School Choice Under Imperfect Information*

Kehinde Ajayi† Modibo Sidibé‡

March 29, 2023

Abstract

As in many school districts worldwide, high-school students in Ghana are assigned through a centralized system. We build and estimate a model where students engage in costly search to acquire information on school characteristics. Schooling decisions are exerted without full examination of all available options, which may lead to sub-optimal choices. Using administrative data on choices and a survey on beliefs, we document substantial welfare loss: distance traveled to schools could be divided by 3. We propose a new ranked-attributes-list mechanism, which collects preferences over school attributes rather than actual choices and show it recovers most of the lost welfare.

Keywords: school choice, search, uncertainty.

JEL Classification Number: C53, D61, I20.

*This paper has benefited from helpful discussions with Attila Ambrus, Pat Bayer, Jacob Leshno, Adam Rosen and Ran Shorrer as well as seminar and conference participants at Duke University, University of Virginia, University of Wisconsin, UCLA, PeGnet, Yale and Princeton. We would like to thank Chao Fu for walking us through the steps of their algorithm.

†Center for Global Development, 2055 L Street NW, Washington, DC 20036. Email: kajayi@cgdev.org.

‡Department of Economics, Duke University, 213 Social Sciences Building, 419 Chapel Drive, Box 90097, Durham, NC 27708. Email: modibo.sidibe@duke.edu.

Introduction

Over the last 30 years, many school districts have adopted coordinated mechanisms to assign students to schools. In 2005, Ghana introduced the national Computerized School Selection and Placement System (CSSPS) to match prospective students to high schools. The stated goal was to increase equity and access to quality senior high schools. The matching is based on the serial dictatorship algorithm, where priorities are determined by the student's score on the Basic Education Certification Examination (BECE). Every spring, several hundred thousands of students submit a wish list of schools and gain admission into one of the more than 2,000 schools/programs at the end of summer, making it *de facto* one of the largest matching systems in the world.

However, throughout the process, logistical considerations outweigh efficiency concerns. First, the timing introduces uncertainty, as rank-order lists (ROLs) are submitted before the examination determines priority scores. Then, students have to choose between more than 2,000 schooling options, which guarantees that some choices will be overlooked. Finally, constraints were imposed on the length of rank-order lists (3 in 2005, 4 in 2007, then 6 in 2008), which prompted agents to strategize over their submitted lists. The short history of the program, combined with the low potential involvement of parents, may have worsened the potential welfare consequences of these implementation issues.

Three years after introducing the CSSPS, data on applications demonstrates that more than 24% of students do not submit a complete list of choices. Matching outcomes show that more than 13% of students end up unassigned.[1] The application behavior of students led to substantial variation in admission cutoffs, especially for low and medium selectivity schools, and more than 30% of schools end up with at least one vacancy (including the very best schools). While exacerbated in Ghana, many nations faced similar problems.[2]

This paper shows that even coordinated centralized allocations are not a miracle solution against coordination frictions. We show that students may engage in a costly search to acquire information about schooling opportunities in a large matching market. We make several methodological and empirical contributions to the literature.

First, we document that a substantial share of students submits a truncated list, and many students and schools end up unmatched. Second, we formulate the school application process as a dynamic search problem to rationalize the co-existence of vacancies at the school level and unmatched students. Students engage in an iterative and costly search among alternatives to build an endogenous consideration set, and the sequential acquisition of information about schools implies a familiar optimal stopping structure. A key implication of the model is that schooling decisions are exerted without thoroughly examining all available options, which may lead to

[1] According to Abdulkadiroglu et al. (2005), a little less than 30% of students were administratively assigned under the decentralized application system in New York City, which motivated the switch to a centralized system.

[2] Two notable examples are Tunisia and Jamaica, where school officials have adopted "twists" to the standard DA mechanism. In Tunisia, the algorithm runs sequentially, with cohorts being divided into three groups based on priority scores. In Jamaica, when a student fails to match to a school from his ROL, an administrator assigns him to a school with an existing vacancy ahead of students with lower priority who might have applied for that school.

sub-optimal decisions. Third, we assume that students have subjective beliefs about admission chances, and those beliefs are likely to be biased for many schooling options. Information and search frictions lead to welfare losses. Fourth, we overcome the curse of dimensionality, making it impossible to account for sizeable potential choice sets, and limiting standard empirical applications to markets with few choices. Our empirical application, which uses data from more than 160,000 students choosing more than 2,000 schooling options, requires us to specify a search technology. We consider models where the search is random and a more general setting where students can direct their search toward a specific set of schools. We use the model to rationalize the observation that similar individuals make very different choices and produce credible counterfactual simulations. Potentially biased beliefs about admission chances and search frictions lead to an inefficient allocation. Finally, we use the model to propose a policy reform that generates substantial welfare gains.

Our work is related to the recent literature on the role of information in empirical market designs (see for example Luflade, 2017; Kapor et al., 2018).[3] While these papers introduce critical innovations to the analysis of centralized school allocation systems, our setting differs with imperfect information, which results from the size of the market, and compels students to search for information over schools. Recent literature analyzes similar questions (Arteaga et al., 2021) using a theoretical model to derive testable predictions. Our framework allows us to endogenize how consideration sets are formed. As such, our work is related to an extensive literature on search frictions that dates back to Stigler (1961) and complements the active field of consumer search in industrial organization.[4]

The most critical parameters for our analysis are preference parameters for school attributes. These parameters may be affected by the information sets and search behavior of agents, which are not observed in administrative data. We take three complementary approaches to isolate the potential consequences on preference estimates and derive credible counterfactual analysis.

First, using survey data on 12,871 students, that contains extensive information about beliefs and realized test scores, we characterize the belief and information sets of students as in Kapor et al. (2018). Analysis of the survey data shows substantial test score uncertainty, and beliefs about school level cutoffs are generally biased. Ultimately, we apply the K-means clustering technique to derive a mixture of distributions to account for belief heterogeneity in the population. Second, we develop a series of behavioral search models with increasing complexity to illustrate the empirical content of the administrative data. The first model, which is based on exogenous consideration (Manzini and Mariotti, 2014), is shown to be identified using rank-order list data. We provide two additional models where search is sequential and the sampling

[3]The predominant focus in the empirical school choice literature has been on lottery-based admission and the Boston mechanism (see Abdulkadiroglu and Sonmez, 2003). The recent literature tries to quantify the welfare gains associated with changing the allocation mechanism (Agarwal and Somaini, 2018; Fack et al., 2019; Calsamiglia et al., 2020). Related literature also includes Kapor (2016); Abdulkadiroglu et al. (2017); Walters (2018).

