College entry exams: A dynamic discrete choice model

Jose Raimundo Carvalho*and Thierry Magnac†

First version: July 2009
This version: April 30, 2010

Comments welcome

Abstract

A simple mechanism is used in some universities in Brazil to select students at entry and allocate them into various majors. Students first choose a major and then take exams that either select them in the chosen major or select them out. The matching literature analyzing the student placement issue, points out that this mechanism is not fair and is strategic. Pairs of major & students can be made better off and students tend to disguise their preferences. We build up a dynamic model of choice of major and of grading as well as effort exerted to be successful where preferences are carefully modelled. We estimate this model by simulated maximum likelihood using cross-section data about entry exams at Universidade Federal do Ceara in Brazil in 2004. Using the empirical results of the model estimation, we then evaluate changes in the way choices are given to the prospective students. Ex-ante expected utilitarian social welfare indeed increases but it hides very strong distributive effects among students. Strategic effects are found to be very strong.

*CAEN, Universidade Federal do Ceara, Fortaleza, Brazil.

†Toulouse School of Economics (GREMAQ & IDEI), Toulouse, France, magnac@cict.fr

1 Introduction[1]

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

In this paper, we use comprehensive data on the choices of majors by students and the grades that they obtain at the *vestibular* of the Universidade Federal do Ceará (UFC thereafter) in Northeast Brazil in 2004 and we concentrate on the specifics of this case. The main characteristics of that specific *vestibular* is that the student chooses a single undergraduate major before the exams and competes against those students who made the same choice only. Another interesting characteristics is that the exam consists in two stages. The first stage is common to all majors and is comprised of many sub-exams, each one evaluating knowledge in a definite subject, i.e. Mathematics, Portuguese etc.. The second stage is specific to each major and comprises two sub-exams.

What should be the optimal organization of the vestibular? This is known in the literature in economics as the college admission problem. This subject has a long history and a brief survey of the recent literature on one of the most popular solution is given in Roth (2008). The issue at hand is to match students with colleges which are in our case, the schools offering undergraduate majors at the university (medicine, engineering and so on). In the case where college preferences are simple[2] and consist in attracting the students who are the best in their discipline, it boils down to what is called student placement (Balinski and Sönmez, 1999) or one-sided matching. The

[1]We thank CNRS and CNPq for funding (Project 21207). Comments by participants at conferences in Brown, Bristol and Atlanta and seminars in Oxford and CREST, Paris are gratefully acknowledged. The usual disclaimer applies.

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

matching mechanisms that are studied are supposed to satisfy certain properties. First, they could be *stable*, or *fair* in the student placement literature, in the sense that there is no pair (student, college) who would like to block the final allocation in order to improve their lot by matching with another partner. Second, the mechanisms could be strategy proof i.e. every student has an interest to reveal her true preferences. Stable mechanisms are not unique and some of them are better for the students and others are better for the schools.

Among those mechanisms, the Gale Shapley student optimal stable mechanism (hereafter GS mechanism) satisfies the properties of stability and strategy-proofness and is optimal among those mechanisms from the student perspective (Abdulkadiroglu and Sonmez, 2003). In the GS mechanism, the first step consists in each student proposing to her preferred school. Each school tentatively ranks all its proposers (with respect to the possibly specific preference ranking the school has or using a lottery to break ties) and rejects the ones in surplus with respect to the number of seats the school has and which is fixed and publicly known ex-ante. In further steps of the algorithm, every student who was rejected in the previous steps proposes to her next choice. Each school ranks all its proposers (the ones from the previous step AND the new ones) and rejects the ones in surplus with respect to the number of seats. A student who was not rejected in the first step can well be rejected in further steps. This algorithm, also called a deferred acceptance algorithm, terminates when all new student proposals are rejected and nothing can be modified.

We can then compare the *vestibular* at UFC to the Gale Shapley mechanism. As a matter of fact it turns out that the *vestibular* at UFC corresponds to step 1 only of the GS algorithm. Students are allowed to propose to their first choice only. This is why the mechanism loses its two properties: it is not stable (i.e. there exist pairs of student & school which could be made better off by changing the final allocation) and it is not strategy proof. Some students prefer to disguise their preferences for very demanded disciplines (medicine, law,...) in order to improve their probability of being accepted. The present form under which the vestibular is organized at UFC is thus difficult to justify.

Nevertheless, recent research concerned with comparing the result of the GS mechanism with the so-called Boston mechanism questions the use of the former even if it is in a different context. The Boston mechanism consists for the schools in accepting all the students who propose and are ranked first instead of deferring the decision until all the iterations have finished as in GS. This

mechanism is known to be not stable and actors play strategically and disguise their preferences. This recent research questions that the Gale Shapley mechanism leads to a Pareto solution in terms of ex-ante utilities (Abdulkadiroğlu, Che and Yasuda, 2008 in a school choice problem, Budish and Cantillon, 2010 in a multi-unit assignment problem). Those papers exhibit other mechanisms that are not stable and not strategyproof although they are dominating the GS mechanism in terms of ex-ante utility. the main intuition for this result is that the latter mechanism does not allow applicants to reveal the intensity of their preferences, just the ranking of them.

What we do in this paper is to contribute to teh empirical literature on this subject by evaluating this *vestibular* using some counterfactual mechanisms. We start by constructing a dynamic model of choice of majors following the literature about choices of colleges (Arcidiacono, 2006 and Bourdabat and Montmarquette, 2007, for instance). In contrast to this literature though, we cannot use information on wages after school and we model them as undistinguishable from preferences. Choice probabilities of majors thus depend on (1) expected probabilities of success and (2) preferences for the majors.

In addition, we shall consider that exams have both a dimension of selection of the most talented (although the selection is imperfect) but also of those who exert more effort. The tournament literature indeed insists on the double dimension of selection and incentives that exams, or tournaments, have and distinguishing between them is one of the substantive issues studied in applied research (Davies and Stoian, 2007 or Leuven, Osterbeek, Sonnemans and van der Klaauw, 2008). It is also interesting to include effort since it allows students to reveal the degree of preferences that they have for different majors since higher preferences lead to higher effort and thus increases their probabilities of success.

The advantage of these data is that we can carefully model the probability of success at entry of each school using data on performances that we have i.e. the grades at the two stages of the exam as well as an initial measure of talent obtained a year before the exam is taken. We adopt the assumption that expectations are perfect (see Manski, 1992, for a critical evaluation of this assumption) and thus that players are sophisticated. The thresholds above which students are accepted into the programs are the results of the Nash equilibrium of this game in which beliefs are given by what is observed in the data and in which each player is assumed to be small.

We estimate this model of performances, preferences and effort using data made available to us

by UFC that we restrict, for simplicity, to the choice process into three majors in medicine, the most competitive major field. We use parametric models for the success probabilities and preferences although we study non parametric identification of the model. Using these empirical results, we then construct counterfactuals by recomputing the Nash equilibrium under the counterfactual mechanism. Specifically, we analyze the mechanism which allows students to have two choices instead of one and thus play less strategically. We show that indeed, enlarging the choice set has a positive aggregate effect in terms of utilitarist social welfare but has also strong distributive effects. The strategic effects are shown to be very important.

This paper builds upon various literatures and in particular student placement. There are a few papers concerned by the analysis of school choice (Lai, Sadoulet and de Janvry, 2009) and the Boston mechanisms using Chinese data (He, 2009) or the GS mechanism using US data (Abdulkadiroğlu, Pathak and Roth, 2009). In a more theoretical work but oriented towards the analysis of a specific mechanism, Balinski and Sönmez (1999) study the optimality of the placement of students in Turkish universities although the selection there concerns all students & colleges throughout the country. Students first write exams in various disciplines and scores are constructed by each college. Colleges choose the weight that they give to different fields: grades in maths can presumably be given more weight by math colleges. They show that this mechanism is suboptimal with respect to the Gale & Shapley mechanism.

Section 2 describes the set-up and the game that applicants play. The identification and estimation of the econometric model is the object of Section 3. Section 4 reports the empirical analysis and counterfactual scenarii are studied in Section 5. Section 6 concludes.

2 Description & modelling

We start by describing the way the selection of students was organized at Universidade Federal do Ceara in 2004 and we formalize its timing and the choices that the students make. Students first choose one and only one major to dispute.[3] As already mentioned, the exam consists in two stages. The access to the second stage is conditional on the grade obtained at the first stage and the selection is performed among students having chosen a given major. Are accepted all students

[3]The only exception is for Architecture and Urbanization where the student must choose a second option that will be active if she fails the ability test specific to this major.

having a rank above a multiple (usually 4 sometimes 3) of the number of available positions in the chosen major. This ranking procedure at the first stage as well as at the second stage defines thresholds in terms of grades that determine if the exam is passed. Appendix B gives further details on the exam.

We consider a parsimonious theoretical set-up building up from models of college choices and of tournaments. Students are supposed to be heterogenous in talent a single dimensional term and students have preferences over different majors which can be monetary or non monetary. The former include rewards that a degree in a specific major raises in the labor market. Furthermore, we consider that entry is not a matter of talent and preferences alone but depends also on a variable called effort exerted before the exams. Talent and effort are distinguished so that both selection and incentives are the two main operating determinants of success in tournaments.

We analyze the entry exams as a game between students in which information is incomplete. Agents do not know the types of competing students, only their distribution in the population. They do know however their own talent and their own preferences. We shall consider Nash equilibria of this game for all students. We thus write the decision model for any student where we consider that actions, or choice probabilities, taken by any other student and which are described below are fixed.

In dynamic models, assumptions about expectations play a key rôle (Manski, 1992). We assume that expectations about own grades and others' grades, which are uncertain, are perfect, in the sense that the distribution of those random variables are equal to the distribution of those in the data. Even if there are multiple equilibria of the game, a point we shall return to in the section about counterfactuals, we nevertheless assume that everybody coordinates on the equilibrium that is observed in the data. Therefore, all thresholds determining at which grades the two-stage exams are passed are supposed to be perfectly anticipated. In other words, players are sophisticated. The validity of this assumption is questionned by Lai, Sadoulet & de Janvry (2009) and He (2009).

We start by looking at the timing of the extended structure and continue by defining notations and formalizing the sequence of decisions & information arrival times. We then turn to measurements and to the analysis of the solution to the economic model.

2.1 Timing

We begin with setting up the notations and we omit the individual index for simplicity. Variable d is a specific major and D *is* the number of such majors. The outside option is denoted $d = 0$. Observed characteristics of the student are denoted z, an unobserved single dimensional variable, ε, stands for student's talent and various unobserved tastes for every major are piled up in a vector of preferences $u = \{u_d\}_{d=0,.,D}$. Because preferences are written as a reduced form of future rewards on the labor market yielded by a specific major degree, they are likely to be correlated with talent ε.

- **Step 1**: *Information*: A standardized national exam whose nickname is ENEM is organized about a year before the entry exam. Denote, m_0, the grade obtained at this exam and assume that:

$$m_0 = \varepsilon + \eta_0,$$

  where η_0 is a noise scrambling the signal for talent ε. Talent ε is known by each agent ex-ante while η_0 is not. The distribution of both is common knowledge among students.

- **Step 2**: *Decisions*: The student simultaneously chooses one single major, $d \in \{0, ., D\}$ and resources, y, to apply in terms of hours or expenditures in order to prepare the entry exam into the university. These resources or *effort*, y, are written in terms of units of higher grades that they allow the student to obtain at least in expectation. We assume that resources y that are unobserved are written in the same units as talent ε is so what impacts the grades is the sum of talent ε and resources y. Cost of resources is supposed to be quadratic in effort and equal to:

$$c_0(y + cy^2/2).$$

  Parameter c_0 could be heterogenous across agents and potentially correlated with ε although we will show that its distribution is not identified. On the other hand, parameter c is assumed to depend on characteristics z or talent ε via a deterministic function only.

- **Step 3:** *Information*: At the first stage of the vestibular, which is common to all majors, the student gets a grade denoted m_1 that we assume is given by:

$$m_1 = \varepsilon + y + \eta_1,$$

where η_1 is some noise. Students are ranked according to a known weighted combination of grades m_0 and m_1 decided by the University. The first stage is passed and the student proceeds to Step 2 if and only if the rank of the student among its fellow students in major $d \in \{1, ., D\}$ is larger than a certain reference rank. We neglect ties that are broken using a formal institutional rule that has marginal importance here. The number of students who are allowed in is equal to three or four times the number of seats available in that major. As m_0 is observed by everybody and as we are looking at Nash equilibria, the passing rule can be written as:

$$m_1 \geq t_1(d, m_0).$$

Otherwise the exam is failed and the student gets the utility u_0 of the outside option. this is the utility of the best option among; The expected value of investing an additional year so as to try to enter again into the university; The expected value of trying to enter another university, a private or a State university – since both exist in the same town – or in another town; To any other option that the student has, for instance if the student desists once and for all. It is likely that the outside option depends on talent ε. Nevertheless, we suppose for simplicity that the value of the outside option does not depend on resources y expended in the last period. The impact of resources y are supposed to be specific to the exam taking place this year and at this university. this assumption enables us to argue later on that modifying the selection mechanism does not affect the population of students willing to take the exam.

- **Step 4**: *Information*: At the second stage,[4] the student gets another grade, m_2 and we assume that:

$$m_2 = \varepsilon + y + \eta_2,$$

where η_2 is some noise. Again, students are ranked according to a known weighted combination of m_0, m_1 and m_2 and only the higher ranked fraction of students is accepted. Again, we can write the passing rule as a function of observed grades m_0 and m_1:

$$m_2 \geq t_2(d, m_0, m_1).$$

[4] At the second stage, the majors have more freedom to set the exams in the subjects they want and that form the core of high school education in Brazil: Portuguese, Geography, History, Biology, Chemistry, Mathematics, Physics and Foreign Language.

Otherwise, if the grade is smaller than the threshold, the exam is failed and the student gets the outside option u_0. Namely, the access to the second stage does not grant any privilege given unobserved talent and effort. Finally, in case of success, the student gains u_d as a function of future wages, major choice and tastes. We chose this specification because we have no information on wages. Note that u_d and talent ε are generally correlated through unobserved wages.

2.2 Observations, Measurements and Expectations

Measurement of talent As a summary, the grades obtained at the different stages are functions of unobserved talent ε and effort y such that:

$$\begin{cases} m_0 &= \varepsilon + \eta_0 \\ m_1 &= \varepsilon + y + \eta_1 \\ m_2 &= \varepsilon + y + \eta_2 \end{cases}$$

where η_i are noises affecting grades. Note that talent and effort have the same effect at the two stages of the exam something that we could try to generalize by using information coming from different exams (i.e. mathematics, portuguese etc) although it is crucial in this set-up since it allows the identification of the distribution of effort under conditions that we study below.

Expectations The student knows her own talent ε, tastes & rewards u_d and continuation value u_0 and learns about (η_0, η_1, η_2) at every step. We first assume that measurement errors (η_0, η_1, η_2) are independently distributed and are independent of any other variables. Their distribution is common knowledge. Furthermore, the student is supposed to know the distribution of the structural random errors $(\varepsilon, \{u_d\}_{d=0,.,D})$ and of covariates z in the population:

$$\Pr(\varepsilon, \{u_d\}_{d=0,.,D} \mid z),$$

but not the precise shocks affecting competitors. We assume that the anticipated distribution is equal to the actual distribution of these variables in the data. [5]

[5] We also impose some technical conditions that all distributions are smooth enough and everywhere increasing. Furthermore, in parametric models below, we neglect all constraints coming from the fact that grades are bounded.

2.3 Solving the model backward

We now write the dynamic model of choice. We do not use a discount factor even if this process takes time, as the discount factor is generically not identified in these dynamic discrete choice models (e.g. Magnac and Thesmar, 2004). We solve the model backward given the information that is available to the agent at each stage.

