UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 01-59
Economics Series 19
December 2001

Departamento de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624 98 75

UNDERSTANDING PREFERENCE FORMATION IN A MATCHING MARKET *

Ricardo Mora and Antonio Romero-Medina[2]

Abstract

We analyze the role of formal and informal information gathering in students' preference formation. We analyzed this role in the college admission process using Spanish individual data. We introduce students' risk aversion and information costs on the standard college admission problem. Then, we model the students' list formation as a two-stage procedure. In first stage, students must decide whether they gather information or not about a college. In the second stage, they give their preferred list to the matching office. The observed changes in preferences suggest that information gathering is important in the last two months of the process and that students with less ex-ante information are more affected by these changes.

Keywords: Matching Markets, Preference Formation, College Choice.

JEL Classification: C78, D73, C52.

[1] **R. Mora**, Departamento de Economía, Universidad Carlos III de Madrid; E.mail: ricmora@eco.uc3m.es. Phone: (34) 91 624 9576

[2] **A. Romero-Medina**, Departamento de Economía, Universidad Carlos III de Madrid; E.mail: aromero@eco.uc3m.es. Phone: (34) 91 624 9752

***** Romero-Medina acknowledges financial support from DGI, Grant BEC-2002-02194, and Mora acknowledges financial support from DGI, Grant BEC2000-0170. Both authors acknowledges financial support from Comunidad de Madrid project 1083. We would like to thank Jordi Massó, Ignacio Ortuño, Barbara Petrongolo, Raquel Carrasco, Mª Angeles de Frutos, Andrés Perea, Lars Ehlers and participants to the Microeconometrics and Theory workshops at UCIIIM for their helpful comments.

1. Introduction

The process that matches students and universities resembles a market with some peculiar features. Demand and supply of positions are discrete variables. Prices (academic fees) are not flexible and cannot play the usual role of adjusting supply and demand, although there is evidence that grants and loans can alter the cost of college attendance and that students' decisions are responsive to changes in related prices and subsidies (see Paulsen [8] and [7]). Most importantly, college choice has long-term effects on students' future income. Thus, efficiency considerations make it important not only to take students' preferences into account, but to understand the uncertainties they face and how the information available on college tuition affects preferences.

In this paper, we model and test how the information acquisition process and the cost of gathering information affects preference formation in a centralized model of college admission. The model is an extended version of the two-sided many-to-one matching model which has been extensively studied since the seminal work of Gale and Shapley [2] (two excellent surveys are Roth and Sotomayor [12] and Gusfield and Irving [3].) We introduce risk aversion and cost of gathering information in the student's decision and test the importance of information on preference formation exploiting individual Spanish data from a 1990 survey including both high school and university students.

Students often form images of schools based on inaccurate or limited information (Kotler and Fox [5]). In this paper, students face uncertainty in future returns from education in a specific college. Such uncertainty can nevertheless be reduced by collecting college information. In our model, students first decide whether to gather information about each college. Once this process is completed, they fill in the list that they report to the matching office. Under this setting, the student's behavior is consistent with a model of risk aversion and information costs.

We test the importance of information on preference formation by a two-fold strategy. First, we study the probability that students would always choose the same college as a first option irrespective of their high school results (satiation). Then we proceed to analyze the probability of acceptance in the first choice. Our results are important as they highlight the fact that the availability of information, coming from both formal and informal channels, does affect students' preferences and choices. In this context, lowering the cost of information gathering would be the simplest way to set up an allocation mechanism not affected by unequal

opportunities.

The rest of the paper is organized as follows. We give a brief description of the Spanish college admission process in 1989 in Section 2. Section 3 deals with the model whilst Section 4 contains a description of the data and presents the empirical results. Finally, Section 5 offers some concluding comments.

2. The Spanish System of College Admissions in 1989

In Spain, colleges admit students coming from high school by using a ranking based on a grade, which reflects previous performance, and the student's high school major (sciences, art, and joint studies). Given that each college is assigned a target quota by an external administrative body, the ranking leaves no room for colleges to influence their admission lists. The performance grade is obtained computing an average between the students high school final grade and the result from a regional examination called PAAU[1]. We will refer to this average as the PAAU grade. The PAAU grade is a two-digit number between 0.00, the lowest score, and 10.00, the maximum score that can be achieved. Students who want to go to college must get a minimum score of five. If several students with identical high school majors and the same PAAU grade are tied for the last position available in a given college, then all are admitted in the college.[2] Before taking the exam, students submit an ordered list of colleges to the central matching office. On the list, there is a maximum number t of options which varies from one region to another. At the period of reference, the lowest limit is eight possible declarations. However, for some regions the "limit" is given by the size of the application form, which is, in fact, greater than the number of options available.

Each university offers positions for different degrees, such as physics, history, economics, etc. To keep with the traditional terminology in the matching literature, we refer to each one of these schools as a separate college. High school students are assigned to their closest university for the PAAU exams. Only the absence of a given degree can justify the application for a position in an university different from that originally assigned. In practice, only a very small percentage of students are allowed to study in a region other than the one they live in and most students simply apply to the public universities of their region. At the time

[1]PAAU is the Spanish acronym for University Aptitude Entry Examination.

[2]Thus, target quotas do not neccesarily coincide with achieved quotas. However, the situation only arises in a very small number of cases, as ties are extremely rare and do not usually involve more than two students.

of reference, there are only three districts, Madrid, Catalonia, and the Basque Country, in which students can choose amongst the different Universities within the district. Choice is thus reduced to type of training within the local university for most of the students and type of training and university within the district for students in Madrid, Catalonia, and the Basque Country.

Once regional matching offices have received all lists, a version of the Deferred Acceptance Algorithm (DA) is used to allocate students to colleges within each district or university. The Student-Proposing Deferred Acceptance Algorithm used by the offices can be described by the following iteration:

Step 1: Each student makes an offer to the first college on her preference list of acceptable colleges. Each college rejects the offer of any student who is unaceptable to the college, and each college that receives more than one offer "holds" only the q students with the highest PAAU grade (where q is the quota of students assigned to the college).

Step k: Any student whose offer was rejected at the previous step makes an offer to her next choice (i.e. to her most preferred college among those which have not yet rejected her), so long as there remains an acceptable college to which she has not yet made an offer. If a student has already offered a position to, and has been rejected by all the colleges that she finds acceptable, then she makes no further offers. Each college receiving offers rejects those from unacceptable students, and also rejects all but its q most preferred students among the set consisting of the new offers together with any student that it may have held from the previous step.

Stop: The algorithm stops when no student's offer is rejected. At this step, every student is either being held by some college or has been rejected by every college on her list of acceptable colleges. The output of the algorithm is the matching at which each student is matched to the college where she is held when the algorithm stops. Empty college positions and students who where rejected by all their acceptable colleges remain unmatched.

This matching process leads to a stable matching that is Student-optimal in the sense that all students unanimously consider the result to be the best stable matching for them. At the end of the procedure the grade of the last student admitted to a given college, the so-called "cut-off" grade, is made public. Before the period of reference, entry grades had not historically shown large variations from year to year.[3] It is therefore reasonable to assume that in 1989 students

[3]This system was first implemented in the 1980s. For a number of colleges, quotas were not binding, and "cut-off" grades were systematically the minimum score (5 points). Many

foresee with accuracy the grade that will be required to enter a given college.

There are four reasons to test preference formation in the Spanish college admission system in the period of reference. First, colleges play no role in the process. There is evidence that in a system where the college has an active role, there are factors other than the admittance test that are relevant for college decisions (see Manski *et al.* [6] for an extensive comment on this issue). In the Spanish system, however, this is not the case, and admission in a given college depends only on the admittance test and the student's high school curriculum. Colleges are distributed among students according to the lists given by the students to the public office and do not set policy on admissions or establish additional criteria for admittance. This simplifies the decision that students face and allows us to isolate the analysis on one side of the market.

Second, students' behavior is predictable within our model. Although the Spanish system is potentially open to manipulation in the short run as there are limits in the number of options available (see Romero-Medina [9]), we claim that students were in fact submitting sincere lists to the clearing office in 1989. This situation changed gradually in the years after 1989. As many colleges raised their entry requirements, students started filling their lists completely. For example, by 1999, most students in Madrid were submitting lists with 15 options included, the maximum number allowed.

Third, the source of uncertainty in the process of college choice is identified. There are, *a priori*, three sources of uncertainty which are related to the students' information acquisition process: (a) Students submit the lists before they know their PAAU grade. However, students can anticipate these grades with accuracy and, in fact, they have two days to change their lists after their grade is known. Not surprisingly, very few of them do it. (b) "Cut-off" grades are unknown to the students at the time of submitting the lists, but they had remained low and constant previous to 1989. (c) Finally, there is uncertainty pertaining to the true characteristics of the colleges. We assume that this is the fundamental source of uncertainty that students face in the period of reference.

Finally, gathering college information is an option available to all Spanish students especially since applications are mostly to local universities. This situation has changed in the last years. On the one hand, improvements in the information publicly available through formal channels has been substantial. On the other

colleges, however, did have "cut-off" grades that were greater than 5 and would vary from year to year. Nonetheless, these variations were never greater than 0.34 points, a very small range of variation.

hand, more students have been allowed to apply to non-local universities.