[4]See for example Goeree (2008); Honka et al. (2017); Dinerstein et al. (2018) among others. A recent literature in decision theory analyzes the role of information attention and rational inattention in individual choice (see Masatlioglu et al., 2012; Caplin and Dean, 2011; Sims, 2003).

of schools may be random or directed. Third, we show that if search cost does not include an unobserved component and the probability of considering an alternative is known a-priori, then preference parameters can be identified provided that the econometrician has access to a special regressor, which shifts consideration probabilities without affecting preferences. We perform a Monte Carlo study confirming this result and showing that preference parameters are precisely and accurately estimated. Finally, we show multiple specifications that confirm that preference estimates are remarkably stable across different models and specifications. The results of the three analyses show a clear and consistent pattern: preference parameters are identified.

Our estimated parameters are as expected, indicating individuals' preferences for school quality and its proxies. Under our preferred model, the median student considers only seven choices compared to more than 2,000 schooling options. Comparisons of the different search methods conclude that a model where students direct their search toward specific groups of schools provides the best fit to the data.

We quantify the welfare implications of school choice in the presence of imperfect information on school characteristics. Our welfare analysis shows that only 34% of the efficient allocation is realized. As distance is the numeraire, our results imply that the cost of boarding could be divided by three. Since low ability students incur the most sizable share of the welfare losses, our findings suggest that the initial objective to increase equity in access to secondary education may be negated. Further computations show that 37.4% of the welfare loss can be attributed to the inability of students to gather information about alternatives. Test score uncertainty accounts for 20.5% of the welfare loss and aggregate uncertainty for 14.8%.[5]

Finally, since the planner may be more informed about schooling opportunities than students, we study whether restricting choice could improve welfare. Our intuition is that given the number of choices, some alternatives may be looked over. Yet, counterfactual outcomes will depend mainly on the information set of the planner. On one extreme, if we were to assume that the planner knows all the utility function parameters, efficient allocation is achieved. We propose a simple mechanism, the Ranked-Attributes-List (RAL), which retains the key properties of the serial dictatorship algorithm, as students are asked to resolve the trade-off between different school attributes using a ranking of which characteristics to consider first. We find that school choice under RAL generates substantial welfare gains - recovering most of the welfare loss. RAL sidesteps the effect of uncertainty and search frictions on the application behavior and achieves approximately 87.6% of the efficient allocation with a minimal change to the current system.

The rest of the paper is organized as follows. In Section 1, we describe our data and report several empirical regularities. Section 2 describes the model, while estimation and identification are discussed in section 3. The estimation results are presented in section 4. Section 5 presents a welfare analysis. Finally, section 6 concludes.

[5]Welfare losses do not sum to 1 because of interactions between the various sources of uncertainty and search frictions.

1 Motivation

Our data comes from Ghana, where the national school system consists of six years of primary school, three years of junior high school (JHS), and three years of senior high school (SHS). In contrast to most higher income nations, high-school completion (SHS) is the final qualification for almost 80% of students (Duflo et al., 2021). Starting in 2005, students completing junior high school apply for admission to senior high school through a centralized application system. One may wonder about the rationale for organizing a nationwide education system for teenagers. As we show later, top academic programs are located in few regions; a national education system gives a pathway to elite schools for students in rural locations.[6] We should stress that the school choice setting in Ghana is similar to college choice in higher income nations, with students moving to boarding schools on the other side of the country.

Students apply to specific academic programs within a school and can submit a ranked list of up to six choices. Available programs include agriculture, business, general arts, general science, home economics, visual arts, and several occupational programs offered by technical or vocational institutes.[7]

After submitting their rank-order lists, students take a standardized Basic Education Certification Exam (BECE). The application system then allocates students to schools based on a serial dictatorship where priorities are determined by the BECE score.

Students who are unassigned at the end of the algorithm are administratively assigned to a nearby program with remaining vacancies. Our data, which provides individual choices along with BECE scores as well as admission outcomes, consists of final year (grade 9) students from the universe of junior high schools (grades 6-9) in 2008.

Appendix Section A provides a detailed description of individuals' application behavior as well as admission outcomes. Our final analytical sample consists of 160,869 students, who choose between 2,089 school-programs. In this section, we focus on some key empirical regularities that motivate our analysis.

1.1 Evidence of Search.

First, we present empirical evidence of limited search. Despite having six choices out of more than 2,000 options, many students do not submit all six choices. As documented by Table 1, almost six percent of choices are missing. A choice is defined as missing if the student does not list a choice, lists the same choice more than once, or lists a program that does not exist. The traditional explanation for missing choice is based on the participation margins in school matching - students prefer the outside option to many schooling alternatives.[8]

[6] Student placement before the centralized system was based on yearly regional selection meetings.

[7] For exposition simplicity, we use the term school or program to refer to a bundle school$\times$program. When strictly referring to a school, we use the term senior high school (SHS).

[8] This explanation is unlikely in Ghana. That is, outside options (labor market outcomes, matching outside of the system) are extremely limited for many students.

Table 1: Missing Choices

		Students test score (quartiles)			
ROL	All	0.337 – 0.537	0.537 – 0.603	0.603 – 0.691	0.69 – 1
All	0.9403	0.9283	0.9357	0.9421	0.9558
Choice 1	1	1	1	1	1
Choice 2	0.9993	0.9995	0.9994	0.9992	0.9993
Choice 3	0.9952	0.9935	0.9946	0.9952	0.9976
Choice 4	0.9833	0.9777	0.9812	0.9829	0.9915
Choice 5	0.9047	0.8957	0.9012	0.905	0.9175
Choice 6	0.7595	0.7032	0.7378	0.7705	0.8289

Notes: Table reports the proportion of non missing choices. Proportions are computed for the full sample, and by interval of student test score.

Low test score students are more likely to report truncated lists - that is 29.7% to be compared to less than 17.1% for the highest test score students. More specifically, as higher ability students have presumably better outside options, missing choices in this setting can not be attributed to difference in outside values.

1.2 Biased beliefs

As we use administrative data, we do not have access to the beliefs of students. We use additional data from a field experiment that surveys 12,871 students. The survey took place across 450 junior high schools in the Ashanti Region in 2016 (Ajayi et al., 2020).[9] Several details about the survey are important. First, students are interviewed before submitting applications for choice - as such, the information set of students in the survey mirrors that of students in the administrative data. Second, although the survey sample is from a single region, Appendix Table E shows that in key dimensions such as gender, age and test score, there is almost no difference between the survey sample and the population. It should be clear that the timing of the survey align perfectly with the choice environment with respect to student search.