At Step 4 which is reached in the case of success at the first stage exam, $m_1 > t_1(d, m_0)$, the agent has no decision to take and the history that she conditions on is $h_1 = (\varepsilon, m_0, y, d, m_1)$ comprising talent, ε, initial and first-stage grades, m_0 and m_1, effort y and selected major d. The value of such an history is the sum of what can be obtained in case of either success or failure:

$$V_2(h_1) = \Pr_{\eta_2}\{m_2 > t_2(d, m_0, m_1) \mid h_1\}.u_d + \Pr_{\eta_2}\{m_2 < t_2(d, m_0, m_1) \mid h_1\}.u_0.$$

Given that measurement shocks η_j are independent, the only thing that matters in h_1 are variables $(\varepsilon, y, t_2(d, m_0, m_1) \equiv t_{2d})$. Thus:

$$V_2(\varepsilon, y, t_{2d}) = \Pr\{\eta_2 > t_{2d} - \varepsilon - y\}.u_d + \Pr\{\eta_2 \leq t_{2d} - \varepsilon - y\}.u_0.$$

At the previous step, Step 3, the first stage grade is revealed so that the student gains the value $V_2(.)$ if she passes, i.e. $m_1 > t_1(d, m_0)$ and gains u_0 if she fails.

Going backward, at Step 2, two decisions are to be taken about the selection of a major and about resources y to put up in such an endeavour. History is $h_0 = (\varepsilon, m_0)$, composed by talent and initial grade so that the utility at Step 2 as a function of the two decisions is written as:

$$V_1(y, d; h_0) = -c_0.(y + cy^2/2) + E_{\eta_1}\left[\mathbf{1}\{m_1 > t_1(d, m_0)\}.V_2(h_1)\right] + \Pr_{\eta_1}\{m_1 < t_1(d, m_0)\}.u_0,$$

Denote the overall probability of success in major d as:

$$P_d(y; h_0) = \Pr(\eta_1 > t_1(d, m_0) - \varepsilon - y, \eta_2 > t_2(d, m_0, \varepsilon + y + \eta_1) - \varepsilon - y). \tag{1}$$

Function P_d is derived from the independent distributions of ηs. Note that ressources y unambiguously increase this probability and the derivative of this probability with respect to y, denoted P'_d is positive (see Note 5). Additionally, when y tends to $+\infty$ (respectively $-\infty$ if it was possible), P_d tends to 1 (resp. 0) we can interpret this derivative as a density function.

Regrouping terms, we get:

$$V_1(y, d; \varepsilon, m_0) = -c_0.(y + cy^2/2) + P_d(y; \varepsilon, m_0).u_d + (1 - P_d(y; \varepsilon, m_0)).u_0.$$

Finally, at Step 0, the preferred major is selected as well as exerted effort:

$$V_0(\varepsilon, m_0) = \max_{y,d} V_1(y, d; \varepsilon, m_0).$$

The existence of solutions to this program is easy to argue. Decision d is discrete and y is bounded from below by 0 (i.e. $y \geq 0$). Furthermore if y tends to infinity, the benefit tends to zero because the probability of success is bounded while the cost tends to infinity. Regarding uniqueness arguments are studied below. We shall denote d^*, the selected major, and y_d the optimal solution for effort if major d is chosen.

2.4 Characterization of the solution

As usual in discrete choice models, some normalization of the payoffs are needed. Given the optimal effort y_d, whose determination is analyzed below, the value at time 0 simplifies to:

$$-c_0.(y_d + cy_d^2/2) + P_d(y_d; \varepsilon, m_0).(u_d - u_0) + u_0.$$

As the choice of major d is a discrete decision and as u_0 is independent of y, the continuation value u_0 is irrelevant. The location normalization in discrete choice models is to set $u_0 = 0$ without any loss of generality. The same would apply to any fixed costs involved in the application of ressources (e.g. preparatory courses). The net value of each major is therefore given by:

$$v_d = -c_0.(y_d + cy_d^2/2) + P_d(y_d; \varepsilon, m_0).u_d.$$

Note that the optimal major d^* satisfies the condition $v_{d^*} \geq 0$ since our sample only comprises students willing to take the exam. If $v_{d^*} < 0$, the person does not belong to our population of interest. We will return to the normalization of the level of value functions later on.

Furthermore, the monetary unit in which these values are expressed is not identified in discrete choice and some scale normalization is necessary. We divide these values by cost c_0 (or normalize c_0 to 1) so that:

$$v_d = -(y_d + cy_d^2/2) + P_d(y_d; \varepsilon, m_0).u_d. \tag{2}$$

We now turn to the characterization of the solutions. We first analyze the first order conditions related to the choice of effort for any choice of major d. The first order condition with respect to y yields:

$$1 + cy_d = P'_d(y; \varepsilon, m_0)u_d. \tag{3}$$

As derivative P'_d is positive, effort is positive if and only if $u_d > 0$ so that major d yields more value than exerting the outside option. Besides, using the second order condition, the first order condition corresponds to a maximum if and only if at that point we have

$$P_d" u_d < c \Longrightarrow P_d" u_d y_d < P'_d u_d - 1 \Longrightarrow u_d > \frac{1}{P'_d - P_d".y_d} \text{ where } P'_d - P_d".y_d > 0.$$

for $u_d \geq 0$. There can be multiple solutions to this equation or none although it is easy to argue that they are bounded.

We can also have a corner solution at $y_d = 0$. For instance, if $u_d > 0$ and $u_d < (\max_{y \geq 0} P'_d(y; \varepsilon, m_0))^{-1}$ the cost of effort is too large with respect to the benefit (proportional to u_d) and $y_d = 0$. The general solution is derived from the set of optimal solutions to the first order condition and the comparison of values at those different solutions. This defines a set of regimes that are obtained for different solutions. Figure 1 represents the case of an optimal solution in a diagram when $c = 0$ and P'_d is unimodal so that there is a unique interior solution that can be compared to the corner solution.

Furthermore, the optimal effort function is continuous in u_d when the first order solution remains in the same regime. It can also jump from 0 to a positive solution or from one solution to the next when there is a change in regime. Nevertheless, the value v_d is continuous in u_d and is also increasing in u_d. This is the object of the following:

Lemma 1 *The value v_d given by equation (2) is continuous and increasing in u_d. Furthermore, $v_d(0) = 0$ and function $v_d(u_d)$ can be inverted.*

Proof. Write equation (2) as:

$$v_d = \sum (-\tilde{y}_d + P_d(\tilde{y}_d; \varepsilon, m_0).u_d)\mathbf{1}\{y_d = \tilde{y}_d\} + P_d(0; \varepsilon, m_0).u_d \mathbf{1}\{y_d = 0\},$$

where the sum is taken with respect to all solutions of equation (3). By assumption, $P_d(.)$ is a continuous function of y and within regimes y_d is a continuous function of u_d (see equation (3) above). The only points at which it could be discontinuous are the switching points between regimes but the values v_d in both regimes are equal at these switching points.

Showing that v_d is increasing uses that in every regime the quantity $-\tilde{y}_d + P_d(\tilde{y}_d; \varepsilon, m_0).u_d$ is increasing in u_d, the derivative being equal to $P_d(\tilde{y}_d; \varepsilon, m_0) > 0$. For a corner solution, the derivative is also equal to $P_d(0; \varepsilon, m_0) > 0$. v_d is therefore differentiable except possibly at point $\tilde{u}_d$ where left-hand and right-hand side derivatives may differ. Moreover note that $u_d = 0$ implies that $v_d = 0$. Namely, when $u_d = 0$, we have no investment $y_d = 0$ since they are unproductive and therefore $v_d = 0$. The existence of the reciprocal uses that v_d is an increasing function in u. ■

Some final remarks are in order regarding the structure of the game. The interactions between agents are modelled through the thresholds $t_i(d)$ which are here supposed to be known i.e. are perfectly anticipated by the students. The additional complication (auction) would be to assume that they are the results of these interactions. Imagine that there are N players which are drawn in the distribution of $(\varepsilon, \eta_0, \eta_1, \eta_2)$. Then $t_i(d)$ are determined considering this population of players. As the number of players N is quite large, the sampling variability of the thresholds seems to be negligible with respect to the variability of the measurement errors.

3 The econometric model

3.1 Non-parametric Identification

We here discuss informally some characteristics relative to non parametric identification by contrasting it to the usual estimation of discrete variable models. We assume that we observe m_0, d, y_d, m_1, m_2 and we proceed in several steps. First of all by observing the rank of the auction we can derive $t_1(d; m_0)$ and t_{2m} for each choice d and values of m_0 and m_1.

Second, we analyze the identification of the success probability function $P_d(y; m_0, \varepsilon)$. Last, we study the identification of preferences. We first present the general case and then turn to special cases.

3.1.1 The distribution of measurement errors in grades

To identify the distribution of η_1 and η_2, the main issue arises because of the truncation of the observed sample at the second stage, a truncation that can be written as $m_1 > m_1^*$ where m_1^* is a deterministic threshold, $m_1^* = t_1(d, m_0)$. We have:

$$\begin{cases} m_1 = \varepsilon + y + \eta_1 \\ m_2 = \varepsilon + y + \eta_2 \end{cases}$$

There are two ways to proceed. Either use an argument of identification at infinity by assuming that if m_0 is sufficiently large, the truncation is irrelevant i.e. $m_1^* = t_1(d, m_0) \to -\infty$. We can then use the deconvolution argument of Kotlarski (see for instance, Heckman and Navarro, 2005).

We can also develop identification of the distribution of $\varepsilon + y$ and η_1 in the case in which η_2 is assumed to have a finite number of points of support. The formalization of the identification of mixtures that follows concern mixtures which have two points of support and is thus a special case.

Suppose that m_1 is observed continuously although m_2 is observed continuously only when $m_1 \geq 0$. We assume that:

$$\begin{cases} m_1 = x + \eta_1, \\ m_2 = x + \eta_2. \end{cases} \tag{4}$$

where η_1 (resp. η_2) can take only two values 0 and Δ with probabilities α_1 and $1 - \alpha_1$ (resp. α_2 and $1 - \alpha_2$). We assume that $\alpha_j \in (0, 1)$. In contrast, x is allowed to take a continuum of values and its density function exists and is denoted $\pi(x)$ and for simplicity we shall assume that $\pi(x) > 0$ over the whole real line. Relaxing this assumption is not difficult.

This framework implies that the density of m_1 exists and is equal to:

$$p(m) = \pi(m)\alpha_1 + \pi(m - \Delta)(1 - \alpha_1), \tag{5}$$

which is observed for any m. Second, that the support of the joint density of m_1 and m_2 are 3 straight lines (m, m), $(m, m + \Delta)$ and $(m, m - \Delta)$ and:

$$\begin{cases} p(m, m) = \pi(m)\alpha_1\alpha_2 + \pi(m - \Delta)(1 - \alpha_1)(1 - \alpha_2), \\ p(m, m + \Delta) = \pi(m)\alpha_1(1 - \alpha_2), \\ p(m, m - \Delta) = \pi(m - \Delta)(1 - \alpha_1)\alpha_2. \end{cases}$$

These quantities are observed when $m \geq 0$ and therefore Δ is identified. We now show that all parameters are identified:

Lemma 2 *Parameters α_1, α_2 and $\pi(m)$ are identified*

Proof. Write the last equation as:

$$\pi(m - \Delta)(1 - \alpha_1) = \frac{p(m, m - \Delta)}{\alpha_2},$$

replace in the first:

$$p(m, m) = \pi(m)\alpha_1\alpha_2 + p(m, m - \Delta)\frac{(1 - \alpha_2)}{\alpha_2},$$

and derive:

$$\pi(m)\alpha_1 = \frac{1}{\alpha_2}\left(p(m,m) - p(m,m-\Delta)\frac{(1-\alpha_2)}{\alpha_2}\right).$$

Replace in the second equation to get:

$$p(m,m+\Delta) = \frac{1-\alpha_2}{\alpha_2}\left(p(m,m) - p(m,m-\Delta)\frac{(1-\alpha_2)}{\alpha_2}\right) = a_2\left(p(m,m) - p(m,m-\Delta)a_2\right)$$

where $a_2 = \frac{1-\alpha_2}{\alpha_2} > 0$. This is a second degree equation since $p(m,m-\Delta) \neq 0$ because of our assumptions. It can be rewritten as:

$$(a_2)^2 p(m,m-\Delta) - a_2 p(m,m) + p(m,m+\Delta) = 0,$$

whose discriminant is $D = p(m,m)^2 - 4p(m,m-\Delta)p(m,m+\Delta)$. A necessary condition is therefore that the discriminant is positive and we shall assume this condition. The solution(s) is (are) thus:

$$a_2^{\pm} = \frac{p(m,m) \pm \sqrt{p(m,m)^2 - 4p(m,m-\Delta)p(m,m+\Delta)}}{2p(m,m-\Delta)}.$$

To select one of the solution, consider that:

$$\begin{aligned} D &= \left(\pi(m)\alpha_1\alpha_2 + \pi(m-\Delta)(1-\alpha_1)(1-\alpha_2)\right)^2 - 4\pi(m)\alpha_1(1-\alpha_2)\pi(m-\Delta)(1-\alpha_1)\alpha_2, \\ &= \left(\pi(m)\alpha_1\alpha_2 - \pi(m-\Delta)(1-\alpha_1)(1-\alpha_2)\right)^2, \end{aligned}$$

so that:

$$a_2^{+} = \frac{\pi(m)\alpha_1\alpha_2 + \pi(m-\Delta)(1-\alpha_1)(1-\alpha_2) + |\pi(m)\alpha_1\alpha_2 - \pi(m-\Delta)(1-\alpha_1)(1-\alpha_2)|}{2\pi(m-\Delta)(1-\alpha_1)\alpha_2}.$$

When $\pi(m)\alpha_1\alpha_2 - \pi(m-\Delta)(1-\alpha_1)(1-\alpha_2) > 0$, or equivalently:

$$\frac{\pi(m)\alpha_1}{\pi(m-\Delta)(1-\alpha_1)}\frac{\alpha_2}{1-\alpha_2} > 1 \Longleftrightarrow \frac{\pi(m)\alpha_1}{\pi(m-\Delta)(1-\alpha_1)} > a_2,$$

solution a_2^{+} is equal to:

$$\frac{\pi(m)\alpha_1}{\pi(m-\Delta)(1-\alpha_1)},$$

which is by construction strictly larger than a_2. When $\pi(m)\alpha_1\alpha_2 - \pi(m-\Delta)(1-\alpha_1)(1-\alpha_2) \leq 0$,

$$a_2^{+} = \frac{\pi(m)\alpha_1\alpha_2 + \pi(m-\Delta)(1-\alpha_1)(1-\alpha_2) - \left(\pi(m)\alpha_1\alpha_2 - \pi(m-\Delta)(1-\alpha_1)(1-\alpha_2)\right)}{2\pi(m-\Delta)(1-\alpha_1)\alpha_2} = a_2,$$

which is therefore the generic solution. In the first case, it is straightforward to verify that solution a_2^- is the generic one. To select one of the two solutions reconsider the ratio of the two probabilities which yield:

$$\frac{p(m,m+\Delta)}{p(m,m-\Delta)} = \frac{\pi(m)\alpha_1}{\pi(m-\Delta)(1-\alpha_1)}\frac{1-\alpha_2}{\alpha_2} = \frac{\pi(m)\alpha_1}{\pi(m-\Delta)(1-\alpha_1)}a_2.$$

Therefore:

$$\frac{\pi(m)\alpha_1}{\pi(m-\Delta)(1-\alpha_1)} \gtrless a_2 \Longleftrightarrow \frac{p(m,m+\Delta)}{p(m,m-\Delta)} \gtrless (a_2)^2.$$

The solution is then given as the one which satisfies this latter condition and this identifies α_2. We can now reconsider the expressions:

$$\begin{cases} \pi(m)\alpha_1 = \dfrac{p(m,m+\Delta)}{(1-\alpha_2)}, \\ \pi(m-\Delta)(1-\alpha_1) = \dfrac{p(m,m-\Delta)}{\alpha_2} \Longrightarrow \pi(m)(1-\alpha_1) = \dfrac{p(m+\Delta,m)}{\alpha_2}, \end{cases}$$

to derive that:

$$\frac{1-\alpha_1}{\alpha_1} = a_2\frac{p(m+\Delta,m)}{p(m,m+\Delta)}$$

which (over)-identifies α_1 using observations $m > 0$. Using the equations above identifies $\pi(m)$ for any $m \geq -\Delta$. Finally, using equation (5), we can write:

$$\pi(m-\Delta) = \frac{p(m) - \pi(m)\alpha_1}{1-\alpha_1},$$

which by recursion identifies $\pi(m)$ for any $m < 0$. ■

This proof can be extended to multiple values and might be interesting to extend to the case in which the random shocks are bounded from below.

