The value of information in centralized school choice systems
(JOB MARKET PAPER)

Margaux Luflade*†
Duke University

November 6, 2017
The most recent version is available here.

Abstract

Centralized assignment mechanisms based on the deferred acceptance algorithm (DA) are used by school districts around the world to assign students to schools. Theoretical analyses of the DA consider that students are allowed to list all the alternatives of the choice set in their application rankings. However, in virtually all places where these mechanisms are implemented, students are restricted to list only a small number of choices. As a consequence, students need to take their admission chances into account, and be strategic in their choice. This paper uses administrative data from Tunisia, where high school graduates are assigned to university programs using a sequential variant of the DA, to empirically examine the effect of enabling students to update their expectations about their admissions probabilities. The sequential implementation induces quasi-experimental variation in the information available to students about remaining vacancies, and allows for the identification of students' preferences and expected admission probabilities. When students cannot revise their expectations, and relative to a benchmark situation in which students are given perfect information about which programs would admit them, their average indirect utility is decreased by the equivalent of a 41km-increase in the distance home-university —40% of the median distance traveled by students in the data. While easy to implement, the sequential implementation of the DA procedure reduces this expected utility loss by 67% in Tunisia. The increase in expected welfare is driven by a decrease in the share of students rejected by all their listed choices. Gains disproportionately accrue to low-ability and low-SES students. Counterfactuals suggest that a better targeting of low-priority students by the information provision would increase welfare gains.

*Email address: margaux.luflade@duke.edu

†I am grateful to Arnaud Maurel, Peter Arcidiacono, Joe Hotz, and Matt Masten for help and suggestions, as well as continuous support and encouragement. I thank Atila Abdulkadiroğlu, Federico Bugni, Rob Garlick, Eli Liebman, Seth Sanders, Modibo Sidibé, and participants at the Duke Labor and Microeconometrics workshops, and at the Young Economists Symposium at Yale University for very helpful comments and discussions. I am also extremely grateful to Meryam Zaiem whose insights about the context, help with obtaining the data, and encouragement have been critical to the realization of this project. I also thank the Ministry of Higher Education of Tunisia for providing the data. All errors are mine.

1 Introduction

New York City, Paris, Spain, Finland, Turkey, Chile, Norway, Ghana, Tunisia all use a similar centralized procedure to assign students to public schools or university programs. This mechanism is based on the deferred acceptance (henceforth, DA) algorithm (Gale and Shapley, 1962), and has been recommended to policy-makers by the school choice market design literature (Abdulkadiroğlu and Sönmez, 2003; Balinski and Sönmez, 2003) on the grounds of its desirable theoretical properties. The mechanism involves students submitting to a clearing house an ordered list of schools they would like to attend, schools giving priorities to students over admission offers, and the algorithm processing application lists and priorities to assign students to programs. Not only do all these places use a similar assignment mechanism, but they also implement it with the same departure from the theoretical design. While theoretical analyses of the algorithm are based on students being able to apply to all schools in their choice set, in practice, applicants in all these places are only allowed to list a restricted, and often small, number of programs in their preference report —for instance, twelve in New York, six in Ghana, ten in Tunisia, out of more than six or seven hundreds of alternatives.

This paper empirically examines the students' application portfolio choice problem when they are not able to apply to all academic programs in their choice set, and investigates the effects of enabling applicants to update their expectations about their admission chances. When students are not restricted in the number of applications they can make, mechanisms based on the DA ensure that it is dominant for applicants to simply report schools by order of preference in their application list (Dubins and Freedman, 1981; Roth, 1982). List-size restrictions break this property (Haeringer and Klijn, 2009). When they can only apply to a subset of programs, students face the possibility to be rejected from all their listed choices. To avoid rejection, students need to choose their application portfolio taking into account not only their preferences for academic programs, but also their probability to be admitted to these programs. Students' expectations about their probabilities of admission are then a crucial determinant of where they apply and are ultimately accepted.

Taking restrictions on the number of applications as fixed[1], this paper also investigates how providing students with information can improve the quality of school-student matches. Focusing on guiding the formation of expectations about admission chances, I consider the provision of updated information about programs filling up and vacancies remaining, at various points in time in the assignment process. I examine an information provision design that can be easily embedded in commonly implemented DA-based assignment mechanisms.

This paper uses administrative data from Tunisia, where college applications and assignments are made using a nationwide centralized assignment mechanism based on the DA. A unique institutional feature, the Tunisian mechanism is implemented in a sequential way, and involves different pools of applicants having different information about the available vacancies. This special implementation yields a quasi-experimental setting that enables me to empirically document three facts.

[1]There is evidence that many policy-makers are reluctant to let students submit lists as large as their choice set when the choice set is large, as this quote from Roth (2015) about the New York City match illustrates: "*[I]n my description [...] students can list as many schools as they like. We economists recommended that students be allowed to do just that, but on this important detail we did not prevail. So New York City students today can list only up to twelve programs among the hundreds that the city offers. Students who want to list more than that face a strategic choice of which twelve to list.*" See also Pathak and Sönmez (2013).

First, I use the quasi-experiment generated by the implementation of the mechanism to give evidence that students behave strategically when forming application lists. Despite the increasing number of school systems implementing the restricted-list DA, there is little empirical evidence on the performance of DA-based assignment procedures when the number of applications students can submit is restricted. This paper gives evidence that, under such restrictions, students may not find it optimal to truthfully apply to their most preferred schools.

Second, I show that application list size restrictions can decrease student welfare and increase inequality, relative to a setting in which students are freed from the need to form expectations about their admission chances and to engage in strategic behavior. I consider a context in which students' expectations about their admission chances may not coincide with their true probabilities of admission. I estimate a model of application portfolio choice, and use a counterfactual analysis to compare assignment outcomes resulting from the implementation of the restricted-list DA, to those obtained under a strategy-proof setting. Taking advantage of the quasi-experimental variation in information available to students, and in line with findings in the literature, the model allows expectation formation and the use of public information to differ across socioeconomic status (SES) and related variables (e.g. Hoxby and Turner, 2015).

Third, I find that a simple modification in the implementation of restricted-list DA can improve student welfare. A sequential implementation of the DA, as done in Tunisia, permits the provision of updated information to students about programs filling up and remaining vacancies. The effect of additional information on expected welfare is *a priori* ambiguous. On the one hand, more information about the vacancies that have been taken and remain may prevent students from applying to programs that turn out to be all full, and it may increase the quality of the matches made. On the other hand, it may decrease applicants' ability to signal the magnitude of their preference for the different alternatives (Abdulkadiroğlu, Che and Yasuda, 2015).

Comparing students' outcomes when they under the most common implementation of the restricted-list DA, relative to the strategy-proof benchmark, I find that students' average expected indirect utility is decreased. In magnitude, the average decrease in indirect utility is equivalent to the counterfactual decrease induced by, keeping all other things equal, having students attend a university 41km (25 miles) further away from home —about 38% of the median distance traveled by students in the data. While easy to implement, the 2010 Tunisian three-phase implementation of the restricted-list DA reduces this welfare loss by 67% . The increase in expected welfare is essentially driven by enabling a larger share of students to be assigned to an element of their application list —rather than to assigned students improving their match. Gains disproportionately accrue to low-ability, unsophisticated, and low-SES students. In fact, providing information about vacancies, even through a small number sequential of sequential phases, reduces the expected indirect utility gap existing between high- and low-SES students. Finally, while the 2010 Tunisian implementation of the three-phase procedure does increase welfare and the average match rate, I show that a better targeting of low-priority students by the information provision —through a different sequential partition of the cohort of applicants— could increase gains to students.

I face two main empirical challenges. The first is an identification challenge generally faced by the empirical literature on matching mechanisms. The mapping from students' preferences to their

application choices depends on their expectations about which schools may be available for them. With most school applications datasets, the econometrician cannot separately identify students' preferences and expectations about admission chances (Agarwal and Somaini, 2014). The quasi-experiment induced by the Tunisian sequential procedure helps me circumvent this identification problem. I argue that the sequential design induces a subset of students to truthfully report their most-preferred programs. For this subset of students, perceived admission chances can be ignored, and I can identify and estimate students' preferences for post-secondary programs. In a second step, I characterize students' expectations about their admission chances as those rationalizing other students' observed application lists, given identified preferences. The second challenge is computational. Given a student's preferences and expectations about her admission chances, finding the optimal application portfolio —that is, the expected-utility maximizing ordered list of up to ten programs among more than 600 alternatives— is intractable. The two-step approach, that identifies and estimates separately preferences parameters and expectations about admission chances, partially alleviates this issue.

Related literature

This paper contributes to three branches of the literature. It adds evidence to the small empirical literature on the DA. The theoretical literature on mechanism design is large and influential. In the context of school choice, it is part of an active dialogue between economists and policy-makers that has highlighted strategy-proofness as a way to pursue "transparency, fairness, and equal access to public facilities" (Abdulkadiroğlu, Pathak, Roth and Sönmez, 2006). The use of the strategy-proof DA has been recommended over other mechanisms (e.g. the Boston mechanism) that reward strategic behavior. It avoids penalizing students and families who do not strategize or do not do it well —which has been showed to be correlated with socioeconomic background (Kapor, Neilson and Zimmerman, 2016). Despite the widespread use of the DA, there is little empirical evidence of the consequences of a central feature of its implementation —the restriction imposed on the number of schools students can apply to.[2] Ajayi and Sidibé (2016) is, to my knowledge, the only empirical paper that addresses this question. Using data from Ghana, where the DA is used to assign students to high schools, they quantify the effect of changing the number of programs students are allowed to apply to. Fack, Grenet and He (2015) also document strategic behavior in assignment systems based on the (restricted-list) DA. In their analysis of the Paris high-school match, they test and reject the null hypothesis that students are truth-telling. These two recent analyses deliver an empirical counterpart to the experimental findings in Calsamiglia, Haeringer and Klijn (2010). In contrast, though, Abdulkadiroğlu, Agarwal and Pathak (2017) provide empirical evidence from the high-school match in New York City (NYC) that students may find it optimal to truthfully report their preferences, even when constrained to submit a application list strictly smaller than their choice set.

My paper differs from these empirical papers in two respects. First, it is the only one to document the effects of a practical and simple policy that provides decision-makers with updated information

[2] A number of studies have compared the unrestricted-list DA to alternative mechanisms (e.g. Agarwal and Somaini, 2014; Calsamiglia, Fu and Güell, 2014; Dur, Hammond and Morrill, 2016; He, 2016). A number of studies have also analyzed other less common assignment mechanism (e.g. Carvalho, Magnac and Xiong (2014).

about vacancies, and enables them to update their expectations about their admission chances. Second, the analysis of students' preferences for academic programs and expectations about their admission chances does not *a priori* constrain students to be all strategic, nor to all truthfully apply to their most-preferred programs. Rather, the two-step identification strategy used in this paper allows me to recover the share of students engaging in each type of behavior.

This paper also relates to studies on the role of information and students' imperfect sophistication in the context of centralized school choice systems. The questions tackled in my paper are similar to those Kapor, Neilson and Zimmerman (2016) explore using a survey of about 200 parents of kindergartners and ninth-graders participating in the New Haven school choice mechanism. They show that subjective beliefs about their child's admission chances differ from true admission probabilities, and that the magnitude of the deviation depends on parental effort and demographics. Mitigated empirical evidence about the effect of information on applications was earlier provided by Hasting, Van Weelden and Weinstein (2007) and Hastings and Weinstein (2009) using a field experiment conducted in the Charlotte-Mecklenburg Public School District in 2006.
My work complements these studies, as it considers application in a DA framework, while they focus on variants of an alternative assignment mechanism, the Boston mechanism.

More generally, beliefs about admission chances are part of a larger set of expectations students form about variables entering their application decisions and educational choices in general.[3] Recent studies have acknowledged that the expectations students and families form about the outcomes of their investment and application choices may be inaccurate (for instance on future wages, see, among others, Wiswall and Zafar, 2013; Jensen, 2010; Stinebrickner and Stinebrickner, 2014b). More broadly, studies have shown as well that agents need to form beliefs about the features of their educational decisions they do not have perfect knowledge of, and that providing them with additional information may actually affect their choices —whether it is information about schools and/or curricula characteristics over which the decision-makers may have preferences (e.g., on school quality, see Hasting, Van Weelden and Weinstein, 2007; and Hastings and Weinstein, 2009); or information about one's own ability in the curriculum or taste for these characteristics (see Pistolesi, 2016; Arcidiacono, Hotz and Kang, 2012; Stinebrickner and Stinebrickner, 2014a).

The rest of this paper is organized as follows. The next section reviews the theoretical properties of the deferred-acceptance algorithm, illustrates the inefficiencies likely to arise when it is implemented with application list size restrictions, and presents the alternative sequential procedure. Section 3 introduces the empirical setting of this paper. It presents the post-secondary assignment procedure in Tunisia, describes the data, and highlights reduced-form effects of the sequential information revelation on application behaviors. Section 4 describes my strategy to recover students' preferences for university programs, and shows my estimates. Section 5 describes my strategy to characterize students' expectations about their admission chances, and shows that not all students truthfully list their most-preferred programs. Finally, Section 6 compares students' outcomes under the sequential DA procedure and the standard implementation of the restricted-list DA, and discusses the value of information in a centralized school choice system. Section 7 concludes.

[3] See Altonji, Blom and Meghir (2012), and Altonji, Arcidiacono and Maurel (2015) for a review.

2 Theoretical background

In this section, I review the theoretical properties of the DA; I describe the consequences of its implementation in a school choice context in which applications are constrained or costly; and I present the alternative sequential procedure at the center of this paper. This section serves two main purposes. First, it establishes key properties of the mechanism that will ground the identification strategy in later parts of the paper. Second, it highlights the questions and trade-offs of interest for the policy-maker that will guide the counterfactual analysis presented at the end of this paper.

2.1 The deferred-acceptance algorithm: theoretical properties and tradeoffs.

School choice problems. Formally, a *school choice problem* (Abdulkadiroğlu and Sönmez, 2003) consists of two finite sets: a set of N students, and a set of J schools (or programs). Each school has a finite capacity that determines how many students it can enroll. Students have preferences over schools, while schools rank students by order of priority for admission.[4] Priority orders can be common or differ across schools, may or may not be known to students, and are taken as given. In the empirical setting of this paper, priority is merit-based and determined as a function of past academic performance[5]; it is known to students. A solution to a school choice problem —that is, an allocation in which each student is assigned to at most one school and no school is assigned more students than its capacity— is called a *matching*. A *mechanism* is a systematic rule or procedure that, given any school choice problem, selects a matching. In general, centralized school choice mechanisms involve (1) students simultaneously submitting an ordered list of academic programs to attend; and (2) a central authority assigning students to programs according to a pre-specified rule or algorithm. Because it determines the school or academic program students attend, a centralized mechanism can have significant consequences on students' outcomes such as their academic achievement (e.g. Kapor, Neilson and Zimmerman, 2016). A substantial theoretical literature has been guiding policy-makers in their choices of what mechanism to use by studying their properties, and highlighting some of them as desirable.

Desirable properties for matching mechanisms. Three properties have acquired a central place in the theoretical literature on matching —stability, strategy-proofness and efficiency. Here, I define them and discuss their desirability. In the context of school choice, a matching is *stable* if no student is matched to a school over which she prefers not being matched (it is *individually rational*); and if no student prefers to her assignment a school which has a vacancy in the final match (it is *non-wasteful*), or which admitted a student with lower priority than her (it is *justified-envy free*). A mechanism is stable if it always selects a stable matching. A mechanism being stable means that the outcome will be fair, in the sense that no student will lose a seat at a desired school to a student with lower priority than her at this school. It also means that the implementation of the outcome will be successful, in the sense no student-school pair will be willing to block the final assignment —empirical evidence indeed seems to suggest that failure of stability is a key reason

[4] *School choice* refers to one-sided many-to-one matching problems; while *college admissions* refer to two-sided many-to-one matching problems. In the context of college admissions, students and schools both have preferences over the other side of the market.

[5] In particular, the priority ranking is fine, rather than coarse. When ties occur, they rarely involve more than a handful of students.

why some mechanisms have been abandoned in practice (Roth, 2008).

A mechanism is *strategy-proof* if for all agents, truthfully reporting one's preferences over schools is always a weakly dominant strategy. A mechanism being strategy-proof means that the application game is easy to play for families. Mechanisms in which manipulating one's preferences can be profitable may put at a disadvantage students who are not able to strategize, or do not strategize well (Abdulkadiroğlu, Pathak, Roth and Sönmez, 2006b). Moreover, there is empirical evidence that students' ability to play the game induced by the assignment mechanism depends on demographics, such as their socioeconomic background (Kapor, Neilson and Zimmerman, 2016). A manipulable mechanism can then possibly foster the persistence of inequalities from one generation to the next. By contrast, strategy-proofness enables the policy-maker to purse "transparency, fairness, and equal access to public facilities" (Abdulkadiroğlu *et al.*, 2006b). In addition, application lists submitted by students under a strategy-proof mechanism constitute reliable data on families' preferences, which can inform broader policy-making (Abdulkadiroğlu *et al.*, 2006b).

A matching μ Pareto-dominates another matching ν if every student weakly prefers her assignment under μ over her assignment under ν, and at least one student strictly prefers her assignment under μ over her assignment under ν. A matching is *Pareto-efficient* if it Pareto-dominates all other matchings. A mechanism is Pareto-efficient if the matching it selects always Pareto-dominates the matching selected by other mechanisms. A mechanism being Pareto-efficient means that no welfare is wasted, in the sense that no student could be made better off without hurting someone else.

The deferred-acceptance algorithm. The deferred-acceptance algorithm (DA) is strategy-proof, stable, and Pareto-dominates[6] all other strategy-proof and stable mechanisms (Gale and Shapley, 1962; Dubins and Freedman, 1981; Roth, 1982). Based on these theoretical properties, its use has been recommended over other mechanisms (e.g. the Boston mechanism) in the school choice context (Balinski and Sönmez, 2003; Abdulkadiroğlu and Sönmez, 2003; Abdulkadiroğlu *et al.*, 2006b).

The DA algorithm introduced by Gale and Shapley (1962) takes in two sets of inputs. For each school, a priority ranking of all students over admission offers; and for each student, a preference ranking (the *application list*) of *all* schools of the choice set, from most to least preferred. In the simple case when a unique priority ordering is used by all schools, the DA proceeds as follows, once all students have submitted application lists:[7]

DA

Step 1/ The first-ranked student is assigned to her first-listed program.

Step (k+1)/ For any $k \geq 1$, once the k^{th} student in the priority ranking has been assigned, the student ranked $(k+1)^{th}$ is assigned to the highest-ranked element of her list that still has a vacancy. If all of her listed choices are full at that point, she is left unassigned and the

[6] The matching produced by the DA Pareto-dominates all other stable matches if priorities are strict (i.e. there are no ties). If ties must be broken, the resulting match may not be Pareto-optimal among stable matches (Erdil and Ergin, 2008).

[7] This simple case is the one relevant for the empirical analysis in this paper. When a unique priority ordering is used by all schools, the DA boils down to the so-called *serial dictatorship* algorithm. A more general version of the DA allows for school-specific priorities. It is not directly relevant for the empirical analysis in this paper; it is described in Appendix A.

algorithm proceeds to the next student.
Stop/ The algorithm stops after all students have been processed.

Trade-offs. While Pareto-efficient among stable and strategy-proof mechanisms, DA is not efficient (Abdulkadiroğlu and Sönmez, 2003). Elimination of justified envy requires, when two students have the same ordinal preferences over two seats, the higher-priority student to be assigned his more-preferred school, regardless of the cardinal intensities of students' preferences. If, for instance, the lower-priority student likes more the preferred seat (or dislikes more the less-preferred seat) than the higher-priority student does, elimination of justified envy can create a welfare loss. Hence, when choosing to implement the DA, the policy-maker demonstrates her willingness to pursue elimination of justified envy and strategy-proofness, and to pay the cost of foregoing efficiency.

2.2 List restrictions, uncertainty: implementation constraints and consequences

Versions of the DA are used in many places to assign students to schools (e.g. in NYC, Chicago, Paris but also nationwide in Turkey, Ghana) or colleges (e.g. in Turkey, Taiwan, Tunisia). Most implementations feature one common departure from the theoretical set-up: the size of application list students may submit is restricted to be strictly smaller than the size of the choice set. For instance, in NYC students can list 12 of the 500+ public high schools programs offered in the city; in Ghana, students can apply to 6 of the 1,900+ high school programs in the country.[8] Under such list-size restrictions, the DA *a priori* not strategy-proof. In a restricted-list application setting, students face the possibility of not being assigned to any school, if they get rejected from all the schools they apply to. A student who expects her most-preferred schools to be popular among higher-priority students, may then decide not to submit an application list that truthfully reflects her ordinal preferences over programs, and instead include less-preferred, safer schools. As a consequence of strategic reporting, the final matching may not be stable (with respect to the students' true preferences), and some welfare may be lost. For instance, a student may decide not to apply to a preferred program if she thinks her chances of receiving an offer are low, and then end up being assigned to a less-preferred school, while, *ex post*, the preferred program would have had a seat available for her.

In the paragraphs below, I describe the student's problem in a restricted-list setting, and illustrate consequences of the uncertainty faced by students on their incentives to be truthful and on welfare by an example. As they will be useful in the rest of this paper, I also review a couple of simple theoretical results on truth-telling and dominant strategies in the restricted-list setting.

2.2.1 Uncertainty and strategic incentives

The students' problem. At the time of application, i is assumed to know the flow utility she would derive from any element of her list. The only uncertainty she faces is due to her not knowing which element of her list (if any) she will gain admission to. While she does not know which program she will be offered admission to, she has (subjective) beliefs about her probability to be admitted to the different programs. She maximizes her subjective expected utility —the weighted sum of the flow utilities of elements of her ordered application list, with weights equal to her perceived

[8] Implementations in Boston and Romania are exceptions. For more examples and details on restrictions, see Appendix E in Fack, Grenet and He (2015), or `matching-in-practice.eu`.

admission chances to these programs —within the set of all ordered lists of up to M alternatives:

$$EU_i(\mathcal{L}_i) = \sum_{k=1}^{M} \Big[\pi_i(\mathcal{L}_i(k)) \times u_i(\mathcal{L}_i(k))\Big] + \overline{\pi_i} \times V_i(0) \tag{1}$$

where $\pi_i(\mathcal{L}_i(k))$ denotes i's expectations about her *admission* chances to the k^{th}-ranked element of her application list, $u_i(\mathcal{L}_i(k))$ denotes the flow utility derived from admission to this k^{th}-ranked element, and $V_i(0)$ denotes the option value of being left unassigned.[9] $\overline{\pi_i}$ denotes student i's probability to be rejected from all her listed choices.[10]

Example 1 below shows that when they face uncertainty about their admission chances and can only apply to a subset of their choice set, it may be optimal for students to submit an ordered list that does not coincide with their most-preferred programs.

Example 1. Suppose there are two programs A and B, each with two seats. Suppose there are three students, ranked from 1 to 3 by strict priorities. Students know their priority ranking; and that there are twice two seats to be apportioned. Preferences for programs are private information, but their distribution is common knowledge:

$$u_{iA} = 6.35 + \varepsilon_{iA}$$
$$u_{iB} = 5 + \varepsilon_{iB}$$

where $\varepsilon_{iA}, \varepsilon_{iB} \sim i.i.d.\ N(0,1)$.[11] Suppose students can only apply to one program, and that students who do not get assigned to any program obtain the outside option, that yields a value of 0.

Student 1 knows she has highest priority, and that, hence, she will be assigned by the DA to the program she ranks first in her list. It is strictly dominant for her to (truthfully) list her most-preferred program in her application list. Student 2 knows she is ranked second. She knows she will be assigned to her first-ranked element since no matter which school Student 1 gets assigned to, both schools still have at least one remaining vacancy. It is strictly dominant for her to (truthfully) list her most-preferred program in her application list. Student 3 knows that two seats are taken, and one program may be full by the time the algorithm processes her list. If she happens to list a program that is full, she will be left unassigned and get utility 0. She solves the maximization problem:

$$u_3 = \max_{s \in \{A,B\}} \{p_{3A} \cdot u_{3A};\ p_{3B} \cdot u_{3B}\}$$

[9] If a student who fails to be assigned to any element of her list is left unassigned, the value of unassignment $V_i(0)$ simply corresponds the value of the outside option. Alternatively, if students who fail to be assigned to any element of their list then participate in a secondary application procedure, the option value $V_i(0)$ is equal to the i's expected utility to be derived when participating to this secondary procedure.

[10] $\overline{\pi_i}$ corresponds to the joint probability of $i$not clearing, *ex-post*, the admission cutoff of all her listed choices. I call Student i (*ex-post*) *eligible* to program ℓ if program ℓ has at least one open seat when it is i's turn to be considered for assignment by the DA algorithm —that is, after all students with higher priority score than i have been assigned (or kept aside for a leftover spot), and none of the students with lower priority score than i has been considered for assignment. When assignments are made via the DA algorithm, the *admission* of Student i to program ℓ requires that (i) i has listed ℓ in her application ranking; (ii) i is *eligible* to program ℓ; (iii) i is *non*-eligible to all programs ranked above ℓ in i's ordered application list. Hence, the need to distinguish between between *eligibility* and *admission* probabilities.

[11] Note that $6.35 = 5 + 2 \times \alpha_N(.75)$, with $\alpha_N(.75)$ such that: $Pr(X < \alpha_N(.75)) = .75$ if $X \sim N(0,1)$.

where her eligibility chances to A and B respectively, are given by the distribution of preferences: $p_{3A} = 1 - .75 \times .75 = .4375$; and the probability that there is a seat available in Program B is $p_{3B} = 1 - .25 \times .25 = .9375$. Denote s_3^* the school Student 3 applies to:

$$s_3^* = \begin{cases} A & \text{if } \varepsilon_{3B} \leq \frac{p_{3A}}{p_{3B}} \times (6.35 + \varepsilon_{3A}) - 5 \\ B & \text{otherwise} \end{cases}$$

There is no general dominant strategy in the application game when one can only apply to $M < J$ alternatives, and little can *a priori* be said about students' behavior in such setting. The next two propositions give partial characterizations of students' behavior that will be useful in the rest of the paper. While reporting on one's list a vector of programs that differs from one's most-preferred vector may be dominant, Proposition 1 (Haeringer and Klijn, 2009) establishes that one never benefits from ranking the reported alternatives differently than by decreasing order of preference.

Proposition 1. [Haeringer and Klijn (2009)] (a) If a student finds at most M schools acceptable, then she can do no better than submitting her true preferences.
(b) If a student finds more than M schools acceptable, then she can do no better than employing a strategy that selects M schools among the acceptable schools and ranking them according to her true preferences.

The next proposition establishes a sufficient condition for truth-telling to be a dominant strategy. A proof is given in Appendix A.

