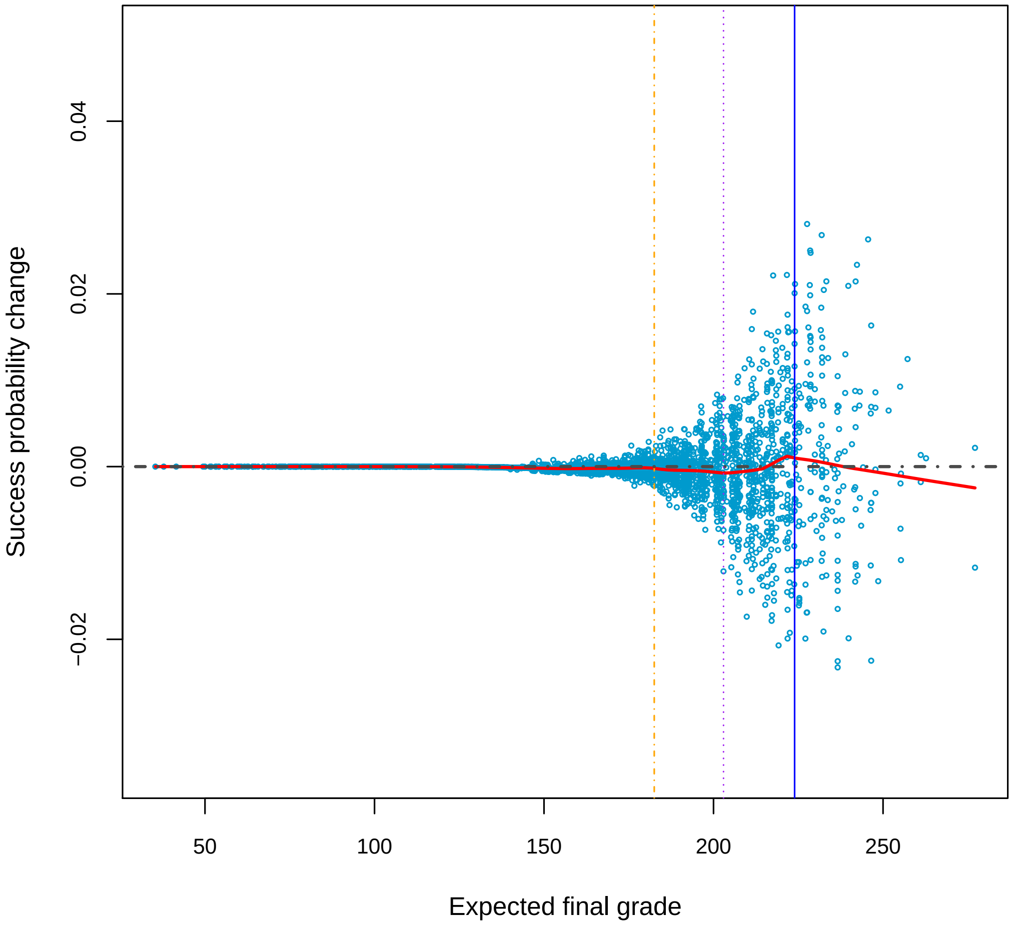
Notes: See notes of Figure 2

Figure S.vii: Two choices: Expected utility changes



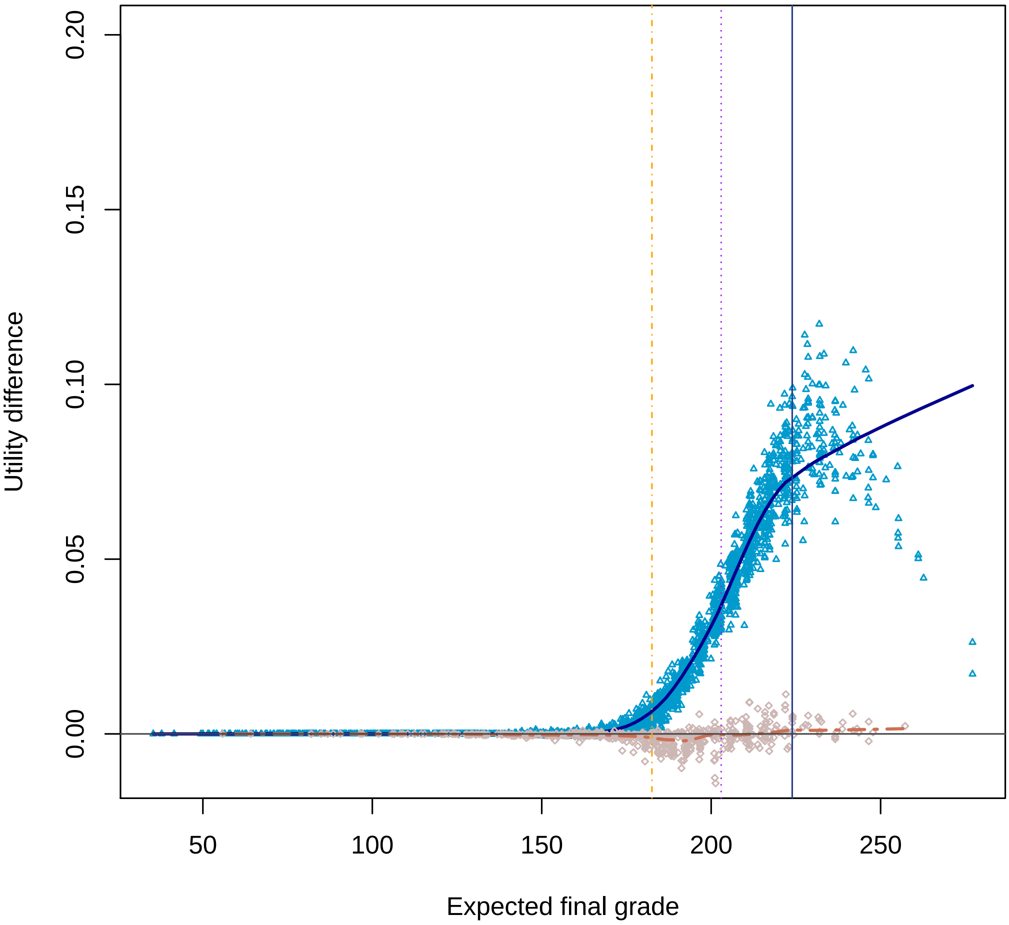
Notes: See notes of Figure S.v

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