Last but not least, we can thus write for any $y \geq 0$ the functions $P_d(y;\varepsilon,m_0)$ and $P'_d(y;\varepsilon,m_0)$ using the distribution of (η_1,η_2).

3.1.2 Preferences for majors

We want to exploit the choice model to write and identify the underlying structure of preferences from the distribution of:

$$\Pr(d \mid m_0).$$

For simplicity, assume that d is a binary variable for a two-state model where 2 is the alternative to 1. To proceed, let us fix ε and utilities u_d. This delivers the optimal value of effort y_d and the optimal value functions $v_d(u_d; \varepsilon, m_0)$ so that the choice probabilities become:

$$\Pr(d = 1 \mid m_0, \varepsilon) = \Pr(v_1(u_1; \varepsilon, m_0) > v_2(u_2; \varepsilon, m_0) \mid m_0, \varepsilon).$$

where v_d are increasing functions of u_d by Lemma 1. The functional forms of these functions are identified since they are functions of functions P_d which are supposed to be identified.

We therefore get:

$$\Pr(d = 1 \mid m_0) = \int \mathbf{1}\{v_1(u_1; \varepsilon, m_0) > v_2(u_2; \varepsilon, m_0)\} f(u_1, u_2 \mid \varepsilon, m_0) f(\varepsilon \mid m_0) du_1 du_2 d\varepsilon. \quad (6)$$

The only unknowns is this expression are the densities $f(u_1, u_2 \mid \varepsilon, m_0)$ because the distribution of ε is identified using the proof above as well as functions v_d.

Moreover, note that this is a choice based sample since only $d = 1$ or $d = 2$ is observed. It is straighforward to show that the density function for negative values of u_1 or u_2 remains non identified. Consider the previous expression and write:

$$\begin{aligned}\Pr(d = 1 \mid m_0) = &\int_{u_1>0, u_2>0} \mathbf{1}\{v_1(u_1; \varepsilon, m_0) \quad > \quad v_2(u_2; \varepsilon, m_0)\} f(u_1, u_2 \mid \varepsilon, m_0) f(\varepsilon \mid m_0) du_1 du_2 d\varepsilon \\ &+ \int_{u_1>0, u_2<0} f(u_1, u_2 \quad | \quad \varepsilon, m_0) f(\varepsilon \mid m_0) du_1 du_2 d\varepsilon.\end{aligned}$$

since in the second case we have necessarily $v_1(u_1; \varepsilon, m_0) > v_2(u_2; \varepsilon, m_0)$. Call the second term on the RHS $P_1(m_0)$ which by construction is lower than the LHS. Proceed in the same way for $\Pr(d = 2 \mid m_0)$ and call the second term $P_2(m_0)$. As this is a choice based sample $\Pr(d = 1 \text{ or } 2 \mid m_0) = 1$, summing the choice probabilities yields:

$$\int_{u_1>0, u_2>0} f(u_1, u_2 \mid \varepsilon, m_0) f(\varepsilon \mid m_0) du_1 du_2 d\varepsilon = 1 - P_1(m_0) - P_2(m_0).$$

The claim of non identification can now be phrased as follows. If $(f(u_1, u_2 \mid \varepsilon, m_0), P_1(m_0), P_2(m_0))$ is a solution to equation (6) (and the corresponding expression for $d = 2$) then $(\frac{f(u_1, u_2 \mid \varepsilon, m_0)}{1 - P_1(m_0) - P_2(m_0)}, 0, 0)$ is also a solution to equation (6).

We will thus argue the distribution of negative values for u_1 and u_2 can be set arbitrarily provided that they respect the constraint that $\Pr(d = j \mid m_0) > P_j(m_0)$.

A special case We simplify the model and assume that $m_0 = \varepsilon$ so that the identification of the distribution of ε is simpler. the expression above, we thus have:

$$\Pr(d = 1 \mid m_0) = \int \Pr(v_1(u_1; m_0) > v_2(u_2; m_0) \mid m_0, u_1, u_2).f(u_1, u_2 \mid m_0)du_1du_2. \tag{7}$$

It is well known that from binary choice models additively linear in the unobservables, it is impossible to identify the distribution of the value of one alternative. The situation is slightly different in this non linear-setting.

First remember that when $d = 1$ necessarily $u_1 > 0$. We shall first study one-to-one increasing mappings for u_1 from $[0, \infty)$ to $[0, \infty)$, denoted $T(.)$. They might depend on m_0 that we drop for simplicity since it can applied for any m_0. Since v_d are invertible, we have:

$$\begin{aligned}
\Pr(d &= 1) = \Pr(v_1(u_1) > v_2(u_2)) = \Pr(u_1 > v_1^{-1} \circ v_2(u_2)) \\
&= \Pr(T(u_1) > T \circ v_1^{-1} \circ v_2(u_2)) = \Pr(v_1(T(u_1)) > v_1 \circ T \circ v_1^{-1} \circ v_2(u_2)) \\
&= \Pr(v_1(T(u_1))) > v_2 \circ v_2^{-1} \circ v_1 \circ T \circ v_1^{-1} \circ v_2(u_2)) \\
&= \Pr(v_1(w_1) > v_2(w_2))
\end{aligned}$$

where $w_1 = T(u_1)$, $w_2 = v_2^{-1} \circ v_1 \circ T \circ v_1^{-1} \circ v_2(u_2)$. It proves that the distribution of u_1 is not identified on $[0, \infty)$. As it can neither be identified on $(-\infty, 0)$, we can thus normalize the distribution of u_1 to any known continuous distribution, for instance $N(0, 1)$.

Furthermore, in the absence of any other restriction, the distribution of u_2 remains underidentified. The only restriction that we have is:

$$\Pr(d = 1 \mid m_0) = \Pr(v_1(u_1) > v_2(u_2) \mid m_0) = \Pr(u_2 < v_2^{-1} \circ v_1(u_1) \mid m_0)$$

We informally discuss the final identification issues; We use exclusion restrictions since we shall assume that conditional on talent, effort and some other variables, grades will not depend on some variables (the education level of the parents in the empirics below). On the other hand, we shall assume that conditional on talent and some other characteristics, the utility of each major does not depend on resources invested beforehand in private schooling, preparation and repetition of the exams. We can thus consider that the success probabilities depend on other variables than the utility of each major which allows us to disentangle these two determinants of choices.

3.2 Parametric estimation

3.2.1 The distribution of grades

We adopt the assumption made earlier in the special case of the previous section that m_0 is the talent measured without error i.e. $\eta_0 = 0$. The initial grade is thus given by:

$$m_0 = \mu_0 + \sigma_0.\varepsilon$$

where, abusing notations, m_0 is the transformation of the initial grade into a range going from $-\infty$ to $+\infty$ of the form $\log(\frac{x-(x_{\min}-1)}{(x_{\max}+1)-x})$. The estimation of μ_0 and σ_0 is thus straighforward.[6]

The first and second stage grades (also transformed accordingly to make their range be the whole real line) are supposed to depend on $\varepsilon + y$ the additive combination of talent $\varepsilon = \frac{m_0-\mu_0}{\sigma_0}$ and effort y which is unobserved. The distribution of y is itself a result of optimization and is a function of unobserved tastes u_d for major d. Talent ε and effort y are correlated since investments y depend on talent ε. Furthermore, y is truncated so that $y \geq 0$ with a mass point at zero.

We assume that effort y can be written as $y = (\mu_y + \rho\varepsilon + v)\mathbf{1}\{\mu_y + \rho\varepsilon + v > 0\}$ where v is normal variable independent of ε. As y is unobservable, we posit directly that:

$$\begin{cases} m_1 = \mu_1 + s_1\varepsilon + \lambda_1.(\mu_y + \rho\varepsilon + v)\mathbf{1}\{\mu_y + \rho\varepsilon + v \geq 0\} + \sigma_1.\eta_1, \\ m_2 = \mu_2 + s_2\varepsilon + \lambda_2.(\mu_y + \rho\varepsilon + v)\mathbf{1}\{\mu_y + \rho\varepsilon + v \geq 0\} + \sigma_2.\eta_2, \end{cases} \tag{8}$$

where η_i are distributed $N(0,1)$ and are independent between themselves and of v and ε and where we normalize the normal variate v so that $Ev = 0$ and $Vv = 1$. Parameters s_j, λ_j and σ_j are scaling factors and if the specification of the model in previous sections is true we should have (leaving the mean unrestricted):

$$\frac{s_1}{s_2} = \frac{\lambda_1}{\lambda_2}. \tag{9}$$

Denote $m_\varepsilon = \mu_y + \rho\varepsilon$. Using first stage grades, it is possible to estimate μ_1 and s_1 by regressing m_1 on ε by

$$E(m_1 \mid \varepsilon) = \mu_1 + s_1\varepsilon + \lambda_1(m_\varepsilon\Phi(m_\varepsilon) + \varphi(m_\varepsilon)), \tag{10}$$

where $\Phi(.)$ (resp. $\varphi(.)$) is the unit normal cdf (resp. pdf) (Johnson, Kotz and Balaskrishnan, 1994, vol1, p156). We can also derive that :

$$V(m_1 \mid \varepsilon) = \sigma_1^2 + \lambda_1^2\left[(1 + m_\varepsilon^2)\Phi(m_\varepsilon) + m_\varepsilon\varphi(m_\varepsilon) - (m_\varepsilon\Phi(m_\varepsilon) + \varphi(m_\varepsilon))^2\right]. \tag{11}$$

[6] In the case in which m_0 is missing and this concern only a small fraction of the sample (5%), we impute the value obtained by a regression of m_0 on explanatory variables and treat the prediction as if it were the true value for m_0.

Using second stage grades is slightly more difficult since the sample is truncated at value $m_1 > m_1^*$. Note that m_1^* depend explicitly on m_0 and that the equation above implies that:

$$\begin{aligned} E\left(m_2 \mid \varepsilon, m_1\right) &= \mu_2 + s_2.\varepsilon + \lambda_2.E\left((m_\varepsilon + v)\mathbf{1}\{m_\varepsilon + v \geq 0\} + \sigma_2.\eta_2 \mid \varepsilon, m_1\right) \\ &= \mu_2 + s_2.\varepsilon + \lambda_2 E\left((m_\varepsilon + v)\mathbf{1}\{\mu_y + \rho\varepsilon + v \geq 0\} \mid \varepsilon, m_1\right) \\ &= \mu_2 + s_2.\varepsilon \\ +\lambda_2.E\left((m_\varepsilon + v)\mathbf{1}\{m_\varepsilon + v \geq 0\} \mid m_1 - \mu_1 - s_1\varepsilon = \lambda_1.(m_\varepsilon + v)\mathbf{1}\{m_\varepsilon + v > 0\} + \sigma_1.\eta_1, \varepsilon\right), \end{aligned}$$

where the second line obtains because of independence between η_2 and ε, η_1 and v and the last line by a mechanical substitution.

Note that conditioning on the fist stage grade dispenses with looking at the selection bias since we look at all $m_1 > m_1^*$. The rest of the algebra is done in Appendix A.1 where the following Lemma is proven:

Lemma 3 *We have:*

$$\left\{ \begin{array}{ll} E\left(m_2 \mid \varepsilon, m_1\right) = & \mu_2 + s_2.\varepsilon + \lambda_2.A_1A_2, \\ V\left(m_2 \mid \varepsilon, m_1\right) = & \sigma_2^2 + \lambda_2^2.\left[A_1B_1 - (A_1A_2)^2\right], \end{array} \right.$$

where A_1, A_2 and B_1 are defined in the proof as a function of the parameters.

The parametric model consists therefore in the two equations (10) and (11) and by the restrictions given in Lemma 3. We used a psseudo likelihood function based on the normal distribution to estimate these restrictions. We can also impose restriction (9) which is overidentifying in this parametric model.

This parametric setting can be extended easily to a semi parametric setting. First, it is immediate to realize that we can dispense with any distributional assumption about η_2 in the previous argument. It is also easy to consider that y is an unrestricted spline function of ε:

$$y = \mu_y + \sum_{k=1}^{K} b_k(\varepsilon) + v,$$

where $b_k(\varepsilon)$ are quadratic spline functions, for instance. It is doable but less easy to rela the normality assumption on η_1 but much less easy to relax the normality assumption on v. We let these investigations for future research.

3.2.2 The choice model

As shown in equation (7), we have that:

$$\Pr(d = 1 \mid m_0, z) = \Pr(v_1(u_1; m_0) > v_0(u_0; m_0) \mid m_0)$$

where we assume that $u_1 \sim N(0, 1)$ and where $u_0 \sim N(z\beta_0, \sigma_0)$ where (β_0, σ_0) are to be estimated.

The first step of the algorithm detailed in the Appendix consists in computing the various functions v_d. It is then given by:

$$v_d = -y_d + P_d(y_d; m_0, \varepsilon).u_d$$

where the result of Lemma 1 is used solving for:

$$1 = P'_d(\tilde{y}_d).u_d$$

and the corner solutions. We do that for each d and evaluate the result using simulation following the lines of the GHK simulator.

4 Empirical analysis

The complete database comprises 41377 students who took the exam in 2004. The list of variables consists in:

- grades at various stages (the initial national exam, the first and second stage of the *vestibular*)
- gender, age by discrete categories (16, 17.5, 21 and 25), the education levels of father and mother.
- the public/private choice at the primary and high school levels: it is a discrete variable taking values 0,1/3,2/3,1 according to the fraction of time spent in a private school.
- the number of repetitions and the undertaking of a preparatory course

In total there are 58 majors that the students may consider. Table 1 reports the list of majors and the grouping of majors that we performed acccording to the composition of the departments within the University and the contents of the second stage exam. For instance, a major in Medicine or

Pharmacy and other medical fields as well, requires taking specific exams in biology and chemistry at the second stage while an Engineering major requires taking mathematics and either physics or chemistry at the second stage. On the other hand, Law requires the same specific exams as Literature but belongs to a different department. We used these guidelines to group the majors into 4 groups – Business & Law, Mathematics, Medicine and Humanities – which are themselves differentiated into 13 subgroups. The complete tree appears in Table 1. This decomposition makes it also simpler to report descriptive information for majors.

4.1 Descriptive results

Specifically, Table 2 reports the number of student applications, the 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 very 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. Nevertheless, there are other differentiations in terms of quality.

We now look in more detail to teh differences in terms of grades across major fields and we justify the restiction of our analysis to a specific subsample containg three medical majors.

4.1.1 The distribution of grades

Tables 3 and 4 report summary statistics in each major field concerning the grades obtained first at the national examination (Table 3) and at the first stage of the college exam.[7] We report statistics on the distribution of the initial and first stage grades in three samples:[8] 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 programs. Major fields are ranked according to the median grade among those who passed the final exam in that major field.

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

[8] 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.

From Table 3, we can conclude that applications do not differ across majors in the tails of the distribution of initial grade since all minima are around 20 and all maxima are close to the top grade 63. The fact that applicants do self-select by talent when choosing their majors is captured by the medians of initial grades of applicants in column 4 of Table 3. Medians are quite constant around 34 in the 6 first major fields yet then increase to attain the grade level of 44 for Law and 51 for Medicine. A second conclusion from Table 3 is that as expected the initial grades of those students who have access either to the second stage or pass the exam, are larger and are ordered as would be the first stage grades. Medians in the selected samples are now ranging from 42 in agrosciences to 58 in medicine. What strikes in this table is the proximity of the initial grades of those who pass and those who fails at the second stage which expresses that initial grade is an imperfect proxy for the first-stage grade. The range of medians shrinks to 46 to 58.