Proposition 2. (a) Condition 1 (below) is a sufficient condition for students not to have a strict incentive to misreport their preferences over their choice set.
(b) Under Assumption 1 (below), Condition (1) is a sufficient condition for students not to misreport their preferences over their choice set.

Condition 1. Student i has a perceived *eligibility* probability 1 for (at least) one of her M most-preferred programs.

Assumption 1. When indifferent between doing so or not, a student does not mis-represent her unconstrained preference ranking. In other words, a student does not report her most-preferred programs in her application list only when it is *strictly* profitable to do so.

2.2.2 Uncertainty and welfare

Example 1 illustrates the way in which uncertainty can generate inefficiencies *ex post*. Student 3 may not be assigned to her most-preferred program *ex post* available if she does not apply to it —even though not applying to it may be optimal *ex ante*. For instance, consider a case in which $-2.037 < \varepsilon_{3B} - \varepsilon_{3A} < 1.35$,[12] and $\varepsilon_{2B} \geq 1.35 + \varepsilon_{2A}$. Student 3 prefers A to B, but finds it *ex ante* optimal to apply to B. Student 2 prefers B to A, therefore applies to B and gets in. In this case, Student 3 gets either assigned to B (if Student 1 got in A) or is left unassigned (if Student 1 got in B) while she prefers Program A to both these alternatives.

[12] that is, $\varepsilon_{3B} < 1.35 + \varepsilon_{3A}$, and $\varepsilon_{3B} > -5 + \frac{p_{3A}}{p_{3B}} \times (6.35 + \varepsilon_{3A})$

2.3 Sequential implementation and information revelation

In this subsection (and in the rest of this paper), I take as given and fixed any restriction on the size M of the list to be submitted. I describe a simple alternative implementation of the DA in which information about available seats is regularly publicly updated. I explain how this version of the DA, while easy to implement, may partially restore incentives for truthful reporting and increase welfare, relative to the standard single-phase, restricted-list DA.

Sequential implementation of the DA. The standard (one-phase) implementation of the DA involves the whole cohort of N students simultaneously submitting their application lists, and then being assigned via the DA. In contrast, a sequential implementation involves first dividing the cohort in $K \leq N$ assignment groups that successively submit lists and are assigned. In the case in which the same priority order is used by all schools, as is relevant for the empirical analysis in this paper, the division of the cohort can be straightforwardly made along this priority order. Suppose the N students are ranked by a strict priority order from 1 to N. Let K_1, $K_2, \ldots, K_K$ be the sizes of each of the K groups to be created. Assign students with priority ranks 1 to K_1 to Group 1, students with priority ranks $K_1 + 1$ to $K_1 + K_2$ to Group 2, etc. Priority order is preserved within groups. Given these groups, the assignment procedure goes as follows:

> **K-phase DA**
> *Phase 1/* The number of seats open in each program is publicly revealed. Group 1 students submit application lists, and are then assigned using the DA algorithm.
> *Phase (n+1)/* For any $1 \leq n \leq (K-1)$, vacancies remaining after the assignment of Group n students are publicly revealed. Unassigned Group n students are added at the top of Group $n+1$ according to their initial priority order. Group $n+1$ students submit application lists, and are then assigned using the DA algorithm.

In the limit, if K is equal to the number of students, sequentially implementing the DA puts students in a perfect information setting and is equivalent to allowing for the submission of an unrestricted list.

2.3.1 Information can restore incentives for truthful reporting

Example 2 illustrates the sequential implementation of the restricted-list DA and how it can restore, for some students, incentives for truthful reporting.

Example 2. Consider again the setting of Example 1, keeping unchanged the programs, vacancies, priority order and preferences. Suppose however that the pool of applicants is divided into two groups $\{1, 2\}$ and $\{3\}$. The application procedure is ran sequentially, in two phases, and the information about vacancies is updated between the two phases. In Group 1, Student 1 and Student 2 face the exact same application problem as in Example 1. They both submit their list as described in Example 1. Before Student 3 submits her application list, the information about vacancies is updated, and Student 3 knows she will be assigned first in Group 2. It is therefore dominant for her to truthfully report in her application list her most-preferred program *among those that have not been publicly declared full*.

In each group, the student ranked first faces perfect information when applying. She knows exactly which programs have available seats, and therefore optimally apply by truthfully listing her most-preferred programs among those that have not been publicly declared full. Proposition 2 shows

that other students at the top of each group, beyond the very first student, may also face incentives to be truthful after information is revealed.

2.3.2 Information and welfare

Example 2 illustrates one way in which revelation of information, via a sequential implementation of the assignment procedure, can improve welfare. It restores, for some students (e.g. Student 3), a choice situation similar to perfect information. Thereby, it ensures that these students always apply and get assigned to their most-preferred programs among those with remaining seats, rather than possibly being assigned to *ex post* suboptimal programs, or failing to be assigned.

More generally, the revelation of information can (weakly) increase welfare even if it does not fully restore incentives for truth-telling. From a revelation of information about remaining seats, rational students update their expectations about their eligibility chances to the true conditional (on the information received) distribution governing unobservables. Expected utility maximization based on this conditional distribution allows them to choose an application list that is better *ex ante* (given the pre-information revelation realizations of unobservables) than the list they would choose without the information. This is illustrated by Example 3.

Example 3. Consider again the setting of Example 1, keeping unchanged the programs, vacancies, priority order and preferences. Suppose however that the pool of applicants is divided into two groups $\{1\}$ and $\{2,3\}$. The application procedure is ran sequentially, in two phases, and the information about vacancies is updated between the two phases. In Group 1, Student 1 faces the exact same application problem as in Example 1, and applies truthfully. Student 2 faces a similar application problem as in Example 1, and applies truthfully. The information revelation allows Student 3 to update her beliefs, but does not fully restore incentives for her to be truthful. Student 3 chooses her application list by solving:

$$\tilde{V}_3 = \max_{s\in\{A,B\}} \{\tilde{p}_{3A}\cdot u_{3A};\ \tilde{p}_{3B}\cdot u_{3B}\}$$

where $\tilde{p}_{3A}$ and $\tilde{p}_{3B}$ are Student 3's expected eligibility chances to A and B, conditional on the information she received about Student 1's assignment:

$$(\tilde{p}_{3A},\tilde{p}_{3B}) = \begin{cases} (.25,1) & \text{if Student 1 chose A, i.e. } u_{1A}>u_{1B} \\ (1,.25) & \text{otherwise.} \end{cases}$$

Denote $\tilde{s}_3^*$ the school Student 3 applies to, conditional on the information she received about Student 1's assignment:

$$\tilde{s}_3^* = \begin{cases} \tilde{s}_3^{*(a)} & \text{if Student 1 chose A, i.e. } u_{1A}>u_{1B} \\ \tilde{s}_3^{*(b)} & \text{otherwise.} \end{cases}$$

Ex ante welfare under the information scenario of Example 1 writes:

$$\begin{aligned} W &= \int_{\mathbf{e}} \Big[\max\{u_{1A};\ u_{1B}\} + \max\{u_{2A};\ u_{2B}\} + u_3(s_3^*)\Big] d\mathbf{e} \\ &= \int_{\mathbf{e}_1} \max\{u_{1A};\ u_{1B}\} d\mathbf{e}_1 + \int_{\mathbf{e}_2} \max\{u_{2A};\ u_{2B}\} d\mathbf{e}_2 + \int_{\mathbf{e}} u_3(s_3^*) d\mathbf{e} \end{aligned}$$

Ex ante welfare under the present information scenario writes:

$$\begin{aligned}\tilde{W} &= \int_{\mathbf{e}} \Big[\max\{u_{1A};\ u_{1B}\} + \max\{u_{2A};\ u_{2B}\} + u_3(\tilde{s}_3^*) \Big] d\mathbf{e} \\ &= \int_{\mathbf{e}_1} \max\{u_{1A};\ u_{1B}\} d\mathbf{e}_1 + \int_{\mathbf{e}_2} \max\{u_{2A};\ u_{2B}\} d\mathbf{e}_2 + \int_{\mathbf{e}} u_3(\tilde{s}_3^*) d\mathbf{e}\end{aligned}$$

To see why $W \leq \tilde{W}$, decompose the expected indirect utility of Student 3 in each case:

$$\begin{aligned}\int_{\mathbf{e}} u_3(s_3^*) d\mathbf{e} &= Pr(u_{1A} > u_{1B}) \int_{\mathbf{e}|(u_{1A}>u_{1B})} u_3(s_3^*) d[\mathbf{e}|(u_{1A} > u_{1B})] \\ &\quad + Pr(u_{1A} \leq u_{1B}) \int_{\mathbf{e}|(u_{1A}\leq u_{1B})} u_3(s_3^*) d[\mathbf{e}|(u_{1A} \leq u_{1B})] \\ \text{and} \quad \int_{\mathbf{e}} u_3(\tilde{s}_3^*) d\mathbf{e} &= Pr(u_{1A} > u_{1B}) \int_{\mathbf{e}|(u_{1A}>u_{1B})} u_3(\tilde{s}_3^{*(a)}) d[\mathbf{e}|(u_{1A} > u_{1B})] \\ &\quad + Pr(u_{1A} \leq u_{1B}) \int_{\mathbf{e}|(u_{1A}\leq u_{1B})} u_3(\tilde{s}_3^{*(b)}) d[\mathbf{e}|(u_{1A} \leq u_{1B})]\end{aligned}$$

By $\left(\tilde{s}_3^{*(a)}, \tilde{s}_3^{*(b)}\right)$ being solution to the conditional optimization problem faced by Student 3:

$$\begin{aligned}&\int_{\mathbf{e}|(u_{1A}>u_{1B})} u_3(\tilde{s}_3^{*(a)}) d[\mathbf{e}|(u_{1A} > u_{1B})] \geq \int_{\mathbf{e}|(u_{1A}>u_{1B})} u_3(s_3^*) d[\mathbf{e}|(u_{1A} > u_{1B})] \\ \text{and} \quad &\int_{\mathbf{e}|(u_{1A}\leq u_{1B})} u_3(\tilde{s}_3^{*(b)}) d[\mathbf{e}|(u_{1A} \leq u_{1B})] \geq \int_{\mathbf{e}|(u_{1A}\leq u_{1B})} u_3(s_3^{*(b)}) d[\mathbf{e}|(u_{1A} \leq u_{1B})].\end{aligned}$$

On the other hand, the revelation of information, and the possibly induced incentives for truthfullness may have an negative effect on welfare. By restoring incentives for truthful-reporting, the revelation of information increases the probability that any student gets assigned to her most-preferred program among those that are available *ex post*. As explained in Section 2.1, elimination of justified envy can conflict with welfare maximization.

2.4 The value of information: an empirical question

Beyond perfect knowledge of true admission probabilities. The existence of gains from information revelation, both in terms of incentives for truthfulness and welfare, crucially depends on students' ability to understand the information they are given, to update their expectations about their admission chances given this information. Example 3 shows that the revelation of information in the sequential DA (weakly) increases welfare when students are *perfectly rational*, that is, able to perfectly update their beliefs to their true conditional eligibility chances, from the information revealed about remaining seats. While straightforward in this three-student, two-school example, in which mean utilities and the distribution of unobserved preferences are common knowledge, the expectations-formation problem can get hard as the choice set gets large. Previous studies in the empirical school choice literature have recognized that, even given the distribution of preferences, deducing one's probability of admission to all programs is a hard problem, which high-school students and their families may not be able to solve (e.g. Agarwal and Somaini, 2014; Calsamiglia, Fu and Güell, 2014; Ajayi and Sidibé, 2017; Kapor, Neilson and Zimmerman, 2016). In practice, assessing the effect on students' behaviors of providing information about vacancies requires testing whether students understand the information they are given, and characterizing

the way they use it to update their expectations about their admission chances.

Magnitude of gains. Students' preferences over the programs in their choice set are a crucial determinant of the magnitude of gains, both in terms of incentives for truthfulness and welfare. Proposition 2 shows that whether a student may find optimal or not to be truthful in a restricted-list application scenario depends on her expectations about her admission chances to her most-preferred programs. The change in welfare induced by the change in students' application behaviors and assignments when more information is provided naturally depends on students' utility for the alternatives. Quantifying the change in welfare induced by the revelation of information therefore requires recovering students' preferences for programs.

Heterogeneity in gains. From a policy perspective, and given the *a priori* ambiguous effect of information revelation on expected welfare, it is important to characterize who may win or lose from the sequential implementation of the restricted-list DA. In addition, an empirical analysis allows to investigate differential effects of information across ability and demographic groups, which have been documented in other settings (e.g. Hoxby and Turner, 2013).

Frequency of informational updates. Finally, when it comes to implementation of the sequential DA, the policy-maker needs to decide on how many phases to implement —that is how frequently to reveal information. If the revelation of information may have benefits, the sequential implementation certainly has its costs too. A fully sequential implementation, with a number of phases equal to the number of students, would give applicants perfect information, and restore the desirable properties of the unrestricted-list DA. When the number of students is large, though, updating the number of vacancies after every single assignment can take a prohibitive amount of time. In addition, as shown by Example 2, a fully sequential implementation may not be needed to induce perfect information. The characterization of the optimal information revelation structure is beyond the scope of this paper, and the data does not allow to identify implementation costs of the assignment mechanism. However, the empirical analysis can provide evidence on the marginal effects of an extra revelation of information, as a function of the information already revealed. It can also inform the cost-effectiveness of information updates as a function of the position, in the priority ranking, at which they are provided.

3 The university match in Tunisia

This section introduces the empirical setting of this paper. It describes the data, and presents the practical features of the sequential implementation of the DA in Tunisia, to assign high-school graduates to universities at the nationwide level. Importantly, taking advantage of the cutoffs generated by the division of the applicant pool in group, it provides reduced-form evidence of the consequences of the sequential implementation on students' application behaviors and assignments.

3.1 Institutional background

Every June, high-school seniors in Tunisia take the national end-of-high-school exam. Passing this exam —that is, scoring at least 10 out of 20 on average over the eight to ten tests of the exam— is a sufficient and necessary condition to graduate from high school and gain access to public post-secondary education in Tunisia. Tunisia counts fourteen public universities, each delivering a wide

range of degrees. Degrees are field-specific; each of them requires the completion of a standard curriculum approved by the Ministry of Higher Education, which generally involves undergraduate students specializing in one field of study as soon as their first semester. Hence, when deciding on her post-secondary education, a student needs to choose a university *and* a field of study. In this paper, I refer to such pair (university, field) as a *program* or *track*. While graduating from high school guarantees Tunisian students access to public higher education (graduating seniors are automatically registered in the centralized post-secondary application system), the particular program they will be allowed to enroll in is determined by a central assignment mechanism. Assignment is made according to a sequential variant of the DA algorithm, similar to that described in Section 2.3. Context-specific implementation features are presented now.

Priority score. In year 2010, a common priority ranking of students was used by all programs. A student's priority score was determined as a function of the student's grades at the various tests of the national end-of-high-school exam, which can be viewed as a standardized test. A student with a higher score is given priority over a student with a lower score for admission offers to the post-secondary programs. Students know their priority ranking.

Application groups. The application-assignment process is split into three successive phases. Namely, the cohort of applicants is divided into three groups based on their priority score —in this particular case: the top 30% of students ("Group 1" students), the middle 40% ("Group 2"), and finally the bottom 30% ("Group 3").

Public information. All high school graduates are given a handout containing information about the available post-secondary programs over the country. The handout indicates, for each existing program, the number of vacancies open for the next academic year and the past-year admission cutoff, that is, the priority score of the marginal student admitted in the previous year. After each group has gone through the assignment algorithm, the number of vacancies in each program is publicly updated, so next-group students are told which vacancies remain before submitting their application list.

Application lists. Students may submit an ordered list of up to 10 post-secondary programs.

Unmatched students. Application lists are processed using the DA. Students who fail to be admitted to any of their listed choices are pooled on top of the next application group —if there is one— and proceed to submitting a new application list after the information about vacancies is publicly updated.[13] If there is no next application group, unmatched students are administratively assigned to left-over seats.

3.2 Sample description

I use administrative data from the Tunisian Ministry of Higher Education and Scientific Research. The database contains the ordered application lists and assignment of all students applying to post-secondary programs in public institutions in Tunisia in 2010, as well as an identifier of the high

[13] The new list is formed based on the programs available at the time it is submitted, and not based on the programs available when the student submitted her initial list. In the data, only the very last list submitted by each student is recorded.

Table 1: Descriptive statistics: students

	All		Group 1		Group 2		Group 3	
	Mean	S.dev.	Mean	S.dev.	Mean	S.dev.	Mean	S.dev.
Demographics								
Female	.53	.50	.52	.50	.54	.50	.52	.50
Birth year	1990.8	.91	1991.2	.39	1990.9	.68	1990.2	1.23
High SES	.60	.49	.78	.41	.58	.49	.47	.50
From Tunis	.30	.46	.33	.47	.30	.46	.27	.44
From Coast (excl. Tunis)	.48	.50	.53	.50	.49	.50	.43	.49
From West/Interior	.19	.39	.13	.36	.18	.39	.26	.47
From South	.03	.17	.01	.11	.03	.17	.05	.22
Priority and academic perf.								
Raw priority score	123.16	28.98	160.21	13.23	119.30	11.0	91.1	6.9
Stdized priority score	0	1	1.28	.46	-.13	.38	-1.10	.34
STEM high-school	0	.85	1.04	.39	-.09	.40	-.92	.33
non-STEM high-sch. perf.	0	.79	.75	.54	-.07	.58	-.66	.59
Applications								
List 10 choices	.70	.46	.67	.47	.76	.43	.65	.48
Number of choices listed	9.02	1.8	8.86	1.9	9.3	1.5	8.86	1.9
List all programs in same field	.06	.24	.03	.16	.09	.28	.07	.26
List $\geq$ 75% prog. in same field	.20	.40	.11	.32	.25	.43	.22	.41
List all programs in STEM	.36	.48	.19	.40	.46	.50	.40	.49
List all prog. in same university	.03	.18	.01	.09	.04	.20	.05	.21
List $\geq$ 75% prog. in same univ.	.08	.27	.02	.13	.09	.29	.12	.33
List all prog. in same region	.26	.44	.10	.31	.35	.48	.31	.46
Assignments								
Admitted to 1st listed prog.	.39	.49	.45	.50	.39	.49	.35	.48
Admitted to 2nd listed prog.	.15	.36	.16	.36	.16	.37	.12	.33
Admitted to 3rd listed prog.	.10	.30	.11	.31	.11	.31	.07	.26
Admitted to 4th listed prog.	.07	.26	.08	.28	.07	.26	.06	.24
Admitted to 5th listed prog.	.05	.22	.06	.23	.05	.22	.05	.21
Admitted to 6th listed prog.	.04	.19	.05	.20	.04	.20	.03	.18
Admitted to 7th listed prog.	.02	.16	.02	.13	.03	.16	.04	.19
Admitted to 8th listed prog.	.02	.12	0	.06	.02	.13	.03	.16
Admitted to 9th listed prog.	.01	.10	0	.05	.01	.10	.02	.14
Admitted to 10th listed prog.	.01	.09	0	.03	.01	.09	.02	.13
Administratively assigned	.02	.14	0	.06	0	.07	.06	.23
Admitted in later round	.02	.14	.02	.13	.04	.19	0	0
Sample size								
	10,935		3,299		4,384		3,252	

Note: In the second panel, STEM (resp. non-STEM) high-school performance is the unweighted average of the student's standardized scores at the Math, Physics, Natural Sciences, and Comp. Sci. (resp. English, French, Arabic, and Philosophy) tests of the end-of-high-school national exam.

school they graduated from and the grades obtained at the various tests of the national end-of-high-school exam. It also contains a limited number of demographic characteristics, such as gender, date and region of birth, and a category indicator of father's occupation. In this subsection, I describe the student sample and students' choice set of post-secondary programs, as well as patterns in their

applications and assignments.

Students. In 2010, 82,748 students graduated from high school and were registered in the centralized post-secondary application system. The assignment procedure is run in parallel for each of the six majors students may graduate from high school with. Here, I focus on students graduating with the Math major.[14] They were 11,029 in 2010. Among them, I drop students for whom high school information and/or all end-of-high-school test scores are missing,[15] as well as the 65 students recorded in the data as not having submitting any application list[16]. Table 1 describes the 10,935 students in the final sample, as well as the division of the sample into the three application groups. The ratio of sexes is roughly constant across groups; and slightly more than half of sample is female. High-SES students represent 60% of the student sample; and their share decreases along the priority ranking. The represent 78% of Group 1 students, 58% of Group 2, and 47% of Group 3. A similar pattern is observed for geographical origin. Students from Tunis, and from the dynamic coastal regions account for 30% and 48% of the sample, respectively; their respective shares are highest in Group 1, and decrease along the priority ranking.

Table 2: Descriptive statistics: programs

	Phase 1		Phase 2		Phase 3	
	Mean	Std.dev.	Mean	Std.dev.	Mean	Std.dev.
Filling up in 2010	.84	.36	.83	.38	.67	.47
Filling up in 2009	.90	.31	.89	.32	.80	.40
2009 cardinal cutoff	-.47	.79	-.62	.63	-1.07	.44
2009 ordinal cutoff	.61	.25	.66	.21	.81	.15
# of seats	22	45	17	27	16	19
# at least 1 applicant	.73	.44	.98	.50	.98	.15
# of applicants	77	251	80	123	96	78
# of applicants/seat	2.91	4.9	8.7	16.5	10.0	10.6
Sample size						
Total # of programs	616		562		290	
Total # of seats	13,580		9,574		4,516	

Note: In the second panel, 2009 marginal admission score are shown conditional on programs filling up in 2009 –hence the change in sample size from 616 to 552. In the rest of the paper, for programs which did not fill up in 2009, the marginal admission score is set to the score of the very last student in the priority ranking (1 in percentiles terms).

Programs. In 2010, 616 post-secondary programs had seats available for students who graduated high-school with a Math major. 54 of them filled up by the end of the first assignment phase; 326

[14] More detail about high school and high school majors in Tunisia is provided in Luflade and Zaiem (2016). The Math high school major is among those allowing students to pursue the widest range of fields of study in their post-secondary career. A similar and separate analysis could be made for the other high-school majors. The comparative analysis of application and updating behaviors across students who graduated from high school with different majors is part of a future project.

[15] High school information and/or all end-of-high-school test scores are missing for 25 students. In addition, I drop 32 students, whose application lists comprise only programs out of their choice set (that is, publicly declared full before these students submit their list) and/or some of the six programs I drop because they did not exist, nor have an equivalent existing program, in the previous year.

[16] Among the 65 students recorded not to submit an application list, 3 are in Group 1, 18 in Group 2, and 44 in Group 3.

Table 3: Descriptive statistics: programs

	Phase 1		Phase 2		Phase 3	
	Share of Programs	Share of Seats	Share of Programs	Share of Seats	Share of Programs	Share of Seats
Field						
Field: Humanities	.14	.04	.14	.05	.08	.04
Field: Arts	.08	.11	.08	.13	.08	.17
Field: PE & Educ.	.02	.05	.01	.01	0	0
Field: Social sciences	.05	.01	.05	.01	.07	.01
Field: Economics & Mgmt	.15	.11	.02	.01	.20	.02
Field: Law	.02	.01	.02	.01	.01	.01
Field: Health & Life sciences	.10	.09	.08	.05	.01	.02
Field: Earth sciences	.05	.03	.05	.04	.06	.04
Field: Math & Comp. sci.	.12	.11	.12	.16	.15	.20
Field: Physics, Chem., Engineering	.28	.44	.28	.41	.32	.34
Degree						
Degree: Bachelor equiv. (LA)	.67	.36	.70	.48	.77	.59
Degree: Bachelor equiv. (LF)	.27	.29	.26	.33	.22	.39
Degree: advanced	.07	.34	.04	.19	.01	.03
Location						
In Tunis	.30	.30	.27	.21	.07	.07
In coastal regions (excl. Tunis)	.51	.52	.53	.58	.60	.56
In western/interior regions	.16	.16	.17	.21	.29	.34
In southern regions	.02	.01	.02	.01	.04	.02
Abroad	.01	0	0	0	0	0

by the end of the second phase; and 100 (16%) did not get assigned as many students as allowed by their capacity. Table 2 shows programs characteristics, and illustrates the changes in the choice set faced by students as the sequential assignment procedure moves from one phase to the other. Programs are offered in 10 fields of study,[17] four of them in STEM. Seats in STEM represent 67% of initially offered seats; 66% of the seats still available at the beginning of the second phase; and 60% of the seats remaining at the beginning of the third phase. About two thirds of initially offered seats are in programs preparing to the equivalent of a Bachelor degree (*Licence*), to be earned after the successful completion of three years of classes. There are two types of Bachelor degree: '*Licence appliquée*' (LA), which prepares students who plan to enter the labor market after graduation; and '*Licence fondamentale*' (LF), which prepares students who plan to pursue their education (in a Master's program) after graduation. The remaining third of initially offered seats are in programs preparing students to more advanced degrees, essentially in engineering and medical fields. By the end of the first and second phases, seats in advanced-degree programs represent only 19 and 3% of the seats still available, respectively. About a third of the seats initially available are offered in Tunis; they represent only 7% of the seats still available at the end of Phase 2. In contrast, 18% of initial seats are in western and southern regions of the country; while they constitute 36% of the seats that remain vacant at the beginning of Phase 3.

[17] Each post-secondary program is associated a six-digit code that reflects the fields classification defined by the Ministry of Higher Education. I use this classification. The ten fields are: Humanities; Arts; Education (incl. Physical Education); Economics/Business/Management; Social Sciences; Law; Health and Life Sciences; Earth Sciences; Physics/Chemistry/Engineering; and Math/Computer Science.

Application behaviors and assignment patterns. Students included an average of 9.1 programs in their list out of an allowed maximum of 10. 70% of students rank 10 programs.[18] On average, applicants got admitted to their second or third choice (2.6^{th} choice), with 39% of them admitted to their first-listed choice, and 76% of them to one of their five first-listed choices. Although shares vary, the pattern is the same across groups —45% of Group 1 students are admitted to their first-listed choice, against 35% of Group 3; 86% of Group 1 students are admitted to one of their five first-listed choices, against 65% of Group 3. 4% of the students fail to be assigned to any element of their initial list. Half of them (and most of Group 1 and Group 2 unassigned students) were assigned in a later round, to a program they listed when pooled in the next group. The other half (essentially Group 3 unassigned students) ended up being administratively assigned.

3.3 Local effects of the sequential implementation of the DA on application behaviors and assignments

Figure 1 illustrates the changes in behaviors observed at the information revelation cutoffs. I plots, as a function of students' priority, the rank in their application list of the choice they are assigned to. Top ranked students, at the left, are assigned to their first-listed choice. As priority goes down in Group 1, and as popular programs fill up, students get assigned to increasingly lower choices in their application portfolio. When updated information about vacancies is provided, as the limit between Groups 1 and 2, application behaviors change in such a way that students get assigned to their top-listed choice again. The cycle starts again until the next revelation of information, at the limit between Groups 2 and 3.