The survey contains detailed information about beliefs (admission chances, aspirations, and expectations) as well as detailed socio-demographic attributes. Specifically, the survey asks all students for a discrete measure of their likely BECE, which allows us to create a mapping between beliefs and realizations. Then, we consider uncertainty regarding cutoffs. Ahead of their application, the survey asks students to list four choices, and their beliefs about the cutoffs of these choices. In order to link beliefs about test score, which are stated once, and those about cutoffs, which are stated for multiple schools, we proceed as follows. A student has correct beliefs about a school when her beliefs about admission chances align with realized chances. Stated

[9]As the survey and choice data are collected different years, our main assumption is that uncertainty is constant across periods.

differently, belief are correct as long as the ordering between realized test score and cutoff of the school is maintained. As such, the definition of correct belief is a little restrictive as pessimism does not constitute a mistake. Similarly, when beliefs do not align with realizations, we refer to these beliefs as incorrect. Beliefs may be incorrect because of cutoffs, test score uncertainties and both. We perform a decomposition of belief errors in Table 2.

Table 2: Biased Beliefs

	All	Students test score (quantiles)			
Beliefs	Students	0.337 − 0.537	0.537 − 0.603	0.603 − 0.691	0.69 − 1
Correct	0.602	0.418	0.576	0.638	0.779
Incorrect Beliefs					
Both	0.146	0.195	0.167	0.142	0.080
Cutoff	0.079	0.030	0.078	0.105	0.103
Test Score	0.172	0.356	0.179	0.115	0.037

Notes: Table reports percentage of choices for which students have correct beliefs about their admission chances and then decomposes incorrect beliefs into mistakes about cutoffs and test scores,

We find that 60.2% of all choices are exerted with correct beliefs. Conversely, almost 40% of choices are based on incorrect beliefs. Exposure to these biased beliefs varies by test score. That is, almost 78% of high test score students have correct beliefs, while only 41.8% of low test score have correct beliefs. Yet, it would be incorrect to argue that low test score students are more affected by uncertainty. The underlying reason is based on the differential effect of uncertainty on students as illustrated by figure C in the appendix. High test score students tend to be pessimistic about their admission chances, while low test score students are generally too optimistic. While pessimism implies biased beliefs, its consequences in terms of admission probabilities are not as severe as optimism.

We also quantify the instances where individuals have incorrect beliefs. On average, test score and cutoff uncertainties account for respectively 17.2% and 7.9% of all mistakes. In 14.6% of all choices, biased beliefs about both cutoffs and test scores lead to incorrect beliefs about admission chances. Analyzing beliefs by test score quartiles reveals that test score uncertainty accounts for a substantial part of mistakes for low test score students, while uncertainty regarding cutoffs is more prevalent among high test score students.

1.3 Placement Outcomes.

Next, we consider the outcome of students' application behavior. Table 3 reports the placement outcome of all students. Placement outcomes are then derived by student's test score. Finally, we document the placement outcomes of students depending on whether or not one of their choices is missing.

The vast majority of students gain admission into their first three choices. Altogether, 25.7% of individuals are admitted into their first choice. Interestingly, not only high test score students are placed into their first choice, approximately 11.2% of the low test score students are assigned to their first choice as well, which speaks to potential non-diversification in the ROLs of some students.

The value of the fifth and sixth choices appears relatively limited, as respectively 7% and 3.2% of students get admitted into those choices. This observation is at odds with the share of unassigned students (administrative assignment), which is 13.3%.[10]

We should stress that administrative assignment is a dreadful prospect for anyone participating in the match. The uncertainty and panic over this outcome are summarized by press articles stating "Over 60,000 students' fate in limbo as they can not be placed under CSSPS". Since 2018, the government has adopted a double track system, whereby individuals initially placed are assigned to the green track with school starting in September, while administratively assigned students are placed on a Gold Track that starts in November. As a consequence, *one should not expect any student to purposely get administratively assigned in the hope of a better assignment later*.

Table 3: Placement

		Students test score (quantiles)				Truncation	
Placement	All	0.337 − 0.537	0.537 − 0.603	0.603 − 0.691	0.69 − 1	Complete	Missing
1	0.257	0.112	0.199	0.273	0.448	0.242	0.301
2	0.211	0.117	0.181	0.252	0.297	0.212	0.207
3	0.166	0.131	0.176	0.201	0.159	0.171	0.154
4	0.131	0.143	0.156	0.152	0.072	0.138	0.109
5	0.070	0.085	0.089	0.085	0.022	0.080	0.040
6	0.032	0.058	0.052	0.017	0.000	0.042	0.000
99	0.133	0.354	0.148	0.021	0.002	0.115	0.190

Notes: Table reports the placement of students, including administrative assignment. The placement outcome is also reported by student test score. Placement "99" refers to administrative assignment.

As expected, administrative assignment is closely related to test score - inasmuch as 35.4% of lowest ability students are administratively assigned, while only 80 high test score students end up unmatched. In Appendix A.2.2, we study the application behaviour of students and show that some application lists contain multiple selective schools and no safety options leading to administrative assignments. Finally, the truncation panel shows that while students who have

[10]Recent press articles state that "About Sixty-seven thousand three hundred and eighty-two (67,382) students who qualified for senior high schools (SHS) could not be placed under CSSPS this year (2018)", and "we could not place 122,706 students" in 2019. The problem is still present.

missing choices are more likely to be assigned to their first ranked choice, they are also more likely to be administratively assigned. As we document in Appendix A.2.5, similar schools face very heterogenous numbers of applications.

1.4 Vacancies.

Finally, we consider the vacancy rate. Table 4 shows that only 71.5% of the schools end-up at capacity.

Table 4: Vacancies

	Full sample	At-capacity	Below-capacity
Share		0.715	0.285
Total Seats	78.434	79.532	75.748
Number of Vacancies			40.938
Historic	0.083	0.113	0.010
Boarding	0.562	0.660	0.319
Programs			
Agriculture	0.146	0.120	0.211
Business	0.172	0.169	0.181
General Arts	0.212	0.223	0.186
General Science	0.118	0.142	0.057
Home Economics	0.174	0.176	0.168
Technical	0.064	0.052	0.094
Visual Arts	0.113	0.118	0.102

Notes: Table reports the occurrence of vacancy at the school level and illustrates the characteristics of vacant and non vacant schools.

Not surprising, the vacancy rate is decreasing in school selectivity. However, vacancies are not confined to low selectivity schools.[11] Almost 32% of vacant schools have boarding facilities, and all academic tracks are represented.

The observation that almost a quarter of students do not submit the required number of choices combined with two digits administrative assignment and vacancy rates suggest a deeper problem, which we posit to be the existence of imperfect information. Under imperfect information, students engage in a costly search process to learn about schooling options, similar to standard consumer search models. Although beliefs about admission chances are not a central part of our setting, we use survey data to account for the possibility that individuals may have biased beliefs about their admission chances.

[11] That is, only 78% of the 25% most selective schools are at capacity, a ratio that increases to 79.5% when we consider the 5% most selective schools. While most vacant schools have one remaining seat, the least selective schools have more vacancies.