3. A Behavioral Model of Matching

We consider a market with n students and l colleges. Let $S = \{s_1, \ldots, s_n\}$ and $C = \left\{c^1, \ldots, c^l\right\}$ be the set of students and colleges, respectively.

Students have some prior, possibly incomplete, information about the quality of colleges. With this information student s_i forms her *ex-ante* preferences. We shall assume that these preferences are based on the expected monetary payments student s_i will receive for attending college c^j. In particular, her preferences depend on earnings which are conditional on the success of her training at each college.

Let $\overline{c_i}^j$ be the monetary earnings, net of the monetary value of time, effort and fees spent in college, for student s_i of attending and successfully finishing her training in college c^j in case c^j is a "good option". Similarly, let $\underline{c}_i^j$ represent the monetary earnings of attending (and perhaps dropping out from) college c^j in case c^j is a "bad option". The function

$$C_i : \left\{\bar{c}_i^1, \underline{c}_i^1, \bar{c}_i^2, \underline{c}_i^2, ..., \bar{c}_i^l, \underline{c}_i^l\right\} \to \mathbb{R}$$

takes into account both the monetary net earnings of attending a college and the fact that those earnings depend on the affinity between students' expectations and performances. Thus C_i does not refer only to college quality, but also to the complementarity between the college and the student.

We assume that students have a binomial probability distribution over each college. Thus, c^j can be a good option for student s_i with probability $\alpha_i^j \in [0, 1]$, and a bad option with probability $(1 - \alpha_i^j)$. Then each college becomes a lottery such that:

$$c_i^j \equiv \alpha_i^j \overline{c_i}^j + (1 - \alpha_i^j)\underline{c}_i^j$$

Students have preferences $\succsim_i$ on a lottery space $\pounds_i$ satisfying the Von-Neuman-Morgenstern axioms. Let $u_i(c_i^j) : \pounds \to \mathbb{R}$ be the utility function of student s_i,

$$c_i^j \succ_i c_i^k \longleftrightarrow u_i(c_i^j) > u_i(c_i^k)$$

Therefore, $u_i(c_i^j)$ is the utility that s_i has for studying at college c^j. Let us call $\bar{c}_i^j$ the lottery $\alpha_i^j u_i(\overline{c}_i^j) + (1 - \alpha_i^j)u_i(\underline{c}_i^j)$ when $\alpha_i^j = 1$, and $\underline{c}_i^j$ the lottery

$\alpha_i^j u_i(\overline{c}_i^j) + (1-\alpha_i^j)u_i(\underline{c}_i^j)$ when $\alpha_i^j = 0$. A student who is not admitted to any college reaches a utility level $u_i(\emptyset) = 0$. That is, $c = \emptyset$ represents the situation in which the student is not admitted by any college. We also require students' preferences to be strict, i.e., for any two colleges $c^k, \sigma \in C \cup \emptyset$, $c^k \neq \sigma$, either $u_i(c_i^k) > u_i(c_i^\sigma)$ or $u_i(c_i^\sigma) > u_i(c_i^k)$. If $u_i(\emptyset)$ is greater than the utility of attending c^k, then c^k is called an unacceptable college for s_i in the sense that s_i will remain unmatched rather than attend college c^k.

Let us assume that $u_i(c_i^j)$ is strictly-concave, i.e., the students are risk averse.

The utility for each college depends on the set of students it admits, say $S^j \subseteq S$. Let $u^j : 2^S \longrightarrow \mathbb{R}$ be the utility function of college c^j. A college which does not admit any student obtains $u^j(\emptyset)$. Each college c^j has a predetermined quota q^j representing the maximum number of students it can admit. Any set of students $\hat{S}$ containing more than q^j students is such that $u^j(\emptyset) > u^j(\hat{S})$.

A matching describes which college (if any) admits each student and vice-versa. More precisely, a *matching* μ is a correspondence that maps $S \cup C$ into $S \cup C \cup \emptyset$ such that (a) for each $s_i \in S$, if $\mu(s_i)$ does not belong to C, then it is the empty set; (b) for each c^j in C, $\mu(c^j)$ is contained in S or is equal to $\emptyset$, and (c) for each pair $(s_i, c^j) \in S \times C$, $\mu(s_i) = c^j$ if and only if s_i belongs to $\mu(c^j)$.

We are interested in college allocations that are stable. A matching is stable if the following two conditions are satisfied. The first one is individual rationality: each agent weakly prefers the payoff resulting from matching rather than being unmatched. The second one is collective rationality: it is not possible for a college and a group of students to reallocate in such a way that both, the college and the students find the new situation more profitable.

A matching is stable if it is robust to deviations by a coalition of one college and a group of students, or by a college or a student alone. These coalitions are the only essential coalitions in this environment.

3.1. The Student's Decision Process

We assume that students have a subjective *ex-ante* idea of the probability of being admitted to a given college, in other words, the probability of achieving a given grade in the admittance test. We denote the probability of the student s_i to be admitted to college c^j as p_i^j.

The model follows the argument in Kotler and Fox ([5]) by assuming that students form images of schools based on inaccurate or limited information. In particular, students face uncertainty on future returns from completing college

education in a particular college. However, this uncertainty can be reduced by collecting information about each specific college. This process is costly so that to collect information about college c^j the student s_i must pay a cost ζ_i^j. Once this cost is paid, the student knows with certainty whether attending that college is a good or bad option.

Students submit an ordered list of colleges to the matching office. We shall say that a student s_i submits a list according to her true preferences P_i when she ranks all the acceptable colleges according to her utility from best to worse. Otherwise she is misrepresenting her preferences. In this list $P_i = c^1, c^2, ..., s_i, c^l, c^{l+1}, ..., c^m$ the college c^1 is s_i's preferred college, c^2 is her second choice, and so on. In this list, all colleges after s_i are unacceptable options for s_i. From now on, we will not include s_i, or any unacceptable college, on P_i for notational convenience.

Given P_i, the Students-optimal DA Algorithm is used to allocate students into colleges. Given that the probabilities p_i^j are known by student s_i[4], the expected utility of the list P_i is:

$$U_i(P_i) = \left(\sum\nolimits_{h=1}^{l} p_i^h u_i(c_i^h) \prod\nolimits_{j=1}^{h} (1 - p_i^{j-1}) \right).$$

where, for notational convenience and without loss of generality, we add up only those options acceptable according to P_i and c_i^1 is the preferred college for student s_i, c_i^2 is her second choice, an so on.

Note that we are assuming that if a student obtains a given grade on her exam she will be accepted to a given college. That is, we are assuming that entry grades are fixed. If we assume a probability distribution on college preferences, which is what we have before the PAAU exams take place, our framework becomes closer to Roth and Rothblum [11]. However, their model is not totally equivalent to ours, as they consider a model where the Colleges-optimal DA Algorithm is used and colleges can admit only one student.

In order to eliminate uncertainty over the quality of college c^j, a student s_i must incur a cost ζ_i^j. As an example, if s_i has *ex-ante* complete information on college c^j, then α_i^j is either 1 or 0. We say that student s_i is informed about college c^j and $\zeta_i^j = 0$. The maximum amount that a student s_i is willing to pay for the information on the quality of a given college c^j is defined by the following condition:

$$U_i(P_i \mid \zeta_i^{j*}) = U_i(P_i^{j*})$$

[4]These are subjective assessments made by student s_i based on the known behaviour of the college and the distribution of preferences and PAAU grades.

That is, the maximum fee ζ_i^{j*} that a person would pay, in advance of receiving the information on the true characteristics of college c^j, is such that it makes the expected utility of the best informed action exactly equal to the expected utility of the best uninformed action on college c^j, $U_i(P_i^{j*})$. Based on student s_i true preferences, P_i, let us denote as $\dddot{P}_i(c^j, \underline{c}_i^j) \equiv \dddot{P}_i(\underline{c}_i^j)$ the list of colleges ordered from best to worse according to s_i preferences where $u_i(c^j)$ has changed to $u_i\left(\underline{c}_i^j\right)$. Let us also denote as $\ddot{P}_i(\bar{c}^j)$ the list of colleges ordered from best to worse according to s_i preferences where $u_i(c^j)$ has changed to $u_i\left(\bar{c}^j\right)$, and the student's information about the other colleges remains unchanged. Clearly, $\dddot{P}_i(c^j, c^j) \equiv P_i$. The value of information on college c^j quality expressed in utility terms is $U_i(P_i) - U_i(P_i^{j*})$. Therefore the maximum fee ζ_i^{j*} that a student would pay in advance of receiving the complete information on college c^j is such that:

$$\left[\alpha_i^j U_i(\dddot{P}_i(c_i^j, (\bar{c}_i^j - \zeta_i^{j*}))) + (1-\alpha_i^j) U_i(\dddot{P}_i(c_i^j, (\underline{c}_i^j - \zeta_i^{j*})))\right] = U_i(P_i^{j*})$$

To model the *ex-post* preference formation process, we propose the following two-step procedure. Each student s_i has an *ex-ante* ranking of alternatives in terms of the expected utility of each of the colleges. In the first step, the student s_i must decide whether or not to collect information on each college. Let us call $\check{P}_i$ the *ex-ante* preference list for s_i. In the second step, the student s_i chooses the list that maximizes her expected utility. Let us call $\hat{P}_i$ the *ex-post* preference list for s_i. This list sorts colleges from best to worse and may contain both colleges where she has collected information and colleges where she has not.