In this subsection, I further documents local application and assignment changes at the information revelation cutoffs. The division of the applicant pool into application subgroups creates cutoffs (henceforth *information revelation* cutoffs or *group* cutoffs) that I use in a sharp regression discontinuity (RD) design in order to document the local effects of providing updated information to applicants.[19]

The effect of a change in groups on any outcome Y of interest is estimated by local linear regression

$$\min_{\alpha,\beta,\tau,\gamma} \sum_{i=1}^{N} \mathbf{1}_{[c-h\leq T_i\leq c+h]} \cdot \Big(Y_i - \big[\alpha + \beta(T_i - c) + \Delta \mathbf{1}_{[T_i<c]} + \gamma(T_i - c)\mathbf{1}_{[T_i<c]}\big]\Big) \tag{2}$$

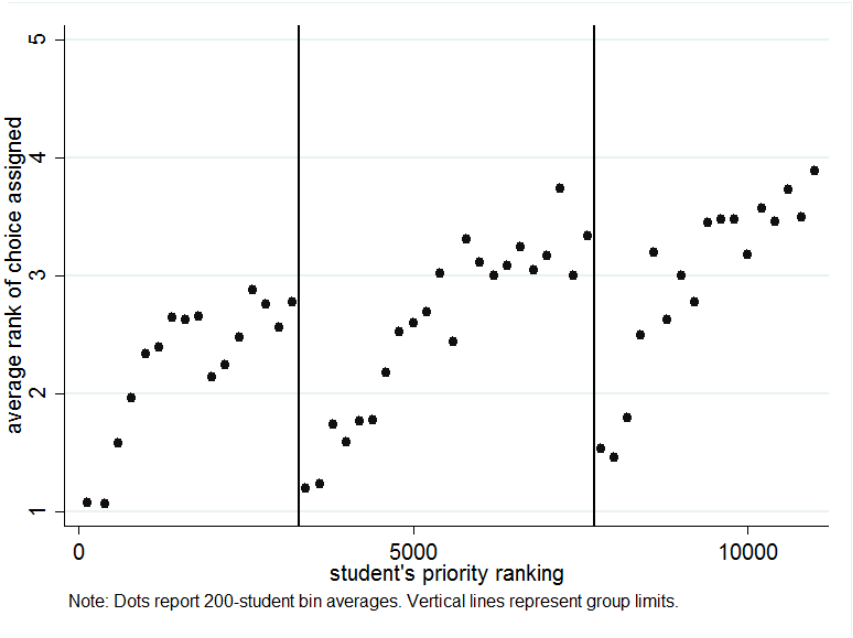
where T_i is student i's priority score (running variable), c denotes the group cutoff, and h is the estimation bandwidth.[20] $\mathbf{1}_{[T_i<c]}$ is a indicator of i being assigned to the 'informed group' (that is, Group 2 at the Group 1/ Group 2 cutoff; and Group 3 at the Group 2/ Group 3 cutoff) or not, $(T_i - c)$ is the distance of i's score to the group cutoff, and $(T_i - c)\mathbf{1}_{[T_i<c]}$ is an interaction term that allows the slope (of the outcome as a function of distance to the cutoff) to differ on either side of the group cutoff. Δ is the coefficient of interest, it measures the change outcome Y induced by

[18] The small and selected share of students listing strictly fewer choices than they are allowed to prevents from using an identification argument similar to the one used by Abdulkadiroğlu, Agarwal and Pathak (2017) when trying to recover students' preferences for programs.

[19] Standard graphical evidence supporting the sharpness and validity of the RD design can be found in Appendix C.

[20] For each outcome and subsample, the 'optimal' bandwidth is chosen using the Imbens and Kalyanaraman (2011) method and may vary from one outcome to another.

Figure 1: Choice assigned as a function of priority



the revelation of information.

I estimate (2) using as Y various application list characteristics (e.g. length, selectivity measures), and assignment outcomes (e.g. probability of assignment, rank of the choice assigned). In addition, to understand further how changes in application rates at the cutoffs correlate with the information revealed about programs, I estimate, for each program j, the change in application rate at the group cutoffs, using the binary indicator of whether or not student i ranked the track in her list as the dependent variable $Y_i^{(j)}$ in equation (2). I then regress the estimated change in application rate on various program characteristics. Results are shown in Tables 4 and 5, and summarized below.

Application behaviors. (1) *Marginally informed students submit shorter and less selective application lists than marginally uninformed students.* The top panel of Table 4 shows that, as compared to students that are marginally non-informed, marginally informed students list slightly fewer choices —.26 fewer at the Group 1/ Group 2 cutoff, and .63 fewer at the Group 2/ Group 3 cutoff, with only the latter difference being statistically different from 0 at conventional levels. Marginally informed students apply to programs that are less selective than their marginally uninformed counterparts. For instance, the most selective of their choices, has a past-year cutoff that is about .4 standard deviation lower, which corresponds to 9 percentiles of the priority distribution at the Group 1/ Group 2 cutoff, and 11 percentiles at the Group 2/ Group 3 cutoff. Interestingly, the same is true for safe choices as well. The least selective choice listed by marginally informed students has a past-year cutoff that is about .7 percentiles lower in the priority distribution than that of their marginally informed counterparts.

(2) *Marginally informed students increase their application rate to safer and more popular programs among those with remaining vacancies.* The top panel in Table 5 shows that a program being declared full (for the first time) at the group cutoff is correlated with a drop in application rate —

Table 4: Reduced-form effects of informational updates on application behaviors and assignment patterns

	Groups 1/2 cutoff		Groups 2/3 cutoff	
	Change	Base level	Change	Base level
Application behaviors				
# listed choices	-0.264	9.081	-0.627***	9.289
	(0.256)	(1.754)	(0.239)	(1.506)
Obs.	727		917	
PY cutoff of most selective choice	-0.389***	1.239	-0.405***	0.239
	(0.083)	(0.516)	(0.063)	(0.470)
Obs.	702		910	
PY cutoff of least selective choice	-0.413***	-0.301	-0.211***	-1.135
	(0.106)	(0.524)	(0.054)	(0.437)
Obs.	347		917	
Avg. PY cutoff over listed choices	-0.314***	0.481	-0.270***	-0.456
	(0.058)	(0.344)	(0.046)	(0.376)
Obs.	487		944	
PY (ordinal) cutoff of most selective choice	-0.094***	0.845	-0.110***	0.582
	(0.019)	(0.101)	(0.019)	(0.131)
Obs.	535		848	
PY (ordinal) cutoff of least selective choice	-0.073***	0.437	-0.078***	0.132
	(0.024)	(0.180)	(0.027)	(0.180)
Obs.	819		944	
Avg. (ordinal) PY cutoff over listed choices	-0.092***	0.653	-0.091***	0.363
	(0.017)	(0.099)	(0.016)	(0.116)
Obs.	478		820	
Assignment patterns				
Proba. to be assigned	0.025	0.961	0.092***	0.909
	(0.022)	(0.195)	(0.031)	(0.288)
Obs.	676		683	
# of listed choices eligible to	2.784***	5.406	4.206***	4.079
	(0.318)	(2.397)	(0.500)	(2.243)
Obs.	767		356	
% of listed choices eligible to	0.322***	0.589	0.583***	0.444
	(0.036)	(0.236)	(0.047)	(0.236)
Obs.	487		323	
Rank of choice assigned	-1.888***	2.200	-2.412***	3.057
	(0.365)	(2.287)	(0.455)	(2.630)
Obs.	397		360	

The '*Change*' column gives the estimated average change in outcome at the group cutoff. Std. errors are reported in parentheses below estimates. The '*Base level*' column gives control-group statistics about the outcome. Std. deviations are reported in parentheses below mean values. Cardinal cutoffs correspond to standardized scores of past-year marginally admitted students. Ordinal cutoffs are expressed in percentiles ($\times 100$) of the previous-year priority score distribution, with 0 indicating the lowest scores, and 1 the highest scores. *Read*: The most selective program listed by marginally uninformed students at the Group 1/2 cutoff has, on average, a standardized past-year (PY) cutoff of 1.2 —in ordinal terms, this corresponds to the 85th percentile of the priority score distribution. The most selective program listed by their marginally informed counterpart has a standardized past-year cutoff that is .4 std. dev. lower —in ordinal terms, this corresponds to a decrease of 9 percentiles in the priority score distribution. Estimation bandwidth is 1/8 of Imbens and Kalyanaraman (2011) optimal bandwidth. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

by 14 and 5 percentage points at the Group 1/ Group 2 and Group 2/ Group 3 cutoffs, respectively. The middle and bottom panels show that, for programs that are declared full, the magnitude of the

drop in application rates increases with the program's initial number of vacancies, and its selectivity level. Symmetrically, for programs that are not full, a larger number of remaining vacancies and a higher past-year cutoff are correlated with a larger surge in application rates.

Assignment patterns. (3) *Marginally informed students are more likely to be assigned to an element of their list than their marginally uninformed counterparts.* The bottom panel of Table 4 shows that students' probability to be actually assigned to one of their listed choices, rather than being rejected from all of them, is increased by 9 percentage points at the Group 2/ Group 3 cutoff, and by 2 percentage points at the Group 1/ Group 2 cutoff (but this latter effect is not statistically different from 0).

(4) *Marginally informed students are assigned to higher-ranked elements of their lists than marginally uninformed students.* The bottom panel of Table 4 also shows that marginally informed students end up clearing the *ex-post* admission cutoff of a larger share of their listed choices (+32% at the Group 1/ Group 2 cutoff, and +58% at the Group 2/ Group 3 cutoff) that marginally uninformed students do. This induces them to be assigned to a higher-listed element of their list —1.9 and 2.4 ranks higher at the Group 1/ Group 2 and Group 2/ Group 3 cutoffs, respectively.

Table 5: Correlations between local change in application rates and information

	Groups 1/2 cutoff	Groups 2/3 cutoff
Regression 1		
Just full	-.1400***	-.0492***
	(.0079)	(.0022)
Constant	.0122***	.0204***
	(.0011)	(.0013)
R-sqr	0.208	0.257
Obs.	616	616
Regression 2		
Remaining vacancies (10s) × Not just full	.0006	.0005***
	(.0001)	(.0003)
Earlier vacancies (10s) × Just full	-.0018***	-.0013***
	(.0001)	(.0001)
Constant	-.0007	.0037
	(.0180)	(.0012)
R-sqr	0.440	0.186
Obs.	616	616
Regression 3		
Ordinal past-year cutoff × Just full	-.1617***	-.0705***
	(.0094)	(.0047)
Ordinal past-year cutoff × Not just full	.0303***	.0867***
	(.0066)	(.0138)
Constant	0.026	.0071**
	(.0022)	(.0017)
R-sqr	–	–
Obs.	616	616

For all regressions, the outcome variable is '*Estimate change in application rate at the group cutoff*'.
Bootstrap std. errors in parentheses, account for two-step estimation.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Discontinuities in application behaviors at the information-revelation cutoffs are evidence of students' lack of perfect foresight and use of information. Validity of the RD design means that the assignment to students in one group or the next is, locally, as good as random: students on either side of a group cutoff have, on average, the same observable and unobservable characteristics. In particular, they also have, on average, the same preferences for post-secondary programs. As a consequence, if students were not using the information they are given, or if students had perfect foresight (in which case they would be able to predict the information they are given, which would then be redundant), application behaviors would not to change discontinuously at the cutoffs. In addition, Table 5 shows that the changes in application rates are consistent with students understanding and using the information they are given at the group cutoffs.

However, this reduced-form analysis does not inform whether students actually *benefit* from the informational updates. It does not inform either about the behavior and gains of students located further away from the information-revelation cutoffs in the priority ranking. Conducting a welfare evaluation of the effects of information provision requires comparing how students fare (here, in terms of indirect utility, which students derive from the program they are assigned to) under alternative counterfactual scenarios of information revelation. Performing the comparison requires simulating students' applications and assignment under the alternative scenarios. Generating students' application lists in turn requires to know the flow utility they associate with each program, and to understand how they derive beliefs about their admission chances from the available information. In the next section, I recover students' preferences for post-secondary programs. In Section 5, I turn to characterizing students' perceived admission chances. Finally, in Section 6, I present the results of the counterfactual analysis.

4 Recovering students' preferences for post-secondary programs

In this section, I present my approach to recover students' preferences for post-secondary programs. The identification strategy takes full advantage of local incentives for truth-telling induced by a sequential implementation of the DA. Using standard discrete choice methods, I am then able to estimate utility parameters without taking a stand on the way students form their expectations. At the end of this section, I present estimation results.

4.1 Identification strategy

When observed choices are the result of expected utility maximization, the econometrician who does not know the agents' expectations generally faces an identification problem (in the context of school choice, see Agarwal and Somaini, 2014). Without additional variation or assumption, it is not possible to disentangle the extent to which students' decisions are driven by what they like and the extent to which they are driven by what they think they can get. The quasi-experimental design induced by the sequential implementation of the DA enables me to circumvent this identification problem. My strategy to recover students' preferences for post-secondary programs directly builds on the three-group structure of the Tunisian mechanism. First, I show that the particular structure of information revelation embedded in the Tunisian assignment mechanism gives incentives to a subset of students to truthfully report their most-preferred programs in their application list (i.e. to 'be truthful'). The choices made by these students can be used to recover their preferences

without characterizing their expectations about their admission chances. Then, I argue that this subset of truthful students is informative about, and identifies, the utility parameters governing the preferences of the population of students.

4.1.1 A local discrete choice setting

To fix ideas, I first explain how, given a set of truthful students, the preferences of this set of students can be recovered. In the next paragraph, I characterize such set of students. A student being truth-telling (or truthful) means that the alternatives listed in her application ranking coincide with her most-preferred programs among those that have not been declared full. In what follows, I call student i's *choice set* (denoted $\mathcal{J}_i$) the subset of all post-secondary programs that have not been publicly declared full at the time student i chooses and submits her application list. Then, precisely, student i being truthful means that (i) her first-listed choice has higher flow utility than any alternative in her choice set, (ii) her second-listed choice has higher flow utility than any alternative in her choice set *but* her first-listed choice, and so on, until her last-listed choice.
For a given subset of truthful students, maximizing Problem (1) is equivalent (in the sense that the solution sets of the two problems coincide) to the following discrete choice problem:

$$\begin{cases} \ell_i(1) & = arg\max_\ell \left[u_i(\ell) \mid \ell \in \mathcal{J}_i\right] \\ \ell_i(2) & = arg\max_\ell \left[u_i(\ell) \mid \ell \in \mathcal{J}_i \backslash \left\{\ell_i(1)\right\}\right] \\ & \vdots \\ \ell_i(M_i) & = arg\max_\ell \left[u_i(\ell) \mid \ell \in \mathcal{J}_i \backslash \left\{\ell_i(1), \ell_i(2), \ldots, \ell_i(M_i-1)\right\}\right] \end{cases} \tag{3}$$

where $M_i \leq 10$ is the length of the application list submitted by student i,[21] and $\ell_i(k),\ k = 1, \ldots, M$ denote the *ordered* elements of i's application list.

4.1.2 Truthful reports

'Top' students. It is natural to assume that the very first-ranked student in each group, who knows she is ranked first, truthfully reports her most-preferred programs in her application list (among those that have not been publicly declared full). Indeed, given the information revelations publicly made before each group submits applications, the very first student in each group is faced with making a choice under perfect information. She knows she has probability one to be assigned to the first-ranked element of the list she submits (as long as she lists a program that has not been publicly declared full). It is therefore strictly dominant for her to list her most-preferred program first in her application ranking. And it is weakly dominant for her to also list her second to tenth most-preferred programs.[22]

Because students may apply to up to ten programs and because most programs have more than one vacancy, not only the first-ranked student, but a subset of applicants at the top of each group have incentives to truthfully report their most-preferred programs. Proposition 2 and Condition 1 stated in Section 2.2 give a sufficient condition for students to truthfully report their most-preferred programs. They imply that, when going down along the priority ranking within a group, students

[21] I do not model the choice of M_i in $\{1, 2, \ldots, 10\}$, nor do I control for the change of M_i across students.
[22] The possibility to be tied in the priority order may encourage students to list choices beyond the very first rank.

will truthfully report their preferences as long as they think they have probability one to have a seat in one of their 10 most-preferred programs.

'Short-list' students. Regardless of their priority rank, students who submit application list with strictly less than the allowed 10 programs (henceforth, 'short-list students') can be interpreted as truthfully reporting their most-preferred programs (Abdulkadiroğlu, Agarwal and Pathak, 2017[23]). On the one hand, submitting a list of size smaller than 10 is a (weakly) dominant strategy for students who like less than 10 schools. Furthermore, this list will coincide with their preferences. Indeed, Proposition 1(a) (Haeringer and Klijn, 2009) in Section 2 establishes that it is dominant for students interested in strictly less than 10 programs to truthfully report their preferences. On the other hand, it is (weakly) dominant for students who like 10 schools or more to submit a full list of 10 programs. Indeed, it is always (weakly) profitable for such a student to add a program to an application list of less than 10 programs.

4.1.3 Extrapolation

I argue that each of the two subsets of students described in Section 4.1.2 is sufficient to identify preferences parameters representative of those of the whole population of applicants. This result partially relies on assuming that conditional on observables, students all have the same mean flow utility for a given program. The available data however allows for enough flexibility in the specification of the flow utility function to make this assumption reasonable.

Utility specification. I assume flow utilities have the additively separable form:

$$u_i(\ell) = \delta_\ell + v_{i\ell} + \varepsilon_{i\ell}$$

δ_ℓ is a program fixed effect; it corresponds to the mean flow utility students derive from program ℓ. $v_{i\ell}$ is the part of i's demeaned (across i) flow utility for program ℓ that depends on individual characteristics observable to the analyst. $\varepsilon_{i\ell}$ is an individual- and program- specific utility shock which is privately known to the student at the time of decision-making, but remains unobserved to the analyst. I assume the program fixed effect δ_ℓ can be written as a linear (in the parameters) combination of program characteristics. I further assume that $v_{i\ell}$ can be written as a linear (in the parameters) combination of individual and individual-program characteristics.[24] Specifically:

$$\delta_\ell = Z'_\ell \gamma$$
$$v_{i\ell} = W'_{i,\ell} \beta$$

where Z_ℓ are program-specific attributes; $W_{i,\ell}$ are characteristics specific to the (individual, program) pair; and γ, β are utility parameters of interest, assumed to be invariant across programs and individuals. All program and individual characteristics Z, W are thought of as observed by the student at the time of decision-making, and by the analyst. In the empirical part, program attributes consist in the field of study, the degree to be received upon completion of the program, and, as a proxy for selectivity/quality/popularity, the admission score of the marginally admitted

[23] In the setting of Abdulkadiroğlu, Agarwal and Pathak (2017), 80% of students submit an application list with strictly fewer schools than the 12 allowed in the NYC high-school match. They use this subset of students to identify the preferences of NYC eighth-graders for high schools.

[24] This specification rules out preferences that depends on identified peers' assignments.

student in the previous year. Individual and program-student characteristics include distance between the student's home (as proxied by her high school) and the university hosting the program;[25] the student's high-school performance in the field of the program and outside this field; interactions between distance traveled and SES as well as program quality; and interactions between gender and field of study, as well SES and terminal degree.

The distribution of unobservables $\varepsilon_{i\cdot} := (\varepsilon_{i\ell})_\ell$ is assumed to be known, and independent of programs' and students' observable characteristics.[26] To facilitate estimation, I later assume $\varepsilon_{i\cdot}$ are i.i.d. type-1 extreme value. Normalizing to zero the coefficient on a reference field and on a reference terminal degree then fully identies the model. This means that, for each student, the value of every post-secondary program is interpreted as relative to the mean value of a local (in that distance traveled is 0) program that is not selective (past-year cutoff is 0 for programs that did not fill to capacity in 2009[27]), and upon completion of which the student would earn an 'LA' Bachelor degree (the reference degree) in Humanities (the reference field of study)[28].

The specified function controls flexibly for the two main determinants of post-secondary choices: distance between the university and the student's home, and the student's academic performance (Altonji, Arcidiacono and Maurel, 2015). Distance from home enters in a quadratic way. It is interacted with the student's socioeconomic status, to account for the fact that traveling may be more costly to economically disadvantaged students. It is also interacted with the program's selectivity level, to account for the fact that students may be willing to travel more to have better peers. The data, which contain students' scores at the national exam in eight different subjects, allows me to control separately for the student's high-school performance in STEM fields and non-STEM fields. The student's high-school performance in each field is also interacted with the program's field of study to account for individual comparative advantages in studying one subject vs. another, and for the fact that studying a given field may require more effort from students with lower high-school performance in the field.

Representativeness of truthful students and their choices. The particular structure of the subset of truthful students is crucial to the identification of such a flexible utility function, which renders the extrapolation credible. Tables 13 and 14 in Appendix D show descriptive statistics for key student and choice characteristics, comparatively for three samples of interest: the whole population (for which we would like to recover utility parameters); and each of the two truthful subsamples.

Table 13 shows that students' characteristics have, in each of the truthful subsamples, similar variation and support as they have in the population. In that sense, each of the truthful sam-

[25] As a proxy for student's i distance to university j, I use the distance between the capital city of their respective regions. Hence, student i is at distance 0 of any university in her home region. The distance between regions capitals is provided to students in the application handout made available by the Ministry of Higher Education.

[26] This rules out students sorting based on unobservable preferences —for instance, students systematically choosing their geographical residence at the time of high school to be next to the university programs they like. This guarantees that coefficients on school attributes identify the students' valuation for that attribute, and does not capture correlated variation with unobservable tastes. However, note that programs' and students' observable characteristics will be taken as given and fixed in the counterfactual analysis and welfare evaluations in this paper.

[27] Past-year cutoffs are expressed in percentiles of the distribution of priority scores; non-selective programs have cutoff at the 0^{th} percentile.

[28] In other words, student i derives flow utility $u_{i,\ell} = \beta_{\text{SES}_i} \times \text{distance}_{i,\ell} + \gamma_{\text{SES}_i} \times \text{past-year cutoff}_\ell + \varepsilon_{i,\ell}$ from a program ℓ preparing her to receive an 'LA' Bachelor degree in Humanities. If program ℓ is local ($\text{distance}_{i,\ell} = 0$) and non-selective ($\text{past-year cutoff}_\ell = 0$), then i's utility for ℓ is $u_{i,\ell} = \varepsilon_{i,\ell}$.

ples is representative of the population. There is one main exception: high-school performance variables have smaller support in the 'top' sample than they have in the population. This is a consequence of 'top' students being sampled from three points in the priority (a deterministic function of high-school performance) distribution. When using 'top' students to recover population utility parameters, the relationship between preferences and high-school performance in the population is therefore extrapolated from the relationship between preferences and high-school performance among 'top' students, via the continuity of the utility function. The validity of such extrapolation would be a strong assumption if 'top' students were sampled from *one* point of the priority distribution. However, Table 13 shows that the three-point sampling allowed by the Tunisian design ensures sufficient range and variance in 'top' students performance to allow for a reasonable extrapolation by continuity.[29]

Table 14 shows that the characteristics of the choices made by students in each of the truthful subsamples described in Section 4.1.2 span the full support of programs' characteristics in the initial choice set. In that sense, students in each of the truthful samples express preferences over all relevant tradeoffs existing in the choice set. This is a consequence of the choice set restrictions imposed by the informational updates. Students in low-priority groups are induced to express preferences over and solve tradeoffs involving programs others than the programs most popular among high-priority students and publicly declared to be full.[30]

4.2 Estimation

4.2.1 A local discrete choice estimation procedure.

I assume that the unobservable components $\varepsilon_{i,j}$ are i.i.d. type-1 extreme value. Estimation proceeds by maximum likelihood, using the sample of truthful students.[31] Independence of unobservables across individuals allows to write the sample likelihood L as the product of individual likelihoods p_i; independence of unobservables across alternatives further allows to write individual likelihoods as

[29] The argument is made clear by comparing 'top' students characteristics in the Tunisian design with the characteristics of those who would be 'top' students in a single-phase implementation of the assignment mechanism —that is, students at the very top of Group 1. Descriptive statistics for these students, also provided in Table 13, show that there is very little variation in 'top Group 1' students' high-school performance, and the variable has a very small support relative to its support in the population. This is unsurprising given the very selected nature of the top of Group 1 data. As a consequence, extrapolating to the population the utility function recovered from the top of Group 1 data would require unreasonable assumptions about the homogeneity of students' preferences across the range of high-school performance. As shown by Table 13, The Tunisian design, which gives incentives to be truthful to students at three point of the priority distribution rather than one, crucially alleviates this issue.

[30] As an illustration, Table 14 shows that there is also little variation in the characteristics of the programs chosen by 'top Group 1' students. Only 10% of the existing programs are listed by top of Group 1 students in their application lists. There is very little variation in the selectivity level of the listed programs by students at the top of Group 1, relative to what is observed in the population. Moreover, some program characteristics do not have full support in 'top Group 1' students' choices. This is the case not only for some fields of study (no program in Social Sciences and Law, while in the population students do express preferences regarding these fields), but also for key interaction variables such as *distance from home* $\times$ *program selectivity*. As a consequence, important aspects of students' preferences, such as the way they solve trade-offs between traveling further from home, attending more selective institutions, and studying a field they like, could not be identified by the sample of 'top Group 1' students.

[31] In Appendix D, I discuss an alternative estimation strategy that could yield more precise utility parameter estimates by using all students' application lists, rather than only those from truthful students.

the product of logit probabilities, yielding the well-known ranked-ordered (or exploded) logit form:

$$L = \prod_{a=1}^{N_E} p_{i_a}$$

$$\text{with} \quad p_i = \frac{\exp(u_{i\ell_i(1)})}{\sum_{k\in J_i}\exp(u_{ik})} \times \frac{\exp(u_{i\ell_i(2)})}{\sum_{k\in J_i\setminus\{\ell_i(1)\}}\exp(u_{ik})} \times \cdots \times \frac{\exp(u_{i\ell_i(M_i)})}{\sum_{k\in J_i\setminus\{\ell_i(1),\ldots,\ell_i(M_i-1)\}}\exp(u_{ik})} \tag{4}$$

where $M_i \leq 10$ is number of programs included by i in her application list; and $(i_a)_{a=1,\ldots,N_E}$ denote the estimation sample.

While the subset of 'short-list' students is readily observable from the data and can straightforwardly be used for estimation, the subset of 'top' students is *a priori* unobserved. Indeed, Condition 1, which ensures that students who *think* they have probability 1 to clear the *ex-post* admission cutoff of (at least) one of ten most-*preferred* programs (among those not declared to be full) truthfully report their preferences, cannot be used in practice as students' preferences and expectations are unknown at this stage. In the rest of this subsection, I explain how I select the 'top' estimation sample.