Initial grades are a predictor of talent and of effort in the model. This is why the same statistics using first stage grades reported in Table 4 should be more informative. Indeed, even the minima tend to be ordered as the median of students who pass (column 6) from 70 to 90 in column 1. The first columns also reveal that some groupings might be somewhat artificial. The whole distribution is for example scattered out in mathematics from a minimum of 70 to a maximum of 222 while in medicine the range is 189 to 224. Other details are worth mentioning. The minimum grade in medicine to pass to the second stage is close to the maximum that was obtained by a successful students in Other fields and somewhat less than in Agrosciences.

In conclusion, Medicine and Law are ranked the highest, as a matter of fact by a large amount of difference with other major fields. For instance, in Table 4, the first stage grade among those who passed in Medicine (resp. Law) has a median of 206 (resp. 189) while the next two are Pharmacy, Dentist and Other (175) and Engineering (171) and the minimum is for Agrosciences at 142.

4.1.2 Restricting the sample

For computational simplicity, the empirical analysis will be performed using a sub-sample of applicants to this college entry exam. The form of the exam consisting in only one choice allows us to simply restrict the sample without modifying the argument developed in the economic model. All other majors are now summarized by the outside option. In the rest of the analysis, we shall consider only individuals who take exams in the majors that are part of Medicine, the most com-

petitive major field as shown above. There are three majors in this group corresponding to three different locations in the state of Ceará: Barbalha, Sobral and Fortaleza. The first two majors are small and offer 40 positions only while the last one, Fortaleza, is much larger since it offers 160 seats. As shown in the empirical analysis below, this assymetry turns out to be important tp prove the importance of strategic effects.

Table 5 repeats the analysis performed in Table 4 at the disaggregated level of those majors. Fortaleza is the most competitive one since the median of the first-stage grade of those who passed is equal to 208.57 while for the two others, it remains around 200. nevertheless, the pass rate as shown in Table 5 relating the number of applicants and the number of positions is about the same in Sobral and Fortaleza (7%) while it is slightly lower in Barbalha (5%). At the same time, Barbalha receives applications from the weakest students as shown by the median grades in the sample of all applicants to this major.

The list of variables and descriptive statistics in the pools of applicants to the three different majors appear in Table 6. The number of applicants taking the first exam are in total 3606 and are decomposed into respectively 739 (Barbalha), 542 (Sobral) and 2325 (Fortaleza). The number of seats after the first-stage is four times the number of final seats and is thus respectively equal to 160 for the small majors and 600 for Fortaleza. Note also that only two applicants in the pool of Fortaleza applicants and none in the others 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.[9]

Apart from statistics on grades that we already reviewed, the three subsamples are somewhat different. More women and individuals whose father and mother's education level is higher apply to the main major in Fortaleza. There are also some differences in terms of private high school attendance or the number of repetitions although they are not striking. Finally, Figures 2 and 3 report the estimated density functions of the first-stage and second-stage grades in the three subsamples of applicants. These distributions are unimodal and the distributions of second-stage grades are very similar in the three subsamples. Selection performed at the first stage seems to be quite uniform. In contrast, the distributions of first stage grades are quite different. First, they have a long-tail on the left concerning the weakest applicants. It seems also that the distribution

[9]The failure of two students to take the exam out of 920 might be put onto the account of sickness or other accidents even if these events are not modelled here.

for Fortaleza first-order stochastically dominates the distribution of first-stage grades applicants. Fortaleza seems to be selected by better applicants.

4.2 Estimation of the dynamic model

4.2.1 Grade equations

We first estimate parametric grade equations for m_1 and m_2 as developed in Section 3.2.1 by pseudo-maximum likelihood in which we use the first and second order moments of the grades and where the pseudo-distribution is normal. Results, using robust standard errors, are reported in Table 7. Table 7a reports the results using a simple specification including the grade at national exam as the only covariate while Table 7b reports results for a more complete specification. These tables report three sets of results corresponding to the coefficients in equation (8). The first two columns report the estimated coefficients of variables entering directly the specification of first stage grades, $\mu_1 = x\beta_1$ and s_1 (resp. second-stage grades, $\mu_2 = x\beta_2$ and s_2). The last column reports the estimates for the variables appearing in the common component, $\mu_y = z\beta_y$ and ρ. Finally, the estimated coefficients of effort, λ_j, and standard errors, σ_j, of each equation in (8) are reported at the bottom of the first two columns.

In Table 7a talent as measured by the initial grade is influencing positively the first and second stage grades. It is very significant at the first stage but not at the second stage. As expected also, the effect of effort, as described by parameters λ_j, are positive and highly significant. Overall, restriction (9) that says that unobserved effort has the same effect at both stages of the exam is frankly rejected (Student = 6.54). It might due to substantive differences or it might be due to the too restrictive nature of the parametric model.

This is why we considered the complete specification in Table 7b. The restriction that says that unobserved effort has the same effect at both stages of the exam is now not rejected at the 5% level (p-value = 6.8%). There are two points to note before commenting the complete results. First, there does not seem to be any difference for males and females (p-value = 8.2%). Second, the variables concerning the education levels of the father and mother fail to appear in a direct way in this specification (p-value = 31.2%) while they affect the common component concerning effort (p-value = 3.5%). As developed briefly in the identification section, these variables will provide an important identification leverage since they will be assumed to affect utility but not probabilities

of success. On the other hand, the variables that are supposed to affect grades but not utilities are the number of repetitions, the attendance of a preparatory course and a private high school. They are jointly significant in the terms μ_1 and μ_2, as reported in the first two columns of table 7b.

In terms of the variables, results are quite expected. Regarding the direct effects on grades, the older the applicant is, the lower grades at the two stages are. Attending a private high school increases first stage grades significantly but not second stage grades while attending a preparatory course does the reverse by increasing second-stage grades significantly. It conforms with the intuition that the first-stage content is general while the second-stage is specific. The number of repetitions increases at the 10% level the second-stage grade. Talent as described by the grade obtained at the national exam is unambiguously positive and significant. Turning to the effect of these variables on effort (last column), we find again that age decreases effort at least at age 25. The number of repetitions increases effort significantly and it might be that effort expanded in the previous exams might find a way to express itself here. Parents' education unambiguously increases effort so that the utility of the majors unambiguously increases when these variables increase. Finally, talent also increases significantly effort.

4.2.2 Preference estimates

Second we estimated preferences using simulated maximum likelihood. Random preference heterogeneities are assumed to be independent normals and we use the GHK simulator. The only explanatory variable in the simple specification that is reported here is the Grade at the national exam and the results, using robust standard errors, are reported in Table 8. By assumption, the main parameter is normalized to 0 and the standard error to 1 for the reference major (Fortaleza). As for the other two majors, the negative and significant coefficients for the intercepts indicate that the choice of the small majors (Barbalha and Sobral) are dominated by Fortaleza and that talent attracts less students at Barbalha than at the other two schools something that we already spotted using descriptive statistics. The estimate of standard errors of tastes for the latter major are nevertheless larger and a significant fraction of the population have preferences for Barbalha. This will have an impact on the results for some counterfactuals that we study now.

[Include full specification results here]

5 Evaluation of policy changes

As we estimated the model using candidates to the exam only, it is not immediately clear that we can evaluate the impact of policy changes on the extensive margin i.e. how it modifies the composition of the population of candidates. If the assumptions of the economic model are correct it does not as a matter of fact. Indeed, changing the selection mechanism modifies the success probabilities but it does not modify preferences and the key position of the outside option in the preference list. Indeed, if a major yields utility above the outside option, it will always deliver a value of this major above the outside option whatever the selection mechanism. The same argument applies to majors yielding utility below the outside option. Therefore, the population of interest remains the same .

Two possible changes among many others, are interesting to study:

- Students could choose more than one major before the first stage.
- Students could choose between the two stages and not before the first stage.

We develop these two cases in this section. For welfare, we use an utilitarist social welfare function where students get their ex-ante expected utility.

5.1 Enlarging choices

Suppose now that the choice set is composed by pairs (d^*, d^{**}) instead of a single choice d^*. The timing of the game remains the same, choices and investment being made before the first stage. After the first stage, there are now three possibilities:

- $m_1 > t_1^A(d^*, m_0)$: the student qualifies for the second stage of major d^*.
- $m_1 < t_1^A(d^*, m_0)$ and $m_1 > t_1^A(d^{**}, m_0)$: the student qualifies for the second stage of major d^{**}.
- $m_1 < t_1^A(d^{**}, m_0)$: the student fails.

Note that the first and third stage are as in the original game whereas the second regime is original. It is also obvious the limit grades $t_1^A(d, m_0)$ varies with respect to the original experiment.

Note also that because of perfect expectations, choosing a d^{**} such that $t_1^A(d^{**}, m_0) > t_1^A(d^*, m_0)$ implies that the second regime disappears and is thus equivalent to make a single choice $(d^*, \varnothing)$ as in the original experiment. It happens in all cases where only a single choice is valued positively by the agent. This is why we also allow for this choice possibility where $t_1^A(\varnothing, m_0) = \infty$.

The solving of the Nash equilibrium is slightly more difficult that in the original game. We follow the Gale Shapley student optimal stable mechanism to do that. Specifically, let us denote the common parameter controling limit grades as a vector t :

$$t_1^*(d, m_0) = t_1(d, m_0, t^0), \quad t_1^A(d, m_0) = t_1(d, m_0, t^A)$$

where t^0 is the original set of limits in the exam as it works currently and t^A is the counterfactual set of grades which will describe the Nash equilibrium in the new game. This will apply sililarly to the second stage limit grades, $t_2(d, m_0, t^A)$.

Set up the individual model as follows. For those succesful at the first stage, we have before the second stage:

$$\begin{aligned} V_2(h_1) &= \Pr_{\eta_2}\{m_2 > t_2(d^*, m_0, t^A)\}u_{d^*} \text{ if } m_1 > t_1(d^*, m_0, t^A) \\ &= \Pr_{\eta_2}\{m_2 > t_2(d^{**}, m_0, t^A)\}u_{d^{**}} \text{ if } m_1 \in [t_1(d^{**}, m_0, t^A), t_1(d^*, m_0, t^A)) \\ &= 0 \text{ if } m_1 < t_1(d^{**}, m_0, t^A) \end{aligned}$$

where it should be understood that $V_2(h_1) = 0$ when the interval on the second line is empty.

The value function at the first period becomes:

$$\begin{aligned} V_1(d^*, d^{**}, y, m_0) &= -y + E_{\eta_1} V_2(h_1) \\ &= -y + E_{\eta_1} \Pr_{\eta_2}\{m_2 > t_2(d^*, m_0, t^A)\}\mathbf{1}\{m_1 > t_1(d, m_0, t^A)\}u_{d^*} \\ +E_{\eta_1} \Pr_{\eta_2}\{m_2 &> t_2(d^{**}, m_0, t^A)\}\mathbf{1}\{m_1 \in [t_1(d^{**}, m_0, t^A), t_1(d^*, m_0, t^A))\}u_{d^{**}} \\ &= -y + P_{d^*}(y, m_0, t^A)u_{d^*} + P_{d^{**}}^{(2)}(y, m_0, t^A)u_{d^{**}}. \end{aligned} \tag{12}$$

where $P_{d^*}(y, m_0, t^A)$ is the overall probability of success for major d^* as defined above. The second probability $P_{d^*}^{(2)}(y, m_0, t^A)$ is:

$$P_{d^{**}}^{(2)}(y, m_0, t^A) = E_{\eta_1} \Pr_{\eta_2}\{m_2 > t_2(d^{**}, m_0, t^A)\}\mathbf{1}\{m_1 \in [t_1(d^{**}, m_0, t^A), t_1(d^*, m_0, t^A))\}.$$

Define $\tilde{t}^A$ such that all limit grades remain the same except:

$$t_1(d^{**}, m_0, \tilde{t}^A) = t_1(d^*, m_0, t^A).$$

Therefore:

$$P^{(2)}_{d^{**}}(y, m_0, t^A) = P_{d^{**}}(y, m_0, t^A) - P_{d^{**}}(y, m_0, \tilde{t}^A), \tag{13}$$

as a function of the previous success probabilities.

From equation (12), we can define $y_{(d^*, d^{**})}$ and therefore the optimal value function as a function of $(u_{d^*}, u_{d^{**}})$. We then define:

$$(d^*, d^{**}) = \arg \max_{(d_1, d_2, y)} V_1(d_1, d_2, y, m_0).$$

In order to compute the counterfactual limit grades t^A we proceed as follows. We predict choice and success at both stage probabilities for all individuals:

$$\begin{aligned} p_{i1}(d, t^A) \quad = \quad & \Pr\{\text{Choosing } (d, \tilde{d}) \text{ and success at Stage 1 for } d \text{ or} \\ & \text{Choosing } (\tilde{d}, d) \text{ and success at Stage 1 for } d \text{ and not for } \tilde{d}\} \end{aligned}$$

Anlogously we can define $p_{i2}(d, t^A)$ describing choice and full success at both stages.

We then solve the non-linear D equations with D unknowns:

$$\sum_{i=1}^{N} p_{ij}(d, t^A) = N_j(d),$$

where $N_j(d)$ are the number of offered seats at Stage j for major d.

5.2 Changing the timing of choices

We can also change the timing of the game in the following way. The individual is supposed to choose his/her major after full revelation of the first stage grade. Note that it is not equivalent to the game where the list of preferences over all majors is as long as the individual wants since the choice can be made dependent upon the revelation of the first stage grade m_1.

Let $C(m_1, m_0, t^B)$ be the choice set left after full revelation of the first stage grade:

$$C(m_1, m_0, t^B) = \{d; m_1 > t_1(d, m_0, t^B)\}$$

where t^B is any set of equilibrium limit grades in this new setting. The value before the second stage is:

$$\begin{aligned} V_2(h_1) &= \Pr_{\eta_2}\{m_2 > t_2(d, m_1, m_0, t^B)\}u_{d^*} \text{ if } d \in C(m_1, m_0, t^B) \\ &= 0 \text{ if not.} \end{aligned}$$

The individual chooses d^* such that:

$$d^* = \arg \max_{d \in C(m_1, m_0, t^B)} V_2(d, m_1, m_0)$$

The value function at the first period becomes:

$$V_1(y, m_0) = -y + E_{\eta_1} V_2(d^*, m_1, m_0).$$

and we maximize this quantity in order to derive the optimal effort, y.

There is no simple way of writing this maximization program since it corresponds to the inversion of the expectation and the maximization operators. Indeed, the dynamic program that was solved before correponds to:

$$\max_{d,y} \left(-y + E_{\eta_1} V_2(d, m_1, m_0)\right)$$

to compare with the current one:

$$\max_{y} \left(-y + E_{\eta_1} \max_{d} V_2(d, m_1, m_0)\right).$$

The algorithm that could be used to solve this program is by simulation. Let η_1^s a draw in the distribution of η_1. We can thus compute $\max_d V_2(d, m_1, m_0)$ as a function of y and specifically, the derivative of this function with respect to y. We repeat this computation over S simulations and get the evaluation of the second term $E_{\eta_1} \max_d V_2(d, m_1, m_0)$ as a function of y. We can then solve for the optimal y. As y is bounded from below by 0 and the return to y is bounded if y tends to ∞, a solution exists. It might not be given by a first order condition though depending on the characteristics of the function $E_{\eta_1} \max_d V_2(d, m_1, m_0)$.

5.3 Discussion of the uniqueness of equilibrium

In each of these experiments, including the one which is the current scheme of selection, remains the pending question of the uniqueness of the equilibrium. This property should be proven in each

set-up and we do not have any general result on uniqueness, to our knowledge. Nevertheless, it is possible to prove uniqueness in a simple context. We assume that the scheme is the current selection scheme with heterogeneity across agents in preferences only (equal talent) and in which there is no effort. We first look at the equilibrium at the second stage of the exam, given some probability of success, $\{p_d\}_{d=1,.,D}$, at the exam and given some choice probabilities $\{\pi_d\}_{d=1,.,D}$. We pile up these objects into vectors p and π.