2 A Model of School Application Under Imperfect Information.

In this section, we develop an empirical model, which is consistent with the key facts presented in section 1. To that end, we introduce frictional search in the standard school application problem.[12] We formulate the school application process as a search problem, where students iteratively and sequentially acquire information about schools. The assumption that search is sequential implies that the decision to stop acquiring information depends on the "luck" of the student. This assumption helps us match the observation that some high ability students do not list six choices. Also, the search framework allows us to generate mismatch on the extensive (administrative assignment for students and vacancy at the school level) and intensive (matching into more desirable school) margins. Finally, search costs may compel students to consider only a subset of choices, leading to sub-optimal decisions.

2.1 Framework.

The school choice problem is summarized as follows. A finite set of students indexed by $\mathcal{I} = i \in \{1, \ldots, I\}$ apply to a finite set of schools indexed by $s \in \{1, 2, \ldots, S\}$. We denote by $\mathcal{S}$ the full set of schooling options. Each school has a positive and exogenous capacity. Total capacity is given by $\chi = \sum_s^S \chi_s$.

2.2 Utilities and Preferences.

A student i is characterized by a vector of observed attributes $\mathbf{x}_i$ and a test score t_i^r, which is unknown when she submits a rank order list (ROL). The latter defines individual admission priorities while the former captures preferences. School s is characterized by a vector of observed attributes $\mathbf{z}_s$, and a cutoff q_s. We describe our assignment mechanism later.

A student with characteristics $\mathbf{x}$ derives utility $u(\mathbf{x}, \mathbf{z}, \xi, \epsilon)$ from an assignment to a school with observed and unobserved characteristics $\mathbf{z}, \xi$.[13] The parameter ξ_s captures choice-specific unobserved attributes, which may be similarly valued across students, but unobserved to the econometrician. Examples of ξ include the reputation of a school and the quality of teachers. Finally, ϵ_{is} captures individuals' and schools' match-specific shocks. For example, a student may have an unobserved taste for her parents' alma mater.

Students can opt out of the matching system and enjoy an outside utility u_{i0}. All other high school admissions will go through the matching process except for international private high schools. Therefore, u_{i0} is best interpreted as the inclusive value of remaining unassigned. In our empirical application, we specify the indirect utility of individual i at school s as:

[12] A recent literature provides a theoretical foundation to search as originating from endogenous consideration sets under the notion of rational inattention. Our approach is related to theoretical models that study the implications of rational inattention for choices using search technology (Masatlioglu and Nakajima, 2013; Caplin and Dean, 2011).

[13] Implicitly, this formulation rules out preferences for peers that are not captured through observed and unobserved school attributes.

$$
\begin{aligned}
u_{is} &= u(\mathbf{x}_i, \mathbf{z}_s, \xi_s, \epsilon_{is}) = \sum_{j=1}^{J}\sum_{k=1}^{K}\beta_{kj}\mathbf{z}_{ks}\mathbf{x}_{ij} - d_{is} + \xi_s + \epsilon_{is}, \qquad (1)\\
u_{i0} &= 0.
\end{aligned}
$$

where d_{is} is the distance as the crow flies between student i's JHS school district and school s district. There are K school covariates and J student characteristics. Further, we assume that

$$
\xi_s \sim \mathcal{N}(0, \sigma_\xi^2), \qquad \epsilon_{is} \sim \mathcal{N}(0, \sigma_\epsilon^2). \qquad (2)
$$

Since over 99% of programs are public schools, we use distance as our numeraire. This scale normalization is embedded in the distance coefficient being -1. The set of parameters to be estimated consists of β, σ_ξ and σ_ϵ. Consequently, the parameters β measure students' willingness to travel for school attributes and interactions between student and school characteristics.

Arguably, the distance variable is not the best numeraire as most students do not commute daily in this setting. One would have preferred a measure such as value-added, however, we do not have access to any such variable.[14] In addition to the fixed lodging expenses (boarding fee), there are multiple ways in which the distance between the parental home and the school may decrease the student's utility. Longer distance implies lengthy and expensive trips when students go back home. As a consequence, we use distance as our numeraire.

2.3 Assignment Mechanisms.

The assignment mechanism is a serial dictatorship, which uses submitted ROLs and students' priority to determine final assignments. Test scores determine student priorities. Let $R \in \mathcal{R}$ be a ROL. Students submit a ROL that does not have to reflect their true preferences over schools. In our current setting, students can submit up to six choices, a constraint that makes it even more likely that ROLs may not reflect true preferences. The payoffs depend not only on which schools are listed, but also on the order in which they are listed. We assume that students understand this. We also assume that students act as price takers, taking admission probability as given.[15]

Finally and following the literature, we denote by q_s, the cutoff of school s. Formally, the cutoff of school s, q_s, is the minimal test-score required for admission at school s and formally defined as $\mathbb{1}\{\mathcal{D}_s(q_s) \leq \chi_s\} \quad \forall s \in \mathcal{S}$ where $\mathcal{D}_s()$ is the demand for a school s with capacity χ_s.

[14] School quality could be an alternative, but using this variable as a numeraire would hinder our ability to fit the data.

[15] See, for example, Agarwal and Somaini (2018); Azevedo and Leshno (2016), among others, for a similar assumption.

2.4 Agent Beliefs.

At the time of submitting lists, priorities and cutoffs are not known. Priorities are based on individual test scores obtained from a national exam that will take place four months later. We assume agents submit optimal rank-order lists given a set of beliefs about the assignment probability.[16]

We assume that agents have private information about their preferences u_i, and are endowed with beliefs regarding test score and cutoffs. Specifically, we assume that individual beliefs about test score are denoted by $\mathcal{B}_i^t$ and follow $\mathcal{B}_i^t \sim \mathcal{N}(\mu_i^t, (\sigma_i^t)^2)$. The parameters μ_i^t and σ_i^t, which are individual-specific, introduce the idea that individuals do not know their test score, but the underlying distribution.

We adopt a similar approach for beliefs about cutoffs $\mathcal{B}_{is}^q$ and follows $\mathcal{B}_{is}^q \sim \mathcal{N}(\mu_s^q, 1)$. In that sense, $\mathcal{B}_{is}^q$ is a prior over school cutoffs. Due to data limitations, we do not let the parameters that characterize cutoffs' beliefs vary across individuals. In addition, the variance of $\mathcal{B}^q$ is normalized for reasons that will become clear later.