4.2.2 Estimation from 'top' students: choosing the 'top' sample in practice

A standard bandwidth choice problem. The choice of the 'top' estimation sample is akin to the choice of the estimation bandwidth in any local estimation procedure (e.g. local linear regression, as in Section 3.3). Selecting the 'top' estimation sample involves solving a trade-off between bias and variance of the estimator. The sample should be sufficiently large to have identification power and for the estimates to be precise. However, the sample should be small enough not to include any non-truthful students, whose inclusion would bias the estimates.

In practice, I include in the 'top' estimation sample all students with priority in the top 200 ranks within each group.[32] In the next paragraph, I present empirical evidence that the selected 'top' students are aware of the incentives they face and behave truthfully. In Section 4.3, I discuss the robustness of my results to changes in the estimation bandwidth.

Empirical validation of the bandwidth choice. I present three pieces of empirical evidence suggesting that, within the chosen 'top' bandwidth, students truthfully report their most-preferred programs —that is, that my estimates are unlikely to be biased by the presence of students misreporting their preferences. First, given Condition 1, it is crucial that students use and understand the public information about vacancies. In particular, it must be reasonable to assume that they *understand* that, given their priority ordering and the number of vacancies said to be remaining, they do have eligibility probability 1 to a range of programs. The RDD analysis in Section 3.3 strongly suggests that this is indeed true. In particular, Table 5 suggests that students understand

[32] The size of the subset of truthful students is increasing in the size of the application list students are able to submit, and in programs' capacities. Indeed, keeping preferences and everything else fixed, the number of students who perceive to be eligible with probability 1 to one of their $M' > 10$ most-preferred programs is weakly larger than the number of students who perceive to be eligible with probability 1 to one of their $M = 10$ most-preferred programs. Similarly, all other things equal, if the number of available seats to all programs (weakly) increases, the number of students who perceive to be eligible with probability 1 to one of their 10 most-preferred programs weakly increases as well.

the incentives they face. In addition, marginally informed students submitting shorter lists that their marginally uninformed counterparts (Table 4) suggests that students at the top of each group recognize that their eligibility to some programs is certain. It is then natural to infer that they understand that a truthful report of their most preferred alternatives (in their choice set) is dominant.

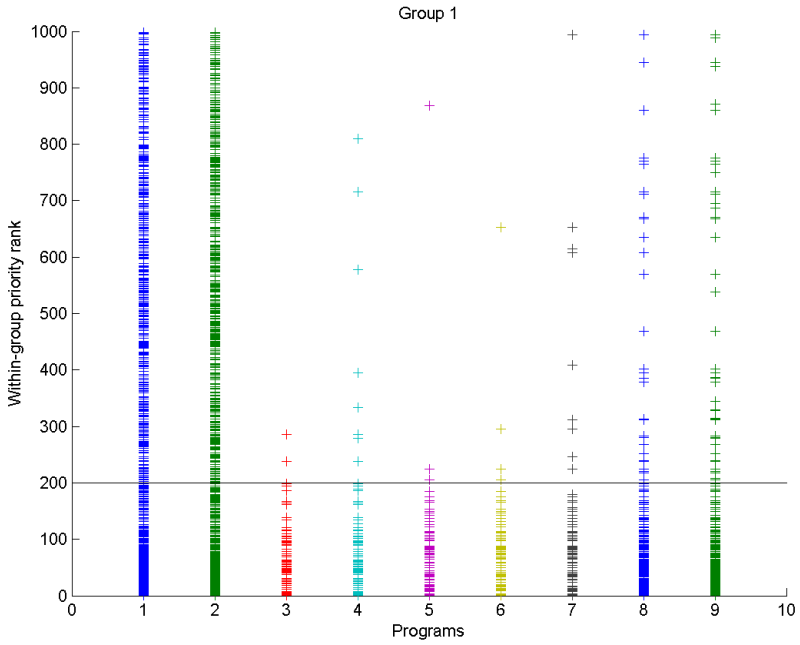
Second, I show that suggestive evidence of students censoring themselves (in the sense that they do not apply to their most-preferred but very popular alternatives) can be found only among students outside the chosen bandwidth. It is uncontroversial that the very first student at the top of each group is certain about her eligibility chances and is truth-telling. Ambiguity about whether other applicants think that reporting truthfully is dominant for them increases as within-group priority decreases. When students stop being truthful, we expect to observe a decreased frequency of application to programs listed by the uncontroversially truthful students. Figure 2 shows that such frequency decrease does not happen within the chosen 'top' estimation bandwidth. Figure 2 considers the ten programs listed the first three students at the top of Group 1,[33] and shows the frequency at which these programs are listed by Group 1 students as a function of students' priority. Programs are represented on the x-axis, priority on the y-axis. A dot in position (a, b) in the graph means that student ranked b in Group 1 included program a in her list. The vertical line at rank 200 represents the limit of the 'top' estimation bandwidth —so students with priority rank lower than 200 are included in the estimation sample. The frequency of application to the programs listed by the top three students does not decrease within the 'top' estimation sample. Passed the bandwidth limit, it then decreases more or less abruptly for some programs—suggesting that omission or censoring start occurring.

Third, I show *ex-post* evidence that students in the 'top' estimation sample submit a truthful report of their preferences. For this, I rely on another unique feature of the implementation of the DA algorithm in the Tunisian context: a *reassignment round*. After students of all three groups have been assigned by the DA algorithm[34] but before the new academic year starts, students are invited to express any dissatisfaction about their assignment. Precisely, students may submit a new ordered list of four programs they would prefer to attend over their assigned match. Any program can be included in this new list, irrespective of whether it was part of the student's initial ranking or not, of whether it has vacancies left or not, and of the program's realized admission cutoff relative to the student's priority score. Importantly, students do not have to forgo their initial assignment to participate in the reassignment round: unless their reassignment request is approved, they keep their initial match. No precise procedure is explicitly defined regarding the processing and approval of requests. It is generally understood that priority within students is preserved in the reassignment round and that approval depends on the ability of the requested track to welcome an additional student. The top panel in Table 6 shows the shares of top and non-top students applying for reassignment. The bottom panel describes further the behavior of students applying for reassignment. Students in the 'top' bandwidth apply for reassignment at a significantly lower rate than other students (17.6 vs 24.6%) and when they do, they submit fewer requests (2.7 vs 3 programs included in the reassignment list). Most importantly, about 84% of the programs students in the 'top' bandwidth request are outside their choice set, that is, had already been declared full when the students applied. On the contrary, the majority (54%) of requests submitted by other students are within their choice set. The large difference between the shares

[33] Figure 12 in Appendix D shows similar evidence for Groups 2 and 3.

[34] or administratively for those not eligible to any element of any of their application lists

Figure 2: Persistence of the top-ranked students' listed choices over the priority ranking



Legend: This graph considers the ten programs listed the first three students at the top of Group 1, and shows the frequency at which these programs are listed by Group 1 students as a function of students' priority. Programs are represented on the x-axis, priority on the y-axis. A dot in position (a, b) in the graph means that student ranked b in Group 1 included program a in her list. The vertical line at rank 200 represents the limit of the estimation bandwidth —students with priority rank lower than 200 are included in the estimation sample.

of students in and out of the 'top' bandwidth who request reassignment within their choice set (2 vs 11%[35]) suggests that students in the estimation sample did not censor themselves —that is, indeed reported their most-preferred programs. Indeed, students out of the 'top' bandwidth reveal by their reassignment requests that they prefer some of the alternatives in their choice set that they did not inially list over the ones they initially applied to.[36]

4.3 Results

Main estimates

Tables 7 and 8 show maximum likelihood (ML) estimates obtained from pooling both subsets of truthful students ('top' and 'short-list'). Results are shown for each of the two estimation samples are similar to the one shown here, and displayed in Tables 15 and 16 in Appendix D. Results

[35] Shares: $.02 = .176 \times .13$ and $.11 = .246 \times .43$.

[36] Calsamiglia, Fu and Guell (2014) and Kapor, Neilson and Zimmerman (2016) rationalize students "changing their mind" by them receiving a post-assignment, pre-enrollment utility shock. The contexts in these two papers are different from the one here: they consider families applying for seats in public schools and kindergarten in Barcelona and Cambridge, MA respectively (both cities use a variant of the Boston mechanisms). In their context, a student "changing her mind" is a student who is matched to her first choice by the centralized mechanism but ends up not enrolling in the school —and supposedly enrolling in a private school instead. Despite this difference of settings, a post-assignment, pre-enrollment utility shock could be used here as well to justify students requesting reassignment. However, there is no reason *a priori* why students 'at the top' would be induced to change their mind at such a much lower rate relative to other students. At the very least, the share of non-top students asking for reassignment in excess of the share of students at the top can reasonably be attributed to forecasting errors on their part.

Table 6: Students at the top reveal to be satisfied with their assignments

	In the 'top' estimation sample	Out of the 'top' estimation sample
# students	636	10,368
% requesting reassignment	17.61	24.60
Conditional on request		
Average # requests per student	2.71 (1.28)	3.02 (1.22)
% requesting w/in choice set	13.39	43.39
% of all requests w/in choice set	16.17	54.32

Standard deviations in parentheses, next to sample means.

being similar across the two subsets of truthful students supports the validity of the extrapolation argument made above. As expected, estimates obtained from the larger 'top + short-list' sample are more precise —most standard-errors are reduced by half relative to the sample of 'top' students. Results of a sensitivity analysis regarding the bandwidth choice for the 'top' sample are also presented in Appendix D. Tables 17 and 18 in Appendix D show results derived with alternative bandwidth sizes (twice, five and ten times the original bandwidth size, namely). Utility coefficients derived under the original bandwidth and a bandwidth of twice its size are not identical, but, for many of them, similar. This suggests that results are robust to small changes in the bandwidth size. As the bandwidth size increases, estimated coefficients are increasingly different from the original estimates from Tables 7 and 8, illustrating the increasing bias affecting estimates as more non-truthful students get included in the estimation sample.

Interpretation is easier when distance is used as a numeraire. Column (2) in each table shows ML estimates obtained for a utility function with no quadratic term on distance. While Column (1) shows preferred estimates that will be used in later parts of the analysis (Sections 5 to 6), I use the linear-in-distance utility estimates to comment on parameters in this paragraph.

A positive coefficient on *Past-year median admit* means that students value program quality, as measured by the priority ranking of the past-year median admit of the program —an increase in *Past-year median admit* means that the 2009 median admit had higher priority, hence corresponds to a increase in the program quality. A positive coefficient on squared *Past-year median admit* means that the marginal value of an increase in program selectivity increases with the program level of selectivity. In other words, students' willingness to travel to attend a program with marginally higher quality increases as quality gets higher. The magnitudes estimated with specification (2) suggest that low-SES students are willing to travel 3.0km ($\approx$ 1.875 miles[37]) for a 1-percentile increase of in program quality, all other things equal, against 4.3km for high-SES students..

A positive coefficient on a field dummy means that, on average, students prefer the field to the reference field (Humanities). All field-dummy coefficients being positive means that Humanities is the least-preferred field for average-performing[38] male students. Comparison of field-dummy coefficients show that STEM fields are preferred over non-STEM fields, and that Math/Comp.Sci. is

[37] 3.015×.01/(1.009×.01)=3.0; (3.015+.713)/(1.009-.137)=4.3

[38] High-school performance in STEM and non-STEM are standardized to have mean 0 and standard-deviation 1 in the population of students graduating from high-school with a Math major.

Table 7: Utility parameter estimates (1/2)

	(1) Main	(2) Lin. in distance
Distance (100km)	-2.010***	-1.009***
	(0.07)	(0.05)
× high SES	0.026	0.137*
	(0.04)	(0.06)
Distance (100km) sq.	0.221***	
	(0.01)	
Past-year marginal admit	2.282***	3.015***
	(0.30)	(0.30)
× high SES	0.534	0.713
	(0.39)	(0.39)
Past-year marginal admit sq.	-1.029***	-0.712*
	(0.31)	(0.33)
× high SES	0.958*	0.580
	(0.38)	(0.38)
Distance (100km) × Past-year marginal adm.	0.884***	
	(0.07)	
Degree: Bachelor (LF)	0.547***	0.548***
	(0.06)	(0.06)
× h-s perf.	0.379***	0.384***
	(0.05)	(0.05)
× high SES	0.033	0.009
	(0.07)	(0.07)
Degree: Adv. degree	2.544***	2.529***
	(0.08)	(0.08)
× h-s perf.	1.838***	1.849***
	(0.08)	(0.08)
× high SES	-0.160	-0.193*
	(0.09)	(0.09)
Program location: Tunis	0.520***	0.748***
	(0.08)	(0.10)
Program location: Coast	0.352***	0.395***
	(0.07)	(0.08)
Program location: Abroad	-9.626***	-10.676***
	(1.43)	(1.51)
× STEM h-s perf.	3.448***	3.754***
	(0.72)	(0.77)
× non-STEM h-s perf.	2.254***	2.371***
	(0.35)	(0.38)
× high SES	-0.116	0.177
	(0.36)	(0.37)
Sample	Top + Short	Bdw
PseudoObs.	24,961	24,961
Obs.	3,629	3,629

Std. errors in parentheses, clustered at the high school level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Distance (100km) gives the distance (in 100km) between the program's region and the region of the student high school (as a proxy for home); *Distance (100km) sq.* is the square of this distance. *Past-year (PY) marginal admit* gives the priority score (in percentiles of the priority scores distribution) of the program's marginally admitted student in the past year. A higher score number means a higher priority —for instance, a value of *Past-year marginal admit* of .01 means that the program's 2009 marginally admitted student was at the bottom 1% of the 2009 priority distribution. *Degree:'...'* are indicators of whether the program prepares to the degree considered; these coefficients are allowed to differ continuously × *h-s perf.*, where *h-s perf.* is the student's (standardized) unweighted average score at the end-of-high-school exam. LA (*Licence appliquée*) is used as the reference group for degree dummies. *Program location: '...'* are indicators of whether the program is located in the region considered. '*Southern and western regions*' is used as the reference group for program location dummies. *Field:'...'* (see Table 8) are indicators of whether the program is in the field considered; these coefficients are allowed to differ across sexes, and continuously with high-school performance in STEM and non-STEM fields via the interactions × *female*, × *STEM h-s perf.* and × *non-STEM h-s perf.*. *STEM h-s perf.* and *nonSTEM h-s perf.* are the student's (standardized) unweighted average score at the STEM and non-STEM tests taken in the end-of-high-school exam. 'Humanities' is used as the reference group for field dummies.

Table 8: Utility parameter estimates (2/2)

	(1)	(2)
Field: Arts	2.788***	2.787***
	(0.28)	(0.28)
× STEM h-s perf.	1.672***	1.672***
	(0.18)	(0.18)
× non-STEM h-s perf.	-1.261***	-1.245***
	(0.23)	(0.22)
× female	-0.949**	-0.944**
	(0.31)	(0.31)
Field: Educ.	2.042***	2.073***
	(0.39)	(0.39)
× STEM h-s perf.	1.287***	1.242**
	(0.37)	(0.39)
× non-STEM h-s perf.	-0.725	-0.743
	(0.51)	(0.53)
× female	-1.568**	-1.560**
	(0.51)	(0.52)
Field: Soc. Sc.	0.911**	0.935**
	(0.35)	(0.35)
× STEM h-s perf.	0.820**	0.841**
	(0.32)	(0.32)
× non-STEM h-s perf.	-0.912**	-0.901**
	(0.34)	(0.33)
× female	-1.337**	-1.336**
	(0.41)	(0.41)
Field: Eco/Mgmt	3.647***	3.632***
	(0.29)	(0.29)
× STEM h-s perf.	1.168***	1.200***
	(0.19)	(0.18)
× non-STEM h-s perf.	-1.194***	-1.186***
	(0.23)	(0.23)
× female	-1.056***	-1.050***
	(0.30)	(0.30)
Field: Law	2.189***	2.177***
	(0.40)	(0.39)
× STEM h-s perf.	0.434	0.481
	(0.35)	(0.34)
× non-STEM h-s perf.	-0.408	-0.415
	(0.34)	(0.33)
× female	-1.187**	-1.190**
	(0.42)	(0.42)
Field: Math/Comp.Sci.	4.296***	4.284***
	(0.28)	(0.28)
× STEM h-s perf.	1.317***	1.341***
	(0.17)	(0.17)
× non-STEM h-s perf.	-1.551***	-1.542***
	(0.23)	(0.22)
× female	-1.230***	-1.220***
	(0.30)	(0.30)
Field: Phys./Chem./Engin.	4.074***	4.054***
	(0.27)	(0.27)
× STEM h-s perf.	1.291***	1.311***
	(0.17)	(0.16)
× non-STEM h-s perf.	-1.598***	-1.586***
	(0.22)	(0.22)
× female	-1.391***	-1.383***
	(0.29)	(0.29)
Field: Health/Life Sc.	3.491***	3.445***
	(0.28)	(0.28)
× STEM h-s perf.	1.143***	1.177***
	(0.17)	(0.17)
× non-STEM h-s perf.	-1.310***	-1.286***
	(0.23)	(0.23)
× female	-0.346	-0.324
	(0.30)	(0.30)
Field: Earth Sc.	2.169***	2.173***
	(0.28)	(0.28)
× STEM h-s perf.	0.293	0.301
	(0.18)	(0.18)
× non-STEM h-s perf.	-1.633***	-1.626***
	(0.24)	(0.23)
× female	-0.863**	-0.870**
	(0.30)	(0.30)

the most-preferred field of study —which is not surprising given that students in the sample graduated from high-school with a Math major. All other things equal, low-SES (resp. high-SES) male students are willing to travel 23km (resp. 26km[39]) to study Math/Comp.Sci. rather than Physics/ Chemistry/ Engineering (the second most-popular STEM field among male students), and 65km (resp 75km[40]) to study Math/Comp.Sci. rather than Economics/ Business/ Management (the most-popular non-STEM field among male students). Female students prefer Math/Comp.Sci. over Physics/ Chemistry/ Engineering by the same magnitude as males, and their most-preferred field is also a STEM field —again, as to be expected from high-school Math-majors. In contrast to males, though, female students strongly prefer Health and Life Sciences over Math/Comp.Sci.. Economics/ Business/ Management is also female students' most preferred non-STEM field (at the mean performance levels in STEM and non-STEM), and they generally dislike less non-STEM fields than male students. For students of both sexes, Earth Sciences is the least preferred STEM field. Preferences for field of study are correlated with high-school performance in STEM and non-STEM fields, although most coefficients on field-performance interactions are not statistically significant. Notably though, in my sample of Math-high-school graduates, preference for Earth Sciences strongly decreases as high-school performance in STEM increases. As for non-STEM fields, preference for Humanities strongly increase as high-school performance decreases in STEM subjects and/or increases in non-STEM subjects.

A positive coefficient on a degree dummy means that, on average, students prefer the degree in question over the reference degree ('*licence appliquée*', a type of Bachelor degree designed for students likely to enter the labor market upon graduation). On average, students prefer both the '*licence fondamentale*' (a type of Bachelor degree designed for students likely to pursue a graduate studies upon graduation) and the more advanced degrees to the reference Bachelor-equivalent. Preference for these two types of degree increases with high-school performance —especially for the latter, as expected given the higher academic level certified by this type of degree. On average, students with high-school performance equal to the mean are willing to travel about 54 to 64km, depending on SES, to work towards a *licence fondamentale* rather than a *licence appliquée*; and between 251 and 268km to work towards an advanced degree (e.g., Masters' and M.D.) instead of a *licence appliquée*.[41]

5 Expectations about admission chances

Students' application lists depend not only on their preferences, but also on their expectations about their admission chances. When studying the effect of provision of information about program vacancies on students' chosen application lists, it is important to have a sense of how their expectations about their admission chances relate to the information available to them. Taking preferences as known from the previous section, this section characterizes students' expectations about their admission chances.

A student's admission to a given university program is the result of all students' application deci-

[39] (4.284-4.054)/1.009=.23; and (4.284-4.054)/(1.009-.137)=.26

[40] (4.284-3.632)/1.009=.65; and (4.284-3.632)/(1.009-.137)=1.75

[41] LF: .548/1.009=.54; and (.548+.009)/(1.009-.137)=.64. Adv. degrees: 2.529/1.009=2.51; and (2.529-.193)/(1.009-.137)=2.68.

sions. Indeed, whether or not a seat in that program will be offered to the student by the algorithm depends not only on whether or not the student listed the program in her portfolio, but also on whether or not the program was already filled up as a consequence of other students' assignments and applications. It has been acknowledged in previous studies (e.g., Agarwal and Somaini, 2014; Calsamiglia, Fu and Güell, 2014; Ajayi and Sidibé, 2016; Kapor, Neilson and Zimmerman, 2016) that the complexity of forming expectations about admission chances in such a game setting and solving the expected-utility maximization problem (1) likely exceeds the computational capacity of high-school students and their families, especially as the number of programs students can choose from and the size of the application list they may submit get large. Hence, students have been allowed to behave with limited sophistication, viewing the application decision as a single-agent problem rather than a game. Previous analyses have also established the possible existence of differences in expectations formation and use of public information across socioeconomic status (SES) and related variables (e.g., Hoxby and Turner, 2015).

The approach taken in this section is in line with these concerns. I specify *types* of expectations-formation behavior for students who take their application choice as a single-agent problem. My analysis then proceeds to recovering the *share* of each type in the student population (conditional on students' observables). The identification strategy takes advantage of the fact that utility parameters were recovered from a strict subsample and without taking a stand on students' expectations about their admission chances. Perceived admission probabilities are sought to rationalize, given utility parameters, the application lists submitted by the popstudents who can *a priori* be truthful or strategic. Given utility parameters, I use maximum simulated likelihood (MSL) to estimate the share of various types of expectations formation behaviors in the population.

5.1 Forming expectations about one's admission chances: a model

I allow for two main types of application behavior among students —sophisticated and unsophisticated. I describe these types now, along with the effect of informational updates about programs filling up and vacancies remaining on each of them.

Unsophisticated students. Unsophisticated students simply report in their application list their most-preferred alternatives in their choice set (Agarwal and Somaini, 2014; Calsamiglia, Fu and Güell, 2014). No matter their position in the priority ranking, they are truthful. The provision of information about vacancies, made through the sequential implementation of the application procedure, enables them to update their choice set: they only consider programs that have not been publicly declared full at the time they submit their portfolio.

Sophisticated students. Sophisticated students maximize expected-utility (1) to choose their application portfolio. I assume that sophisticated students form their expectations about their eligibility chances on the grounds of the programs' past-year admission cutoffs and their own priority score, rather than based on a model for other students' behavior. This is a natural approach given the availability of past-year marginal admission scores to students. Specifically, to report the expected-utility-maximizing list, such students derive their expectations assuming marginal admission scores follow, from one year to the next, some AR(1) process of the form:

$$\text{cutoff}_{j,2010} = a + b \times \text{cutoff}_{j,2009} + \eta_j \quad \text{with } \eta_j \sim N(0, \sigma^2). \tag{5}$$

Taking the parameters a, b, σ of the relationship (5) as given, student i's expectation about her probability to clear the admission cutoff for program j is:

$$\mathbb{P}\left(\text{priority}_i \geq \text{cutoff}_{j,2010}\right) = \Phi\left(\frac{1}{\sigma}(\text{priority}_i - a - b \times \text{cutoff}_{j,2009})\right) \tag{6}$$

In this framework, the effect on eligibility chances of an informational update, such as those provided in the Tunisian mechanism, is to reset to 0 the perceived probability of admission to programs which are declared to be full. In addition, when a student's priority ranking within her group is such that there are fewer students in the group to be assigned before her than vacancies publicly declared to be remaining in the program, the student's perceived probability of admission to this program is reset to 1 by the information revelation.

In the Tunisian sequential design, Group 1 and Group 2 students who fail to be assigned to any of their listed choices are pooled at the top of the next group and allowed to participate again time in the application process (see Section 3.1). At that time, they can only pretend to alternatives still available in the choice set of this next group. This 'second chance' affects students' option value of being rejected from all their listed choices —in Equation (1), $V_i(0)$. In the present model of sophisticated behavior, I assume that a student i in Group 1 or 2 computes $V_i(0)$ as follows. Let $groupCutoff_i$ denote the position in the priority ranking of the division between i's application group an the next. Student i can form expectations the choice set she would face at the time of her 'second chance' if she were to use it by using $groupCutoff_i$ instead of $priority_i$ in Equation (6). Then, $V_i(0)$ is the value of the program with highest expected utility, based on these 'second-chance' expectations. The probability that i will use that second chance ($\overline{\pi_i}$ in Equation (1)) is determined by her admission probabilities to her listed choices.

Estimation of the model of expectations formation proceeds in two parts, both described in the next subsection. First, I fix the AR(1) parameters characterizing the sophisticated type; then I recover the respective shares of sophisticated and unsophisticated students in the population. The second part uses data on students' application behaviors; the first part uses panel data on admission cutoffs. While I fix AR(1) parameters and do not recover them from students application choices *per se*, I maintain flexibility in the model by allowing for different sophisticated (sub)types. Hence, different sophisticated students may use different sets of AR(1) parameters to form their expectations about their admission chances.

5.2 Using observed choices to recover types shares

Paragraphs 5.2.1 and 5.2.2 present the second part of the estimation strategy. I explain how, when types are taken as fixed and known, their population shares can be recovered. Paragraph 5.2.3 turns to the first part of the estimation strategy, and describes the way I fix AR(1) parameters for the sophisticated (sub)types.

In the next two paragraphs, it is useful to view each sophisticated expectations-formation type t as a *known* function that takes as inputs the student's priority score and the information about programs that is publicly available at the time she submits her application list, and returns a vector of J probabilities, which is interpreted as the student's perceived eligibility probabilities to

all university programs. A type-t student then uses the J probabilities outputed by the type-t function, along with her preferences, as inputs in the expected-utility maximization problem (1). (Consistently, the unsophisticated type can be viewed as the function that, given the student's preferences and the information about programs that is publicly available at the time she submits her application list, returns the programs with highest flow utility among those that have not been declared full).

5.2.1 A maximum simulated likelihood approach

Suppose each student has one of T discrete expectations-formation types. The probability P_i to observe the actual application list $\mathcal{L}_i$ submitted by student i given her mean flow utilities for all programs $(\bar{u}_{i,j})_j$, and her observable characteristics X_i, writes

$$P_i = \sum_{t=1}^{T} \underbrace{\mathbb{P}\left(\mathcal{L}_i \mid (\bar{u}_{i,j})_j, X_i, \theta_i = t\right)}_{:=p_i(t)} \times \underbrace{\mathbb{P}\left(\theta_i = t \mid (\bar{u}_{i,j})_j, X_i\right)}_{:=\rho(t,X)} \tag{7}$$

where, for types $t = 1, \ldots, T$, $p_i(t)$ is the probability to observe the actual application list $\mathcal{L}_i$ submitted by student i conditional on her being of type t, and $\rho(t, X)$ is her probability to be of type t given her observable characteristics. Since, conditional on observable characteristics, a student's expectations-formation type is independent of her vector of mean flow utilities,[42] $\rho(t, X) = \mathbb{P}(\theta_i = t \mid X_i)$.