The choice probabilities are given by the comparison between value functions $\{v_d(p_d)\}_{d=1,.,D}$ where each value function v_d depends on the success probability p_d only and where it is strictly increasing, i.e. $p_d > p'_d \Longrightarrow v_d(p_d) > v_d(p'_d)$. We assume that for all d and all p_d, we have $\pi_d > 0$. An additional interesting property is that:

$$\forall p; \sum_{d=1}^{D} \pi_d(p) = \alpha \text{ independent of } p.$$

Without loss of generality, we will assume that $\alpha = 1$ in the following.

Let $\{\lambda_d\}_{d=1,.,D}$ be the fraction of seats in the population attributable to each major. The equilibrium relationships can then be written as:

$$\lambda_d = \Pr(\text{Choosing } d, \text{ Success in } d) = \Pr(\text{Choosing } d) \Pr(\text{Success in } d) = p_d \pi_d = z_d(p),$$

since choices and realizations are independent because effort and talent are absent. We pile up the elements $z_d(p)$ into $z(p)$. The probability of failing is:

$$\sum_{d=1}^{D} (1 - p_d)\pi_d = 1 - \sum_{d=1}^{D} \lambda_d,$$

and is satisfied by construction as an accounting identity.

The following Lemma ensures the uniqueness of equilibrium:

Lemma 4 *For any (p, p'), $p \neq p'$ and no elements of p is equal to zero, we have $z(p) \neq z(p')$.*

Proof. By contradiction, assume that $z(p) = z(p')$ so that for any d, $p_d \pi_d = p'_d \pi'_d$.

Consider first that (i) $p'_d \leq p_d$ for all d and the inequality is strict for at least one d. We thus have:

$$p_d \pi_d = p'_d \pi'_d \leq p_d \pi'_d$$

and for one d at least the inequality is strict since for all d, $\pi_d > 0$. Thus $\pi_d \leq \pi'_d$ and one inequality at least is strict. It is a contradiction with $\sum_{d=1}^{D} \pi_d = 1$. Case (i) can obviously be extended to the case where $p'_d \geq p_d$ and one inequality is strict.

Second, consider (ii): for all $d \in I$, $p'_d < p_d$ and for all $d \in J, p'_d \geq p_d$ and where I is not empty. The case where I is empty is the complement of case (i). We have:

$$d \in I, p_d \pi_d = p'_d \pi'_d \Longrightarrow \pi_d = \frac{p'_d}{p_d} \pi'_d < \pi'_d,$$

since $\pi'_d > 0$. It implies that:

$$\sum_{d \in I} \pi_d < \sum_{d \in I} \pi'_d.$$

Yet, by definition:

$$\sum_{d \in I} \pi_d = \Pr(\max_{d \in I} v_d(p_d) \geq \max_{d \in J} v_d(p_d)), \sum_{d \in I} \pi'_d = \Pr(\max_{d \in I} v_d(p'_d) \geq \max_{d \in J} v_d(p'_d)).$$

As for all $d \in I$, $p'_d < p_d$, $\max_{d \in I} v_d(p'_d) < \max_{d \in I} v_d(p_d)$ since the value functions are increasing, and as for all $d \in J$, $p'_d \geq p_d$, $\max_{d \in J} v_d(p'_d) \geq \max_{d \in J} v_d(p_d)$, we have:

$$\Pr(\max_{d \in I} v_d(p_d) \geq \max_{d \in J} v_d(p_d)) \geq \Pr(\max_{d \in I} v_d(p'_d) \geq \max_{d \in J} v_d(p'_d)) \Longrightarrow \sum_{d \in I} \pi_d \leq \sum_{d \in I} \pi'_d,$$

a contradiction with the inequality above. ■

This Lemma ensures that the equilibrium is unique in terms of probabilities p. These equilibrium values are obtained as a fucntion of the thresholds:

$$p^*_d = \Pr(m_1 > t_1(d), m_2 > t_2(d)). \tag{14}$$

Using the fact that first stage and second stage probabilities are fixed and known, we have:

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

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

$$\Pr(m_1 > t_1(d)) = \frac{p^*_d}{\lambda}.$$

The second threshold $t_2(d)$ is then obtained by solving equation (14).

The general case is more difficult to tackle since it consist in solving equilibrium relationships such as:

$$E_u \left[p_d(u) \pi_d(u) \right] = z_d(p) = \lambda_d.$$

5.4 Results

We computed the equilibrium thresholds in the first counterfactual developed in the subsection 5.1 above and using values reported in Tables 7a and 8. We computed these counterfactuals using the population of applicants at UFC by using that even success probabilities change, only students who have at least one positively valued major take exams at this University. The population of reference does not change as a result.

Table 9 reports the current and counterfactual thresholds. Quite surprisingly, the effect is strong. The first school, Barbalha, becomes a very competitive place since the thresholds at the first and second stage are now the highest of all three majors. We attribute this to the very large dispersion of tastes for Barbalha in the population and to the fact that students have less incentives to censor themselves when they declare their first choices. They can "try" at Barbalha and as an insurance device select Fotaleza second, a thing that they would not do in the current system because of the small number of seats at that school (40). This is an illustration of the strategic efffect that the current system has on students. Furthermore, what Barbalha gets, the largest one Fortaleza loses it and at a lesser degree the other small one, Sobral. The counterfactual tends to create an elitist small medicine school at Barbalha while the elite big school was before in Fortaleza. This is not neutral for school managers and this could be evaluated.

If we adopt a pure utilitarist viewpoint by summing the ex-ante expected values for all students, the counterfactual is slightly preferred ($Ev_1 = 1406.750$) to the current system ($Ev_0 = 1353.776$). Figure 4 report the estimated current and counterfactual distribution of expected values in the whole population and Figure 5 reports what we found in terms of differences of expected values. The distribution of differences is slightly assymetric. The ones who gain to the counterfactual change, gain more that the ones who lose. The distributive effects are thus quite strong.

6 Conclusion

Our main result is that strategic effects are indeed very strong when the matching mechanism demands that the choice of majors should be before the first stage exam and that only one choice is offered to every student. Second, there are some gains to have more choices in the list although the distributive effects might be strong.

These results need to be extended in various directions. We should be able to extend the experiment using three majors only to the whole set of majors. It would also be interesting to perform semi-parametric estimation of grades to see if our results are robust to this change in specification. Other counterfactuals like the one developed in the second subsection could also be analyzed.

On the theoretical side, there are also large margins for improvement. For instance, there is a complication related to the two-stage aspect of the exam. The first stage is general and the second stage is chosen by the major. Some students do not undertake all specific exams (biology is unnecessary to go into physics although needed when entering medicine). The vestibular system thus allows a more refined selection and that is an interesting theoretical object of study.

REFERENCES

Abdulkadiroğlu, A., Y., K., Che and Y. Yasuda, 2008, "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. and T., Sonmez, 2003, "School Choice: A Mechanism Design Approach", *American Economic Review*, Vol. 93, No. 3, pp. 729-747

Arcidiacono, P., 2004, "Ability Sorting and the Returns to College Major," *Journal of Econometrics*, 121, 343-375.

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

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

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

Budish, E. and E. Cantillon, 2010, "The Multi-unit Assignment Problem: Theory and Evidence from Course Allocation at Harvard", CEPR WP 7641

Carneiro, P., K. Hansen and J.J. Heckman, 2003, "Estimating Distributions of Counterfactuals with an Application to the Returns to Schooling and Measurement of the Effects of Un certainty on Schooling Choice," *International Economic Review*, 44, 361-422.

Davies, T. and A. Stoian, 2007, "Measuring the Sorting and Incentive Effects of Tournaments Prizes", unpublished manuscript.

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

He, Y., 2009, "Gaming the School Choice Mechanism", unpublished manuscript, Columbia University.

Heckman J.J., and S., Navarro, 2007, "Dynamic Discrete Choice and Dynamic Treatment Effects", *Journal of Econometrics*, Volume 136, Issue 2, Pages 341-396.

Lai, F., E., Sadoulet and A., de Janvry, 2009, "The Adverse Effect of Parents' School Se-

clection Errors on Academic Achievement: Evidence from the Beijing Open Enrollment Program", *Economics of Education Review*, 28:485-496.

Leuven, E. H. Osterbeek, J. Sonnemans and B. van der Klauuw, 2008, "Incentives versus sorting in tournaments: Evidence from a field experiment", unpublished manuscript.

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.

Magnac, T. and Thesmar, D. 2002, "Identifying Dynamic Discrete Decision Processes", *Econometrica*, 70, 801-816.

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.

Manski, C., and D. Wise, 1983; *College Choice in America*. Cambridge, MA: Harvard University Press.

Montmarquette C, K. Cannings and S. Mahseredjian, 2002, "How do young people choose college majors?", *Economics of Education Review* 21 543–556.

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

A Statistical appendix

A.1 Proof of Lemma 3

Denote :

$$\begin{aligned} M_1 &= m_1 - \mu_1 - s_1\varepsilon, \\ X &= m_\varepsilon + v \end{aligned}$$

so that the term on the last line is proportional to:

$$A \equiv E\left(X.\mathbf{1}\{X \geq 0\} \mid \lambda_1.X + \sigma_1.\eta_1 = M_1, \varepsilon\right) = \int\int_{\lambda_1.X+\sigma_1.\eta_1=M_1} X.\mathbf{1}\{X \geq 0\}\varphi(X - m_\varepsilon)\varphi(\eta_1)dXd\eta_1$$

since X and η are independent. We obtain:

$$A = \int\int_{\lambda_1.X+\sigma_1.\eta_1=M_1, X\geq 0} X.\varphi(X - m_\varepsilon)\varphi(\eta)dXd\eta = \int_{X\geq 0} X.\varphi(X - m_\varepsilon)\varphi(\frac{M_1 - \lambda_1 X}{\sigma_1})dX.$$

We can write that:

$$\varphi(X - m_\varepsilon)\varphi(\frac{M_1 - \lambda_1 X}{\sigma_1}) = A_1.\frac{1}{\sigma}\varphi(\frac{X - \mu}{\sigma})$$

where A_1, μ and σ are constant to determine. The left hand side is equal to:

$$\frac{1}{2\pi}\exp(-\frac{1}{2\sigma_1^2}\left[\sigma_1^2(X - m_\varepsilon)^2 + (M_1 - \lambda_1 X)^2\right]) \tag{15}$$

and the argument between square brackets in the exponential function is:

$$\begin{aligned} & \sigma_1^2X^2 + \sigma_1^2m_\varepsilon^2 - 2\sigma_1^2m_\varepsilon X + M_1^2 + \lambda_1^2X^2 - 2M_1\lambda_1 X \\ = \; & X^2(\sigma_1^2 + \lambda_1^2) - 2X(\sigma_1^2m_\varepsilon + M_1\lambda_1) + \sigma_1^2m_\varepsilon^2 + M_1^2 \\ = \; & (\sigma_1^2 + \lambda_1^2)(X - \mu)^2 + 2X((\sigma_1^2 + \lambda_1^2)\mu - (\sigma_1^2m_\varepsilon + M_1\lambda_1)) \\ & -(\sigma_1^2 + \lambda_1^2)\mu^2 + \sigma_1^2m_\varepsilon^2 + M_1^2 \\ = \; & (\sigma_1^2 + \lambda_1^2)(X - \mu)^2 - (\sigma_1^2 + \lambda_1^2)\mu^2 + \sigma_1^2m_\varepsilon^2 + M_1^2, \end{aligned}$$

if we set μ to:

$$\mu = \frac{\sigma_1^2m_\varepsilon + M_1\lambda_1}{\sigma_1^2 + \lambda_1^2}.$$

Let set $\sigma^2 = \frac{\sigma_1^2}{\sigma_1^2+\lambda_1^2}$, and replace in equation (15) to get :

$$\frac{\sigma}{2\pi}\frac{1}{\sigma}\exp(-\frac{1}{2\sigma^2}(X-\mu)^2).\exp(-\frac{1}{2\sigma_1^2}\left(\sigma_1^2 m_\varepsilon^2 + M_1^2 - (\sigma_1^2+\lambda_1^2)\mu^2\right)$$

so that:

$$A = \frac{\sigma}{\sqrt{2\pi}}\exp(-\frac{1}{2\sigma_1^2}\left(\sigma_1^2 m_\varepsilon^2 + M_1^2 - (\sigma_1^2+\lambda_1^2)\mu^2\right)\int_{X\geq 0} X.\frac{1}{\sigma}\varphi(\frac{X-\mu}{\sigma})dX = A_1 A_2$$

where:

$$A_1 = \frac{\sigma}{\sqrt{2\pi}}\exp(-\frac{1}{2\sigma_1^2}\left(\sigma_1^2 m_\varepsilon^2 + M_1^2 - (\sigma_1^2+\lambda_1^2)\mu^2\right)$$

Because:

$$A_2 = \int_{X\geq 0} X.\frac{1}{\sigma}\varphi(\frac{X-\mu}{\sigma})dX = E(X\mathbf{1}\{X\geq 0\}) = \mu\Phi(\frac{\mu}{\sigma}) + \sigma\varphi(\frac{\mu}{\sigma})$$

we thus get:

$$E\left(m_2 \mid \varepsilon, m_1\right) = \mu_2 + s_2.\varepsilon + \lambda_2 A_1\left[\mu\Phi(\frac{\mu}{\sigma}) + \sigma\varphi(\frac{\mu}{\sigma})\right].$$

In the truncated sample, we can also use that:

$$\begin{aligned}
V\left(m_2 \mid \varepsilon, m_1\right) &= \lambda_2^2.V\left((m_\varepsilon+v)\mathbf{1}\{m_\varepsilon+v\geq 0\} + \sigma_2.\eta_2 \mid \varepsilon, m_1\right) \\
&= \lambda_2^2.V\left((m_\varepsilon+v)\mathbf{1}\{m_\varepsilon+v\geq 0\} \mid \varepsilon, m_1\right) + \sigma_2^2, \\
&= \lambda_2^2.V\left(X\mathbf{1}\{X\geq 0\} \mid \varepsilon, m_1\right) + \sigma_2^2, \\
&= \lambda_2^2.(E\left(X^2\mathbf{1}\{X\geq 0\} \mid \varepsilon, m_1\right) - (E\left(X\mathbf{1}\{X\geq 0\} \mid \varepsilon, m_1\right))^2) + \sigma_2^2,
\end{aligned}$$

where $X = m_\varepsilon + v$ as before. It remains to evaluate:

$$B \equiv E\left(X^2\mathbf{1}\{X\geq 0\} \mid \varepsilon, m_1\right)$$

which by the same argument as above leads to:

$$B = A_1\int_{X\geq 0} X^2.\frac{1}{\sigma}\varphi(\frac{X-\mu}{\sigma})dX.$$

Furthermore:

$$\begin{aligned}
B_1 &= \int_{X\geq 0} X^2.\frac{1}{\sigma}\varphi(\frac{X-\mu}{\sigma})dX = \sigma^2.\int_{Y\geq 0} Y^2.\varphi(Y-\frac{\mu}{\sigma})dY, \\
&= \sigma^2((1+\frac{\mu^2}{\sigma^2})\Phi(\frac{\mu}{\sigma}) + \frac{\mu}{\sigma}\varphi(\frac{\mu}{\sigma})) = (\sigma^2+\mu^2)\Phi(\frac{\mu}{\sigma}) + \mu\sigma\varphi(\frac{\mu}{\sigma}).
\end{aligned}$$

which proves the equations in the Lemma:

$$\left\{\begin{array}{ll} E\left(m_2 \mid \varepsilon, m_1\right) = & \mu_2 + s_2.\varepsilon + \lambda_2.A_1A_2, \\ V\left(m_2 \mid \varepsilon, m_1\right) = & \sigma_2^2 + \lambda_2^2.\left[A_1B_1 - (A_1A_2)^2\right]. \end{array}\right.$$

A.2 The Success Probability and its Derivative

We start from:

$$\begin{cases} m_1 = \mu_1 + s_1.\varepsilon + \lambda_1 y + \sigma_1.\eta_1, \\ m_2 = \mu_2 + s_2.\varepsilon + \lambda_2 y + \sigma_2.\eta_2. \end{cases}$$

The selection through the first stage is given by the condition, varying across majors:

$$FS_{FSE} \geq FS^0_{FSE}$$

or equivalently as for the first stage grade:

$$m_1 = FS_{FSE} - 120 * ENEM/63 \geq m_1^0,$$

whose threshold m_1^0 varies between individuals because their grade at ENEM varies.