Finally, we assume that agents may update their beliefs about cutoffs through search, which is described in section 2.7. Specifically, upon searching a school s, a student may update her beliefs about the cutoff of that school. As, it will become clear in the section 2.7, students update beliefs regarding only the discovered schools. Similarly, there is no update on beliefs regarding test score. Under this setting, perceived admission chances are respectively given by

$$\varphi_{is} = \Pr(\mathcal{B}_i^t \geq \mathcal{B}_{is}^q). \tag{3}$$

Given preferences $u(\cdot)$, and perceived admission chances $\varphi(\cdot)$, the expected utility from reporting R is denoted by $\mathcal{U}(R)$. We should note that $u(\cdot)$ captures the utility of a school, while $\mathcal{U}(\cdot)$ denotes the expected utility from a ROL. Optimality conditions impose:

$$\mathcal{U}(R) \geq \mathcal{U}(R') \quad \forall \quad R' \in \mathcal{R}^{\mathcal{S}}. \tag{4}$$

Where $\mathcal{R}^{\mathcal{S}}$ is the set of possible ROLs from choice set $\mathcal{S}$.

2.5 Search Models

We consider the problem of an individual searching for a set of schools, taking the characteristics of schools, and the behavior of other students as given. Agents are risk-neutral and maximize the expected utility of the submitted ROL. In contrast to standard consumer search models, students are searching for an optimal list of schools. We consider three alternative models.

[16]The standard approach to modeling beliefs is to assume agents have correct beliefs about the probabilities of assignments given their priorities and the distributions of preferences and priorities in the population. The evidence presented in section 1 on the share of administrative assignment strongly rejects rational expectations.[17]

2.5.1 Exogenous Consideration Sets.

In this section, we consider a model where the consideration set of students is a random draw from the list of schools as in Manzini and Mariotti (2014) and Barseghyan et al. (2019).

In practice, the consideration set forms in two steps. First, with probability ϕ_n, a student draws a consideration level n. The consideration level takes discrete values in $\{1, \ldots, S\}$ and determines how the individual considers many schooling options. In the second step, a student with consideration level n, draws a consideration set $\mathcal{K}$, sampling from $\mathcal{S}$ without replacement. Standard approaches assume that both the consideration levels and sets are independent from preferences. We maintain this assumption.

The exogenous consideration set model is a first step toward a theory of choices where all alternatives are not considered. However, such a model is unable to match differences across individuals when it comes to the application behavior: more than 29% of sixth choices for low ability students are missing, compared to around 17% for the high ability students. As a consequence, we propose an alternative model where students endogenously build consideration sets.

2.5.2 Endogenous Consideration Sets

We propose a sequential search process, where students build and expand their endogenous consideration set. The key insight of this model is that students expand their consideration set until the expected benefit of search is inferior to the potential cost of searching. As we do not observe the search sequence, the model requires several assumptions to discipline the behavior of individuals. In the rest of this section, we describe the search problem.

Agents search over schools to build and expand their consideration set, denoted by $\mathcal{K} \subset \mathcal{S}$. At the beginning of time, the consideration set is empty. The assumption of an empty initial consideration, which is easy to relax, is first and foremost motivated by the empirical observation that more than 1,000 students submit less than three choices, and almost 25% of the sample do not submit six choices. Search is random, an assumption that we relax in the next section.

At search step n, which corresponds to the norm of the consideration $\|\mathcal{K}\|$, upon paying a search cost $c(n)$, a student draws a school with probability λ_s. Given these elements, we can provide a recursive formulation to the search problem of the agents. The initial value of search is given by:

$$\mathcal{V}_i^{\emptyset} = \max\left\{u_0, \mathbb{E}\mathcal{V}_i^1\right\}. \tag{5}$$

where $\mathbb{E}\mathcal{V}_i^1$ depends on utilities $u(\cdot)$, perceived admission chances $\varphi(\cdot)$, and sampling probabilities $\lambda(\cdot)$.[18] To be clear, the definition of $\mathbb{E}\mathcal{V}_i^1$ implies a one-period lookahead simplifying assumption for the initial search. Still, continuation values may differ across individuals be-

[18] The $\mathbb{E}$ notation in the $\mathbb{E}\mathcal{V}$ captures the idea of expected value as students are uncertain about the school draw of the future period.

cause of observed and unobserved heterogeneity. However, we assume that the first draw is free such that all students whose participation constraint is satisfied (at least one school with positive utility) keep searching after the initial step.

At search step n, the value of continuing search to a student i with consideration set $\mathcal{K}$ of length n is given by $\mathcal{V}_i(\mathcal{K})$:

$$\mathcal{V}_i^{\mathcal{K}} = \max\left\{\mathcal{U}(\rho_i^{\mathcal{K}}), \mathbb{E}\mathcal{V}_i^{\mathcal{K}} - c(\mathcal{K})\right\}. \tag{6}$$

where the value of search $\mathbb{E}\mathcal{V}_i(\mathcal{K}) = \sum_{s\in\mathcal{S}\setminus\mathcal{K}} \lambda_s \mathcal{U}(\rho_i^{\mathcal{K}_s})$, and $\mathcal{K}_s = \mathcal{K} \cup s$. The notation $\mathcal{U}(\rho_i^{\mathcal{K}_s})$, which captures the utility of a ROL, implicitly includes the role of admission chances, and belief updating. The value of search is the utility gain associated with the remaining schools, which is weighted by the sampling probability. $\rho_i^{\mathcal{K}}$ is the ROL that yields the highest utility, which is formally defined as $\mathcal{U}(\rho_i^{\mathcal{K}}) \geq \mathcal{U}(R_i') \quad \forall \quad R_i' \in \mathcal{R}_i^{\mathcal{K}}$.

The object $\mathbb{E}\mathcal{V}(\cdot)$ is the expected value of a consideration set, which is a scalar. As such, it is weakly increasing and concave in the consideration set size. Consequently, the model implies that there exists an optimal consideration set. Also, higher order beliefs students will have larger consideration sets. The intuition follows that the expected gain from search $\mathbb{E}\mathcal{V}$ increases with test score when $\varphi(\cdot)$ increases with beliefs. Finally, the term $\mathbb{E}\mathcal{V}(\cdot)$ captures not only the expected gain of search, but it also accounts for the fact that through search, students acquire a signal about the cutoff.

2.6 Directed Search

A strong simplifying assumption of the two previous approaches is related to undirected search as students are equally likely to consider desirable and undesirable schools. This section considers a version of the model where students can direct their search toward more desirable schools. Although, we show in the Appendix Table C that there is no feature of geographical search, individuals may use other characteristics to direct their search.

We follow the direct cognition model developed by Gabaix et al. (2006). The intuition is that students classify schools into a finite number of groups and search over groups instead of schools. There is still a random search component as students can not direct their search toward a specific school.

Formally, assume there is a finite group of school types denoted by $\mathcal{J}$. The set of schools belonging to group $\mathcal{J}$ is denoted by $s_{\mathcal{J}}$. The strategy is similar to Avery et al. (2014), where students are advised to group schooling options into 3 categories: reach, match and safety. Each of those types is characterized by a distribution of schools known to the students. Then, we iteratively let students choose an optimal type that depends on its characteristics and current consideration set, upon which she receives a draw of a specific school/program.