The type functions being known means that, assuming a student has type t, and given her priority score, I can straightforwardly derive her perceived probabilities of admission to all university programs using Equation (6). Hence, for any fixed type t, and for each student i, since utility parameters are known, $p_i(t)$ can be estimated by simulation (of choices, over preferences unobservables). Given the set of possible types, and estimated conditional choice probabilities $\hat{p}_i(t)$ for all $(i, t) \in \{1, \ldots, N\} \times \{1, \ldots, T\}$, the type shares $\rho := (\rho(t, X))_t$ can then be recovered by maximizing the (simulated) sample likelihood:

$$\mathbf{L}(\rho) = \prod_{i=1}^{n} P_i = \prod_{i=1}^{n} \left(\sum_{t=1}^{T} \hat{p}_i(t) \times \rho(t, X_i) \right).$$

5.2.2 Identifying variation

If two expectations-formation types t_1 and t_2 are such that $p_i(t_1) = p_i(t_2)$ for all $i = 1, \ldots, N$ such that $X_i = X$, their respective shares $\rho(t_2, X)$ and $\rho(t_1, X)$ cannot be separately identified. The estimation strategy exploits the fact that, given her (known) mean utility, a student does not have the same probability to submit the application list she did submit (which is observed in the data) conditional on being of two different (sub)types. Precisely, the identification of conditional type shares $\rho(t, X)$ in (7) for the specified types $t = 1, \ldots, T$ relies on $p_i(t)$, the (simulated) likelihood to observe the data conditional on X_i, varying across t. Figure 14 in Appendix E illustrates the identifying variation in the data, and the way it is enters in the estimation strategy.

[42] This follows from the fact that, conditional on observable characteristics, utility parameters are independent of student's expectations and level of sophistication. This holds under the assumptions made in Section 4.

In practice, given the size of the choice set and programs' observable characteristics, there is more variation across expectations-formation types in the likelihood (given preference parameters and individual characteristics) of observing the *characteristics* of students' chosen programs (shown in Figure 14) than in the likelihood of observing the *identity* of their chosen programs. Hence in the implementation of the MSL, rather than the definition given in Equation (7), I use

$$p_i(t) = \mathbb{P}\left(\mathbf{Y}(\mathcal{L}_i) \mid (\bar{u}_{i,j})_j, X_i, \theta_i = t\right)$$

where $\mathbf{Y}(\mathcal{L}_i)$ is a vector of program characteristics of the programs listed by i. For instance, $\mathbf{Y}(\mathcal{L}_i)$ may include the selectivity level (in terms of past-year admission cutoff), the distance home-university for student i, the number of vacancies publicly announced to be remaining a the beginning of i's application group.[43]

5.2.3 Fixing types

Fixing the types can be done in a very flexible way. The set of possible types *a priori* allowed can be large, and the estimation procedure allows the data to dictate which types have positive probability in the population. The only limitations on the set of possible types are imposed by identification requirements. If two sets of expectations-formation parameters induce the same application behavior for all students, their shares in the population cannot be separately identified —regardless of whether they induce the same expectations for all students or not.

The results shown in Section 5.3, are based on allowing for six sophisticated (sub)types, in addition to the unsophisticated type. Each sophisticated (sub)type corresponds to a different specification of the AR(1) equation (5). Specifications differ from one another in the level of observable heterogeneity across programs accounted for in (5). I characterize these specifications more in detail in Section 5.3, as I describe and interpret results. The choice of AR(1) coefficients for these specifications is based on data on programs' 2009 and 2010 marginal admission scores. Namely, using this cutoffs data, I estimate the AR(1) processes characterizing each sophisticated type. Estimated coefficients, which I use in simulations to recover conditional choice probabilities (7), are showed in Appendix E.[44]

5.3 Results

5.3.1 Estimated types.

Table 9 shows types shares estimated conditional on socioeconomic status (SES) and for six sophisticated subtypes. Estimated shares of unsophisticated students are robust to changes in the

[43] Because the estimator does not use all the available data, it is not efficient.

[44] One may think that the ideal approach would allow to recover not only the shares of each type, but the parameters characterizing the types as well. Such approach would however face two pitfalls. The first is the identification issue already mentioned at the beginning of paragraph 5.2.3. The second is a computational issue. Consider the following (already restrictive) framework. Suppose each student is of one of $T < +\infty$ discrete types of expectations formation processes. Suppose a student of type t forms her expectations about her admission chances by assuming that the changes in admission cutoffs from one year to the next are normally distributed so that: $\text{cutoff}_{j,2010} = \text{cutoff}_{j,2009} + \nu_j$ where $\nu_j \sim \mathcal{N}(0, \sigma_t^2)$. The variances $(\sigma_t^2)_t$, along with the shares $(\rho)_t$, are to be recovered. *Each* evaluation of the likelihood $\mathbf{L}(\sigma^2, \rho)$ requires simulating choices for *all* students under type t to estimate the conditional choice probabilities $p_i(t)$. Optimization over a set of values for σ^2 quickly gets very demanding. Fixing the types, simulations and estimation of the conditional choice probabilities $p_i(t)$ can be done once and for all outside the optimization routine, and hence greatly simplify estimation. In this case, an evaluation of the likelihood $\mathbf{L}(\rho)$ simply involves reweighing this estimated conditional choice probabilities $p_i(t)$.

set of AR(1) specifications characterizing sophisticated subtypes, and conditioning on other demographics, such as sex, region of origin, or interactions of these demographics with SES. Shares for the main two types —sophisticated and unsophisticated– are shown in bold font, along with a breakdown of the sophisticated type into its subtypes. Slightly less than two thirds of the low-SES students are estimated to behave naively, against half of high-SES students. These large shares of unsophisticated behaviors are not necessarily surprising. In the context of the NYC high school match, which is based on a single-phase DA with a restricted list of 12 choices, Abdulkadiroğlu, Agarwal and Pathak (2017) show evidence that at least 80% of applicants (and very possibly all of them) truthfully report their most preferred programs in their application list. Focusing on an alternative assignment mechanism that highly rewards sophisticated behavior (a variant of the Boston mechanism), Agarwal and Somaini (2014) estimate that a third of families participating in the elementary school match in Cambridge, Massachusetts behave in a unsophisticated way, the other two thirds behaving as if they knew their true admission probabilities. Consistent differences in expectations-formation, sophistication, and application behaviors across SES and related variables have also been documented in other studies (e.g., Hoxby and Turner, 2015).

Among sophisticated students, most students, regardless of SES, form expectations about their admission chances using an AR(1) whose parameters differ (at least) for programs in different fields of study. About 12% of both low- and high-SES students use an AR(1) whose parameters differ only across programs' fields of study. In addition, 26% of high-SES students and 16% of low-SES use a finer AR(1) process allowing different parameters not only across field of study, but also across programs' capacity filling status in the previous year or (quantiles) of selectivity level. 6% of high-SES students and 3% of low-SES students use an AR(1) whose parameters differ only along either or these two dimensions —programs' capacity filling status in the previous year or (quantiles) of selectivity level. Finally, 4% of students, regardless of SES, are estimated to form expectations about their admission chances using an AR(1) specification that does not account for any kind of heterogeneity across programs.

Table 9: Estimated shares of expectations formation behaviors

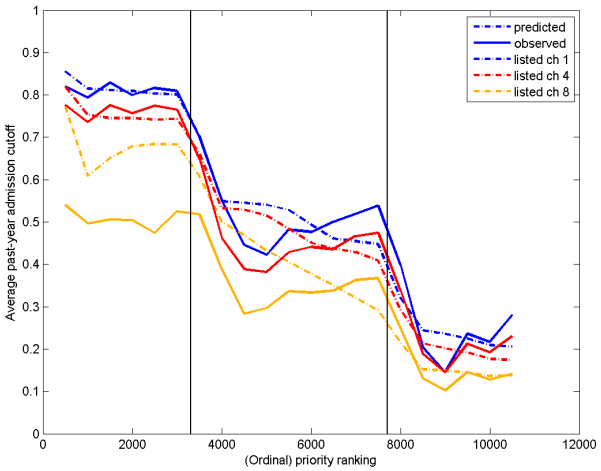
	low-SES stud.	high-SES stud.
Unsophisticated	**0.64**	**0.50**
Sophisticated	**0.36**	**0.50**
Homogeneous AR(1)	0.04	0.04
AR(1) w/ heterog. across fields	0.12	0.13
AR(1) w/ heterog. across capacity filling status in 2009	0.02	0.03
AR(1) w/ heterog. across selectivity levels	0.01	0.03
AR(1) w/ heterog. across field, and capacity filling in 2009	0.08	0.12
AR(1) w/ heterog. across field, and selectivity level	0.08	0.14

5.3.2 Model fit

Figure 3 plots the selectivity level of students' choices as a function of their priority ranking. For clarity, it focused on students' first, fourth, and eighth-listed choices. Solid lines represent choices observed in the data; dotted lines represent choices predicted given utility parameter estimates

from Section 4 and estimated expectations-formation types shown in Table 9. Figure 3 suggests that the estimated model is able to reproduce the variation in the selectivity level of students' listed choices as a function of their priority rank; and predicts reasonably well the diversification of students' application portfolios in terms of selectivity level.

Figure 3: Selectivity level of predicted vs. observed choices



Note: This graph plots the selectivity level (in terms of past-year admission cutoff) of students' choices as a function of their priority ranking —focusing on students' first, fourth, and eighth-listed choices. Solid lines represent choices observed in the data; dotted lines represent choices predicted given utility parameter estimates from Section 4 and estimated expectations-formation types shown in Table 9.

6 Understanding the value of information

In this section, I evaluate the effects of informational updates in a restricted-list DA mechanism. To evaluate these effects, I use simulations, and compare students' outcomes under the standard single-phase implementation of the DA and alternative multiple-phase implementations of the mechanism, in which information about available vacancies is publicly updated between phases.

Welfare. I measure average student welfare as the uniformly weighted sum of utilities that students derive from their assignment (indirect utilities). Given the distribution of preferences, a simulation exercise allows me to compute the expected average student welfare induced by a given mechanism. Welfare comparisons made in this section are based on expected average student welfare:

$$W = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[u_{i,\mu(i)} + \varepsilon_{i,\mu(i)}\right]$$

where $\mu(i)$ denotes student i's assignment, and the expected value $\mathbb{E}$ is estimated by simulations over ε.

Counterfactual scenarios. I comparatively evaluate the effects of informational updates by considering four multiple-phase scenarios: dividing the cohort in two, three, four, or five groups by order of priority. When simulating applications in the three-phase mechanism, I divide the cohort into groups as was done in the 2010 Tunisian mechanism (i.e. top 30, middle 40, and bottom 30%).

When simulating applications in the two-, four-, and five-phase mechanisms, I divide the cohort in equally-sized groups. As a benchmark, I also simulate applications in a perfect information setting, publicly updating vacancies after every single assignment. This corresponds to a limit N-phase scenario, where N is the total number of students in the population. The number of phases is the only difference between the scenarios I simulate.

I first show that, when applying under the single-phase restricted list DA, and relative to the perfect-information benchmark, their average expected indirect utility is significantly decreased. While easy to implement, the 2010 Tunisian three-phase implementation of the restricted-list DA reduces by 67% the welfare loss induced by the implementation of a standard (single-phase) restricted-list DA, relative to the perfect information benchmark. Investigating the mechanisms underlying these changes in indirect utility, I show that expected indirect utility gains essentially accrue to students who fail to be admitted to any of their listed elements under a single-phase mechanism and who gain assignment because of the informational updates —rather than to assigned students improving their match. Exploring heterogeneous effects across students with different ability, sophistication, and socioeconomic backgrounds, I find that gains disproportionately accrue to low-ability, unsophisticated, and low-SES students. In fact, providing information about vacancies, even through a small number sequential of sequential phases, reduces the expected indirect utility gap existing between high- and low-SES students. Finally, while the 2010 Tunisian implementation of the three-phase procedure does increase welfare and the average match rate, I show that a better targeting of low-priority students by the information provision —through a different division of the cohort into three groups— could increase gains to students.

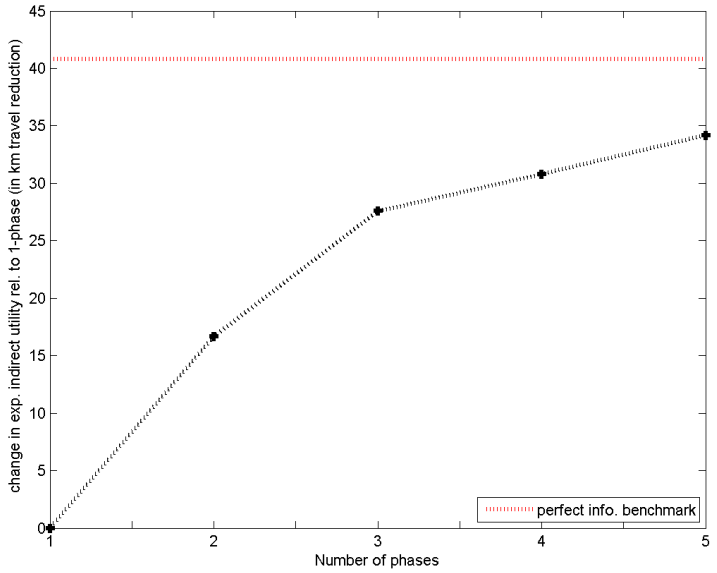
6.1 Effects of information-revelation on expected welfare and assignment rates

6.1.1 Welfare

Average welfare gains. Figure 4 shows, as a function of the number of phases implemented, the difference in student welfare relative to the single-phase scenario. The horizontal dotted red line shows the difference in expected average student welfare between perfect information and to the one-phase implementation. A positive difference in welfare means that, on average, students derive more utility from their assignment under a multiple-phase mechanism than under the standard single-phase DA. Under the perfect-information benchmark, the average indirect utility is higher than in the single-phase DA by the equivalent of a 41km-reduction in distance traveled. As a reference, distance to actual assignments has mean 145km ($\approx$ 90 miles), median 107km, and standard deviation 200km in the data. Comparing alternative multiple-phase scenarios suggests that welfare gains increase as the number of information revelations made increases, although the marginal value of an extra phase seems to be decreasing. Under the two-, three-, four-, and five-phase scenarios, indirect utility gains average an equivalent of about 17km, 28km, 31km, and 34km travel-distance reductions, respectively. In particular, this means that a three-phase implementation of the DA, as done in Tunisia, reduces by about 68% the loss in average utility induced by using a single-phase restricted-list DA in a environment where students face uncertainty about their admission chances, relative to the perfect-information benchmark.

Distribution of indirect utility gains. Table 10 reports selected quantiles of the distributions of indirect utility gains under each scenario, relative to the single-phase implementation. Under

Figure 4: Change in expected average student welfare relative to single-phase implementation of the restricted-list DA



each multiple-phase scenario, the range of gains is pretty large. Under the 2010 three-phase mechanism for instance, the first and last percentiles of indirect utility gains are equivalent to a 41km *in*crease in distance traveled, and a 192km *de*crease in distance traveled, respectively. Although individual losses can be significant, under each multiple-phase scenario, the share of students hurt by the multiple-phase mechanism (relative to the standard single-phase DA) is relatively small as compared to the share of students who (weakly) benefit from it (null gains correspond to the 15th, 11th, 10th, and 9th percentile under the two-, three-, four-, and five-phase scenario, respectively). Interestingly, the perfect information setting does not constitute a Pareto improvement relative to the single-phase implementation of the DA. Indeed, 8% of students derive less utility from the assignment they obtain under perfect information than under the single-phase DA. The next section sheds light on the mechanisms underlying these facts.

Table 10: Distribution of utility gains in *ex-post* flow utility (in km): selected quantiles

	1st pct	5th pct	25th pct	50th pct	75th pct	95th pct	99th pct
Two-phase	-67	-22	0	5	25	93	147
Three-phase	-41	-5	0	7	38	134	192
Four-phase	-28	-4	1	9	43	143	200
Five-phase	-16	-3	1	9	49	152	207
Perfect Info.	-8	-2	1	11	62	174	230

6.1.2 Underlying mechanisms

There are two margins through which a student's indirect utility may change from the single-phase implementation of the DA to a multiple-phase implementation. The first is a change in assignment status; the student fails to be assigned to any element of her list under the single-phase

implementation (and is therefore administratively assigned to a left-over seat) while she managed to be assigned to one of the programs she listed under the multiple-phase implementation —or vice-versa. The second is a change in assignment, holding the assignment status fixed; under both implementations, the student is assigned to one of her listed elements, but the program she is assigned to changes —or symmetrically, she is administratively assigned in both cases but her assigned left-over program changes. The former can be seen as a change in indirect utility at the extensive margin, and the latter as a change at the intensive margin. Table 11 decomposes the welfare gains shown in Figure 4 into these two mechanisms. It shows the average share of students who, under the multiple-phase scenarios and relative to the single-phase implementation, switch assignment status, and the share of those who do not. Table 11 also shows the average change in indirect utility experienced by students with each of the four assignment-status pairs. Figure 15 in Appendix F provides additional information by showing the distribution of indirect utility changes (relative to the single-phase DA) within each assignment-status-pair group.

Table 11 shows that an increase in the match rate is the main mechanism underlying the increase in average indirect utility induced by the revelation of information. Under the single-phase implementation, 9.1% of students end up administratively assigned. These students all gain assignment (i.e. match) under the perfect information benchmark. As a consequence, they experience a large average expected indirect utility gain —equal to more than 10 times the population-average expected indirect utility gain. By contrast, the 90.9% of students who are assigned under both the single implementation and the perfect information setting experience on average little expected indirect utility changes. As Figure 15 shows, students who are assigned under both scenarios may experience an increase or decrease in expected indirect utility when information in revealed. Increases in indirect utility are due to some students failing to apply to a desired program under the single-phase implementation, because they expect their admission chances to be low, while the program would actually have had a seat for them (which they are able to claim under perfect information). The slight worsening of some assignments among 'always-assigned' students is the result of equilibrium effects, given programs' finite capacities. Some students with higher-priority improving their assignment or gaining assignment takes away from lower-priority applicants spots that are available under the one-phase implementation. That is, some students are better off under the single-phase mechanism than under a perfect-information setting (last row of Table 10) because they benefit from other students' misplacement in the absence of informational updates.

While under perfect information no one is administratively assigned, a few students fail to be admitted to any element of their list under the other multiple-phase scenarios. Among the 9.1% of students administratively assigned under the single-phase scenario, 82% (hence 7.5% of the population) switched to being assigned to an element of their list when applying under the 2010 three-phase mechanism. These students experience large welfare gains. The other 18% of the students administratively assigned under the single-phase scenario keep this assignment status under the three-phase scenario. On average, these students experience a decrease in expected indirect utility —in magnitude lower than the average gain. This is again the consequence of equilibrium effects. As more students get assigned to desired programs, leftover programs that remain available for administrative assignment get worse[45] (as suggested by the larger loss experienced by 'never-assigned'

[45] No explicit rule is provided relative to how administrative assignments are made. In simulations, for each student who fail to be assigned to any element of her list, I randomly set the administrative assignment using a uniform distribution over the 50% closest seats that are leftover at the end of the DA. Random administrative assignments

students when moving from one to five phases, than to three phases). Finally, a few students fail to be assigned to any element of their application portfolio under the multiple-phase mechanism while they were under the single-phase mechanism —less than 1% (resp. .5%) of students when switching from one to three (resp. five) sequential phases.

Table 11: Changes in assignment status and associated changes in utility

	Changes on the extensive margin				Changes on the intensive margin			
In 1-phase, students are ...	*... admin. assigned*		*... matched*		*... matched*		*... admin. assigned*	
In multi.-phase, students are ...	*... matched*		*... admin. assigned*		*... matched*		*... admin. assigned*	
	%	ΔW	%	ΔW	%	ΔW	%	ΔW
Perfect info.	9.1	510	0	–	90.9	-6	0	–
Three-phase	7.5	544	0.9	-648	90	-8	1.6	-31
Five-phase	8.5	523	0.4	-635	90.4	-8	0.6	-38

6.2 Heterogeneous effects by ability, sophistication and SES

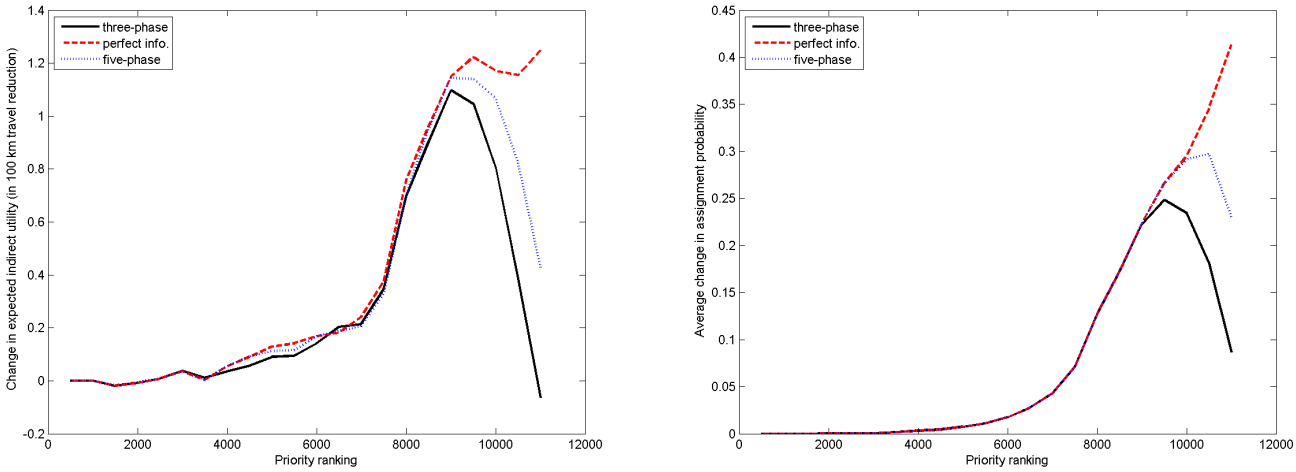
6.2.1 Gains and priority ranking

The left panel of Figure 5 plots indirect utility gains as a function of students' priority ranking under the three- and five-phase DA (relative to the single-phase DA), and under the perfect-information benchmark. Other multiple-phase scenarios yield similar graphs. Close to all welfare gains accrue to students in the second half of the priority distribution. The right panel of Figure 5 plots students' probability to be assigned to one of their listed choices as a function of their priority ranking under the various multiple-phase implementations. It echoes the findings of the previous section that the larger gains in indirect utility accrue to students with larger drops in their probability to be administratively assigned. These graphs highlight the fact that, when students do not know their true admission probabilities, the accuracy of their expectations decreases as the state of the world to be forecast gets further from the one they have information on. Under the single-phase restricted-list DA, the lower a student's priority ranking, the larger the number of students to be assigned before her. In other words, the larger the number of random events (assignments) to alter the initial state of the world before she gets to be assigned. The revelation of information, as done in the Tunisian mechanism, increases low-priority students' average expected indirect utility by bringing them 'closer' to up-to-date information.

As mentioned earlier, the sequential implementation of the DA may affect applicants' behaviors through two channels. The provision of information about programs filling up enable (later groups) students to update their expectations about their admission chances. In addition, the introduction of a 'second chance' to (early group) students if they fail to be assigned to any of their listed choices increase students' option value of being rejected. Gains being small in the first part of the priority ranking (and in particular in what becomes Group 1 under the three- and five-phase scenario) suggests that the latter channel has a small effect on welfare relative to the former.

using a uniform distribution over all leftover seats yield similar results.

Figure 5: Changes in indirect utility and assignment probability as a function of priority ranking



Under the three-phase scenario, average welfare gains are negative at the very end of the priority ranking; the very last students are the most likely to have their assignment worsened as a result of equilibrium effects, while at the same time remaining relatively far from the information and so relatively likely to fail to be assigned to any of their choices.

6.2.2 Gains and sophistication

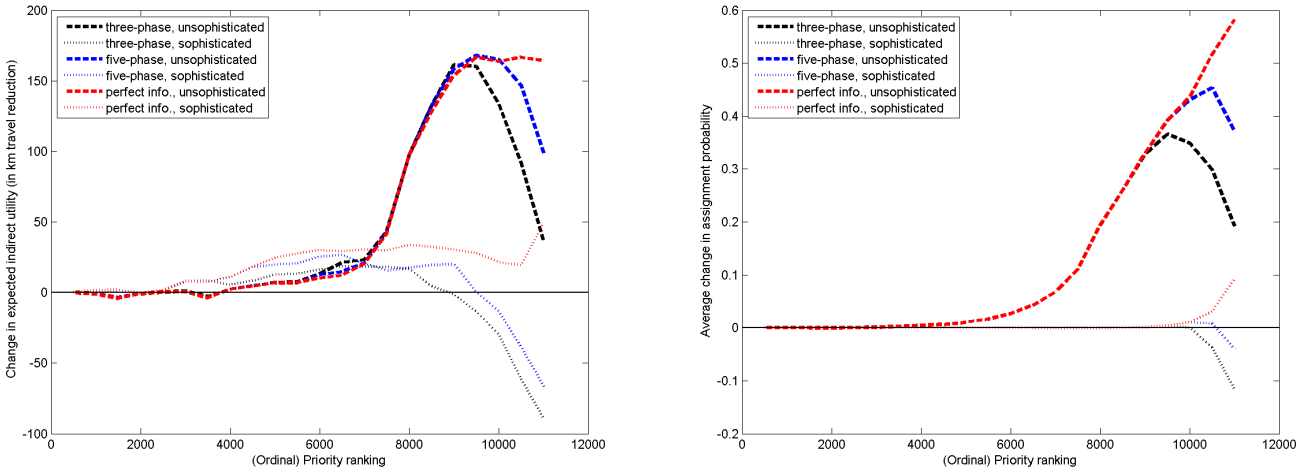
Figure 6 is analogous to Figure 5 but shows average welfare gains as a function of priority separately for sophisticated and unsophisticated students. The left panel shows that increases in the match rate are mostly experienced by unsophisticated students. Most sophisticated students maintain their assignment status relative to the single-phase scenario. At the very end of the priority ranking though, sophisticated students, on average, experience a decrease in their match rate. This is the result of equilibrium effects described earlier; as higher-priority students who end administratively assigned under the single-phase implementation manage to gain a match with the sequential implementation, fewer seats become available at the end of the priority ranking, increasing students probability to be rejected from all their listed choices. This also explains why the increase in match rate dips down at the end of the priority ranking for unsophisticated students.