- First step:
  - If $\left[\alpha_i^j U_i(\dddot{P}_i(c_i^j, (\bar{c}_i^j - \zeta_i^j))) + (1-\alpha_i^j) U_i(\dddot{P}_i(c_i^j, (\underline{c}_i^j - \zeta_i^j)))\right] \geq U_i(P_i^{j*})$, she will collect information for c^j, and
  - if $\left[\alpha_i^j U_i(\dddot{P}_i(c_i^j, (\bar{c}_i^j - \zeta_i^j))) + (1-\alpha_i^j) U_i(\dddot{P}_i(c_i^j, (\underline{c}_i^j - \zeta_i^j)))\right] < U_i(P_i^{j*})$, she will not collect information.
- Second step: she chooses the list of colleges $\tilde{P}_i$ that maximizes her expected utility given the information collected during the first stage.

$$\tilde{P}_i \equiv \arg\max \left(\sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^{j-1}) \right).$$

3.2. Strategic Behavior and the PAAU

It is well known that under the mechanism associated to the *DA Algorithm*, it is a dominant strategy for students to declare their true preferences (Roth [10]). However, limits on the number of options that can be declared can change this property. In this case, the *DA Algorithm* generates stable allocations according to the preferences declared by the students, but is no longer a dominant strategy for students to declare their true preferences.

In practice, Spanish students at the period of reference usually did not declare as many options as they had available. Therefore, they were either discarding most of their feasible options or truncating their preferences. In theory, two types of truncation are found: First, students declare truncated strategies as defined by Roth and Vande Vate [13] and Roth and Rothblum [11].

Definition 1. *Given P_i containing k acceptable colleges a* truncation (from below) *of P_i is a strategy P_i' containing $k' \leq k$ acceptable colleges such that the k' elements of P_i' are the first k' elements of P_i, in the same order.*

Second, students remove non-feasible alternatives from the top of their list. This strategic behavior was described in Romero-Medina [9] as a way to restore stability when only a limited number of options can be declared.

Definition 2. *A* truncation from above *of a preference list P_i containing k acceptable colleges is a list P_i' containing $k' \leq k$ acceptable colleges such that the k' elements of P_i' are the* last *k' elements of P_i, in the same order.*

In our model, there is no benefit for students from either including unaceptable colleges or changing the order of acceptable colleges in the list submitted to the matching office. Moreover, there is no gain derived from using truncation of any kind. The result can be summarized in Theorem 3.1.

Theorem 3.1. *If students can submit their preference lists without restrictions, then no student can gain from a misrepresentation of preferences.*

Thus, introducing risk aversion and costs of gathering information enriches the model but does not alter its properties relating to strategic behavior. As in the mechanism without risk aversion and cost of gathering information (Roth [10]),

it is a dominant strategy for students to declare their true preferences when the student-optimal *DA Algorithm* is applied.

However, given that students do not usually declare as many options as they have available, it may be the case that they are not including their complete list of acceptable colleges. A justification of such behavior is presented in Corollary 1. Truncation from below arises as natural behavior for students when they are certain that they will be admitted to one of the colleges in the submitted list.

Corollary 1. *Let P_i^k be a list P_i of s_i acceptable colleges truncated to its first k elements. Let all the colleges in P_i be such $p_i^h > 0$. Then $U_i(P_i) = U_i(P_i^k)$ if and only if the probability to obtain at least one of the colleges in P_i^k is equal to one.*

Truncation from above arises in our two-step setting for the student's decision model because true preferences change as a result of the process of information gathering. Therefore, once we rule out the possibility of strategic behavior in the application process, the only factor that can lead to changes in the preferences of risk-averse students is the costly acquisition of information.

Two factors will contribute to the change in preferences. On the one hand, the fact that a student decides to acquire information about some colleges and not about others might produce changes in the ranking. Colleges that have greater *ex-post* utility will rise in the ranking and colleges with lower *ex-post* utility will fall in the *ex-post* ranking. On the other hand, colleges with *ex-ante* greater variance are ranked lower than alternatives with less variance but the same expected value.

Let us illustrate these results with a particular example.

Example 1. *For a student s_i, let $\check{P}_i = c^1, c^2, c^3, c^4$ be the ex-ante preference list, $\hat{P}_i = \bar{c}^2, \bar{c}^4, \bar{c}^3, c_i^1$ be the ex-post preference list, and $p_i^4 = 1$. Finally, let $\tilde{P}_i = c^2, c^4$ be the final list submitted to the matching office.*

In the example, student s_i has decided to gather information about colleges c_2, c_3 and c_4. College 1 is preferred *ex ante* but not *ex post*. Due to the result on Corollary 1, although c^1 is the most preferred *ex-ante*, it will not be on s_i final list $\tilde{P}_i = c^2, c^4$. Based on a comparison of $\check{P}_i$ and $\tilde{P}_i$ without taking into account the information acquisition process, an independent observer could falsely conclude that student s_i is acting strategically.

3.3. Sources of Information

Students have two sources of information, namely institutional and informal sources of information. Whilst students acquire informal information through a wide range of channels, they are exposed to institutional information only in high school centers. In particular, institutional information is provided by the educational institutions freely and unconditionally. According to the nature of the information transmitted, institutional information can be either "menu" or "personal". Students receive "personal" information at interviews with psychologists and through IQ tests. This helps students to focus on a smaller number of options that are likely to be good matches. Thus, "personal" information changes parameters α_i^j and may totally reduce the uncertainty on future returns from completing a particular college education. "Personal" information can be exhaustive in the sense that it satisfies the student's informational demand and no further information acquisition process should be expected. In contrast, "menu" information is given to the student by means of written materials, oral presentations and audiovisual information. "Menu" information opens the spectrum of possibilities for the student and reduces the costs of acquiring further information. Thus, when "menu" information is provided, an intensification of the information acquisition process may be expected.

It is important to distinguish between the provision of "menu" information, which takes place at educational centers, from the acquisition of information using sources obtained through "menu" information. When teachers hand out leaflets to the students, they are providing "menu" information that dramatically reduces costs ζ_i^j. Later on, when students decide to read a selected subset of these leaflets at home, they are acquiring information "informally". Informal sources of information can be of a quite varied nature, ranging from social networking to reading leaflets. However, in all these methods of information acquisition, the student willingly decides to acquire the information and incurs a cost for doing so.

Another distinction between institutional and informal information arises from considering the timing of their provision. It is natural to think of institutional information as one-off events, like the IQ test, or the interview with the school psychologist. These events will likely change the students's preferences and may potentially affect the acquisition of informal information. Friendship and family networks should also play a crucial role in the way preferences change. However, information flows through these channels continuously, and it is plausible that students get information through informal channels until the last day before filling in the lists.

3.4. Testable Hypotheses in the Student's Decision Process

Differences in $\check{P}_i$ and $\hat{P}_i$ are due to the acquisition of information before submitting the preference list to the matching office. When neither $\check{P}_i$ nor $\hat{P}_i$ are available to the researcher, it is still possible to assess the importance of the process of information acquisition by looking at features in students' preferences that change with the arrival of new information and are observable to the researcher.

In this paper, we study two such features. The first one is the probability of being satiated, i.e. the probability of always choosing the same college regardless of the grade at the PAAU examinations. The second feature is the probability that a student is not restricted by the matching algorithm, i.e. her best option is achievable.

It is of interest to know which students are most likely to change their preferences. According to our model, the greater the chances to be admitted in a college, the higher the chances that a student will seek information about a particular college. This can be seen by assuming that student s_i acquires information about college c^j. If p_i^j increases, student s_i will still acquire information about college c^j. However, there is always a value for p_i^j below which student s_i will not longer seek information on college c^j. Likewise, the greater the cost of acquiring information on a particular college, the lower the chances that students will seek information on that college. To illustrate this point, consider that student s_i does not acquire information about c^j at cost ζ_i^j and $\tilde{\zeta}_i^j > \zeta_i^j$, then student s_i will not acquire information about c^j at cost $\tilde{\zeta}_i^j$. Costs of acquiring information will be individual-specific since individual characteristics, such as, for example, the parent's educational levels, affect them. Thus, individual characteristics will influence the likelihood of information acquisition and, therefore, the likelihood of a change in preferences. In addition, institutional information will also affect both preferences and the following process of information acquisition. All types of information, both formal and informal, may change preferences directly. In addition, when information is of a "menu" type, reducing costs leads to new information and possibly a change in preferences through informal information acquisition.

4. The Data Set

The information we use is in CIS [1], "Los jóvenes ante el sistema educativo". This is a national survey carried out in April 1990 by the *Centro de Investigaciones Sociológicas* (CIS) with the cooperation of the Spanish Government. It consists

of an initial sample of 7993 university students and 3770 high school students. We restrict our analysis to ensure that the students in the sample are under the same admission process by selecting last-year high school students that are considering entering college in the future and first-year university students coming from high school. We also restrict the sample to include students who will choose or have chosen colleges within the same local area. After applying these filters, the resulting sample size is 1163 high school students and 3177 college students.