The selection into the second stage is given by the condition, varying across majors:

$$FG = 0.4FS_{FSE} + 0.6FS_{SSE} \geq FG^0$$

or equivalently, if we set $m_2 = FS_{SSE}$:

$$\begin{aligned} 0.4m_1 + 0.6m_2 \quad &\geq \quad FG^0 - 0.4 * 120 * ENEM/63 \\ &\iff \quad m_2 \geq \frac{FG^0 - 0.4 * 120 * ENEM/63 - 0.4m_1}{0.6} \\ &\iff \quad m_2 \geq m_2^0 - \omega m_1, \end{aligned}$$

which varies across individuals because grades m_1 and at ENEM vary.

The (full) success probability is given by:

$$\begin{aligned} p(y) &= \int_{m_1 > m_1^0, m_2 > m_2^0 - \omega m_1} f(\eta_1) f(\eta_2) d\eta_1 d\eta_2, \\ &= \int_{m_1 > m_1^0, \eta_2 > \frac{m_2^0 - (\mu_2 + s_2\varepsilon + \lambda_2 y) - \omega m_1}{\sigma_2}} f(\eta_1) f(\eta_2) d\eta_1 d\eta_2 \\ &= \int_{m_1 > m_1^0} f(\eta_1)(1 - F(\frac{m_2^0 - (\mu_2 + s_2\varepsilon + \lambda_2 y) - \omega m_1}{\sigma_2}))d\eta_1, \\ &= \int_{\eta_1 > \eta_1^0} f(\eta_1)(1 - F(\frac{m_2^0 - (\mu_2 + s_2\varepsilon + \lambda_2 y) - \omega m_1}{\sigma_2}))d\eta_1, \end{aligned}$$

where $\eta_1^0(y) = \frac{m_1^0 - (\mu_1 + s_1\varepsilon + \lambda_1 y)}{\sigma_1}$. We get:

$$\begin{aligned} p'(y) &= \frac{\lambda_1}{\sigma_1} f(\eta_1^0)(1 - F(\frac{m_2^0 - (\mu_2 + s_2\varepsilon + \lambda_2 y) - \omega m_1^0}{\sigma_2})) \\ &\quad + (\frac{\lambda_2 + \omega\lambda_1}{\sigma_2}) \int_{\eta_1 > \eta_1^0} f(\eta_1) f(\frac{m_2^0 - (\mu_2 + s_2\varepsilon + \lambda_2 y) - \omega(\mu_1 + s_1\varepsilon + \lambda_1 y + \sigma_1\eta_1)}{\sigma_2}))d\eta_1. \end{aligned}$$

Write the second integral on the RHS as:

$$I_2 = \int_{\eta_1 > \eta_1^0} f(\eta_1) f(-\frac{\omega\sigma_1}{\sigma_2}\eta_1 + \delta) d\eta_1$$

where:

$$\delta(y) = \frac{m_2^0 - (\mu_2 + s_2\varepsilon + \lambda_2 y) - \omega(\mu_1 + s_1\varepsilon + \lambda_1 y)}{\sigma_2}.$$

In the normal variate case:

$$\begin{aligned} f(\eta_1) f(-\frac{\omega\sigma_1}{\sigma_2}\eta_1 + \delta) &= \frac{1}{2\pi}\exp(-\frac{1}{2}(\eta_1^2 + (-\frac{\omega\sigma_1}{\sigma_2}\eta_1 + \delta)^2) \\ &= \frac{1}{2\pi}\exp(-\frac{1}{2}\left[\frac{(\eta_1 - \mu_\eta)^2 + 2\eta_1\mu_\eta - \mu_\eta^2}{\sigma^2} - 2\frac{\omega\sigma_1}{\sigma_2}\delta\eta_1 + \delta^2\right] \end{aligned}$$

where:

$$\sigma^2 = (1 + (\frac{\omega\sigma_1}{\sigma_2})^2)^{-1}.$$

Canceling the terms in η_1 yields:

$$\frac{\mu_\eta(y)}{\sigma^2} \equiv \frac{\omega\sigma_1}{\sigma_2}\delta$$

and the integrand of I_2 becomes:

$$\frac{\kappa(y)}{\sqrt{2\pi}.\sigma}\exp(-\frac{1}{2}\left[\frac{(\eta_1 - \mu_\eta)^2}{\sigma^2}\right]) \text{ where } \kappa(y) = \frac{\sigma}{\sqrt{2\pi}}\exp(\frac{1}{2}(\frac{\mu_\eta^2}{\sigma^2} - \delta^2)).$$

Thus:

$$I_2 = \kappa(y)(1 - \Phi(\frac{\eta_1^0 - \mu_\eta}{\sigma})),$$

and:

$$p'(y) = \frac{\lambda_1}{\sigma_1} f(\eta_1^0)(1 - \Phi(\frac{m_2^0 - (\mu_2 + s_2\varepsilon + \lambda_2 y) - \omega m_1^0}{\sigma_2})) + (\frac{\lambda_2 + \omega\lambda_1}{\sigma_2})\kappa(1 - \Phi(\frac{\eta_1^0 - \mu_\eta}{\sigma})).$$

The second derivative is obtained as:

$$\begin{aligned} p"(y) &= \frac{\lambda_1}{\sigma_1} f(\eta_1^0)\left[\frac{\lambda_1}{\sigma_1}\eta_1^0(1 - \Phi(\frac{m_2^0 - (\mu_2 + s_2\varepsilon + \lambda_2 y) - \omega m_1^0}{\sigma_2})) + \frac{\lambda_2}{\sigma_2} f(\frac{m_2^0 - (\mu_2 + s_2\varepsilon + \lambda_2 y) - \omega m_1^0}{\sigma_2})\right] + \\ &\quad (\frac{\lambda_2 + \omega\lambda_1}{\sigma_2})\left[\kappa'(y)(1 - \Phi(\frac{\eta_1^0 - \mu_\eta}{\sigma})) - \frac{\kappa}{\sigma}\frac{d(\eta_1^0 - \mu_\eta)}{dy} f(\frac{\eta_1^0 - \mu_\eta}{\sigma})\right] \end{aligned}$$

where:

$$\kappa'(y) = (\frac{\mu'_\eta\mu_\eta}{\sigma^2} - \delta'\delta)\kappa(y), \mu'_\eta = \sigma^2\frac{\omega\sigma_1}{\sigma_2}\delta', \delta' = -\frac{\lambda_2 + \omega\lambda_1}{\sigma_2}, \frac{d\eta_1^0}{dy} = -\frac{\lambda_1}{\sigma_1}.$$

Notice that:

$$p(y) = \int_{\eta_1 > \eta_1^0} f(\eta_1)\Phi(\frac{\mu_2 + s_2\varepsilon + \lambda_2 y + \omega m_1 - m_2^0}{\sigma_2}))d\eta_1,$$

that can be obtained easily by simulation.

A.3 Construction of the structural objects

We consider the approximation of the value function written as:

$$v(u) = \max_y \left[P_d(y).u - (y + c\frac{y^2}{2}) \right].$$

The algorithm is designed in such a way that computing $v(u)$ or its inverse can be performed using a grid parameterized by values of y. We start by approximating the integral $P_d(y)$ of $P'_d(y)$. Our objective is to make the approximation as simple as possible so that we could experiment with different degrees of approximation.

The approximation will be controled by one index K which is inversely related to the degree of thinness of the grid. Another index say K_1 controls for the precision of the approximation of the integral and we arbitrarily set that $K_1 = 5K$.

A.3.1 Approximation of $P_d(y)$

Let $f(y) = P'_d(y)$ and define $y_0 = \max(0, y^*)$ where y^* is the mode of the distribution:

$$y^* = \arg\max_y P'_d(y).$$

Arbitrarily, set the bounds on integration as:

$$y_+ = P'^{-1}_d(P'_d(y_0)/2K_1), y_- = \min(0, 2y^* - y_+).$$

Define the Simpson's rule of approximation for the integral over the whole range as:

$$A = \frac{\delta_A}{3}\left[P'_d(y_-) + 4\sum_{i=1}^{2K_1} P'_d(y_- + (2i-1)\delta_A) + 2\sum_{i=1}^{2K_1-1} P'_d(y_- + 2i\delta_A) + P'_d(y_+) \right],$$

where $\delta_A = \frac{y_+ - y_-}{4K_1}$ and the "nodes" are equally spaced.

We then approximate $P_d(y)$ at value 0 as:

$$G(0) = \frac{\delta_-}{3A}\left[P'_d(y_-) + 4\sum_{i=1}^{K_1} P'_d(y_- + (2i-1)\delta_-) + 2\sum_{i=1}^{K_1-1} P'_d(y_- + 2i\delta_-) + P'_d(0) \right]$$

where $\delta_- = -\frac{y_-}{2K_1}$. Note that $G(0) = 0$ if $y_- = 0$. We can also compute on the grid between 1 and K_1 the values:

$$g(2i\delta_+) = \frac{\delta_+}{3(1 - G(0))}\left[P'_d(2(i-1)\delta_+) + 4P'_d((2i-1)\delta_+) + P'_d(2i\delta_+) \right],$$

where $\delta_+ = \frac{y+}{2K_1}$ so that the approximation of $P_d(y)$ on the fine grid would be:

$$G(2k\delta_+) = G(0) + \sum_{i=1}^{k} g(2i\delta_+), k = 1, ., K_1.$$

In fact, we extract from the fine grid above $\{0, y_+\}$, a coarse grid according to $K = K_1/5$:

$$G(10k\delta_+) = G(0) + \sum_{j=1}^{k} \sum_{i=5(j-1)+1}^{5j} g(2i\delta_+), k = 1, ., K.$$

A.3.2 Computation and inversion of value functions

The likelihood function as defined in the next subsection depends on the evaluation of inequalities such as:

$$v_{d^*}(x\beta_{d^*} + \sigma_{d^*}\eta_{d^*}) > v_d(x\beta_d + \sigma_d\eta_d) \Longleftrightarrow \frac{v_d^{-1}(v_{d^*}(x\beta_{d^*} + \sigma_{d^*}\eta_{d^*}))}{\sigma_d} - x\frac{\beta_d}{\sigma_d} > \eta_d$$

so that we have to compute v_d^{-1} for any $d \neq d^*$ and be able to compute v_{d^*}.

The latter computation is easy since the relationship between value and utility can be obtained by maximizing for any value u:

$$v_{d^*}(u) = \max_y \left[P_{d^*}(y).u - y - c\frac{y^2}{2} \right],$$

over the coarse grid $\{0, y_+\}$.

As for the former computations, we use the following algorithms. We first compute a grid values of utilities such that:

$$u_d = \sigma_d(\frac{x\beta_{d^*}}{\sigma_d} + \text{Quantiles}(\varepsilon))$$

where ε follows a truncated normal distribution, truncated in such a way that $u_d > 0$. This is due to the fact that inverting v_d at the value v_{d^*} necessarily yields a positive value. We adjoin on the left (respectively on the right) the value 0 (respectively a large value).

At each of these new nodes, we again solve:

$$v_k(u_k) = \max_y \left[P(y).u_k - y - c\frac{y^2}{2} \right],$$

over the usual grid for y. We then use the grids in u and v to do a linear interpolation of v_d^{-1}.

A.3.3 Random drawings and simulated choice probabilities

We write choice probabilities (say of major 1) as:

$$\begin{aligned}\Pr(d^* &= 1 \mid \text{Application to UFC}) = \Pr(v_1 > v_j, \forall j \in \{2,.,D\} \mid \max_{k\in\{1,.,D\}} v_k > 0) \\ &= \frac{\Pr(v_1 > v_j, \forall j \in \{2,.,D\}, \max_{k\in\{1,.,D\}} v_k > 0)}{\Pr(\max_{k\in\{1,.,D\}} v_k > 0)} \\ &= \frac{\Pr(v_1 > v_j, \forall j \in \{2,.,D\}, v_1 > 0)}{\Pr(\max_{k\in\{1,.,D\}} v_k > 0)} \\ &= \frac{\Pr(v_1 > v_j, \forall j \in \{2,.,D\} \mid v_1 > 0)\Pr(v_1 > 0)}{1 - \Pr(v_j < 0, \forall j \in \{1,.,D\})}.\end{aligned}$$

Since the events $\mathbf{1}\{v_j > 0\} = \mathbf{1}\{u_j > 0\}$ (see text), we have that:

$$\Pr(d^* = 1 \mid \text{Appl. to UFC}) = \Pr(v_1 > v_j, \forall j \in \{2,.,D\} \mid u_1 > 0)\frac{\Pr(u_1 > 0)}{1 - \Pr(u_j < 0, \forall j)}.$$

We thus simulate two objects $\Pr(v_1 > v_j, \forall j \in \{2,.,D\} \mid u_1 > 0)$ and $\Pr(u_j < 0, \forall j)$.

For the first, we draw η_1^s in the distribution function of η_1 truncated by the condition that $u_1 = x\beta_1 + \eta_1^s > 0$. We then draw successively in the distribution function of η_2 given that $v_1^s > v_2$ and η_1^s and so on and so forth using the strategy of Geweke, Hajivassiliou and Keane. We obtain:

$$\begin{aligned}\Pr(v_1 &> v_j, \forall j \in \{2,.,D\} \mid u_1 > 0) = \\ \frac{1}{S}\sum_{s=1}^{S} \Pr(v_1^s &> v_2 \mid u_1^s).\Pr(v_1^s > v_3 \mid u_1^s, u_2^s)...\Pr(v_1^s > v_D \mid u_1^s, u_2^s, ., u_{D-1}^s).\end{aligned} \tag{16}$$

We do the same for the second object using the same underlying uniform draws though adapting to different truncation thresholds.

In the case of independence between random terms affecting utility levels, the likelihood function can be written as:

$$\frac{1}{S}\sum_{s=1}^{S} \Pr(v_1^s > v_2).\Pr(v_1^s > v_3)...\Pr(v_1^s > v_D).$$

When random terms η_d are dependent we build up each term in equation (16) using for any $d \neq d^*$:

$$\eta_d = \sum_{d'\in\mathcal{D}} \rho_{d',d}\eta_{d'} + \sqrt{1 - \sum_{d'\in\mathcal{D}} \rho_{d',d}^2}.\varepsilon_d$$

where coefficients ρ are obtained by the Choleski decomposition of the covariance matrix and ε_d has unit variance. We consider that $\mathcal{D} = \{d^*\} \cup \{d' \text{ lexicographically before } d\}$. The condition:

$$\eta_d < \frac{v_d^{-1}(v_{d^*}(x\beta_{d^*} + \sigma_{d^*}\eta_{d^*}^s)) - x\beta_d}{\sigma_d}$$

becomes:

$$\varepsilon_d < \frac{1}{\sqrt{1-\sum_{d'\in\mathcal{D}}\rho^2_{d',d}}}\left[\frac{v_d^{-1}(v_{d^*}(x\beta_{d^*}+\sigma_{d^*}\eta^s_{d^*}))-x\beta_d}{\sigma_d}-\sum_{d'\in\mathcal{D}}\rho_{d',d}\eta_{d'}\right].$$

Remark: In the case of linear multinomial Probit, the assumption on the covariance matrix of $(\eta_1,.,\eta_D)$ that insures formal identification (for instance, Keane, 1992) is:

$$\sum = \begin{pmatrix} 1 & c_0 & 0 \\ c_0' & \Sigma_0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

where Σ_0 and c_0 are unrestricted provided that Σ is a definite positive matrix. This normalization comes from setting one utility index to zero i.e. considering $(\eta_1-\eta_D,.,0)$ We saw in the text that we cannot use this normalization in the non linear case and that we would have to adapt this identifying assumption to a framework where η_D has unit variance. It is easy since it could consist in adding to the previous vector a unit-variance random normal variate η_D independent of $(\eta_1,.,\eta_{D-1})$ (since this dependence cannot be identified) so that we get:

$$\Sigma' = \begin{pmatrix} 2 & c_0+1 & 1 \\ c_0'+1 & \Sigma_0+J & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

where J is a matrix of ones.