The right panel shows that largest increase in average indirect utility accrue to unsophisticated students. Given that unsophisticated students are those who benefit from an increase in their match rate, this is consistent with the mechanisms established in Section 6.1.2. Among sophisticated students, significant increases in average expected indirect utility are essentially experienced by mid-priority students (ranks 3,000–9,000 under three phases; ranks 3,000–10,000 under five phases). The decrease in average expected indirect utility experienced by sophisticated students at the bottom the priority ranking experience follows the same causes as the decrease in their match rate.

6.2.3 Gains by demographics

The previous two paragraphs have established that the extent to which a student gains or not from the revelation of information depends on her position in the priority ranking, and the level of sophistication with which she forms expectations about her admission chances. Statistics from Table 1 and estimates from Table 9 show that a student's position in the priority ranking and

Figure 6: Changes in indirect utility and assignment probability as a function of priority ranking by sophistication type



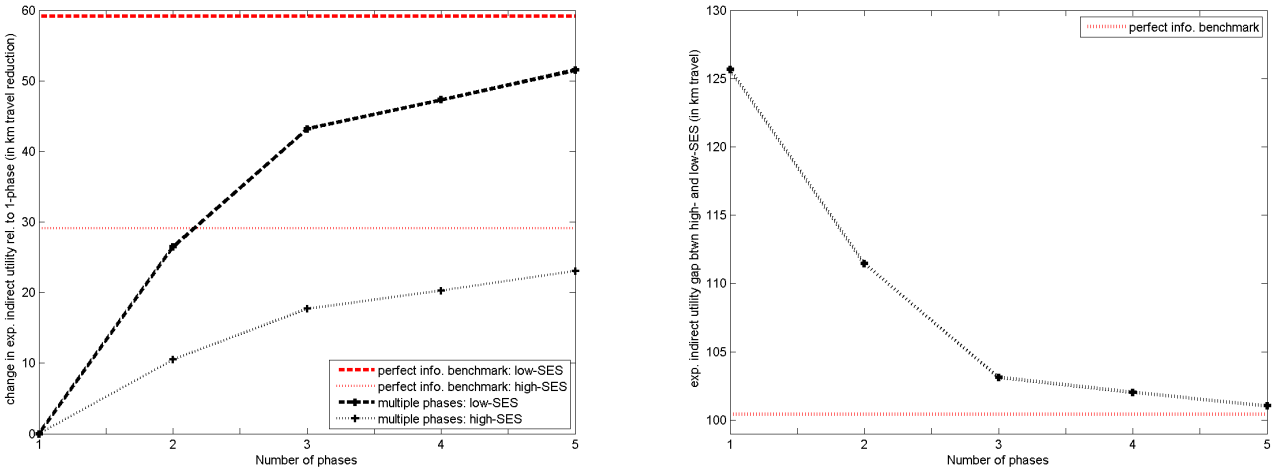
the level of sophistication of her beliefs are correlated with her socioeconomic background. As a consequence, welfare gains from information revelation differ across SES.

The left panel of Figure 7 shows, as function of the number of sequential phases implemented, the average expected indirect utility gains experienced relative to the single-phase scenario, separately for low-SES (thicker dashed black line) and high-SES (thinner dotted black line) students. The horizontal red lines show, separately for low- (thicker dashed red line) and high- SES (thinner dotted red line), the difference in expected average student welfare between the perfect-information and one-phase settings. On average, and in terms of expected indirect utility, low-SES students benefit more from the information revelations than high-SES students do —in other words, low-SES students are more hurt by the single-phase implementation of the DA than their high-SES counterparts are. Switching from the single-phase implementation to the perfect information setting increases average expected indirect utility for low-SES students by an equivalent of a 59-km reduction in travel distance, against 29km for high-SES students. Switching from the single-phase implementation to the 2010 Tunisian three-phase setting increases average expected indirect utility for low-SES students by an equivalent of a 43-km reduction in travel distance, against 18km for high-SES students.

As a consequence of low-SES students being more hurt by the single-phase implementation of the DA than their high-SES counterparts are, the provision of information, through sequential implementation of the mechanism, reduces the welfare gap existing across low- and high-SES students. The right panel of Figure 7 shows, as function of the number of sequential phases implemented, the difference in average expected indirect utility between high- and low-SES students. The horizontal dotted red line show the level of this welfare gap in the perfect-information setting[46]. Switching from the perfect information setting to the single-phase implementation increases the welfare gap across SES by 25%. Switching from the single-phase implementation to the 2010 Tunisian

[46]The welfare gap across SES that persists under the perfect information benchmark is due to low-SES students being matched with lower quality programs (because they are more heavily distributed at the bottom of the priority ranking), and having different preferences for program characteristics. In particular, low-SES students have a larger disutility from traveling (see Table 7); and, due to lower average high-school performance in both STEM and non-STEM fields, derive on average less utility from the different fields of study (see Table 8).

Figure 7: Changes in average expected indirect utility by SES, and changes in the average expected indirect utility gap across SES as a function of the number of sequential phases



three-phase procedure reduces this increase by 88%.

6.3 How much information to give? vs. whom to give information to?

The previous subsection established that, even under perfect information or with information being provided early on as in the three- and five- phase scenarios, most of the welfare gains are generated in the second half of the priority ranking. This suggests that, beyond the amount of information provided (i.e. the number of sequential phase implemented), the points in the priority rankings at which revelations are crucial determinants of welfare gains. In this section, I test this hypothesis by comparing student welfare and match rate under different three-phase scenarios. While I hold the number of sequential phases constant, the division of the cohort in application groups differ across scenarios. I compare the 2010 Tunisian design, in which groups correspond to the top 30%, middle 40% and bottom 30% on the priority distribution (denoted '30/40/30'), to divisions that allow to focus information provision on the lower end of the priority ranking —50/25/25 and 50/37.5/12.5.

The right panel of Figure 8 reproduces Figure 4. The horizontal dotted red line shows the difference in expected average student welfare under perfect information, relative to the one-phase implementation. The black dotted curve shows welfare gains for the multiple-phase scenarios documented in Section 6.1. On this curve, the black dot at the three-phase mark of the horizontal axis represents welfare gains, relative to the single-phase scenario, achieved under the 30/40/30 three-phase implementation. The blue and green dots at the three-phase mark represent welfare gains, relative to the single-phase scenario, achieved under the 50/25/25 and 50/37.5/12.5 three-phase implementations, respectively. The colored dots being above the mark-three black dot means that, relative to the single-phase scenario, average expected indirect utility is increased more under the latter two implementations than by the 2010 Tunisian three-phase procedure. Switching from the single-phase restricted-list DA to the 50/25/25 (resp. 50/37.5/12.5) three-phase scenario achieves 76% (resp. 90%) of the average expected indirect utility increase generated by switching from the single-phase restricted-list DA to a perfect information setting. That is, the 50/25/25 scenario achieves as much as the four-phase implementation documented in Section 6.1; and the 50/37.5/12.5 scenario achieves more than the five-phase implementation. (To ease comparison, the two horizontal thinner black lines show the levels welfare gains achieved by the equally-spaced four-

and five-phase scenarios.)

The left panel of Figure 8 shows that, again, the main mechanism under the increase in average expected indirect utility is an decrease in the share of students administratively assigned. It shows, as a function of the number of phases implemented, the share of students administratively assigned. The blue and green dots at the three-phase mark show the administrative assignment rate under the 50/25/25 and 50/37.5/12.5 three-phase implementations, respectively —which is in both case smaller than under the 30/40/30 implementation (black dot at the three-phase mark).

Figure 8: Change in average expected indirect utility relative to single-phase scenario and assignment rate, as a function of sequential phases

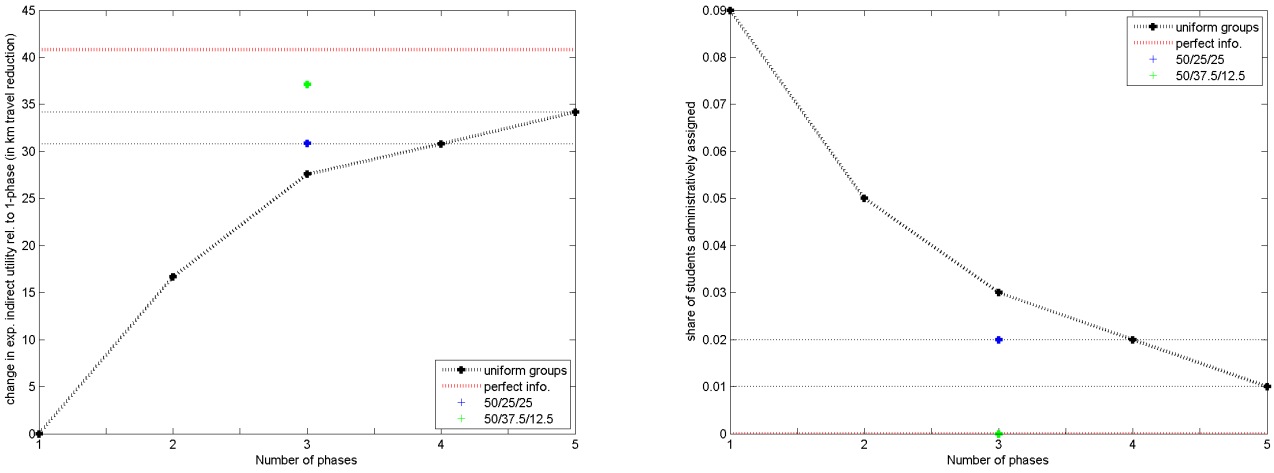
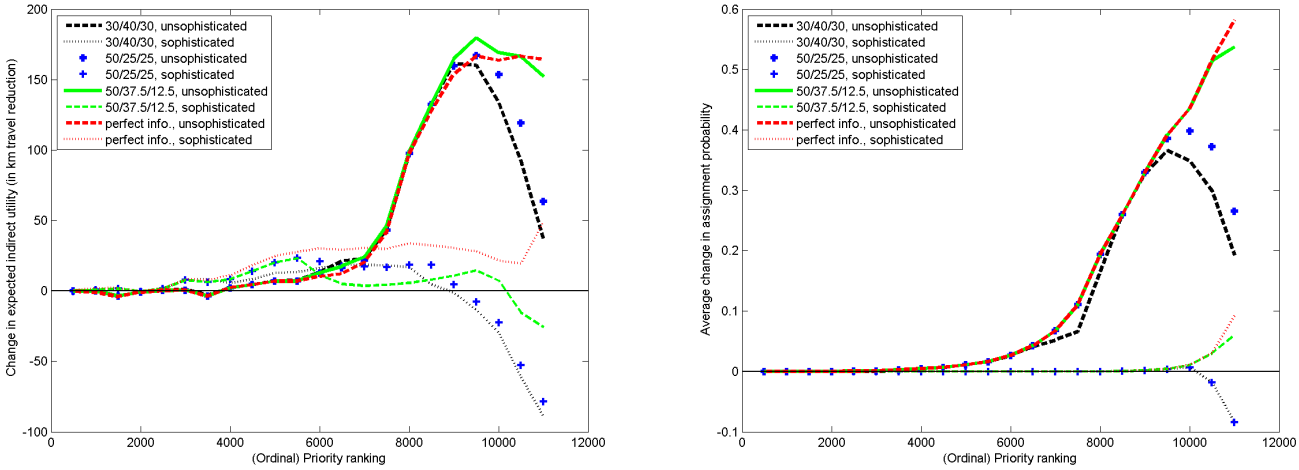


Figure 9 is analogous to Figure 6. It shows average welfare gains (relative to the single-phase implementation of the DA) and match rate as a function of priority separately for sophisticated (thinner plots) and unsophisticated (thicker plots) students. Plots are showed for the perfect-information benchmark (red dotted lines), as well as for the three alternative three-phase scenarios. The black plots are the same as those in Figure 6; they show outcomes for the 30/40/30 implementation. The blue dotted plots and green line plots show outcomes for the 50/25/25 and 50/37.5/12.5 three-phase implementations, respectively. The left panel shows that the 50/37.5/12.5 implementation allows to maintain the upward-sloping increase in matching rate through the entire priority ranking, where the increase dips under the 30/40/30 and 50/25/25 scenarios (and under the five-phase implementation, see the left panel of Figure 6). The increase in matching rate in particularly high for unsophisticated students, but it also benefits very-low-priority sophisticated students —by contrast, these students experience a decrease in match probability (relative to the single-phase implementation) under the other multiple-phase scenarios documented here.

As shown in the right panel of Figure 9, the upward-sloping increase in match rate translates, for unsophisticated students, into a larger increase in average expected indirect utility than under the other multiple-phase scenarios. Just as they do in other multiple-phase implementations, very-low-priority sophisticated students experience, under the 50/37.5/12.5 implementation, a decrease in average expected indirect utility relative to the single-phase scenario. However, this average decrease starts much later (about 1,500 ranks later) in the priority ranking than it does in the other three-phase scenarios, meaning that a larger share of students experience increases in expected indirect utility than under other multiple-phase scenarios. Furthermore, the existing decreases in expected indirect utility (relative to the single-phase DA) are, on average, smaller in magnitude un-

der the 50/37.5/12.5 implementation than under other implementations. In fact, the late provision of information (at the 87.5 percentile of the priority distribution) mitigates the negative equilibrium effects affecting low-priority sophisticated students by enabling them to have more accurate expectations about available seats at the very end of the assignment process.

Figure 9: Changes in indirect utility and matching probability as a function of priority ranking by sophistication type



7 Conclusion

This paper quantifies the welfare effects of enabling students to update their expectations about their admission chances to academic programs in a setting where they cannot apply to all the alternatives in their choice set. It documents a simple way to enable this updating in the context of DA-based assignment mechanisms, which are extensively used around the world to assign students to schools. I estimate a model of application portfolio choice and perform a counterfactual analysis to compare students' application and assignments under scenarios with different levels of updating. I take advantage of a rich administrative data set from Tunisia, where a variant of the DA is used nationwide to assign high-school graduates to university programs. Building on the quasi-experimental design induced by the Tunisian procedure, I am able to recover students' preferences for university programs without taking a stand on their expectations, hence circumventing a common identification challenge in the empirical literature on school choice. I then take preferences as given and characterize students' expectations. This two-step approach partially alleviates the computational intractability of the application portfolio choice problem —the other main challenge faced by the empirical literature on school choice.

I combine preferences estimates and findings about expectations to show that, while easy to implement, a sequential version of the DA can reduce the welfare loss and inequality induced by the standard restricted-list implementation. The 2010 Tunisian three-phase implementation of the restricted-list DA reduces the welfare loss induced by the implementation of a standard (single-phase) restricted-list DA by 67% , relative to the perfect information benchmark. Gains disproportionately accrue to low-ability, unsophisticated, and low-SES students; so providing information about vacancies, even through a small number sequential of sequential phases, reduces the expected

indirect utility gap existing between high- and low-SES students. My results suggest that, while the 2010 Tunisian implementation of the three-phase procedure does increase welfare and the average match rate, a better targeting of low-priority students by the information provision —through a different division of the cohort into three groups— could increase gains to students.

The findings of this paper show that a simple twist in the implementation of the DA can effectively mitigate the consequences of imperfect knowledge of admission chances on welfare and inequality. No data is available on the implementation costs of the sequential procedure, and it is therefore not possible to rigorously compare the costs and benefits of such a policy. However, the small number of sequential phases needed to restore a large share of the loss suggests that the benefits of a sequential implementation are likely to exceed its costs. The Tunisian example also demonstrates that the information revelation intervention can be implemented at a large scale.

The application problem faced by Tunisian high school graduates is no different from the one faced by students in other places. The inefficiencies that can arise with the single-phase implementation of the restricted-list DA, such as a large rate of administrative assignment, have been documented in other instances (e.g., Ajayi and Sidibé, 2016). It is likely that a sequential implementation of the DA would have the same benefits there as in Tunisia —at a limited cost since fixed costs are already being paid in the non-sequential implementation of the DA. Beyond school and college choice, the DA is used as an assignment mechanism in other contexts where information is imperfect and application is costly. A sequential implementation may improve outcomes there as well.

This paper takes as given and fixed the restrictions placed, in virtually every school choice implementation of the DA, on the number of alternatives students can list in their application portfolio. There is evidence that while they value the strategy-proofness implied by the DA when no constraint in imposed on the size on the application portfolio, policy-makers have been unwilling to lift list size restrictions (Pathak and Sönmez, 2013; Roth, 2015). An interesting question for future research is to understand the reasons behind this reluctance. In settings where the choice set faced by students is large (such as in Tunisia, but also for instance in NYC where students can choose from 700+ high schools), a natural hypothesis is that it is costly for applicants to process information, learn about, and precisely assess their preferences for all existing alternatives. In this case, allowing students to downsize their choice set, by revealing which programs are full by the time they get to apply, may be an additional benefit of the sequential design. I plan to investigate this in future work.

A Appendix: Theory

This appendix provides supplemental information to Section 2.
Part A.1 gives a more general description of the DA. Part A.2 gives a proof of Proposition 2.

A.1 Deferred acceptance algorithm (Gale and Shapley, 1962)

DA (1)
Step 1/ Schools receive applications from students who ranked them first in their list. Schools that received fewer applications than their capacity hold on to these applications. Each school j that received more applications than its capacity q_j sends rejection decision to applicants: it temporarily hold on to the q_j applicants with highest priority, and rejects all others (if any). Students receive the rejection notifications sent.
Step (k+1)/ For any $k \geq 1$, students who received a rejection notification at step k send an application to the school ranked next on their list. Schools received these applications. Schools then consider their total pool of applications –those just received, and those they held on at step k (if any). Schools which have fewer applications than their capacity hold on to these applications. Each school j with excess applications sends rejection decision to applicants: it temporarily hold on to the q_j applicants with highest priority, and rejects all others (if any). Students receive the rejection notifications sent.
Stop/ The algorithm stops after all students who received rejections have exhausted their list of acceptable schools. School formally admit applicants they hold on to at this stage.

A.2 Proof of Proposition 2

Proposition 2. (a) Condition 1 (below) is a sufficient condition for students not have a strict incentive to misreport their preferences over their choice set.
(b) Under Assumption 1 (below), Condition (1) is a sufficient condition for students not misreport their preferences over their choice set.

Condition 1. Student i has a perceived *eligibility* probability 1 for (at least) one of her ten most-preferred programs (among those not declared to be full).

Assumption 1. When indifferent between doing so or not, a student does not left-censor nor mis-order her application list relative to her unconstrained preference ranking (over her own choice set). In other words, a student left-censors or mis-orders her application list relative to her unconstrained preference ranking (over her own choice set) only when it is *strictly* profitable to do so.

Proof. Deviation from truth-telling involve at least one of the following:

- Misrepresenting one's preferences by not reporting one's M most-preferred alternatives –i.e. reporting a subset of alternatives that is left-censored or not consecutive relative to one's unrestricted preference ranking.
- Misrepresenting one's preferences by not reporting alternatives in decreasing order of flow utility –i.e. reporting a subset of alternatives that is not well ordered relative to one's unrestricted preference ranking.

First, reporting a subset of alternatives that is not well ordered relative to one's unrestricted preference ranking is never *strictly* profitable since one needs to be rejected from an higher-ranked choice in order to be considered for admission into a lower-ranked choice.

Now, suppose Condition 1 holds for some student i. WLOG, say student i thinks she has probability 1 to clear the *ex-post* cutoff of program j^*, which is ranked second in her unrestricted preference ranking. Denote j_1 the alternative ranked first in her unrestricted preference ranking.

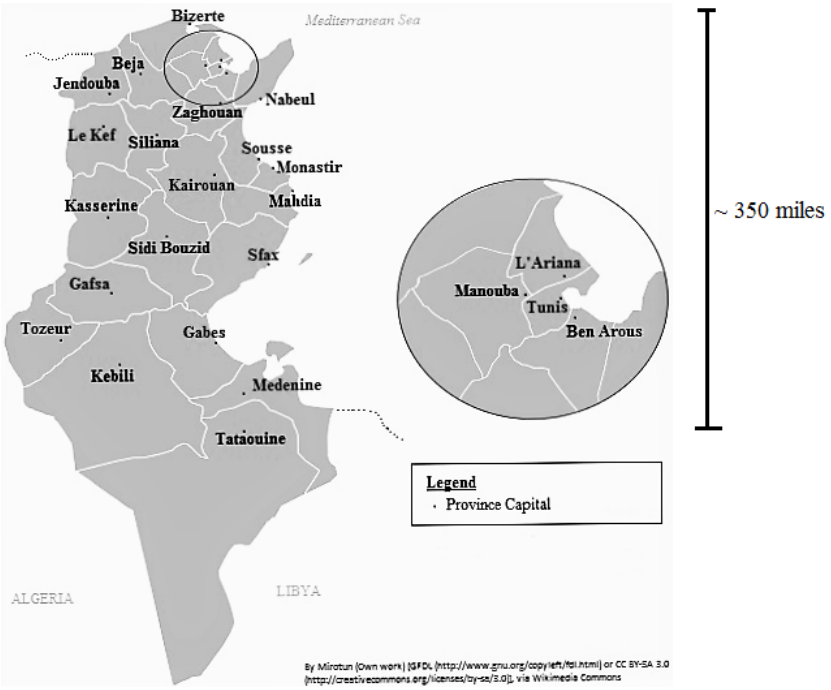
- If i thinks she has probability 0 to clear the cutoff of j_1, she is indifferent between any ordered list starting with: $\{j_1, j^*\}$ or $\{j^*\}$. In that case, it is therefore not strictly profitable to omit j_1 from the list.
- Suppose i thinks she has probability $p_1 > 0$ to clear the cutoff of j_1. Submitting ordered list $\{j_1, j^*\}$ she thinks she will be assigned to j^* unless she is assigned to j_1, which she prefers the j^*, and which occurs with non-0 probability. In expectation, she is then better of submitting $\{j_1, j^*\}$ than $\{j^*\}$, with which she would be admitted to j^* with probability 1.

This shows Part (a) of Proposition 1. Part (b) follows directly from Part (a) and Assumption 1.

B Institutional background & data

To provide some geographical context to the empirical analysis and results in the paper, this appendix shows a map of Tunisia.

Figure 10: Tunisia

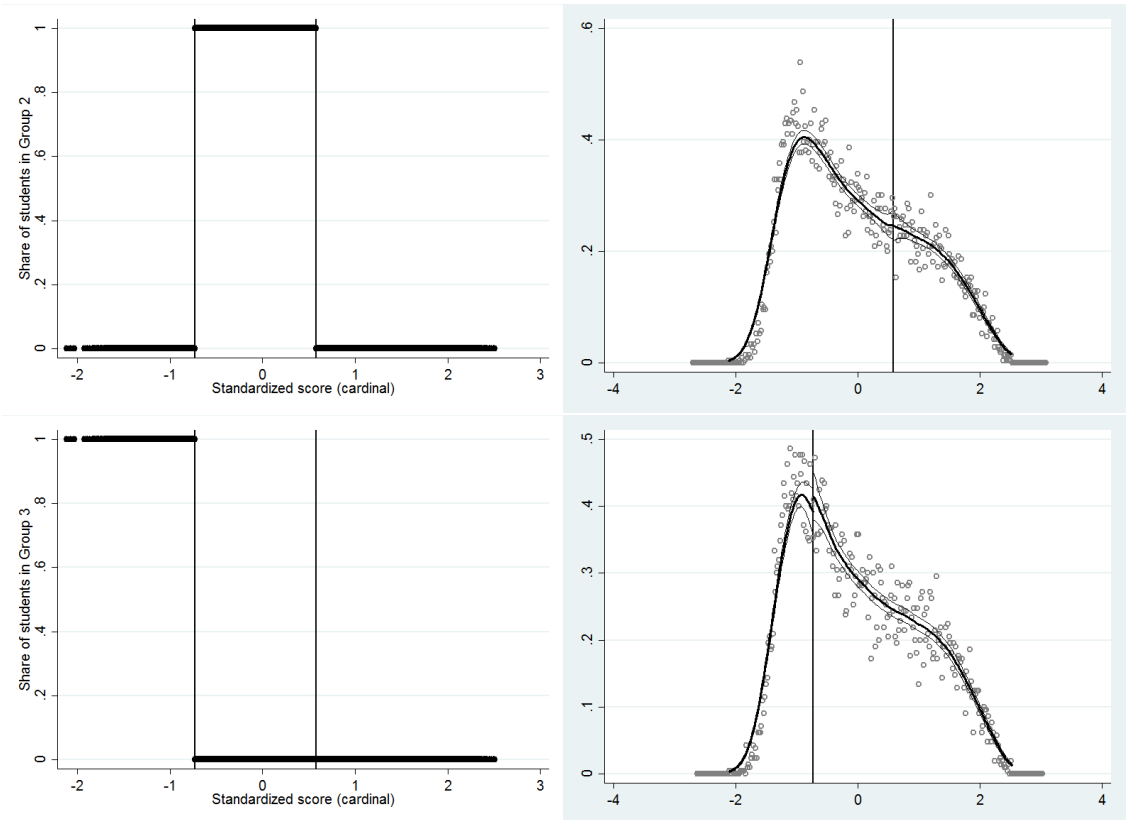


C Sharpness & validity of the regression discontinuity design

This appendix provides standard graphical evidence supporting the sharpness and validity of the regression discontinuity (RD) design used in the analysis of Section 3.3.

The division of the applicant pool in three groups creates assignment cutoffs. The top left panel of Figure 11 displays students' probability to be assigned to Group 2 as a function of the running variable, that is, their priority score (or priority ranking, in ordinal terms). This probability jumps from 0 to 1 at standardized score .578 (score, 3,307 the Group 1/Group 2 cutoff), indicating the sharpness of the discontinuity. The probability drops back from 1 to 0 at standardized score -.726 (rank 7,708, the Group 2/Group 3 cutoff). The bottom left panel of Figure 11 shows students' probability to be assigned to Group 3 as a function of the running variable. It jumps from 0 to 1 at standardized score -.726.
I use McCrary's test as an evidence of the validity of the design, and reject the presence of a discontinuity of the running variable at standardized scores -.578 (estimated discontinuity -.004, with standard error .080), and -.726 (estimated discontinuity .069, with standard error .065). The right panel of Figure 11 illustrates the test, showing the density of the standardized test score.

Figure 11: Sharpness and validity of the RD design: graphical evidence



Note: The top (resp. bottom) left panel shows students' probability to be assigned to Group 2 (resp. 3) as a function of the running variable. It presents a sharp discontinuity at rank 3,307 (resp. 7,709). The top (resp. bottom) right panel shows the density of the running variable, whose continuity at rank 3,307 (resp. 7,709) cannot be rejected (McCrary test).

Table 12 gives the size of the sample as well as its decomposition into assignment groups.