Formally, the search process is as follows. First, students choose the group that yields the highest expected utility at the initial stage:

$$\mathcal{J}' = \arg\max \left\{ \mathbb{E}\mathcal{V}^{\mathcal{K}}_{i\mathcal{J}'} - c_{\mathcal{J}'}(\mathcal{K}) \right\}. \tag{7}$$

Then, given consideration set $\mathcal{K}$ and for any chosen $\mathcal{J}$, the individual randomly draws a school belonging to group $\mathcal{J}$, which enters into its consideration set $\mathcal{K}$. Future search decision is based on the difference between the utility of the current portfolio and the value of the best alternative:

$$\mathcal{V}^{\mathcal{K}}_{i} = \max \left\{ \mathcal{U}(\rho^{\mathcal{K}}_{i}), \mathbb{E}\mathcal{V}^{\mathcal{K}}_{i\mathcal{J}'} - c_{\mathcal{J}'}(\mathcal{K}) \right\}. \tag{8}$$

In order to discipline the model, we need search cost to be alternative-specific. Introducing some directness into the search problem is likely to match observed choices without larger consideration sets.

2.7 Belief update

Through search, agents construct consideration sets but students also learn about schools. We assume that upon "searching" about a school s, a student i receives a noisy information signal $\mathcal{B}^{u}_{is}$ about the selectivity of school s. We let the signal be

$$\mathcal{B}^{u}_{is} = q_s + \epsilon^{u}_{is}. \tag{9}$$

with $\epsilon^{u}_{is} \sim \mathcal{N}(0, (\sigma^{u}_{is})^2)$. At the outset, it should be clear that the discovery of school s does elicit a noisy signal about school s, and only school s. Additionally, it should be clear that there is no updating of individual test score.

As only discovered schools can be listed, we can not separately identify the variance of $\mathcal{B}^{q}_{is}$ from that of $\mathcal{B}^{u}_{is}$. Although the survey data allows us to elicit beliefs about cutoffs, there is no additional information allowing to separate beliefs regarding schools that are listed (subject to $\mathcal{B}^{u}_{is}$) and those that are not listed (not subject to $\mathcal{B}^{u}_{is}$). As a consequence, the variances of $\mathcal{B}^{q}_{is}$ and $\mathcal{B}^{u}_{is}$ can not be separately identified.

Hence, each signal $\mathcal{B}^{u}$ leads to a Bayesian update of beliefs about cutoffs, with the following expressions:

$$\begin{aligned} E(\mathcal{B}^{q}_{is}|\mathcal{B}^{u}_{is}) &= \mu^{q}_{s} + (q_s - \mu^{q}_{s})\frac{1}{1+(\sigma^{u}_{is})^2}. \\ V(\mathcal{B}^{q}_{is}|\mathcal{B}^{u}_{is}) &= \frac{(\sigma^{u}_{is})^2}{1+(\sigma^{u}_{is})^2}. \end{aligned}$$

The sign of $(q_s - \mu^{q}_{s})$ determines the direction of the update, pessimistic students update their belief upward while optimistic students adjust their belief downward. Although every signal generates an update, the magnitude of the change in posterior beliefs depends on individual priors. The posterior variance is independent of the direction of the update, and depends on the variance of the signal. As a consequence, the updated admission chance for a school s is given by

$$\varphi_{is} = \Pr(\mathcal{B}_i^t \geq \mathcal{B}_{is}^q | \mathcal{B}_{is}^u). \tag{10}$$

2.8 Constructing the optimal portfolio

Finally, we should note that constructing the best set of schools given a consideration set involves a complex combinatorial analysis. Since an individual uses the same test score belief to evaluate her admission chance throughout the search process, choices are interdependent. Explained differently, rejection in the first choice conveys additional information on one's test score and the expected distribution of cutoffs. Recently, Shorrer (2019) and Calsamiglia et al. (2020) have proposed a method to construct the best set of schools in that setting. The approach uses dynamic programming to account for the inter-dependence in admission chances across choices. We use this strategy to recover the highest utility ROL.

3 Identification and Estimation

Let θ be a set of parameters to be estimated with parameter space Θ, P is the true distribution of the data, and $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ denotes our model for the distribution of the observed data. The observed data comes from two distinct mechanisms: the formation of consideration sets and the selection of a ROL. As a consequence, the probability of observing a portfolio is given by

$$P_\theta(R|\mathbf{x}, \mathbf{z}, \xi) = \sum_{R \in \mathcal{K}} \mathbb{P}_{\mathcal{K}}(R) P_\theta(R = \underset{R' \in \mathcal{R}^{\mathcal{K}}|\mathbf{x},\mathbf{z}}{\operatorname{argmax}} \, \mathcal{U}_\theta(\mathbf{x}, \mathbf{z}_{R'}, \xi_{R'}, \epsilon)). \tag{11}$$

Where $\mathcal{U}_\theta$ is the utility evaluated under parameter values θ, and $\mathbb{P}_{\mathcal{K}}(R)$ is the probability of considering $\mathcal{K}$, which includes the selected schools R. Assume the model is correctly specified in the sense that there exists $\theta \in \Theta$ such that $P_\theta = P$. Formally, the identification problem consists of determining under which conditions the solution to $P_\theta = P$ is unique.

The standard identification argument in discrete choice with limited attention involves using explicit data on consideration sets (see e.g. Jolivet and Turon, 2014; Honka et al., 2017; and Dinerstein et al., 2018) or restrictions that some determinants of attention are orthogonal to preferences (Barseghyan et al., 2019; Heiss et al., 2016). Abaluck and Adams (2017) show consideration probabilities are identified from asymmetries of cross-derivatives of choice probabilities with respect to attributes of competing products. However, the methods require a regressor that is alternative-specific and enters the indirect utility function linearly.

We develop a strategy to identify and estimate all the parameters of the model. Recognizing the limitations of using administrative data to identify beliefs and preferences separately, we use data from a survey that combines realized test scores and as well as a survey of beliefs. Then, we separate preferences from consideration.

3.1 Identifying Beliefs

In this section, we discuss the identification and estimation of assignment probabilities of students. The environment features two sources of uncertainty. First, at the time of submitting the ROL, the student does not know her test score, as priorities are based on test scores obtained from a national exam that will take place four months later. Second, cutoffs are not known. The substantial share of unmatched students implies that rational expectations may not hold. Also, results from Agarwal and Somaini (2018) and Carvalho et al. (2019) show that beliefs and preferences cannot be separately identified. Under our theoretical model, we have the following distributions characterizing beliefs.

$$\mathcal{B}_i^t \sim \mathcal{N}(\mu_i^t, (\sigma_i^t)^2) \tag{12}$$

$$\mathcal{B}_{is}^q | \mathcal{B}_{is}^u \sim \mathcal{N}\left(\mu_s^q + (q_s - \mu_s^q)\frac{1}{1+(\sigma_{is}^u)^2}, \frac{(\sigma_{is}^u)^2}{1+(\sigma_{is}^u)^2}\right). \tag{13}$$

We use survey data to estimate the parameters of these distributions. As explained in section A.2.4, we use additional data from a field experiment that surveys 12,871 students across 450 junior high schools. Prior to submitting a list of choices, the survey asks students about their beliefs regarding their BECE score as well as beliefs for the cutoffs of four choices. Unfortunately, we do not have enough data to estimate μ^t and σ^t for each individual and beliefs over schools σ^u by individual.