In the general case of dependence we adopt this normalization even though it could be nor sufficient nor necessary in the non-linear case.

A.4 First counterfactual: details

The choice set is now composed of two majors:

$$(d_1,\varnothing),(d_1,d_2),.,(d_1,d_D),(d_2,d_1),(d_2,\varnothing),.,(d_2,d_D),.,(d_D,\varnothing)$$

Note that choices consisting of a single major $(d_i,\varnothing)$ could also be written (d_1,d_1).

First, note that some choices are always (weakly) dominated. Consider that the probability of success at stage 1 is ordered as the natural order $d_1,.,d_D$ i.e.

$$t_1(d_i) > t_1(d_j) \Longleftrightarrow \Pr(m_1 > t_1(d_i)) < \Pr(m_1 > t_1(d_j)) \text{ iff } i<j. \tag{17}$$

Then choices (d_j,d_i) are always dominated by $(d_j,\varnothing)$ since the option d_i can never be exercised. If the candidate does not access the second stage in major d_j, she cannot access the second stage in major d_i. We therefore chose to eliminate all choices (d_j,d_i) where $j>i$.

The value function of any choice (d_i, d_j) (where j can be equal to i) is therefore:

$$v(d_i, d_j) = -y + p_{d_i}(y; t_1(d_i), t_2(d_i))u_i + (p_{d_j}(y; t_1(d_j), t_2(d_j)) - p_{d_j}(y; t_1(d_i), t_2(d_j)))u_j$$

as derived from equation (13) in the text. Note that if $i = j$, the second term is equal to zero and if $i < j$,

$$p_{d_j}(y; t_1(d_j), t_2(d_j)) > p_{d_j}(y; t_1(d_i), t_2(d_j)),$$

because of equation (17). Denote $p_{ij}(y) = p_{d_j}(y; t_1(d_i), t_2(d_j))$ so that:

$$v_{ij}(y) = -y + p_{ii}(y)u_i + (p_{jj}(y) - p_{ij}(y))u_j. \tag{18}$$

The algorithm runs as follows:

- Compute for a grid of values of y and for any $i \leq j$, $p_{ij}(y)$ in the order $(p_{11}, p_{12}, .., p_{1D}, p_{22}, ., p_{2D}, ., p_{DD})$.
- Simulate values for $(u_1, ., u_D)$ drawn in a multivariate normal distribution truncated by $\max_d u_d > 0$.
- Compute for any (d_i, d_j), for the grid of values y and for any simulated u_s of $(u_1, ., u_D)$, $v_s(d_i, d_j; y, u_s)$.
- Maximize $v_s(d_i, d_j; y, u_s)$ on the grid y to obtain $v_s^*(d_i, d_j; u_s)$ as specified in equation (18).
- Compute the frequency estimator of $\Pr((d_i, d_j) = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\{v_s^*(d_i, d_j; u_s) \geq \max_{l,m} v_s^*(d_l, d_m; u_s)\}$.

The expected value is obtained by:

$$\sum_{(d_i,d_j)} \frac{1}{S} \sum_{s=1}^{S} v_s^*(d_i, d_j; u_s).\mathbf{1}\{v_s^*(d_i, d_j; u_s) \geq \max_{l,m} v_s^*(d_l, d_m; u_s)\}.$$

B Data appendix

B.1 Description

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

Roughly, the Vestibular has the following features:

1. The student choose the undergraduate degree before the test, and compete only against those students who made the same choice;

2. It is comprised of many sub-exams, each one evaluating knowledge in Mathematics, Physics, Chemistry, Biology, Portuguese, History, Geography and a Foreign Language;

3. The exams are almost exclusively developed with objective (multiple choice) questions;

4. Different undergraduate courses can weight the sub-exams differently in order to reflect their priorities in terms of required knowledge;

5. More than one stage is allowed during the process of testing.

6. Almost all universities developed their own exam, however its is possible to form groups of universities to develop unified exams;

7. After the exams, students are ranked accordingly to a pre-determined protocol applied to their grades in the exam. Places are filled from top to bottom, and if there are remaining free spots a period of recall of students are made;

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

exam as many times as they want to.

B.2 The Vestibular at UFC

The 2008 Vestibular was taken by 31,304 students who disputed the chance of getting one of the 4,085 available places. This overall student by place ratio of 7.66 is not a good estimator for some specific undergraduate degrees. For instances, Medicine had the highest ratio of 24.1, Law had a ratio of 16.7. In contrast, Information Systems had a ratio of 1.1 and Economics had a ratio of 4.5.

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

The first stage, called General Knowledge (GK), is composed of a unique 66 objective questions (multiple choice, with five alternatives A, B, C, D and E) exam whose content is exactly the core high school curricula, i.e., Portuguese (Grammar and Writing), Geography, History, Biology, Chemistry, Mathematics, Physics and Foreign Language . In order to understand the grading system for this first exam note that there are two types of scores: raw and standardized, respectively. The raw score for each subject is given in Table 1 and the standardized scores in Table 2.

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

1. Get a in each subject appearing in the GK exam;
2. After 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

Table 1

Subject	**Number of Question in the GK Test**	**Value per Question**	**Total Value or x_r**
Portuguese	12	3	36
Geography	8	3	24
History	8	3	24
Biology	8	3	24
Chemistry	8	3	24
Mathematics	8	3	24
Physics	8	3	24
Foreign Language	6	1	6
TOTAL (X_r^{GK})			**186**

Source: Elaborated by the authors.

Table 2

Portuguese	$x_s = 36 + 7.2\left(\frac{x_r - \overline{x_r}}{\sigma_r}\right)$
Geography, History, Biology, Chemistry, Mathematics and Physics	$x_s = 24 + 4.8\left(\frac{x_r - \overline{x_r}}{\sigma_r}\right)$
Foreign Language	$x_s = 6 + 1.2\left(\frac{x_r - \overline{x_r}}{\sigma_r}\right)$

Source: Elaborated by the authors.

calculated based on the following rule: Let N be the number of available places in a specific major previously shown in the Edital. Let r be defined as the ratio of the number of students choosing the major and the number of available seats in the major. If $r < 10$ then the threshold is $3N$, otherwise it is $4N$. Note that the threshold is not known by the candidate when choosing majors. This information is disclosed ex-post the major choice.

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

Table 3

Subject	**Number of Question in the GK Test**	**Number of Question in the SK Test**	**Value per Question**	**Total Value or x_r**
Writing		-	-	80
Specific Knowledge	Specific 1	8	10	80
	Specific 2	8	10	80
TOTAL (X_r^{SK})				**240**

Source: Elaborated by the authors.

The two specific exams are set according to requirements of each major. Again, this list in known ex-ante the choice of major and is given by the following table:

Table 4

Course	Specific Exams
Biblioteconomy; Social Sciences; Social Communication – Journalism; Social Communication - Publicity and Advertising; Law; Musical Education – Graduate; Fashion and Style**;** Philosophy; Letters	Portuguese and History
Geography; History; Pedagogy	History and Geography
Domestic Economy; Physics Education – B. Sc.; Physics Education – Graduate; Psychology	History and Biology
Architecture and Urbanism	History and Physics
Computing; Civil Eng.; Computing Eng.; Mechanical Production Eng.; Tele-informatics Eng.; Electrical Eng.; Mechanical Eng.; Statistics; Physics – B. Sc.; Physics – Graduate; Mathematics – B. S.c.; Mathematics – Teaching.	Physics and Mathematics
Chemistry Eng.; Chemistry – B. Sc.; Chemistry – Teaching	Chemistry and Mathematics
Administration; Actuarial Sciences; Science Accounting; Economic Sciences;.	History and Mathematics
Agronomy; Food Engineering.	Biology and Mathematics
Executive Secretary; Information Systems	Portuguese and Mathematics

Source: Elaborated by the authors.

The standardized scores are calculated according to the following formulas:

Table 5

Writing	$x_s = 80 + 16\left(\frac{x_r - \overline{x_r}}{\sigma_r}\right)$
Specific 1	$x_s = 80 + 16\left(\frac{x_r - \overline{x_r}}{\sigma_r}\right)$
Specific 2	$x_s = 80 + 16\left(\frac{x_r - \overline{x_r}}{\sigma_r}\right)$

Source: Elaborated by the authors.

The sum of all standardized scores taken in the second stage gives . The sum of all first stage standardized scores and all second stage standardized scores gives the final grade (FG = +). All students are ranked again and the available places are allocated to the best ranked students.

Group	Subgroup	Subsubgroup	Majors
Humanities & Social Sciences	Humanities	Biblioteconomia	Biblioteconomia
		Comunicacao Social (Jornalismo)	Comunicacao Social (Jornalismo)
		Comunicacao Social (Publ e Prop)	Comunicacao Social (Publ e Prop)
		Filosofia (Noturno)	Filosofia (Noturno)
		Letras (Portugues)	Letras (Portugues)
		Letras (Portugues-Alemao)	Letras (Portugues-Alemao)
		Letras (Portugues-Espanhol)	Letras (Portugues-Espanhol)
		Letras (Portugues-Frances)	Letras (Portugues-Frances)
		Letras (Portugues-Ingles)	Letras (Portugues-Ingles)
		Letras (Portugues-Italiano)	Letras (Portugues-Italiano)
	Other	Economia Domestica	Economia Domestica
		Educacao Fisica	Educacao Fisica
		Estilismo e Moda	Estilismo e Moda
	Social Sciences	Ciencias Sociais	Ciencias Sociais
		Geografia	Geografia
		Historia	Historia
		Pedagogia	Pedagogia (Diurno) Pedagogia (Noturno)
		Psicologia	Psicologia
Law & Business	Accountancy	Ciencias Atuariais (Noturno)	Ciencias Atuariais (Noturno)
		Ciencias Contabeis	Ciencias Contabeis (Diurno) Ciencias Contabeis (Noturno)
	Administration	Administracao	Administracao (Diurno) Administracao (Noturno)
		Secretariado (Noturno)	Secretariado (Noturno)
	Economics	Ciencias Economicas	Ciencias Economicas (Diurno) Ciencias Economicas (Noturno)
	Law	Direito	Direito (Diurno) Direito (Noturno)
Medicine	Medicine	Medicina	Medicina
		Medicina - Barbalha	Medicina - Barbalha
		Medicina - Sobral	Medicina - Sobral
	Pharmacy, Dentist & Other	Ciencias Biologicas	Ciencias Biologicas
		Enfermagem	Enfermagem
		Farmacia	Farmacia
		Odontologia	Odontologia

Group	Subgroup	Subsubgroup	Majors
Sciences	Agrosciences	Agronomia	Agronomia
		Eng. de Alimentos	Eng. de Alimentos
		Eng. de Pesca	Eng. de Pesca
		Zootecnia	Zootecnia
	Engineering	Arquitetura e Urbanismo	Arquitetura e Urbanismo
		Eng. Civil	Eng. Civil
		Eng. Eletrica	Eng. Eletrica
		Eng. Mecanica	Eng. Mecanica
		Eng. de Producao Mecanica	Eng. de Producao Mecanica
	Mathematics	Computacao	Computacao
		Eng. de Teleinformatica	Eng. de Teleinformatica
		Estatistica	Estatistica
		Matematica	Lic. em Matematica (Noturno) Matematica (Diurno)
	Physics & Chemistry	Eng. Quimica	Eng. Quimica
		Fisica	Fisica (Diurno) Lic. em Fisica (Noturno)
		Geologia	Geologia
		Quimica	Lic. em Quimica (Noturno) Quimica - Licenciatura
		Quimica - Bacharelado	Quimica - Bacharelado
		Quimica Industrial	Quimica Industrial

Table 1: The tree structure of majors

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: UFC Vestibular 2004

Table 2: Number of applications and positions and success probabilities

Subgroup	10th percentile	Min	Min	Median			Maximum		
	All	First stage	Pass	All	First stage	Pass	All	First stage	Pass
Other	18	28	30	32	44	46	58	58	58
Physics & Chemistry	22	0	0	37	42	46	62	62	62
Humanities	18	27	29	32	43	47	60	60	60
Social Sciences	18	27	29	34	45	47	61	61	60
Accountancy	23	36	36	38	46	48	59	59	59
Economics	21	31	37	35	44	49	61	61	61
Administration	19	31	35	34	46	49	62	62	62
Mathematics	21	0	0	39	48	50	62	62	62
Engineering	24	33	33	43	50	53	63	63	63
Pharmacy, Dentist & Other	20	34	40	38	50	52	62	62	62
Law	21	46	47	44	55	57	63	63	63
Medicine	24	47	51	51	58	58	63	63	63

Table 3: Summary statistics of initial grades in the samples of (1) all, (2) pass after first stage (3) definite pass after second stage

Subgroup	10th percentile	Min	Min	Median			Maximum		
	All	First stage	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

Table 4: 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)

Major	10th percentile	Min	Min	Median			Maximum			Observations
	All	First stage	Pass	All	First stage	Pass	All	First stage	Pass	
Barbalha	66.19	182.05	186.86	152.62	191.67	199.62	214.29	214.29	214.29	739
Sobral	121.57	185.05	186.86	171.76	196.52	200.76	214.38	214.38	214.19	542
Fortaleza	93.05	193.67	193.86	172.95	202.57	208.57	224.29	224.29	224.29	2325

Table 5: Summary statistics of initial grades in the samples of (1) all, (2) pass after first stage (3) definite pass after second stage

(Medicine sample composed by three majors: Barbalha, Sobral and Fortaleza)

Table 6: Descriptive statistics in the three choice-based medical majors

Barbalha:

Variable	Mean	(Std. Dev.)	Min.	Max.	N
Grade: National Exam	45.053	(10.906)	12	62	739
Grade: First stage	140.237	(47.975)	19	214.286	739
Grade: Second stage	240	(34.91)	129.449	322.63	160
Female	0.484	(0.5)	0	1	739
Age	19.574	(2.458)	16	25	739
Private High School	0.633	(0.473)	0	1	739
Repetitions	0.667	(0.818)	0	2	739
Father's education	1.786	(1.053)	0	3	739
Mother's education	1.955	(1.042)	0	3	739

Sobral:

Variable	Mean	(Std. Dev.)	Min.	Max.	N
Grade: National Exam	50.297	(7.278)	18	61	542
Grade: First stage	164.681	(32.894)	35	214.381	542
Grade: Second stage	240	(33.984)	94.3	296.649	160
Female	0.469	(0.499)	0	1	542
Age	19.689	(2.378)	16	25	542
Private High School	0.855	(0.34)	0	1	542
Repetitions	0.987	(0.882)	0	2	542
Father's education	2.085	(1.046)	0	3	542
Mother's education	2.218	(0.994)	0	3	542

Fortaleza:

Variable	Mean	(Std. Dev.)	Min.	Max.	N
Grade: National Exam	49.253	(10.028)	12	63	2325
Grade: First stage	160.925	(42.672)	25	224.286	2325
Grade: Second stage	240	(34.374)	48.301	311.105	598
Female	0.538	(0.499)	0	1	2325
Age	19.246	(2.303)	16	25	2325
Private High School	0.771	(0.405)	0	1	2325
Repetitions	0.691	(0.830)	0	2	2325
Father's education	2.135	(0.999)	0	3	2325
Mother's education	2.153	(0.978)	0	3	2325

Coefficients	First stage	Second stage	Common
Grade at National Exam	15.83	2.66	0.16
	(0.497)	(2.75)	(0.05)
λ	16.65	25.93	
	(2.55)	(4.12)	
Intercept	65.66	234.30	-0.61
	(2.08)	(3.12)	(0.53)
Standard errors	8.65	33.42	
(σ)	(1.47)	(1.16)	
Nobs	3606		
Likelihood	-18210.71		

Table 7a: Estimates of grade functions in medicine (simple specification)

Coefficients	First stage	Second stage	Common
Intercept	37	222	2.76
	(2.32)	(6.9)	(0.346)
Age=16	1.80	8.06	0.184
	(2.16)	(7.56)	(0.255)
Age=21	-2.52	-3.03	-0.00441
	(0.93)	(3.73)	(0.130)
Age=25	-0.691	-33.3	-0.797
	(1.26)	(7.52)	(0.183)
Private High School	3.30	5.72	0.0609
	(1.42)	(6.45)	(0.185)
Preparatory Course	0.938	9.00	0.0293
	(0.798)	(3.76)	(0.108)
Nb of repetitions	-0.138	4.38	0.351
	(1.07)	(2.37)	(0.132)
Mothers' education			0.0254
			(0.0367)
Fathers' education			0.0516
			(0.0333)
Grade at national exam	1.95	2.91	1.64
	(0.973)	(1.22)	(0.165)
Lambda	8.85	5.91	
	(0.577)	(–)	
Standard errors	5.74	33.5	
	(0.802)	(1.07)	
Nobs		3606	
Likelihood		-17975.284	

Table 7b: Estimates of grade functions in medicine (complete specification)

Variables	Barbalha	Sobral	Fortaleza
Enem	-1.38	0.005	
	(0.21)	(0.008)	
Intercept	-5.91	-0.25	0
	(0.84)	(0.10)	(-)
Standard errors	5.28	0.24	1
	(.73)	(.09)	(-)
Nobs		3606	
Likelihood		-3142.83	

Table 8: Preference estimates in medicine

		Barbalha	Sobral	Fortaleza
First stage	Current	182.0476	185.0476	193.6667
	Counterfactual	202.4889	186.2268	186.8603
Second stage	Current	236.0816	237.8046	240.8465
	Counterfactual	264.6613	218.9603	219.6025

Table 9: Current and counterfactual thresholds for passing at first and second stage.