Table 12: Size of each assignment group for high-school graduates majoring in Math

Group	Nb of students	Max score	Min score	Top rank	Last rank
1	3,306	2.51	.578	1	3,306
2	4,402	.577	-.726	3,307	7,708
3	3,296	-.727	-2.12	7,709	11,004

D Preferences

This appendix gathers supporting materials and details to Section 4.
Part D.1 shows descriptive statistics supporting the extrapolation argument made in Section 4.1.3. Part D.2 shows utility parameters estimates separately for both truthful samples; as well as for alternative bandwidth choice for the sample of 'top' students. Complementing Section 4.2.2, part D.3 shows additional empirical validation for the bandwidth choice. Part D.4 discuss an alternative estimation strategy that could yield more precise utility parameters by using all students' application lists, rather than only those for truthful students.

D.1 Extrapolation: tables

Tables 13 and 14 show descriptive statistics for key student and choice characteristics, comparatively for three subsamples of interest: the whole population, as well as each of the two truthful subsamples. In addition, it shows statistics for what would be the top sample in a single-phase implementation of the DA: the top of Group 1 only.

D.2 Estimates: sensitivity analysis

Tables 15 and 16 show utility estimates obtained for each of the truthful subsamples separately. Tables 17 and 18 show estimation results for alternative choices for the bandwidth of truthful students. Columns (1), (2) and (3) show estimates for a bandwidth size equal to twice, five and ten times the original bandwidth size, respectively.

Table 13: Comparative statistics: students

	Sample	Mean	Std. dev.	Median	Min	Max	Obs.
Female	Pop.	.53	.50	1	0	1	10,897
	Top	.55	.50	1	0	1	636
	Top G1	.55	.50	1	0	1	205
	Sh.-list	.53	.50	1	0	1	3,236
High SES	Pop.	.60	.49	1	0	1	10,897
	Top	.73	.45	1	0	1	619
	Top G1	.89	.33	1	0	1	205
	Sh.-list	.63	.48	1	0	1	3,236
From Tunis	Pop.	.30	.46	0	0	1	10,897
	Top	.35	.48	0	0	1	619
	Top G1	.38	.49	0	0	1	205
	Sh.-list	.26	.44	0	0	1	3,236
From Coast (excl. Tunis)	Pop.	.48	.50	0	0	1	10,897
	Top	.48	.50	0	0	1	619
	Top G1	.54	.50	1	0	1	205
	Sh.-list	.53	.50	0	0	1	3,236
From West/Interior	Pop.	.19	.39	0	0	1	10,897
	Top	.15	.36	0	0	1	619
	Top G1	.09	.28	0	0	1	205
	Sh.-list	.18	.38	0	0	1	3,236
From South	Pop.	.03	.17	0	0	1	10,897
	Top	.02	.14	0	0	1	619
	Top G1	0	0	0	0	0	205
	Sh.-list	.03	.17	0	0	1	3,236
STEM high-school performance	Pop.	0	.85	-.10	-2.26	1.96	10,897
	Top	.52	.96	.49	-1.15	1.96	619
	Top G1	1.70	.10	1.70	1.34	1.96	205
	Sh.-list	.04	.93	-.08	-2.15	1.96	3,236
non-STEM high-school performance	Pop.	0	.79	0	-2.62	2.40	10,897
	Top	.47	.91	.42	-1.99	2.40	619
	Top G1	1.48	.32	1.49	.62	2.40	205
	Sh.-list	.03	.85	.01	-2.24	2.40	10,897

Note: 'Top' refers to the subset of students at the top of each group that is used for estimation of utility parameters. 'Sh.-list' refers to the subset of all students listing strictly fewer than 10 programs, also used for estimation of utility parameters. By contrast to 'Top', 'Top G1' refers to students in 'Top' who are also in Group 1. 'Pop.' indicates population statistics. In the second panel, STEM (resp. non-STEM) high-school performance is the unweighted average of the student's standardized scores at the Math, Physics, Natural Sciences, and Comp. Sci. (resp. English, French, Arabic, and Philosophy) tests of the end-of-high-school national exam.

Table 14: Comparative statistics: application behaviors

	Sample	Mean	Std. dev.	Median	Min	Max	Obs.
Distance home-sch. (km): min over list	Pop.	39.4	75.7	0	0	1,800	10,897
	Top	51.9	109.2	0	0	1,800	619
	Top G1	89.1	157.4	65	0	1,800	205
	Sh.-list	37.9	71.6	0	0	1,800	3,236
Distance home-sch. (km): max over list	Pop.	235.9	291.2	191	0	1,800	10,897
	Top	535.7	670.9	235	0	1,800	619
	Top G1	1,209.7	766.9	1,800	0	1,800	205
	Sh.-list	316.3	491.6	163	0	1,800	3,236
Distance home-sch. (km): avg. over list	Pop.	123.1	169.1	82.3	0	1,800	10,897
	Top	340.6	484.7	100.4	0	1,800	619
	Top G1	824.8	573	1,285.7	0	1,800	205
	Sh.-list	161	272.1	77.7	0	1,800	3,236
2009 ordinal marg. adm. score: min over list	Pop.	.33	.23	.29	.002	.991	10,897
	Top	.26	.24	.25	.002	.84	619
	Top G1	.02	.03	.005	.002	.23	205
	Sh.-list	.34	.25	.32	.002	.991	3,236
2009 ordinal marg. adm. score: max over list	Pop.	.71	.24	.72	.012	1	10,897
	Top	.57	.34	.57	.012	1	619
	Top G1	.17	.18	.03	.012	.80	205
	Sh.-list	.67	.28	.70	.012	1	3,236
2009 ordinal marg. adm. score: avg. over list	Pop.	.51	.23	.51	.008	.996	10,897
	Top	.41	.28	.42	.008	.94	619
	Top G1	.08	.08	.01	.008	.34	205
	Sh.-list	.50	.27	.51	.008	.996	3,236
At least one choice in Earth Sc.	Pop.	.21	.41	0	0	1	10,897
	Top	.21	.41	0	0	1	619
	Top G1	.02	.14	0	0	1	205
	Sh.-list	.15	.36	0	0	1	3,236
At least one choice in Soc. Sc.	Pop.	.02	.16	0	0	1	10,897
	Top	.01	.13	0	0	1	619
	Top G1	0	0	0	0	0	205
	Sh.-list	.02	.11	0	0	1	3,236
At least one choice in Law	Pop.	.03	.17	0	0	1	10,897
	Top	.02	.13	0	0	1	619
	Top G1	0	0	0	0	0	205
	Sh.-list	.02	.14	0	0	1	3,236
Total # programs applied to	Pop.	609	n/a	n/a	n/a	n/a	10,897
	Top	429	n/a	n/a	n/a	n/a	619
	Top G1	62	n/a	n/a	n/a	n/a	205
	Sh.-list	609	n/a	n/a	n/a	n/a	3,236

Note: 'Top' refers to the subset of students at the top of each group that is used for estimation of utility parameters. 'Sh.-list' refers to the subset of all students listing strictly fewer than 10 programs, also used for estimation of utility parameters. By contrast to 'Top', 'Top G1' refers to students in 'Top' who are also in Group 1. 'Pop.' indicates population statistics.

Table 15: Utility parameter estimates – truthful samples (1/2)

	(1) Main	(2) Lin. in distance	(3) Main	(4) Lin. in distance
Distance (100km)	-2.083***	-0.931***	-2.019***	-1.026***
	(0.12)	(0.08)	(0.08)	(0.05)
× high SES	0.085	0.181	0.021	0.147*
	(0.08)	(0.11)	(0.04)	(0.06)
Distance (100km) sq.	0.223***		0.222***	
	(0.02)		(0.01)	
Past-year marginal admit	2.503***	3.613***	2.097***	2.794***
	(0.60)	(0.58)	(0.31)	(0.31)
× high SES	-0.952	-0.706	0.467	0.679
	(0.87)	(0.88)	(0.40)	(0.41)
Past-year marginal admit sq.	0.761	1.055	-1.015**	-0.668*
	(0.70)	(0.71)	(0.32)	(0.34)
× high SES	2.394*	1.771	1.017*	0.616
	(0.95)	(0.93)	(0.40)	(0.40)
Distance (100km) × Past-year marginal adm.	1.107***		0.885***	
	(0.12)		(0.07)	
Degree: Bachelor (LF)	0.302*	0.303*	0.595***	0.597***
	(0.12)	(0.12)	(0.06)	(0.06)
× h-s perf.	0.431**	0.449**	0.393***	0.399***
	(0.14)	(0.14)	(0.06)	(0.06)
× high SES	0.206	0.178	0.009	-0.016
	(0.16)	(0.16)	(0.07)	(0.07)
Degree: Adv. degree	2.577***	2.551***	2.603***	2.592***
	(0.17)	(0.17)	(0.09)	(0.09)
× h-s perf.	0.884***	0.919***	1.941***	1.955***
	(0.15)	(0.16)	(0.09)	(0.08)
× high SES	-0.024	-0.062	-0.175	-0.210*
	(0.20)	(0.20)	(0.10)	(0.10)
Program location: Tunis	0.902***	1.128***	0.499***	0.728***
	(0.14)	(0.15)	(0.09)	(0.10)
Program location: Coast	0.913***	0.951***	0.304***	0.348***
	(0.13)	(0.13)	(0.07)	(0.08)
Program location: Abroad	-18.693***	-19.996***	-10.233***	-11.335***
	(2.33)	(2.49)	(1.36)	(1.45)
× STEM h-s perf.	7.492***	7.837***	4.319***	4.671***
	(1.25)	(1.40)	(0.77)	(0.82)
× non-STEM h-s perf.	3.671***	3.874***	1.969***	2.069***
	(0.42)	(0.44)	(0.30)	(0.33)
× high SES	0.064	0.368	-0.261	0.040
	(0.50)	(0.46)	(0.35)	(0.37)
Sample	Bdw	Bdw	Short	Short
PseudoObs.	4,927	4,927	24,961	24,961
Obs.	624	624	3,629	3,629

Std. errors in parentheses, clustered at the high school level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 16: Utility parameter estimates – truthful samples (2/2)

	(1)	(2)	(3)	(4)
Field: Arts	1.751**	1.741**	2.962***	2.963***
	(0.57)	(0.57)	(0.32)	(0.32)
× STEM h-s perf.	1.484***	1.478***	1.676***	1.678***
	(0.41)	(0.40)	(0.19)	(0.19)
× non-STEM h-s perf.	-1.064	-1.067*	-1.273***	-1.256***
	(0.55)	(0.54)	(0.24)	(0.24)
× female	-0.090	-0.060	-1.151**	-1.149**
	(0.55)	(0.55)	(0.36)	(0.36)
Field: Educ.	-0.572	-0.625	2.287***	2.318***
	(1.90)	(1.94)	(0.43)	(0.43)
× STEM h-s perf.	2.727**	2.762**	1.142**	1.084**
	(0.88)	(0.88)	(0.35)	(0.36)
× non-STEM h-s perf.	-0.147	-0.145	-0.883	-0.908
	(0.98)	(1.00)	(0.53)	(0.54)
× female	-2.044***	-2.044***	-1.676**	-1.653**
	(0.57)	(0.58)	(0.58)	(0.58)
Field: Soc. Sc.	-0.120	-0.103	1.033**	1.064**
	(0.87)	(0.88)	(0.37)	(0.37)
× STEM h-s perf.	1.490	1.508	0.775*	0.797*
	(0.97)	(0.98)	(0.32)	(0.32)
× non-STEM h-s perf.	-0.944	-0.946	-0.882*	-0.868*
	(0.94)	(0.93)	(0.36)	(0.35)
× female	-0.640	-0.628	-1.455***	-1.461***
	(0.92)	(0.92)	(0.43)	(0.43)
Field: Eco/Mgmt	2.995***	2.931***	3.775***	3.765***
	(0.56)	(0.55)	(0.33)	(0.33)
× STEM h-s perf.	0.981**	1.050**	1.160***	1.191***
	(0.34)	(0.33)	(0.21)	(0.21)
× non-STEM h-s perf.	-1.045*	-1.050*	-1.189***	-1.177***
	(0.49)	(0.49)	(0.25)	(0.25)
× female	-0.478	-0.466	-1.199***	-1.194***
	(0.51)	(0.51)	(0.36)	(0.36)
Field: Law	0.497	0.454	2.386***	2.378***
	(1.07)	(1.06)	(0.43)	(0.42)
× STEM h-s perf.	0.199	0.296	0.451	0.501
	(0.73)	(0.71)	(0.36)	(0.36)
× non-STEM h-s perf.	-1.689*	-1.755*	-0.161	-0.167
	(0.75)	(0.77)	(0.36)	(0.35)
× female	0.943	0.940	-1.455**	-1.454**
	(1.18)	(1.16)	(0.48)	(0.48)
Field: Math/Comp.Sci.	3.863***	3.801***	4.405***	4.399***
	(0.55)	(0.55)	(0.32)	(0.32)
× STEM h-s perf.	1.113**	1.166**	1.338***	1.366***
	(0.36)	(0.35)	(0.19)	(0.19)
× non-STEM h-s perf.	-1.370**	-1.364**	-1.562***	-1.553***
	(0.48)	(0.48)	(0.25)	(0.24)
× female	-0.837	-0.819	-1.337***	-1.328***
	(0.51)	(0.51)	(0.36)	(0.36)
Field: Phys./Chem./Engin.	3.530***	3.464***	4.187***	4.171***
	(0.55)	(0.55)	(0.31)	(0.31)
× STEM h-s perf.	1.145***	1.180***	1.301***	1.322***
	(0.33)	(0.33)	(0.18)	(0.18)
× non-STEM h-s perf.	-1.244**	-1.243**	-1.626***	-1.612***
	(0.47)	(0.46)	(0.24)	(0.23)
× female	-0.835	-0.813	-1.517***	-1.511***
	(0.51)	(0.51)	(0.35)	(0.35)
Field: Health/Life Sc.	3.371***	3.286***	3.603***	3.562***
	(0.56)	(0.56)	(0.32)	(0.32)
× STEM h-s perf.	0.907**	0.978**	1.102***	1.135***
	(0.34)	(0.33)	(0.19)	(0.19)
× non-STEM h-s perf.	-1.467**	-1.436**	-1.322***	-1.300***
	(0.50)	(0.50)	(0.25)	(0.24)
× female	0.226	0.259	-0.536	-0.517
	(0.54)	(0.54)	(0.36)	(0.36)
Field: Earth Sc.	2.234***	2.204***	2.105***	2.109***
	(0.56)	(0.56)	(0.32)	(0.32)
× STEM h-s perf.	-0.130	-0.134	0.264	0.276
	(0.35)	(0.35)	(0.21)	(0.21)
× non-STEM h-s perf.	-1.278**	-1.256**	-1.732***	-1.730***
	(0.48)	(0.48)	(0.26)	(0.25)
× female	-0.654	-0.632	-0.929*	-0.941**
	(0.53)	(0.53)	(0.36)	(0.36)

Table 17: Utility parameter estimates – truthful samples (1/2)

	(1)	(2)	(3)
Distance (100km)	-2.011***	-1.851***	-1.807***
	(0.10)	(0.08)	(0.06)
× high SES	0.035	0.005	0.031
	(0.06)	(0.04)	(0.03)
Distance (100km) sq.	0.215***	0.194***	0.189***
	(0.02)	(0.01)	(0.01)
Past-year marginal admit	2.735***	2.725***	2.983***
	(0.45)	(0.32)	(0.26)
× high SES	-0.090	0.892*	1.141***
	(0.62)	(0.41)	(0.33)
Past-year marginal admit sq.	0.151	0.090	-0.937***
	(0.45)	(0.29)	(0.24)
× high SES	1.146	-0.026	-0.156
	(0.62)	(0.38)	(0.30)
Distance (100km) × Past-year marginal adm.	1.097***	0.959***	0.841***
	(0.09)	(0.07)	(0.06)
Degree: Bachelor (LF)	0.411***	0.407***	0.463***
	(0.09)	(0.06)	(0.05)
× h-s perf.	0.411***	0.442***	0.354***
	(0.09)	(0.06)	(0.04)
× high SES	0.142	0.181**	0.163**
	(0.11)	(0.07)	(0.05)
Degree: Adv. degree	2.783***	2.768***	2.699***
	(0.11)	(0.08)	(0.06)
× h-s perf.	1.047***	1.107***	1.335***
	(0.10)	(0.07)	(0.06)
× high SES	-0.109	0.018	-0.069
	(0.13)	(0.09)	(0.07)
Program location: Tunis	0.800***	0.779***	0.701***
	(0.11)	(0.08)	(0.06)
Program location: Coast	0.753***	0.695***	0.574***
	(0.10)	(0.07)	(0.06)
Program location: Abroad	-25.303***	-23.147***	-18.036***
	(2.10)	(1.98)	(2.05)
× STEM h-s perf.	10.733***	9.989***	7.545***
	(1.08)	(0.97)	(1.02)
× non-STEM h-s perf.	4.245***	3.745***	3.075***
	(0.43)	(0.40)	(0.43)
× high SES	0.290	0.274	0.347
	(0.50)	(0.43)	(0.38)
Sample	Bdw × 2	Bdw × 5	Bdw × 10
PseudoObs.	10,169	26,384	53,485
Obs.	1,252	3,134	6,250

Std. errors in parentheses, clustered at the high school level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

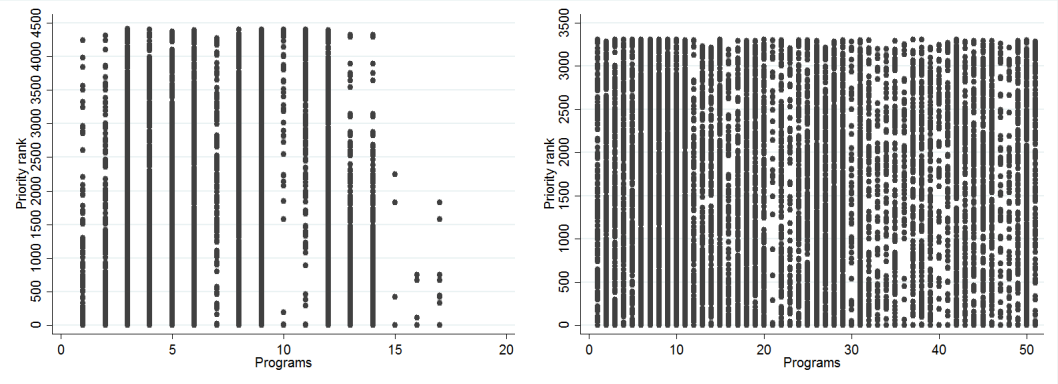
Table 18: Utility parameter estimates – truthful samples (2/2)

	(1)	(2)	(3)
Field: Arts	1.705***	2.211***	2.473***
	(0.35)	(0.22)	(0.17)
× STEM h-s perf.	1.523***	2.145***	2.036***
	(0.26)	(0.22)	(0.16)
× non-STEM h-s perf.	-0.961*	-1.209***	-1.144***
	(0.39)	(0.25)	(0.16)
× female	0.087	-0.264	-0.440*
	(0.41)	(0.26)	(0.19)
Field: Educ.	-0.181	0.328	1.449***
	(1.38)	(0.71)	(0.28)
× STEM h-s perf.	1.344	2.907***	1.798***
	(1.18)	(0.70)	(0.37)
× non-STEM h-s perf.	0.303	-1.531**	-0.868**
	(0.97)	(0.54)	(0.34)
× female	-1.254	-1.177*	-1.122***
	(0.71)	(0.54)	(0.34)
Field: Soc. Sc.	0.136	0.660*	0.637**
	(0.46)	(0.29)	(0.23)
× STEM h-s perf.	1.216*	1.724***	1.476***
	(0.54)	(0.35)	(0.24)
× non-STEM h-s perf.	-0.824	-1.139*	-0.796**
	(0.63)	(0.48)	(0.29)
× female	-1.079	-0.744*	-0.476
	(0.57)	(0.37)	(0.26)
Field: Eco/Mgmt	3.081***	3.355***	3.513***
	(0.34)	(0.21)	(0.17)
× STEM h-s perf.	1.001***	1.618***	1.413***
	(0.25)	(0.21)	(0.15)
× non-STEM h-s perf.	-0.909*	-1.276***	-1.110***
	(0.36)	(0.24)	(0.16)
× female	-0.450	-0.613**	-0.706***
	(0.36)	(0.23)	(0.18)
Field: Law	1.486**	1.922***	2.286***
	(0.47)	(0.34)	(0.22)
× STEM h-s perf.	0.825	0.907	0.815**
	(0.61)	(0.47)	(0.28)
× non-STEM h-s perf.	-0.708	-0.855**	-0.785***
	(0.48)	(0.29)	(0.22)
× female	-0.198	-0.430	-0.525*
	(0.53)	(0.39)	(0.26)
Field: Phys./Chem./Engin.	3.559***	3.777***	3.933***
	(0.35)	(0.21)	(0.16)
× STEM h-s perf.	1.149***	1.710***	1.489***
	(0.24)	(0.21)	(0.14)
× non-STEM h-s perf.	-1.289***	-1.662***	-1.537***
	(0.35)	(0.23)	(0.15)
× female	-0.816*	-0.875***	-1.000***
	(0.38)	(0.22)	(0.17)
Field: Health/Life Sc.	3.493***	3.721***	3.633***
	(0.37)	(0.22)	(0.17)
× STEM h-s perf.	1.010***	1.558***	1.411***
	(0.24)	(0.21)	(0.15)
× non-STEM h-s perf.	-1.389***	-1.590***	-1.370***
	(0.36)	(0.24)	(0.16)
× female	0.074	-0.026	-0.091
	(0.40)	(0.24)	(0.19)
Field: Earth Sc.	2.049***	2.094***	2.079***
	(0.35)	(0.21)	(0.17)
× STEM h-s perf.	-0.207	0.577**	0.404**
	(0.26)	(0.22)	(0.15)
× non-STEM h-s perf.	-1.239***	-1.640***	-1.547***
	(0.36)	(0.23)	(0.15)
× female	-0.391	-0.425	-0.436*
	(0.40)	(0.23)	(0.18)
Field: Math/Comp.Sci.	3.871***	4.029***	4.181***
	(0.34)	(0.21)	(0.17)
× STEM h-s perf.	1.205***	1.901***	1.559***
	(0.25)	(0.21)	(0.14)
× non-STEM h-s perf.	-1.307***	-1.726***	-1.523***
	(0.36)	(0.23)	(0.15)
× female	-0.711	-0.621**	-0.810***
	(0.37)	(0.23)	(0.18)

D.3 Empirical validation of the bandwidth choice: supplementary figures

Figure 12 is analogous to Figure 2 in Section D.3. The right (resp. left) panel considers the programs listed by the ten students of Group 2 (resp. Group 3), and shows the frequency at which these programs are listed by other Group 2 (resp. Group 3) students as a function of their priority ranking.

Figure 12: Persistence of the top-ranked students' listed choices over the priority ranking –Groups 2 and 3



Legend: This graph on the right (resp. left) panel considers the programs listed the first ten students at the top of Group 2 (resp. Group 3), and shows the frequency at which these programs are listed by Group 2 (resp. Group 3) students as a function of students' priority. Programs are represented on the x-axis, priority on the y-axis. A dot in position (a, b) in the graph means that student ranked b in Group 2 (resp. Group 3) included program a in her list.

D.4 An alternative estimation strategy

Proposition 1(b) (Haeringer and Klijn, 2009) in Section 2 shows that while students may not report their *most-preferred* programs, they always report programs in decreasing order of preference. That is, while I identify students' preferences from the choices made by a strict subset of students, all application lists reveal a *partial* preference ranking.

Non-truthful students' application lists generate moment *inequalities*, while truthful students' lists imply moment *equalities*. Indeed, if student i is truthful, then the likelihood of observing her list $\mathcal{L}_i$ in the data corresponds to the likelihood of the listed programs being her most-preferred programs:

$$\mathbb{P}\Big(\mathcal{L}_i = \{\mathcal{L}_i(1), \mathcal{L}_i(2), \ldots, \mathcal{L}_i(M_i)\}\Big) = \mathbb{P}\Big(u_i(\mathcal{L}_i(1)) > u_i(\mathcal{L}_i(2)) > \cdots > \mathcal{L}_i(M_i) > u_i(j),\ \forall\ j \in \mathcal{C}_i \backslash \mathcal{L}_i\Big),$$

where $\mathcal{L}_i$, M_i and $\mathcal{C}_i$ denote i's application list, the length of i's application list, and i's choice set, respectively. By contrast, if student i is (a priori) not truthful, then observing her list $\mathcal{L}_i$ in the data only implies an order on the utilities of listed programs:

$$\mathbb{P}\Big(\mathcal{L}_i = \{\mathcal{L}_i(1), \mathcal{L}_i(2), \ldots, \mathcal{L}_i(M_i)\}\Big) \leq \mathbb{P}\Big(u_i(\mathcal{L}_i(1)) > u_i(\mathcal{L}_i(2)) > \cdots > \mathcal{L}_i(M_i)\Big).$$

Extracting the information included in lists of non-truthful students requires an estimation method that allows for the use of both moment equalities and inequalities (e.g., Andrews and Shi, 2013; Fack, Grenet and He, 2015[47]), which I do not pursue in this paper.

[47] Fack, Grenet and He (2015) implement such a method to recover the preferences for high schools of middle-

E Students' expectations about their admission chances

This appendix provides supporting materials and information to Section 5.
Parts E.1 and E.2 document the estimation of expectations about admission chances in a benchmark setting where students know their true probability of admission to the different programs. Turning to the framework of the main text, in which students may not know their true admission probabilities, Part E.3 shows the parameters characterizing each sophisticated subtype used in the estimated model. As a complement to Section 5.2, Part E.4 illustrates the identifying variation grounding the estimation of types shares.

E.1 True-admission-probability benchmark

As a standard benchmark, I assume that, given the distribution of preferences, students are all able to form expectations that coincide with their true admission chances (are '*perfectly rational*'). In this setting, perfect rationality and the distribution of preferences are common knowledge.

E.1.1 Recovering true admission probabilities

Students having the same priority at all programs allows me compute numerically the joint distribution of each students' admission chances to all programs, in a simple way, given utility parameters and the distribution of preference unobservables.

- *Step 1:* I estimate the distribution of seats left after Student 1's assignment. Given a draw of her unobservable preference term, Student 1 truthfully lists her most-preferred program on top of her application list and get assigned to it. Simulating over her unobservable preference term, I can recover the distribution of her assignment –hence the joint distribution of seats left after her assignment. This distribution gives, for every program, the probability of an available vacancy for Student 2, that is, Student 2's expected admission chances in the rational-expectations benchmark.
- *Step k, $(1 < k \leq N)$:* I estimate the distribution of seats left after Student k's assignment. Student k solves Problem (1) given her preferences, and her expectations about her admission chances –recovered in Step (k-1). For any draw of Student k's unobservables, I solve Problem (1) and deduce Student k's assignment via the DA and the assignment of previous students. Simulating over her unobservable preference term, I can recover the distribution of her assignment –hence the joint distribution of seats left after her assignment. This distribution yields Student $k+1$'s expected admission chances in the rational-expectations benchmark.