As a consequence, our main goal is to devise a strategy to derive heterogeneity in these parameters. An additional difficulty comes from the fact that the survey and administrative data are collected on different students such that any belief parameter needs to be projected on the administrative data. We use discrete types that allow us to construct a finite mixture representation of beliefs in our setting.

Let $\mathcal{G}$ be the number of points of support of student types, and let us denote the types as $g(i) \in \{1, \ldots, \mathcal{G}\}$. A group g is characterized by a distribution of test score beliefs with parameters μ_g^t and σ_g^t such that the test score belief of an individual i, belonging to group $g(i)$ is given by

$$\mathcal{B}_i^t \sim \mathcal{N}(\mu_{g(i)}^t, \sigma_{g(i)}^t). \tag{14}$$

The formulation allows for inter-group variability and intra-group heterogeneity. In the application, we assume $\mathcal{G}$ is known and is the largest number that guarantees inter- and intra-group heterogeneity.

We proceed similarly for schools. Schools can be partitionned into $\mathcal{H}$ groups. Beliefs regarding the cutoffs of those schools follow a same distribution with parameters μ_h^q and σ_h^q such that beliefs about the cutoff of a school s, belonging to group $h(s)$ are given by:

$$\mathcal{B}_{is}^{q} \sim \mathcal{N}(\mu_{h(s)}^{q}, 1) \tag{15}$$

$$\mathcal{B}_{is}^{q}|\mathcal{B}_{is}^{u} \sim \mathcal{N}\left(\mu_{h(s)}^{q} + (q_s - \mu_{h(s)}^{q})\frac{1}{1+(\sigma_{h(s)}^{u})^2}, \frac{(\sigma_{h(s)}^{u})^2}{1+(\sigma_{h(s)}^{u})^2}\right). \tag{16}$$

The expressions for $\mathcal{B}_{is}^{q}$ and $\mathcal{B}_{is}^{q}|\mathcal{B}_{is}^{u}$ are both needed as students evaluate the continuation value of search without updating their beliefs. In order to maintain the basic structure of the model as well as heterogeneity in beliefs, we use a dimension reduction method (k-means) to partition students and schools into classes. For a given number of classes, we find all possible partitions and derive class-sizes (ω^t, ω^q) and class-specific parameters (μ^t, μ^q,σ^t and σ^u). The number of classes is set a-priori $\mathcal{G} = 10$ and $\mathcal{H} = 5$. As we estimate the model by gender, we have 5 heterogeneity groups for male and female students and schools.

3.2 Identifying Preferences

In this section, we analyze the identification of preference parameters. We restrict our attention to the cases where sub-utilities are additive in the unobserved components, such that

$$u(\mathbf{x}, \mathbf{z}, \xi, \epsilon) = m(\mathbf{x}, \mathbf{z}, \xi) + \epsilon. \tag{17}$$

Where $m(\cdot)$ is a known function, which depends on an unknown vector of parameters to be estimated. Matching effects between students and schools are denoted by ϵ. In addition, we assume that the error terms ϵ are i.i.d across students and schools and normally distributed with cdf F_ϵ. We analyze the identification of the model under full consideration, then study how the introduction of limited consideration changes our identification results.

3.2.1 Full consideration

Let $\mathcal{U}(\mathbf{x}, \mathbf{w}, \mathbf{z}_R, \xi_R, \epsilon)$ denote the expected utility of a ROL R. The utility of the ROL depends on $u(\cdot)$ and $p(\cdot)$. The variable $\mathbf{w}$ is a potential shifter that affects admission chances $p(\cdot)$, but does not affect utilities $u(\cdot)$. Conditional choice probabilities can be derived:

$$\Pr(R|\mathbf{x}, \mathbf{z}_R, \xi_R) = \Pr(\mathcal{U}(\mathbf{x}, \mathbf{w}, \mathbf{z}_R, \xi_R, \epsilon) > \mathcal{U}(\mathbf{x}, \mathbf{w}, \mathbf{z}_{R'}, \xi_{R'}, \epsilon')) \quad \forall R' \in \mathcal{R}^{\mathcal{K}}. \tag{18}$$

Because of the portfolio structure of the choice problem, there is no closed-form solution. However, choice probabilities can be derived by simulation.[19] Results from Agarwal and Somaini (2018) and Carvalho et al. (2019), show that $m(\cdot)$ and $F_\epsilon(\cdot)$ are identified. The identification results from Agarwal and Somaini (2018) require either a special regressor, which has large support or variation in choice environments. However, the most intuitive way to think about identification

[19] For example, a simple frequency simulator can recover the choice probabilities.

is to recall that we have access to a ROL. Under our mechanism, the submitted ROL reflects true preference order among the school ranked. As a consequence, a ROL provides a set of moment conditions, which along with additional parametric restrictions guarantee identification. Using the ROL does not preclude the need of a special regressor, as the reported choice of some students may be selected.[20]

In our empirical application, we use test score as a special regressor following Agarwal and Somaini (2018). In practice, we do not include the individual test score as a determinant of utilities, implying that the distributions of (preference) ϵ's do not vary with test scores. We proceed to study how the introduction of limited attention changes these identification results. We first consider the model with exogenous consideration sets, then the sequential search model.

3.2.2 Exogenous consideration sets:

In this section, we consider the identification of the exogenous consideration model presented in section 2.5.1. Recall that the probability of choosing a given ROL is

$$\Pr(R_1|\mathbf{x}, \mathbf{z}, \xi) = \sum_{R_1 \in \mathcal{K}} \mathbb{P}_{\mathcal{K}}(R_1) \Pr(R_1 \mid \mathcal{K}). \tag{19}$$

Where $\mathbb{P}_{\mathcal{K}}(R_1)$ is the probability of drawing a consideration $\mathcal{K}$ that contains R_1, and $\Pr(R_1 \mid \mathcal{K})$ gives the probability that the ROL R_1 is selected from consideration $\mathcal{K}$. Recall that under our exogenous consideration model, individuals draw a personalized set, which is a random subset of the choice set. As the consideration set is unrelated to individual characteristics, the ROL still provides a set of moment conditions, based on the ordering of schools, which can be used to identify individual preferences.