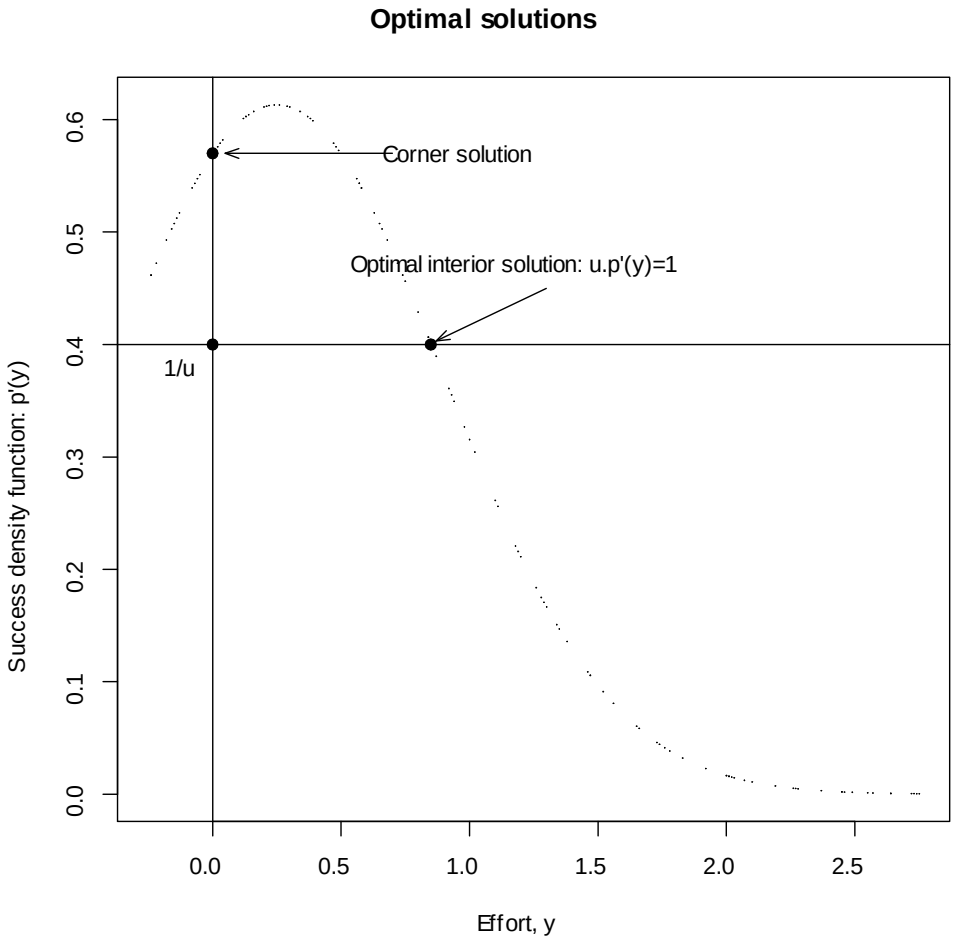

Figure 1: The determination of optimal effort
(Cost of effort is linear, $c = 0$)

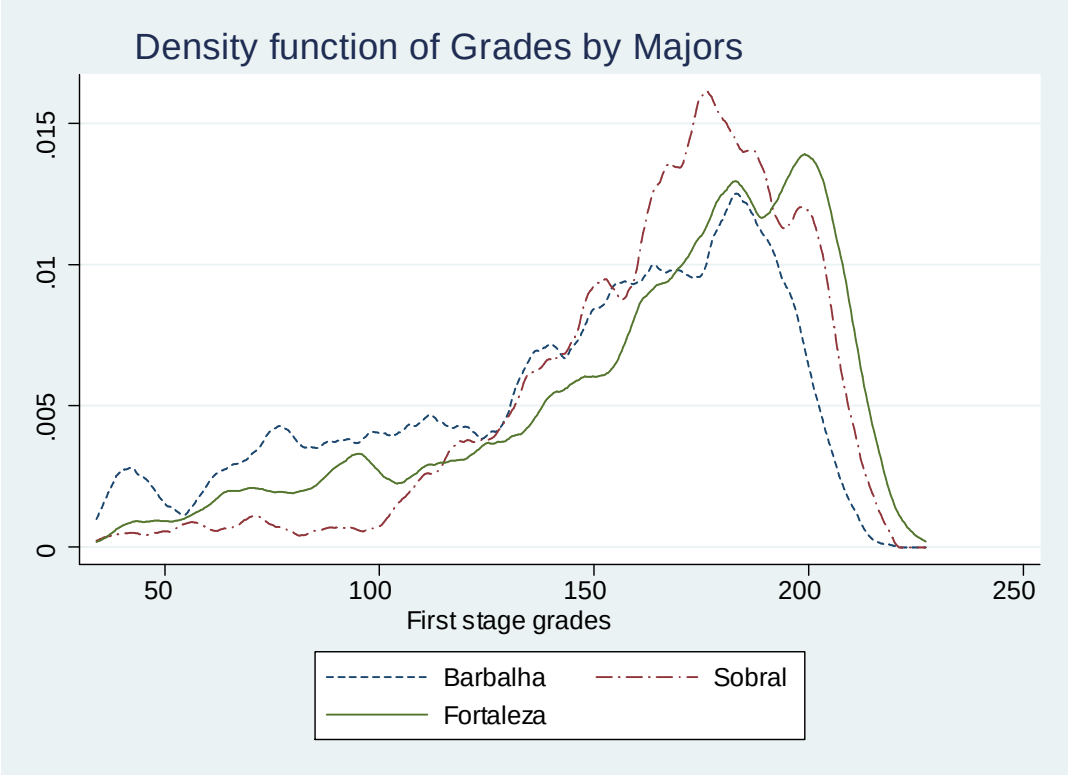

Figure 2: Density function of First Stage Grades by Majors

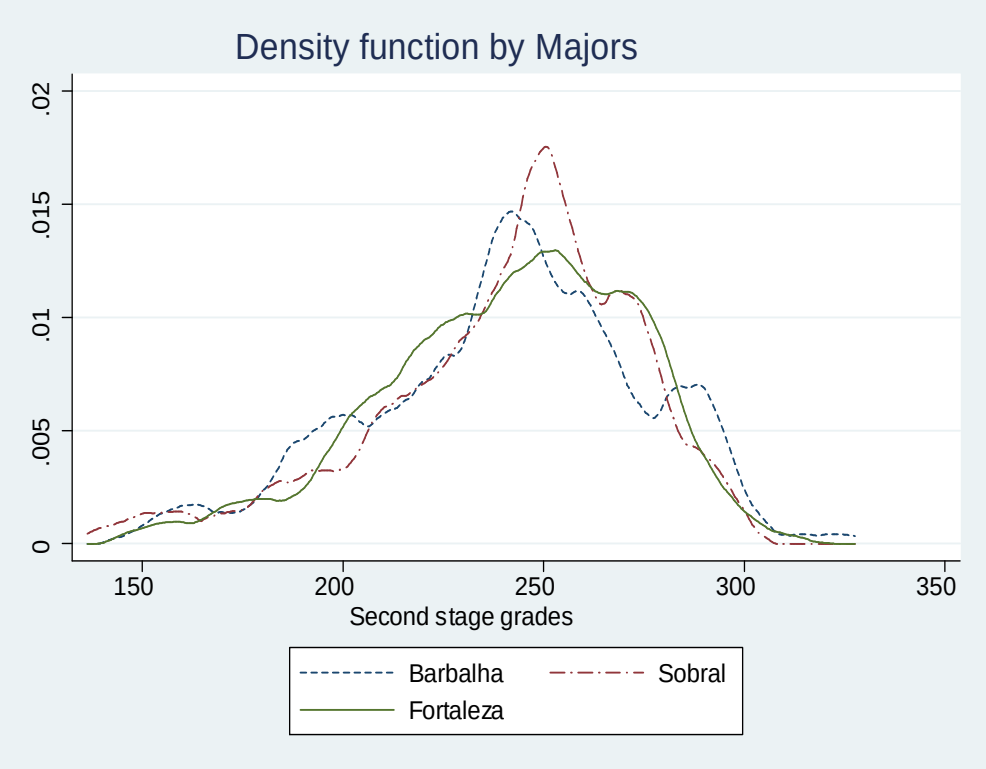

Figure 3: Density function of Second Stage Grades by Majors

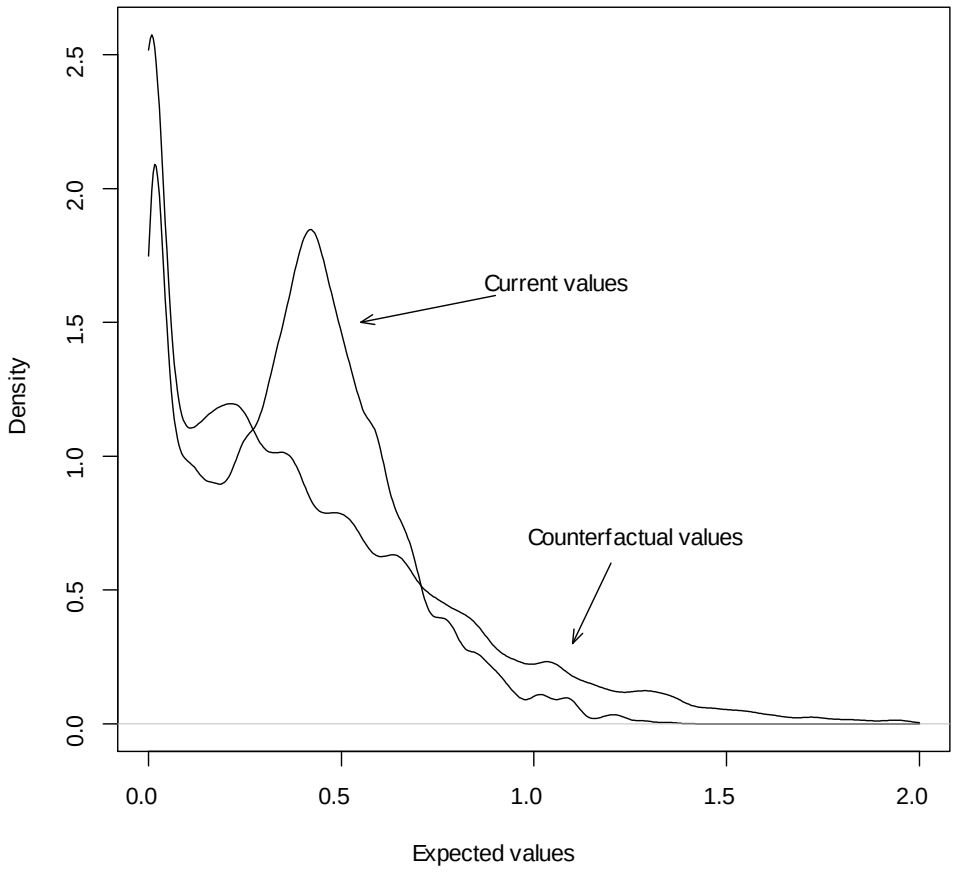

Figure 4: Current and Counterfactual density functions of the value functions (Simple specification)

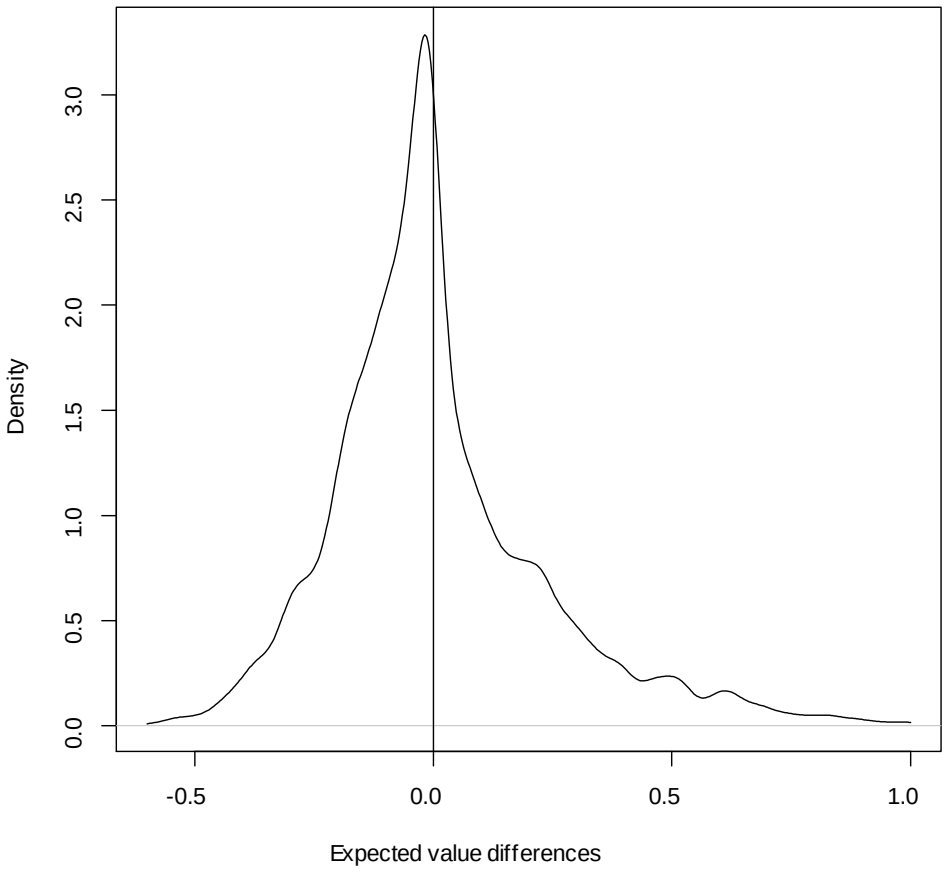

Figure 5: Density functions of the change in the value functions.
(Simple specification)