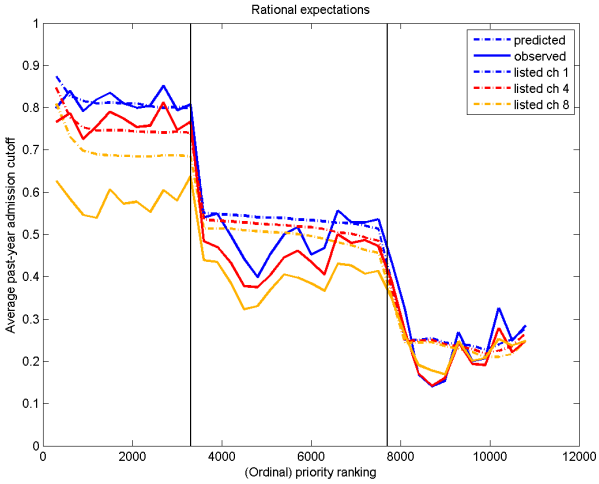
The simulation of optimal application lists is a computational challenge. When the choice set is large, the simulation of application lists of even moderate size is demanding. For instance, in the setting of this paper, finding *one* individual's expected-utility-maximizing ordered list of up to 10 elements among 600 requires evaluating the expected utility function at more than 10^{20} points, and finding the maximum. In practice, to ease computation, I do not solve the optimization Problem (1). Rather, I approximate the solution using the Marginal Improvement Algorithm (MIA) proposed

schoolers in Paris. Their identifying moment equalities differ from those I describe, though. Identification of students' preferences in their analysis relies on the assumption that the match realized in their data is *stable*. They also implement a partial identification approach, using only moments inequalities, without assuming stability.

by Chade and Smith (2006).[48] The MIA starts by first selecting the application list of size 1 (that is, the alternative) with highest expected utility. It then proceeds to finding the best marginal improvement to that list. That is, it selects the alternative that forms, together with the first pick, the application list of size 2 with highest expected utility (among all the lists containing the first pick). This iterative process continues until the desired list size is reached. A detailed description of the MIA is provided in Appendix Section E.2.

E.1.2 Predicted choices

Figure 13: Rational-expectations benchmark –Selectivity level of predicted vs. observed choices



Note: This graph shows the selectivity level (in terms of past-year admission cutoff) of students' choices as a function of their priority ranking. For clarity, it focused on students' first, fourth, and eighth-listed choices. Solid lines represent choices observed in the data; dotted lines represent choices predicted under the rational-expectations benchmark, given utility parameter estimates from Section 4.

Figure 13 plots the selectivity level of students' choices as a function of their priority ranking. For clarity, it focused on students' first, fourth, and eighth-listed choices. Solid lines represent choices observed in the data; dotted lines represent choices predicted under the rational-expectations benchmark, given utility parameter estimates from Section 4. The graph suggests that assuming perfect rationality of students imperfectly captures the variation in the data in two ways. It overshoots the selectivity level of some students' listed choices, and does not reproduce the diversification of students' application portfolios in terms of selectivity levels. In a rational equilibrium, each student understands that when the number of seats in each program is fixed, given the number of students assigned before her, there is a negative correlation between the number of seats remaining available when her turn in the algorithm comes in programs with similar characteristics.[49] In other words, conditional on being rejected from an higher-listed program, they know they have an increased chance to be admitted to a program with similar characteristics if they rank such program lower

[48] When one's eligibility chances are independent across programs, the MIA yields the actual solution of Problem (1) (Chade and Smith, 2006). When these eligibility chances are not independent, however, this result is not guaranteed (Ajayi and Sidibé, 2016).

[49] This is mechanical. Suppose there are only two programs A and B, and that they have identical characteristics. Fix students' tastes and the number of students to be assigned before student i in the algorithm. Statistically, and conditional on the characteristics of A and B, when none of the programs are full, the students assigned to A are the ones whose unobservable utility draw for A is larger than for B. The number of students to be assigned before student i in the algorithm being fixed, the larger the number of students assigned to A before i's turn, the smaller the number of students assigned to B before i's turn.

in their list. Hence the similarity in programs' characteristics across a rational student's listed choices. The diversification of students' portfolios (in terms of selectivity level) in the observed data suggests that students do not fully account for this negative correlation.

Note, from the procedure described in E.1.1, that given utility parameter estimates, predicting choices under the rational-expectations framework does not require the identification of any new parameter. The admission chances expected by perfectly rational students are fully determined once the utility parameters and the distribution of preference unobservables are known. In Section 4, preferences were recovered without taking any stand on expectations, and using a strict subset of the students. In the previous paragraph, other students' observed lists were used to assess the ability of the rational-expectations framework to reproduce patterns in the data (Figure 13). In the next subsection, I use still-unexploited identifying variation in these lists to estimate an alternative framework of expectations formation.

E.2 Marginal Improvement Algorithm (Chade and Smith, 2006)

Marginal Improvement Algorithm

Step 0: Start with the empty list: $\mathcal{L}_i^{(0)} = \emptyset$.
Discard from choice set all alternative with lower flow utility than the outside option.

Step 1: Select the program with highest expected utility: $\mathcal{L}_i^{(1)} = \{s_1\}$

Step k ($2 \leq k \leq 10$)**:** Select the best complement to the current list $\mathcal{L}_i^{(k-1)}$, i.e. solve:

$$\begin{aligned} \max_{s \in \mathcal{J} \setminus \mathcal{L}_i^{(k-1)}} \quad & EU(\mathcal{L}_i') \\ s.t. \quad & \mathcal{L}_i' = \mathcal{O}_i\left(\mathcal{L}_i^{(k-1)} \cup \{s\}\right) \end{aligned}$$

where

- $\mathcal{O}_i$ arranges the elements of $\mathcal{L}_i$ in decreasing order of flow utility for i
- $EU_i(\mathcal{L}_i^{(k)}) = \pi_{i,\ell_1} \cdot u_{i,\ell_1} + \pi_{i,\ell_2|\ell_1} \cdot u_{i,\ell_2} + \cdots + \pi_{i,\ell_k|\ell_1,\ell_2,\ldots,\ell_{k-1}} \cdot u_{i,\ell_k}$

E.3 Types specification

In Section 5, expectations-formation types are specified as AR(1) processes, the coefficients of which are estimated by MLE from 2009-2010 data on marginal admission scores. Types differ from one another in the level of observable heterogeneity allowed in the AR(1) specification:

$$\text{cutoff}_{j,2010} = a_j + b_j \times \text{cutoff}_{j,2009} + \eta_j \quad \text{with } \eta_j \sim N(0, \sigma_j^2).$$

Tables 19 and 20 show estimated coefficients for the different specifications considered.

Table 19: Estimated AR(1) parameters for marginal admission scores (1/2)

No heterogeneity						Log-lik.: 0.54
	All					
cst.	-0.19					
	0.02					
slope	0.82					
	0.02					
σ	0.42					
	0.02					
Log-lik.	0.54					
Obs.	616					
Heterogeneity by selectivity level (in percentile of priority distribution)						Log-lik.: 0.51
	$\leq$ 5th	5–25th	25–50th	50–75th	75–95th	above 95th
cst.	1.73	-0.5	0	-0.31	-0.25	-0.43
	1.57	0.43	0.19	0.07	0.06	0.15
slope	1.66	0.58	1.09	0.65	0.89	1.17
	0.82	0.27	0.19	0.18	0.15	0.07
σ	0.44	0.41	0.39	0.41	0.39	0.38
	0.09	0.05	0.04	0.03	0.04	0.09
Heterogeneity by field of study						Log-lik.: 0.46
	STEM	non-STEM				
cst.	-0.09	-0.32				
	0.02	0.04				
slope	0.89	0.73				
	0.02	0.04				
σ	0.31	0.5				
	0.02	0.03				
Heterogeneity by 2009 filling status						Log-lik.: 0.53
	full	not full				
cst.	-0.21	0.02				
	0.02	0.04				
slope	0.82	0.91				
	0.03	0.02				
σ	0.4	0.45				
	0.02	0.05				
Heterogeneity by field $\times$ 2009 filling status						Log-lik.: 0.44
	STEM full	STEM not full	non-STEM full	non-STEM not full		
cst.	-0.11	0.09	-0.32	-0.63		
	0.02	0.03	0.04	0.42		
slope	0.91	0.92	0.72	0.56		
	0.03	0.01	0.04	0.24		
σ	0.3	0.28	0.48	0.57		
	0.02	0.03	0.03	0.07		
Obs.	616					

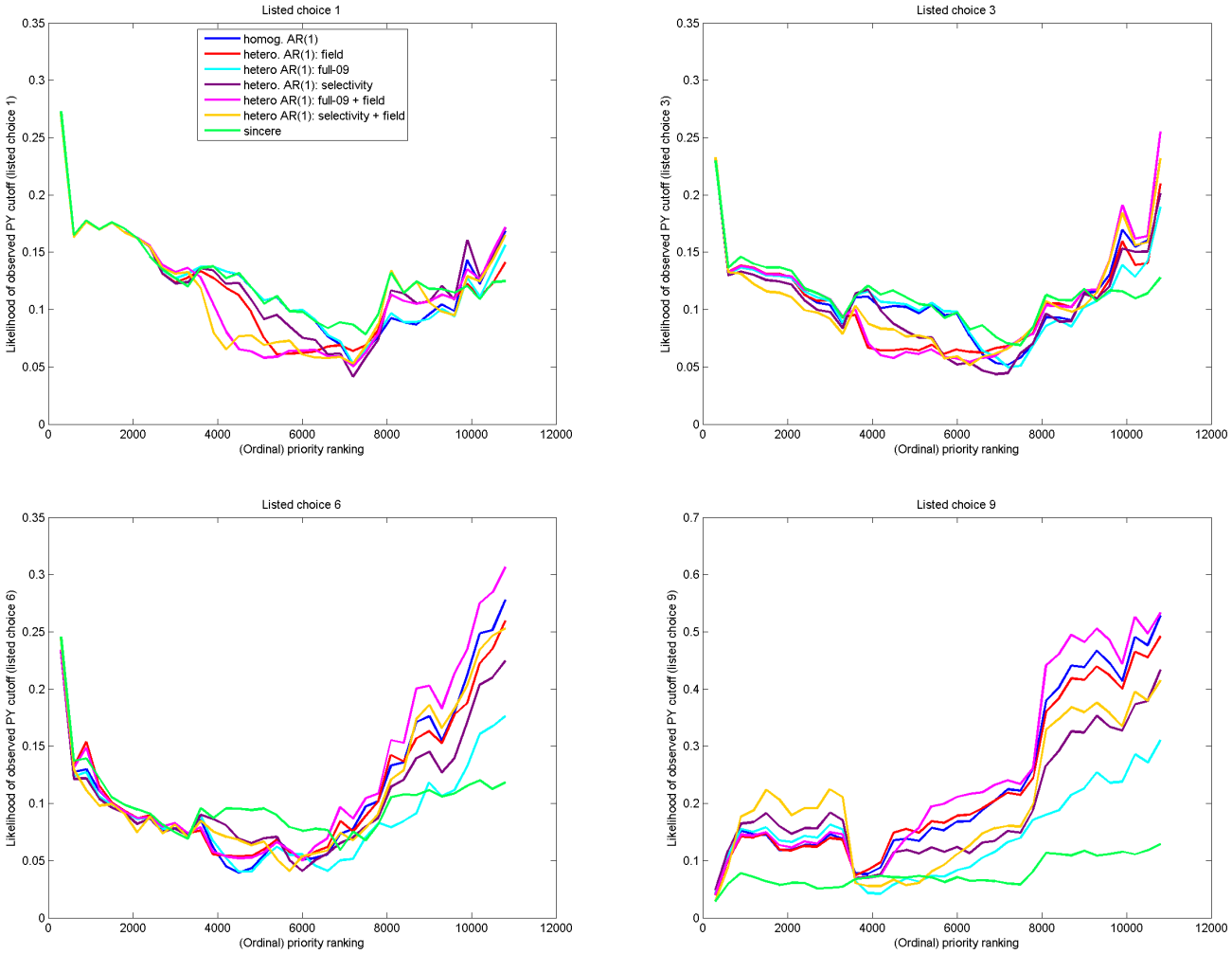
Table 20: Estimated AR(1) parameters for marginal admission scores (2/2)

Heterogeneity by field × selectivity level											Log-lik.: 0.36	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
cst.	-0.76	-0.16	-0.46	-0.27	-0.25	0.08	1.64	-0.95	0.68	-0.37	-0.26	1.86
	2.03	0.26	0.14	0.08	0.04	0.09	1.8	0.83	0.29	0.12	0.1	0.97
slope	0.3	0.8	0.59	0.46	1.21	0.95	1.64	0.3	1.88	0.77	0.67	-1.48
	1.08	0.17	0.15	0.2	0.13	0.05	0.93	0.53	0.33	0.31	0.23	1.02
σ	0.24	0.26	0.31	0.3	0.26	0.11	0.51	0.57	0.45	0.47	0.46	0.25
	0.05	0.03	0.03	0.04	0.03	0.02	0.15	0.08	0.07	0.05	0.07	0.05
Obs.	616											

E.4 Identifying variation

Figure 14 illustrates how, given preferences, the likelihood of observing one's actual choices differs under alternative assumptions about expectations formation process and level of sophistication. This figure illustrates the variation in the data allowing me to characterize students' expectations. It plots, as a function of students' priority ranking, the likelihood (given preferences) of observing the characteristics of one's actual choices under alternative assumptions about the expectations-formation process. Specifically, it focuses on the selectivity level (in terms of past-year admission cutoff) of students' listed programs under the eight distinct scenarios of expectations-formation considered in Section 5.2.3. For the sake of space, I only show plots for students' first, third, sixth, and ninth listed choices –plots for the second, fourth, fifth, seventh, eighth, and tenth listed choices are similar. Lists are simulated assuming that unsophisticated students report their ten most preferred programs among those that have not been publicly declared to be full. Students with any given AR(1) type report the expected-utility-maximizing list, and derive their expectations assuming marginal admission scores follow, from one year to the next, the given AR(1) process.

Figure 14: Density of listed-choice characteristics (past-year cutoff) under alternative expectations formation assumptions



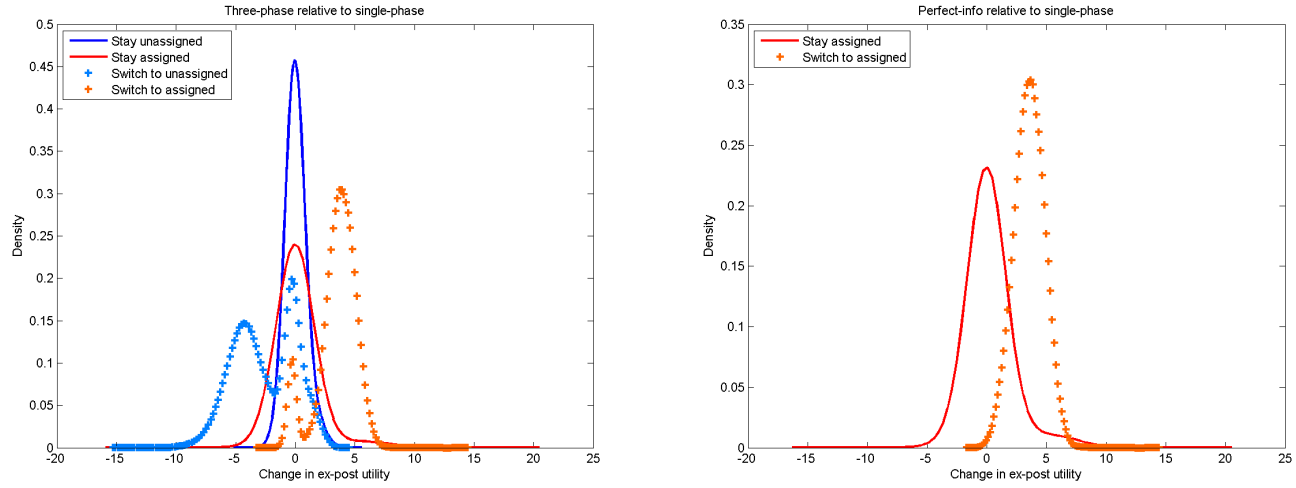
F Counterfactual analysis

This appendix provides supplemental figures and tables to Section 6.
Part F.1 provides an illustration to results in Section 6.1.2. Part F.2 discuss gains from information in a setting where students know their true probabilities of admission to each programs.

F.1 Supplemental figures

As a complement to Figure 11 in Section 6.1.2, Figure 15 plots the distribution of indirect utility changes (relative to the single-phase restricted-list DA) within each assignment-status-pair group.

Figure 15: Distribution of changes in expected indirect utility by assignment status pair



F.2 True-admission-probabilities benchmark

Table 21 is the true-admission-probabilities analogue of Figure 4: it shows the difference in expected average student welfare between scenarios with and without information revelation –in a setting in which students form expectations about their admission chances that coincide with their admission probabilities (see Appendix Section E.1). In this setting, the loss generated by the implementation of the standard (single-phase) restricted-list DA, relative to perfect-information benchmark, is very small. Under the perfect-information benchmark, the average indirect utility is higher than in the single-phase DA by an equivalent of a 0.40km-reduction in distance traveled –a difference 100 times smaller than the one shown in Figure 4. This suggests that, rather than the sole incompleteness of information about which seat are available for them, it is students' inability to form accurate expectations about their admission chances that is responsible for most of the welfare loss induced by the implementation of a single-phase restricted-list DA in an incomplete information setting.

Table 21: Rational expectations benchmark –Change in expected average student welfare relative to single-phase implementation of the restricted-list DA

	Two-phase	Three-phase	Four-phase	Five-phase	Perfect info
equiv. km	0.04	0.22	0.1	0.11	0.39

References

[1] Abdulkadiroğlu, A., N. Agarwal and P. Pathak (2017). The welfare effects of coordinated assignment: Evidence from the New York City high school match, NBER Working Paper No. 21047.

[2] Abdulkadiroğlu, A., Y.-K. Che and Y. Yasuda (2015). Expanding "choice" in school choice. *American Economic Journal: Microeconomics*, 7(1), pp. 1–42.

[3] Abdulkadiroğlu, A., P. Pathak, A. Roth and T. Sönmez (2006a). Changing the Boston school choice mechanism. NBER Working Paper No. 11965.

[4] Abdulkadiroğlu, A., P. Pathak, A. Roth and T. Sönmez (2006b). Changing the Boston school choice mechanism: Strategy-proofness as equal access. Working paper.

[5] Abdulkadiroğlu, A. and T. Sönmez (2003). School choice: a mechanism design approach. *The American Economic Review*, 93(3), pp. 729–47.

[6] Agarwal, N. (2015). An empirical model of the medical match. *The American Economic Review*, 105(7), pp. 1939–78.

[7] Agarwal, N. and W. Diamond (2015). Latent indices in assortative matching models. Working paper.

[8] Agarwal, N. and P. Somaini (2014). Demand analysis using strategic reports: application to a school choice mechanism. NBER Working Paper No. 20775.

[9] Ajayi, K. and M. Sidibé (2016). An empirical analysis of school choice under uncertainty. Working paper.

[10] Altonji, J., E. Blom and C. Meghir (2012). Heterogeneity in human capital investments: high school curriculum, college major, and careers. NBER Working Paper No. 17985.

[11] Altonji, J., P. Arcidiacono and A. Maurel (2015). The analysis of field choice in college and graduate school: determinants and wage effects. NBER Working Paper No. 21655.

[12] Andrews, D. and X. Shi (2013). Inference based on conditional moment inequalities. *Econometrica*, 81(2), pp. 609-22.

[13] Arcidiacono, P., V. J. Hotz and S. Kang (2012). Modeling the College Major Choices using Elicited Measures of Expectations and Counterfactuals. *Journal of Econometrics*, 166(1), pp. 3-16.

[14] Balinski, M. and T. Sönmez (1999). A tale of two mechanisms: Student placement. *Journal of Economic Theory*, 84, pp. 73–94.

[15] Calsamiglia, C., C. Fu and M. Güell (2014). Structural estimation of a model of school choices: the Boston mechanism vs. its alternatives. FEDEA Working Paper No. 2014-21.

[16] Calsamiglia C., G. Haeringer and F. Klijn (2010). Constrained school choice: an experimental study. *The American Economic Review*, 100(4), pp. 1860–74.

[17] Carvalho J.-R., T. Magnac and Q. Xiong (2014). College Choice Allocation Mechanisms: Structural Estimates and Counterfactuals. Toulouse School of Economics Working Paper No. TSE-506.

[18] Chade, H. and L. Smith (2006). Simultaneous search. *Econometrica*, 74(5), pp. 1293–307.

[19] Chen, H. and Y. He (2017). Information acquisition and provision in school choice: an experimental study. Working paper.

[20] Chen, H. and T. Sönmez (2006). School choice: an experimental study. *Journal of Economic Theory*, 127(1), pp. 202–31.

[21] Chiappori, P.-A. and B. Salanié (2014). The econometrics of matching models. Working paper.

[22] Dubins, L. and D. Freedman (1981). Machiavelli and the Gale-Shapley algorithm. *American Mathematical Monthly*, 88, pp. 485–94.

[23] Dur, U., R. Hammond and T. Morrill (2016). Identifying the harm of manipulable school-choice mechanisms, Working paper.

[24] Erdil, A. and H. Ergin (2008). What's the matter with tie-breaking? Improving efficiency in school choice. *American Economic Review*, 98(3), pp. 669–89.

[25] Fack, G., J. Grenet and Y. He (2015). Beyond truth-telling: preference estimation with centralized school choice. Working paper.

[26] Gale, D. and L. Shapley (1962). College admissions and stability of marriage. *American mathematical monthly*, 69, pp. 9–15.

[27] Haeringer, G. and F. Klijn (2009). Constrained school choice. *Journal of Economic Theory*, 144, pp. 1921–47.

[28] Hahn, J., P. Todd and W. van Der Klaauw (2001). Identification and estimation of treatment effects in a regression-discontinuity design. *Econometrica*, 69(1), pp. 201–9.

[29] Hastings, J., C. Neilson, A. Ramirez and S. Zimmerman (2015). (Un)Informed college and major choice: evidence from linked survey and administrative data. NBER Working Paper No. 21330.

[30] Hastings, J., R. Van Weelden and J. Weinstein (2007). Preferences, information, and parental choice behavior in public school choice. NBER Working Paper No. 12995.

[31] Hastings, J. and J. Weinstein (2009). Information, School choice and academic achievement: Evidence from two experiments. *The Quarterly Journal of Economics*, 123(4), pp. 1373–414.

[32] He, Y. (2016). Gaming the Boston School Choice Mechanism in Beijing. Working paper.

[33] Hoxby, C. and S. Turner (2015). What high-achieving low-income students know about college. *American Economic Review: Papers and Proceedings*, 105(5), pp. 514–17.

[34] Huggett, M., G. Ventura and A. Yaron (2011). Sources of lifetime inequality. *American Economic Review*, 101(7), pp. 2923–54.

[35] Imbens, G. and K. Kalyanaraman (2011). Optimal bandwidth choice for the regression discontinuity estimato. *Review of Economic Studies*, 79, pp. 933–59.

[36] Imbens, G. and T. Lemieux (2007). Regression discontinuity designs: a guide to practice. *Journal of Econometrics*, 142, pp. 615–35.

[37] Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics*, 125(2), pp. 515–48.

[38] Kapor, A., C. Neilson and S. Zimmerman (2016). Heterogeneous beliefs and school choice. Working paper.

[39] Keane, M. and K. Wolpin (1997). The career decisions of young men. *Journal of Political Economy*, 105(3), pp. 473–522.

[40] Lee, D. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2), pp. 281–355.

[41] Luflade, M. and M. Zaiem (2016). Do elite schools improve students performance? Evidence from Tunisia. Working paper.

[42] Manski, C. (2004). Measuring expectations. *Econometrica*, 72(5), pp. 1329–76.

[43] McFadden, D. (1978). Modeling the choice of residential location. In: Kralqvist A., et al., eds., *Spatial interaction theory and planning models*, 1st. ed. Amsterdam: North-Holland.

[44] Narita, Y. (2015). Match or mismatch: learning and inertia in school choice. Working paper.

[45] Oreopoulos, P. and R. Dunn (2013). Information and college access: Evidence from a randomized field experiment. *The Scandinavian Journal of Economics*, 115(1), pp. 3–26.

[46] Pallais, A. (2015). Small differences that matter: mistakes in applying to college. *Journal of Labor Economics*, 33(2), pp. 493–520.

[47] Pathak, P. and T. Sönmez (2008). Leveling the playing field: Sincere and sophisticated players in the Boston mechanism. *The American Economic Review*, 98(4), pp. 1636–52.

[48] Pathak, P. and T. Sönmez (2013). School admissions reform in Chicago and England: Comparing mechanisms by their vulnerability to manipulation. *The American Economic Review*, 103(1), pp. 80–106.

[49] Pistolesi, N. (2016). Advising students on their field of study: Evidence from a French university reform. Working paper.

[50] Roth, A. (1982). The economics of matching: stability and incentives. *Mathematics of Operations Research*, 7(4), pp. 617–28.

[51] Roth A. (1984). The evolution of the labor market for medical interns and residents: a case study in game theory. *Journal of Political Economy*, 92(6), pp. 991–1016.

[52] Roth, A. (2008). Deferred acceptance algorithms: history, theory, practice, and open questions. *International Journal of Game Theory*, 36(3-4), pp. 537–69.

[53] Roth, A. (2015). Why New York City's high school admission process only works most of the time. [Blog] Chalkbeat. Available at: https://ny.chalkbeat.org/posts/ny/2015/07/02/why-new-york-citys-high-school-admissions-process-only-works-most-of-the-time/ [Accessed 9/14/2017].

[54] Stinebrickner, R. and T. Stinebrickner (2014a). Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics*, 32(3), pp. 601–44.

[55] Stinebrickner, R. and T. Stinebrickner (2014b). A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout. *Review of Economic Studies*, 81, pp. 426–72.

[56] Tunisian Republic, Ministère de l'Éducation, Secrétariat Général, Direction Générale des Études, de la Planification et des Systèmes d'Information, *Statistiques Scolaires –Année scolaire 2012–13*.

[57] Train, K. (2002). *Discrete choice methods with simulation*. Cambridge: Cambridge University Press.

[58] Wiswall, M. and B. Zafar (2013). How do college students respond to public information about earnings?. Federal Reserve Bank of New York Staff Report No. 516.

[59] Wiswall, M. and B. Zafar (2014). Determinants of college major choice: Identification using an experiment. Federal Reserve Bank of New York Staff Report No. 500.