Although some of the questions in the survey were specific to university and some were specific to high school students, most of them were common or comparable. In particular, high school students were asked questions regarding their preferences and first options at the time, while university students were asked about the declaration they submitted to the matching office and their grades at the entry examination in the year they went into university.

The survey was carried out in April 1990. That was two months before high school students in their last year were required to submit their lists to the matching office. Therefore, we have information on *ex-ante* preferences for high school students and *ex-post* preferences for university students.

The survey includes two questions about the preferences of high school students. The first one, question 22, has three sections.

Question 22: *Which college would you choose if your PAAU grade were...*?

1. (Q2201) *between 5 and 6.5 points.*
2. (Q2202) *between 6.5 and 8 points.*
3. (Q2203) *between 8 and 10 points.*

The second question of interest for our analysis is question 24, where we find information on candidates' expectations about their declarations in the college matching protocol.[5]

Question 24 (Q24): *Which college will you most likely end up choosing as your first option*?

[5]In the survey, there was another question reflecting how preferences would change if the matching algorithm was changed and the PAAU grade was made irrelevant: Question 23. (*If it were just up to you, that is, the PAAU grade were irrelevant, which college would you choose*?). The question thus provides information about the value students assign to the PAAU grade *per-se*.

We also have information about the choices that current university students made when they went through the PAAU matching process. First, for each university student we have information about her first, second and third choice - Questions Q1101, Q1102, and Q1103 respectively. We also know the college in which they entered (Question Q13) and its position in the list of options (Question Q12). Q24 for high school students is naturally associated with Q1101 for university students. Unfortunately, students were not asked the length of their submitted lists in the survey and there are no public figures available on this issue. It is thus impossible to asses directly whether students used all options available or truncated from below.

As it was previously discussed, truncation is not enough to restore stability in the system when limits in the number of options to declare are present. In this context, sophisticated students using "non-reverse" strategies can restore stability. One such strategy is truncation "from above", that is, dropping best choices and moving the "window of options" down to more achievable targets. As already mentioned, we cannot directly test whether students are using non-reverse strategies against declaring their true preferences. However, to the extent that the limit of options available for students to declare was not binding, the significance of non-reverse strategies to obtain stability would become irrelevant in this particular case.

In the Spanish system at the time of the survey, there are two features that make, in our opinion, the limit in the number of options irrelevant. First, the "cut-off" grades are low for a large share of the colleges. Around a third of the total number of colleges accept all students whose PAAU grade is higher than the minimum required to enter any college. Second, the number of options varies from region to region but it is never smaller than eight[6]. In addition, in Table 1, we present the percentage of students in the University sample that declared less than three options. We also show the distribution of Q12, that is, the position in the list of the college where every student was finally admitted. Around 37% of students declared less than three options and around 95% of students entered into

[6]It can be argued that the relevant number of options is not the number of colleges offered by the local university, but the number of different entry grades. This is because students only need to choose the best option within the set of colleges with the same entry grade. Only in 5 out of 22 districts can students have more than eight relevant options and only in three of them is the number of options larger than 12: Madrid, Catalonia, and the Basque Country. More importantly, if we restrict the options that students are considering to those most compatible with the courses that they have taken in high school, then all students actually face less than eight relevant options.

one of their first three options. We have fitted a negative binomial distribution for Q12. The estimated probability of non-admittance in one of the first 8 entries in the list was smaller than 10^{-7}. These results are consistent with the view that students are not constrained by the limit in the number of options in the lists, and that Q24 and Q1101 reflect the true best options of high school and college students respectively.

Insert Table 1 around here

In the survey, differences in the empirical distribution of Q24 and Q1101 arise from three different sources: First, students still get information on colleges in the last two months before submitting their preference lists to the matching office. Second, cohorts might have significant differences in college preferences as these are related to long-term earnings expectations which change with the evolution of labor market conditions. Finally, the list of college options in the survey is not exactly the same as the list of college options in the real process and, in order to answer the questionnaire, students had to adapt the true list codes to the survey codes. We now discuss each of these three problems in detail.

4.1. Measures of Information

In the interviews, students were asked to state whether they had information on colleges in their high school centres from any of the following means: tests, interviews with psychologists, written information, talks and conferences, and videos. For those students already in college, the questions referred to the information they had before taking the PAAU exams. Thus, by comparing the frequencies of college and high school students that were exposed to these means of information, we can assess to what extent institutional sources of information had been employed two months before taking the PAAU exams. In Table 2, we present the percentages of informed students within each group.

Insert Table 2 around here

The timing of the release of formal information to students depends on the channel employed. The data suggest that the most frequent channels of "menu" information, *written information* and *talks*, and to a lesser extent *videos*, had already been offered to students two months before taking the PAAU examinations.

We also construct two measures of institutional informational exposure: *personal* and *menu*. Any student is said to have received "personal" information if she has been exposed to either interviews or tests. In that case, *personal* takes the unity. A student is said to have received "menu" information if she has been exposed to either written information, talks, or videos.

Information obtained from casual or informal exposure is more difficult to measure. In our data, we have potential variables that may help us control for these effects, such as the father's work status and education. Nonetheless, we think that the notion of informal education is broader than these proxies. For example, friendship and family networks should play a crucial role in the way students shape their preferences (Hossler et al. [4].)

4.2. Cohort effects

Cohorts might have significant differences in college preferences as these are strongly linked to long-term expected earnings associated to each degree. We address this heterogeneity by restricting the analysis to high school students who are considering entering college in October 1990, and university students in their first year at college coming from high school. By doing so, we maximize the proportion of students that were born within the same two years, namely 1971 and 1972.

4.3. College codes

Finally, the list of college options in the survey is a simplified version of the list of college options in the real process. Since our aim is to determine the importance of information acquisition in the preference formation process, we assume that all distributional differences between Q24 and Q1101 are due to mismatches in the codings. We set up a restricted clustering algorithm for colleges so that the new codings in Q24 and Q1101 are comparable, i.e. the distribution of the best options in the university sample must be similar to the distribution of the best option in the high school sample . The algorithm works as follows: we assume that high school students encounter classification problems only for colleges within the same area of knowledge and with the same minimum university entry requirements. Thus only these colleges can be clustered into one single college category. The clustering algorithm searches for the two colleges such that once clustered into one category, the overall distribution of college categories in Q1101,the university sample, and Q24, the high school sample, look closer in terms of the χ^2 test

for differences in the frequencies. We proceed recursively until the difference in distributions is not statistically significant at the 99 percent level. Of course, this strategy implies that when we use the new college codings, we bias our analysis by minimizing the impact of the information acquisition process in the last two months before submission of the preference list.[7]

We find that using the original survey codes leads to strongly rejecting the null hypothesis of the equality in the distribution of colleges in the two questions. Once the amalgamation procedure finishes, we still have 27 college categories. Within these college categories, the distribution of questions Q24 and Q1101 -the first option two months in advance and the first option for the cohort of students from the previous year- are not statistically different.

In the following section, we present results using both the original survey codes for colleges and the cluster college categories. The latter will be referred to as "cluster codes" from now on.

5. The results

As indicated by Corollary 1, truncation from below is a natural outcome of utility maximization. As the Spanish system was not binding in the limit of options available, we assume that strategic behavior is irrelevant. In this section, thus, we study the importance of information gathering in the last two months prior to the PAAU exams.

In order to study the process of information in the last two months before submitting the lists, we must compare a feature of the preferences of high school and college students. Our testing strategy consists of a two-step procedure. First, we study the probability that students would always choose the same college as a first option irrespective of their High School results (satiation). Then we proceed to analyze the probability of acceptance in the first choice.

5.1. The empirical model of satiation

A student shows satiation in college preferences whenever she chooses the same college irrespective of her PAAU grade. The survey allows us to identify satiated students amongst the sample of high school students. The logical condition that

[7] A detailed description and the results of the clustering algorithm can be found in Appendix II.

defines a satiated student in the sample is the equality in the answers of questions Q2201, Q2202, Q2203, and Q24.

As similar information is not available for university students, we cannot identify satiated students amongst them. Therefore, it is not possible to study the role of information acquisition in the last two months before submitting the lists for satiation. Nevertheless, it is still possible to assess the importance of the different information channels in the characterization of satiation as a feature of students' preferences. In order to do so, we use a probit specification for a dummy variable which reflects whether or not the high school student is satiated.

The empirical model of satiation includes four socio-demographic variables which potentially affect the students preferences and satiation. Dummy variable *Attended Public School* takes value 1 whenever the student attends a state high school. The variable *Female* refers to the student gender while *Lost at least 1 year* takes value 1 whenever the student had to repeat at least one year. Finally, *Considered training courses* shows whether the student was interested in abandoning high school education in order to enter training courses.

Dummy variables reflecting the student's motivation to choose a particular college are *Vocation*, *Good marks in the field*, *Money*, *To stay in the same city*, and *To leave the city*. The variables *Sciences* and *Arts* are related to two of the three majors students can take in high school, the third one being *Joint studies*.

In order to study the role of information acquisition on preference formation, the most relevant variables in the model of satiation are linked to the student's family background and the official information already received two months before submitting the list. In particular, we include dummy variables for *Family tradition* and *Parents' will* as a reason to attend a college. We also include three variables describing the father's socio-economic status: whether or not he is a manager (*Manager*); whether or not he is a worker with no qualifications (*Father unskilled*); and whether or not he holds a college degree (*Father attended College*). Unfortunately, there is no information in the survey related to the mother's social background. However, given the observed strong tendency for couples to sort themselves according to socio-economic status, variables relating to the father can reasonably be understood to proxy the family's status as a whole.

Finally, we include dummy variables describing whether the student has already received "menu" type information, *Menu*, and whether or not she has already received "personal" type information, *Personal*.

Results of the probit estimates are presented in Table 3. Column 1 presents the point estimates, Column 2 shows the probability changes when the corresponding

dummy variable changes its value from 0 to 1 (evaluated at average values for the other controls), and, finally, p-values are reported in Column 3. We present results using original survey codes.

Insert Table 3 around here

A number of variables do not seem to affect the probability of being satiated. These include *Attended Public School*, *Woman*, *Considered training courses*, *Good marks in the field*, *To stay in the same city*, *Manager*, *Father unskilled*, and the official information variables *Menu* and *Personal*. Some variables decrease the probability of satiation, although the estimated effect is not significant. They are *To leave the city*, *Arts*, and *Parents' will*.

Two dummy variables increase the probability of satiation by at least 20 percentage points each. The first one, *Family tradition*, might be capturing the effect of unofficial information, whilst the second one, *Vocation*, suggests that the probability of satiation is highly influenced by the existence of very strong preferences in favor of a particular college.

Finally, four variables are found to significantly reduce the chances of satiation. Students who previously lost at least 1 year, those who will choose colleges for the money opportunities later on, those whose father had been in college, and finally those whose major is Sciences, are less likely to choose the same college regardless of the PAAU grade that they obtain. The strongest effect is found amongst Science students, who are less likely to be satiated than the reference Joint Studies students by 32 percentage points.

To sum up, our results suggest that satiation is only partly related to information acquisition. Whilst family background may be important in changing the likelihood of the student being satiated, official information, either of the "menu" or of the "personal" type, does not have any effect on it. We also find that some "restricted" students (those who lost at least 1 year and those who are studying Sciences) are less likely to be satiated.

5.2. Information acquisition in the last two months

We have yet not provided a measurement on how important the problem of information acquisition in the last two months is. In our model, differences in $\check{P}_i$ and $\hat{P}_i$ for last-year high school students are due to the acquisition of information. We need to study a feature in the preferences that changes with new information and

it is observable and comparable in the two samples. We choose the probability that a student is not restricted by results in PAAU examinations, i.e. her best option is always achievable.

For university students, we identify unrestricted students as those who entered in their first choice. For high school students, we identify restricted students as those who gave different answers, using cluster codes, to questions Q24 and Q2203. Since Q2203 is their best option if they had the highest mark in the PAAU exam, this is a conservative condition for the definition of restricted students. In particular, so-defined restricted students are only a subset of the real set of restricted students since there will be some individuals with Q24 equal to Q2203 who will end up in a different college than the first option. However, given that students can accurately assess their chances to enter any college and that the proportion of individuals entering into their first option is near 75 percent, the importance of this group is likely to be minor.

Changes between the university sample and the high school sample in the probability of being restricted reflect, *ceteris paribus*, the influence of information acquisition in the last two months of the process. Our testing strategy is twofold. We first present the results of the comparison of the frequencies of unrestricted individuals for specific groups in the two-month period. We also carry out means tests for these percentages and the results are presented in Table 4. Then we model the probability of being unrestricted as a probit model. The results for the probit model are presented in Table 5.

Insert Table 4 around here

The percentage of university students who entered into their first option is around 74.25 percent. In contrast, we see that only around 46.75 percent of high school students reveal that they are not restricted by the exam. Conditioning by type of student yields the same basic result: there is an increase of around 27 percentage points in first option admittances in the last two months. Most importantly, differences are larger for uninformed students, those with low-education parents, and science students. This result suggests that there is a fundamental process of information acquisition that restricts the number of candidate colleges in the last two months. As a result of this process, the probability of an individual entering into the first option must increase.

To test this interpretation, we proceed by estimating a model of the probability of being unrestricted controlling for the sample definition of being unrestricted.

For simplicity, we will present here the results of a probit specification although other specifications, such as the linear probability and the logit models, yield similar results. A useful variable specification, assuming no control variables apart from the information-related dummies, takes the following form:

$$\begin{aligned}\Pr(Unrestricted \mid & Menu, Personal, University) = \Phi(Constant + \\ & + \beta_1 \cdot Menu + \beta_2 \cdot Personal \\ & + \gamma_1 \cdot Menu \cdot Univ + \gamma_2 \cdot Personal \cdot Univ \\ & + \gamma_3 \cdot Menu \cdot Personal \cdot Univ \\ & + \theta \cdot NotInf \cdot Univ)\end{aligned}$$

where $Univ$ takes value 1 for the university sample and $NotInf$ takes value 1 if the student has not been informed through official channels. This variable specification allows us to identify the effect of information on each stage. First, the effect of unofficial information two months in advance is reported in the constant term, together with other factors whose effects are constant within the sample. Note that we do not know whether university students received their official information before or after the two-month limit. However, if we assume that the effect of the provision of official information is the same before or after the two-month limit, then the effect of "menu" information is estimated by the parameter $\beta_1$whilst the effect of personal information is estimated by β_2. It follows that θ shows the effect of unofficial information in the last two months when individuals have not received any official information. When the student receives "menu" information, then the effect of unofficial information in the last two months will be reported in γ_1. Likewise, when individuals receive "personal" information, this effect will be γ_2 and, when they receive both types of official information, it will be γ_3.

In addition to the information variables, the model of restriction includes the same control variables as the model of satiation. Results of the probit estimates are presented in Table 5. As in the case of the model of satiation, standard measures of goodness-of-fit, as the Cramer R^2 or the pseudo-R^2 are low, although the success prediction rate reaches 76 per cent of the cases. However, in spite of this seemingly disappointing low fit, the model clearly outperforms the constant model in the sense that the Wald test for the joint slope coefficients is strongly rejected.

Insert Table 5 around here

A number of socio-economic variables do not have significant effects on the probability of being unrestricted. Amongst them, we find *Attended Public School*, *Female*, *Money*, *To stay in the same city*, and more interestingly, *Family tradition*, *Father informed*, and *Father unskilled*. It could be argued that, according to our model, students with a strong family tradition may have very valuable information that would undoubtedly help them in adjusting to their best option. However, this reading of the results would not be without controversy, as the variable itself reflects whether the student would choose a college because of family tradition, and not whether the family helps in the information process. In contrast, the estimates for both *Father attended College* and *Father unskilled* are of the expected signs, but not significant.

All the other socio-economic variables are significant at conventional levels, perhaps with the exception of *Manager*, with a p-value of 7 per cent. A number of coefficients have negative values, showing that individuals with those characteristics are more likely to be restricted in the sense that they cannot reach their best option. These include *Lost at least 1 year*, *Considered training courses*, *To leave the city*, and *Sciences*. In contrast, *Vocation*, *Good marks in the field*, *Arts*, *Parents' will*, and *Manager* all show positive and significant estimates. Of course, both *Parents' will* and *Manager* are proxies for an influential family background which helps the student obtain information. Thus, their effect should be counted as part of the unofficial information effect on preference formation two months before submitting the lists.

With respect to the information gathering variables, we find some interesting results. First, "menu" information two months before does not have any significant effect, but it does have a very strong negative effect in the last two months, suggesting that individuals who receive "menu" information widen their prospects and tend to include more "unreachable" colleges than students without this information. The effect of "personal" information is almost the opposite: a negative sign two months before, although only significant at the 5 per cent significance level, and a very significant effect in the last two months. The effect of receiving both types of information and the effect of not-receiving any type of information are negligible for the university sample.

We summarize the information results and perform some χ^2 tests in Table 6.

Insert Table 6 around here

The overall effect of the family background (a composite index including *Family tradition*, *Parents' will*, *Manager*, *Father unskilled* -with a negative sign-, and *Father attended College*) is, as expected, positive and significant. The overall effect of official information is not significant. However, the effect of unofficial information differs strongly amongst students. Those students who only received "menu" information are more likely to be restricted. In contrast, those students who have received "personal" information in the last two months of the process experienced a positive and significant increase in their chances to be unconstrained, as the model would suggest. Those who have received both types of information do not show any significant effect from unofficial information, as is the case for those who received no official information at all.[8]

6. Conclusions

In this paper, we consider a many-to-one two-sided matching model with a public office acting as intermediator amongst students who seek a post in the higher educational system and the suppliers of these posts - the universities. We model the process of preference formation in the students by introducing risk aversion in students preferences and a costly process for collecting information. In our model, students first decide whether to gather information about each college. After the process of obtaining information is finished, they fill in the list that they report to the matching office.

To contrast the importance of information gathering, we estimate probit models for both the probability of being satiated and the probability of entering in the first option. We compare college preferences from students in high school with the choices that were submitted to the matching office a year before by university students. Among the results of our empirical analysis, we emphasize the following:

[8]These results seem to conflict with those in the US college choice process (Hossler *et al.* [4]), where they claim that the junior year and the first months of the senior year in high school are the time frame during which most students move from the search stage to the choice stage of students college choice. Of course, the results are not strictly comparable as the processes of college choice and admission differ fundamentally in the timing. In particular, colleges in the U.S. play an important role in the process and it is in the students' interests that they make important decisions earlier than two months before allocation takes places.

Changes in preferences revealed in the data suggest that information gathering is important in the last two months before students must report the list to the matching office. Second, students with less information are more affected by these preference changes.

References

[1] CIS (1990), "Los jóvenes ante el sistema educativo" diseño y redacción, Margarita Latiesa. Centro de Investigaciones Sociológicas, Estudios y Encuestas; 25

[2] Gale, D. y Ll. Shapley, (1962) "College Admissions and Stability of Marriage", American Mathematical Monthly 69, 9-15.

[3] Gusfield, D. and R.W. Irving, (1989) "The Stable Marriage Problem, Structure and Algorithms," The M.I.T. Press Cambridge, Massachusetts.

[4] Hossler, D., Braxton, J. and Coppersmith, G. (1989) "Uderstanding Student College Choice," In *Higher Education: Handbook of Theory and Research* (Edited by J. C. Smart), Vol. 5, 231-288. New York: Agathon Press.

[5] Kotler, P. and Fox, K. (1985). *Strategic Marketing for Educational Institutions.* Englewood Cliffs, NJ: Prentice-Hall.

[6] Manski, C.H. and D.A. Wise (1983) "College Choice in America" Harvard University Press. Cambridge, Massachusetts.

[7] Paulsen, M.B., (1990) "College Choice: Uderstanding Student Enrollment Behavior," ASHE-ERIC Higher Education Report 90-6. Washington DC: The George Washington University.

[8] Paulsen, M.B., (1998) "Recent Research on the Economics of Attending College: Returns on invesment and Responsiveness to Price," Research in Higher Education 39(4), 471-489.

[9] Romero-Medina, A., (1998) "Implementation of Stable solutions in a Restricted Matching Market," Rev. Economic Design, 3, 137-147.

[10] Roth, Alvin E., (1985) "The college Admissions Problem is not Equivalent to The Marriage Problem," Journal of Economic Theory 36, 277-288.

[11] Roth, Alvin E. and G. Rothblum (1999) "Truncation Strategies in matching Markets - In Seach of Advice for Participants," Econometrica 67, 21-44.

[12] Roth, Alvin E. and M. Sotomayor, (1990) "Two-Sided Matching Markets: A Study in Game-Theoretic Modeling and Analysis," Econometric Society Monograph. Cambridge University Press.

[13] Roth, Alvin E. and J.H. Vande Vate, (1991) "Incentives in Two-Sided Matching with Random Stable Mechanism." Economic Theory, 1(1) 31-44.

Figure 1: The Sequence of χ^2 Tests and p-Values in the Clustering Algorithm

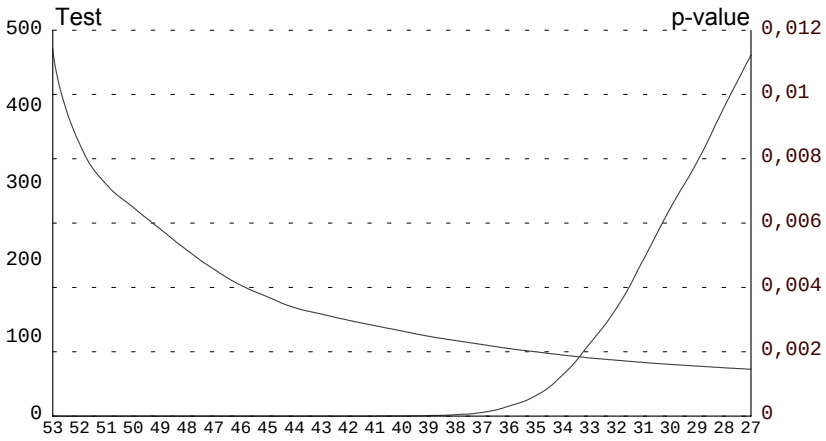

Figure 2a: College Distribution for University and High-School Students with CIS Codes

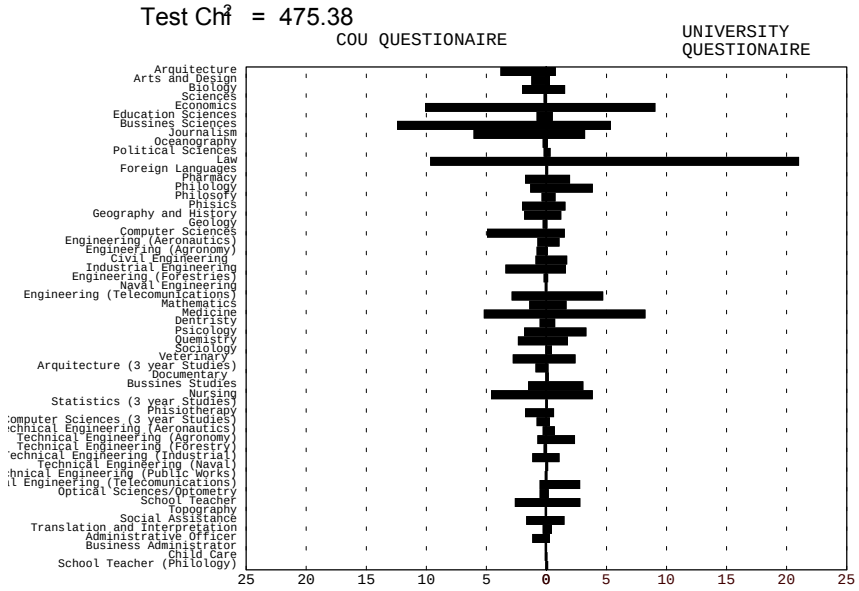

Figure 2b: College Distribution for University and High-School Students with "cluster codes"

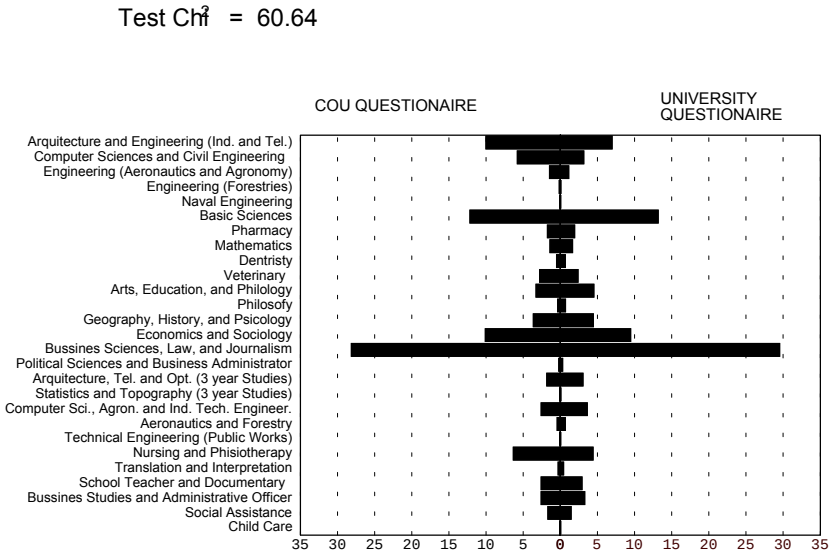

TABLE 1: LISTS SUBMITTED BY COLLEGE STUDENTS[a]

Position	1	2	3	4	5	7	7	8
Lists' Length	19.01	18.01						
(Cumulative)	19.01	37.02						
Entry Position	73.91	14.41	7.14	1.22	0.97	0.50	0.34	0.22
(Cumulative)	73.91	88.32	95.46	96.68	97.65	98.15	98.49	98.71

[a]Distribution (%) of students by list's length and the position of entry College

Table 2: Percentage of Students with official information on colleges[a]

	College Sample	High School Sample
Tests	30.61	19.29
Interviews	21.49	14.02
Written information	59.40	65.75
Talks	74.27	74.41
Videos	17.48	27.48
Personal Information[b]	37.81	26.22
Menu Information[c]	84.95	89.76

[a]For College students, reported percentages refer to the share of students who received information before taking the PAAU exam.
[b]Tests and/or interviews
[b]Written information and/or talks and/or videos

Table 3: The Model of Satiation

	(1)	(2)	(3)
Attended Public School	-0.04	-1.58	0.727
Female	0.10	3.77	0.396
Lost at least 1 year	-0.26	-10.07	0.019
Considered training courses	-0.13	-5.10	0.445
Vocation	0.53	20.73	0.008
Good marks in the field	-0.02	-0.68	0.876
Money	-0.37	-14.48	0.003
To stay in the same city	0.03	1.27	0.853
To leave the city	-0.31	-12.21	0.142
Sciences	-0.85	-32.08	0.000
Arts	-0.23	-9.19	0.253
Family tradition	0.67	24.25	0.006
Parents' will	-0.39	-15.34	0.121
Father unskilled	0.02	0.91	0.911
Manager	0.03	1.35	0.842
Father attended College	-0.24	-9.68	0.053
Personal	0.10	3.85	0.436
Menu	0.09	-3.62	0.626
Constant	0.76		0.020
Log Likelihood	-381.099		
pseudo-R^2	0.1015		
Cramer -R^2	0.1313		
Number of observations	614		
LR tests for joint significance of slopes	86.11		0.000

(1) Maximum Likelihood estimates for probit estimates. The pseudo-R^2 is the scaled value of the likelihood function whereby 100 is perfect fit and 0 is the constant model.
(2) Probability change when the corresponding dummy variable changes its value from 0 to 1 (evaluated at average values for the other controls)
(3) *p*-values

Table 4: Means Tests of Preference Change

	High School[a]		University[b]	Means Test[c]	
	CIS Codes	Cluster Codes		CIS Codes	Cluster codes
All	43.68	46.78	74.25	18.55	16.59
Arts	46.90	49.66	83.48	8.41	7.76
Joint Studies	46.37	47.98	78.01	8.86	8.40
Sciences	42.24	45.92	68.54	12.47	8.49
Uninformed[d]	42.90	45.94	74.01	17.60	15.79
Informed	47.91	50.52	73.96	4.94	5.70
Father unskilled	41.18	42.96	73.95	15.32	14.41
Father attended College	46.36	51.14	75.02	10.96	9.11

[a]Percentage of high school students with Q24 equal to Q2203

[b]Percentage of university students entering in their first option

[c]t-test for equality of percentages. All values are significant at the 99% level

[d]Student did not receive any official information

Table 5: The Model of Non-Restriction.

	(1)	(2)	(3)
Attended Public School	-0.04	-1.30	0.487
Female	0.03	0.88	0.629
Lost at least 1 year	-0.13	-4.00	0.022
Considered training courses	-0.19	-6.26	0.037
Vocation	0.59	20.02	0.000
Good marks in the field	0.18	5.50	0.003
Money	0.02	0.50	0.790
To stay in the same city	-0.10	-3.25	0.235
To leave the city	-0.40	-13.69	0.001
Sciences	-0.42	-12.66	0.000
Arts	0.16	4.80	0.050
Family tradition	-0.01	-0.42	0.918
Parents' will	0.25	7.11	0.036
Father unskilled	-0.13	-4.09	0.195
Manager	0.20	5.60	0.077
Father attended College	0.10	3.16	0.121
Personal	0.09	2.69	0.425
Personal x Univ	-0.47	-15.25	0.097
Menu	-0.40	-10.75	0.037
Menu x Univ	0.25	7.86	0.001
Menu x Personal x Univ	0.30	8.50	0.283
NotInf x Univ	-0.15	-4.69	0.507
Constant	0.68		0.002
Log Likelihood	-1284.643		
pseudo-R^2	0.0764		
Cramer -R^2	0.0848		
Number of observations	2589		
LR tests for joint significance of slopes	212.62		0.000

(1) Maximum Likelihood estimates for probit estimates. The pseudo-R^2 is the scaled value of the likelihood function whereby 100 is perfect fit and 0 is the constant model.
(2) Probability change when the corresponding dummy variable changes its value from 0 to 1 (evaluated at average values for the other controls)
(3) *p*-values

Table 6: The Model of Non-Restriction: Summary and χ^2 Tests			
	Coef.	χ^2	p-value
Social Background	0.416	9.760	0.002
Menu Information before last 2 months	0.089	0.637	0.425
Personal Information before last 2 months	-0.398	4.341	0.037
Official Information before last 2 months	-0.309	4.737	0.094
Unofficial last 2 months if previously Menu	-0.466	2.748	0.097
Unofficial last 2 months if prev. Personal	0.254	10.581	0.001
Unofficial last 2 months if prev. both	0.083	0.542	0.461
Unofficial last 2 months if prev. none	-0.147	0.441	0.507
[a] *Social Background =Family tradition + Parents' will + Manager -Father unskilled + Father informed*			

Appendix I

Lemma 1. *No student can benefit from including unaceptable colleges in the list P_i submitted to the matching office.*

Proof. We proceed by contradiction. Let college c^k be such $u_i(\emptyset) > u_i(c_i^k)$. College c^k will not be ranked on $P_i = c^1, c^2, c^3, ..., c^l$. Let $\hat{P}_i'$ be the list that includes all the elements in P_i, in the same order, $\hat{P}_i'$ also includes c^k. Without lost of generality let us assume that $\hat{P}_i' = c^k, c^1, c^2, c^3, ..., c^l$. Let

$$U(P_i) = \sum\nolimits_{h=1}^{l} p_i^h u_i(c_i^h) \prod\nolimits_{j=1}^{h} (1 - p_i^{j-1})),$$

and

$$U(P_i') = p_i^k u_i(c_i^k) + (1 - p_i^k) \sum\nolimits_{h=1}^{l} p_i^h u_i(c_i^h) \prod\nolimits_{j=1}^{h} (1 - p_i^{j-1})).$$

We will assume that

$$U(\hat{P}_i') > U(\hat{P}_i).$$

Hence,
$p_i^k u_i(c_i^k) + (1 - p_i^k) \sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^{j-1})) > \sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^{j-1}))$. This expression implies that

$$p_i^k u_i(c_i^k) > p_i^k \sum\nolimits_{h=1}^{l} p_i^h u_i(c_i^h) \prod\nolimits_{j=1}^{h} (1 - p_i^{j-1})).$$

Given that $u_i(\emptyset) > u_i(c_i^k)$ then

$$p_i^k \sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^j)) \geq 0.$$

A contradiction. ∎

Lemma 2. *Let P_i be a list that sorts s_i acceptable colleges from best to worse according with s_i preferences. There is no benefit in a change in the order between two colleges $c^t, c_i^k \in C$ in the preference list P_i.*

Proof. Let us assume that $u_i(c_i^t) > u_i(c_i^k)$ let us denote as P_i^{tk} the list where c^t and c_i^k preserve their order and as P_i^{kt} the list where the colleges switch their ranks.

Let τ be the position that t has in P_i^{tk}. Therefore

$$\sum_{h=1}^{\tau} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^{j-1})$$

remains identical in both lists.

Let κ be the position that k has in P_i^{tk}. Therefore

$$\sum_{h=\kappa}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^{j})$$

remains identical in both lists. It remains to be seen that

$p_i^t u_i(c_i^t) \prod_{j=1}^{t}(1 - p_i^{j-1}) + \sum_{h=t+1}^{k} p_i^h u_i(c_i^h) \prod_{j=1}^{h}(1 - p_i^{j-1}) > p_i^k u_i(c_i^k) \prod_{j=1}^{k}(1 - p_i^{j-1}) + \sum_{h=k+1}^{t} p_i^h u_i(c_i^h) \prod_{j=1}^{h}(1 - p_i^{j-1})$.

Let us assume, without lost of generality, that colleges $c^t, c_i^k \in C$ are the first and the last element, respectively in the preference list P_i . Let P_i^{tk} be $c^t, c^{t+1}, ..., c^{k-1}, c^k$ and let P_i^{kt} be $c^k, c^{t+1}, ..., c^{k-1}, c^t$.

$$p_i^t u_i(c_i^t) + (1 - p_i^t) p_i^k u_i(c_i^k) > p_i^k u_i(c_i^k) + (1 - p_i^k) p_i^t u_i(c_i^t),$$

then

$$-p_i^t p_i^k u_i(c_i^k) > -p_i^t p_i^k u_i(c_i^t),$$

Therefore

$$u_i(c_i^t) > u_i(c_i^k).$$

We will proceed iteratively including $c^{t+1}, ..., c^{k-1}$ successively in the list. We will start with c^{t+1} :

$p_i^t u_i(c_i^t) + (1-p_i^t)p_i^{t+1}u_i(c_i^{t+1}) + (1-p_i^t)(1-p_i^{t+1})p_i^k u_i(c_i^k) > p_i^k u_i(c_i^k) + (1-p_i^k)p_i^{t+1}u_i(c_i^{t+1}) + (1-p_i^k)(1-p_i^{t+1})p_i^t u_i(c_i^t)$.

Given that

$$p_i^{t+1}p_i^k(u_i(c_i^{t+1}) - u_i(c_i^k)) \geq p_i^t p_i^{t+1}p_i^k(u_i(c_i^{t+1}) - u_i(c_i^k)),$$

$$p_i^t p_i^k(u_i(c_i^t) - u_i(c_i^k)) > 0,$$

and

$$p_i^t p_i^{t-1}(u_i(c_i^t) - u_i(c_i^{t-1})) > 0,$$

hence,

$p_i^t p_i^k(u_i(c_i^t) - u_i(c_i^k)) + p_i^{t+1}p_i^k(u_i(c_i^{t+1}) - u_i(c_i^k)) + p_i^t p_i^{t-1}(u_i(c_i^t) - u_i(c_i^{t-1})) > p_i^t p_i^{t+1}p_i^k(u_i(c_i^{t+1}) - u_i(c_i^k))$,

Let us introduce c^{t+2}in the list. To maintain the inequality it must be proved that change in the right hand side of the inequality is greater than the change in the left hand side, i.e.

$(1-p_i^t)(1-p_i^{t+1})p_i^{t+2}u_i(c_i^{t+2}) - p_i^{t+2}u_i(c_i^k) > (1-p_i^k)(1-p_i^{t+1})p_i^{t+2}u_i(c_i^{t+2}) - p_i^{t+2}u_i(c_i^t)$.

Therefore

$$(1-p_i^{t+1})p_i^{t+2}(p_i^k - p_i^t)u_i(c_i^{t+2}) + p_i^{t+2}(u_i(c_i^t) - u_i(c_i^k)) > 0. \tag{6.1}$$

Given that

$$p_i^{t+2}(p_i^k - p_i^t)u_i(c_i^{t+2}) - p_i^{t+1}p_i^{t+2}(p_i^k - p_i^t)u_i(c_i^{t+2}) > 0$$

and

$$u_i(c_i^t) - u_i(c_i^k) > 0$$

the expression 6.1 is positive.

Proceeding iteratively with all colleges from c^{t+3} to c^{k-1} we conclude

$p_i^t u_i(c_i^t) + \sum_{h=t+1}^{k} p_i^h u_i(c_i^h) \prod_{j=t}^{h}(1-p_i^{j-1}) > p_i^k u_i(c_i^k) + \sum_{h=t+1}^{k-1} p_i^h u_i(c_i^h) \prod_{j=1}^{h}(1-p_i^{j-1}) + p_i^t u_i(c_i^t) \prod_{j=1}^{k-1}(1-p_i^{j-1})$. ∎

Lemma 3. *Let P_i be a list that sorts s_i acceptable colleges from best to worse according to s_i preferences. There is no benefit in truncating from below the preference list P_i.*

Proof. Let P_i be a preference list and let $P_i^{'}$ be a truncation from below of P_i .

$$U(P_i) = \sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^j) \text{ and}$$

$$U'(P_i^{'}) = \sum_{h=1}^{k} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^j) \text{ for } k \geq 1$$

We shall prove the statement by contradiction. Let us assume that

$$\sum_{h=1}^{k} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^{j-1}) > \sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^{j-1}) \tag{6.2}$$

If $p_i^t = 1$ for $t < k$ the statement 6.2 is not true because in this case both expressions are identical.

Let us assume that $p_i^t \neq 1$ for all $t \geq k$. We will proceed to prove the proposition by contradiction. Let us assume that a truncated list generates a greater expected utility than a complete list. In that case:

$$\sum_{h=1}^{k} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^{j-1}) > \sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^{j-1}) \tag{6.3}$$

Then, if we add another element to the list, the element $k+1$,the expected utility of adding this new element is

$$p_i^{k+1} u_i(c_i^{k+1}) \prod_{j=1}^{k+1} (1 - p_i^{j-1}) \geq 0. \tag{6.4}$$

In this case,

$\sum_{h=1}^{k} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1-p_i^{j-1}) + p_i^{k+1} u_i(c_i^{k+1}) \prod_{j=1}^{k+1} (1-p_i^{j-1}) \geq \sum_{h=1}^{k} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1-p_i^{j-1})$

We can add elements $k+2, k+3$ and so on, until we recompose the complete list. All additional elements generate increments in the expected utility of the resulting list as in expression 6.4. Therefore,

$$\sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h}(1-p_i^{j-1}) > \sum_{h=1}^{k} p_i^h u_i(c_i^h) \prod_{j=1}^{h}(1-p_i^{j-1}). \tag{6.5}$$

However, by transitivity and expressions 6.5 and 6.3:

$$\sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h}(1-p_i^{j-1}) > \sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h}(1-p_i^{j-1}).$$

A contradiction. ∎

Lemma 4. *Let P_h be a list that sorts s_i acceptable colleges from best to worse according with s_h preferences. There is no benefit in truncating from above the preference list P_h.*

Proof. Let P_h be a preference list and let P_h' be a truncation from above of P_h.

$$U(P_h) = \sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h}(1-p_i^{j-1}) \text{ and}$$

$$U'(P_h') = \sum_{h=k}^{l} p_i^h u_i(c_i^h) \prod_{j=k}^{h}(1-p_i^{j-1}) \text{ for } k > 1.$$

We shall prove the statement by contradiction. Let us assume that

$$\sum_{h=k}^{l} p_i^h u_i(c_i^h) \prod_{j=k+1}^{h}(1-p_i^{j-1}) > \sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h}(1-p_i^{j-1})$$

If $p_i^t = 1$ for $t < k$ then the statement is not true because $u_i(c_i^t) > u_i(c_i^k)$ and

$$\sum_{h=k}^{t} p_i^h u_i(c_i^h) \prod_{j=k+1}^{h}(1-p_i^{j-1}) > u_i(c_i^t).$$

Let us assume that $p_i^t \neq 1$ for all $t \geq k$. In that case:

$$\sum_{h=k}^{l} p_i^h u_i(c_i^h) \prod_{j=k+1}^{h} (1 - p_i^{j-1}) > \sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^{j-1})$$

then

$p_i^{k-1} u_i(c^{k-1}) + (1 - p_i^{k-1}) \sum_{h=k}^{l} p_i^h u_i(c_i^h) \prod_{j=k+1}^{h} (1 - p_i^{j-1}) > \sum_{h=k}^{l} p_i^h u_i(c_i^h) \prod_{j=k+1}^{h} (1 - p_i^{j-1})$

This is because $u_i(c^{k-1}) > u_i(c_i^k)$ for all k. If this is the case $p_i^{k-2} u_i(c^{k-2}) + (1 - p_i^{k-1}) \left(p_i^{k-1} u_i(c^{k-1}) + (1 - p_i^{k-1}) \sum_{h=k}^{l} p_i^h u_i(c_i^h) \prod_{j=k+1}^{h} (1 - p_i^{j-1}) \right) \geq p_i^{k-1} u_i(c^{k-1}) + (1 - p_i^{k-1}) \sum_{h=k}^{l} p_i^h u_i(c_i^h) \prod_{j=k+1}^{h} (1 - p_i^{j-1})$

Then

$$\sum_{h=1}^{l} p_i^h u_i(c_i^h) \prod_{j=1}^{h} (1 - p_i^{j-1}) > \sum_{h=k}^{l} p_i^h u_i(c_i^h) \prod_{j=k+1}^{h} (1 - p_i^{j-1}).$$

A contradiction. ∎

Corollary 1. Let P_i^k be a list P_i of s_i acceptable colleges truncated to its first k elements. Let all the colleges in P_i be such $p_i^h > 0$. Then $U_i(P_i) = U_i(P_i^k)$ if and only if the probability to obtain at least one of the colleges in P_i^k is equal to one.
Proof. By Lemma 4 $U(P_i) \geq U(P_i^k)$. If $U(P_i) = U(P_i^k)$ then the expected value of the elements eliminated $U(P_i/P_i^k)$ is zero. Given that all colleges on P_i are all acceptable colleges and, therefore their utility is greater than zero, their probability must be zero. Given that we have assumed that all colleges have positive probability then a college in P_i^k has probability 1. ∎

Theorem 3.1. If the students can submit their preference list without restrictions no student can profit from a misrepresentation of her preferences.
Proof. It follows from Lemmas 1 to 4.

Appendix II

As discussed in the main text, the list of college options in the survey is not exactly the same as the list of college options in the real process. Since our aim in this paper is to determine the importance of information acquisition in the preference formation process, we proceed by the following identification assumption. We suppose that all differences between Q24 and Q1101 are due to mismatches in the codings.

We set up a restricted clustering algorithm for colleges so that the new codings in Q24 and Q1101 are comparable. The restricted clustering algorithm works as follows. We assume that high school students encounter classification problems only for colleges within the same area of knowledge and with the same minimum university entry requirements: Most 5-year courses require passing the PAAU exam whilst for 3-year courses there is no such requirement. However, students do not misclassify between two colleges of different areas of knowledge, or a college that requires passing the PAAU test and a college that does not. Thus, only colleges within the same area of knowledge and minimum entry requirements can be clustered into one single college category. The clustering algorithm searches for the two colleges such that once clustered into one category, the distribution of college categories in Q1101 and Q24 look closer. We make the comparison by computing χ^2 tests for differences in frequencies and cluster the two colleges that minimize the χ^2 test. We proceed recursively until the difference in distributions is not statistically significant at the 99 percent level. The results of this clustering algorithm are presented in Figure 1.

Insert Figure 1 around here

Although the procedure does not guarantee a decrease in the χ^2 test at each step, we can see in Figure 1 that this is what happens. The reduction in χ^2 follows an exponential form, implying that our χ^2 statistic is strongly affected by a few misclassifications. In terms of p-values, we find that using the original survey codes for colleges leads to strongly rejecting the null hypothesis of the equality in the students distributions across colleges in the two questions. Once the amalgamation procedure finishes, we still have 27 college categories. Within these college categories, the distribution of questions Q24 and Q1101 -the expected first option two months in advance and the actual first option for the cohort of students from the previous year- are not statistically different. The discrete

distributions for both surveys and classifications can be compared in Figure 2a and Figure 2b.

Insert Figure 2a around here

Insert Figure 2b around here