However, identifying consideration probabilities requires either additional information or restrictions. In our empirical application, we assume that consideration size in a random draw for the discrete uniform distribution with a-priori set parameters.

3.2.3 Sequential search:

Finally, we consider the sequential version of the model. In contrast to the previous case, agents endogenously decide whether or not to expand their consideration. Consequently, it is not possible to write a generic expression for choice probability.

Consider a simpler example, where a student is choosing between two alternatives. Let λ_1 and λ_2 denote respectively the probabilities of drawing R_1 and R_2. In addition, let c denote the search cost. Then, the probability of selecting R_1 may be written as:

[20] When there is no endogeneity concern (i.e., the error term $\epsilon = (\epsilon_1, \ldots, \epsilon_S)$ is distributed independently from individual observed $\mathbf{x}$ and school unobserved and unobserved $\mathbf{z}, \xi$). However, selection issues occur when for example low test score students apply only to a specific set of schools.

$$\begin{aligned}\Pr(R_1|\mathbf{x}, \mathbf{w}, \mathbf{z}_{R_1}, \xi_{R_1}) = {} & \lambda_1 \Pr(\mathbb{E}\mathcal{V}(R_1) < c(1)) \\ & + \lambda_1 \Pr(\mathbb{E}\mathcal{V}(R_1) > c(1)) \Pr(\mathcal{U}(\mathbf{x}, \mathbf{w}, \mathbf{z}_{R_1}, \xi_{R_1}, \epsilon) \geq \mathcal{U}(\mathbf{x}, \mathbf{w}, \mathbf{z}_{R_2}, \xi_{R_2}, \epsilon)) \\ & + \lambda_2 \Pr(\mathbb{E}\mathcal{V}(R_2) > c(1)) \Pr(\mathcal{U}(\mathbf{x}, \mathbf{w}, \mathbf{z}_{R_1}, \xi_{R_1}, \epsilon) \geq \mathcal{U}(\mathbf{x}, \mathbf{w}, \mathbf{z}_{R_2}, \xi_{R_2}, \epsilon)). \end{aligned} \tag{20}$$

Where $\mathbf{w}$ is the special regressor, $\mathbf{x}$ is individual attributes, while $\mathbf{z}$ and ξ are school observed and unobserved attributes.

The first part of equation 20 accounts for the eventuality that R_1 is drawn first. In this case, R_2 is not considered if the gain of search is lower than c. The second line accounts for the possibility that R_2 may have been considered, in which case, R_1 is selected only if it yields higher utility. Finally, R_2 may have been drawn first, in which case R_1 is selected only if the gain of search outweighs its cost, and R_1 yields higher utility. Equation 20 for choice probabilities is relatively simple in this case because i) search technology is known, and ii) students sample schools at random. Yet identification is more complex as we need to identify not only the parameters of the utility function but also consideration probabilities and search cost.

Making use of the large support condition in choice probabilities equations 20, we have the following system of equations:

$$\Pr(R_1|\mathbf{x}, \underline{\mathbf{w}}, \mathbf{z}_{R_1}, \xi_{R_1}) = \lambda_1 \Pr(\mathbb{E}\mathcal{V}(R_1) < c(1)), \tag{21}$$

$$\Pr(R_1|\mathbf{x}, \overline{\mathbf{w}}, \mathbf{z}_{R_1}, \xi_{R_1}) = \lambda_1 + \lambda_2 \Pr(\mathbb{E}\mathcal{V}(R_2) > c(1)). \tag{22}$$

In essence, we need beliefs to be such that some schooling options can be excluded. While the presence of a special regressor simplifies the problem, preferences can not be identified separately from consideration probabilities. To make further progress, we parametrize consideration probability λ as:

$$\lambda_s = \frac{\chi_s}{\chi}. \tag{23}$$

where χ_s is the number of openings at school s, and $\chi = \sum_s^S \chi_s$. Then, with λ_1 and λ_2 a-priori known, the conditional probabilities identify the optimal stopping probability. Unfortunately, the identification works only in a setting when search cost varies with the size of the consideration. As such, we focus on a setting where $c(\mathcal{K}) = c$, which implies that individuals keep searching until the value of search is equal to c.[21] Our main conclusion is that the sequential search model is identified, provided that sampling probabilities are known a-priori and the researcher has access to a special regressor, which shifts consideration but not preferences.

[21] A recent paper by Agarwal and Somaini (2022) analyzes the identification of a choice model with unspecified consideration sets. Identification in that setting requires two sets of instruments. Identification in our setting requires additional restrictions.

3.2.4 Directed Search Model

The identification of the directed search model follows from the same argument. To see this, let $\mathcal{J}$ tend toward the number of schools S, then the directed search is similar to the sequential search model.[22] In the next section, we confirm these results using Monte Carlo simulations.

3.3 Monte Carlo Experiments

To study the practical identification of these parameters, we performed a Monte Carlo simulation. In the design, the sample size is $N = 160,000$, and the number of choices is 2,089 just like the real data. The true preference parameters are drawn from a uniform distribution $\mathcal{U}(0,1)$, and we assume that the model is correctly specified, i.e., the *the data generating process (DGP) is correct*. In practice, for a set of parameters, we simulate choices for each model and derive a set of empirical moments. We then try to recover the preference parameters that rationalize the empirical moments. Table 5 reports true parameters as well as the 95% confidence intervals from 100 Monte Carlo experiments.

Table 5: Monte Carlo Experiment.

	True	Full Consideration		Exog. Consideration		Endog. Consideration		Directed Search	
		Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Historic	0.069	0.067	0.088	0.049	0.089	0.053	0.074	0.068	0.079
Boarding	0.818	0.785	0.831	0.81	0.847	0.805	0.822	0.81	0.828
Religious	0.943	0.905	0.965	0.925	0.969	0.921	0.965	0.937	0.965
Single-Sex	0.269	0.259	0.292	0.243	0.318	0.231	0.28	0.261	0.304
School Quality	0.169	0.163	0.196	0.16	0.179	0.147	0.192	0.168	0.199
Gen. Science	0.034	0.033	0.067	-0.002	0.076	0.011	0.043	0.031	0.08
Business	0.179	0.172	0.184	0.143	0.184	0.153	0.2	0.172	0.201
Gen. Arts	0.642	0.616	0.653	0.609	0.661	0.609	0.66	0.639	0.66
Home Econ	0.023	0.022	0.062	-0.002	0.025	-0.007	0.034	0.022	0.044
V. Arts	0.008	0.008	0.013	-0.029	0.016	-0.009	0.018	0	0.037
Agri.	0.393	0.377	0.428	0.365	0.431	0.353	0.421	0.389	0.443
σ_ξ	0.814	0.781	0.825	0.78	0.834	0.792	0.838	0.813	0.841
σ_ϵ	0.376	0.361	0.39	0.361	0.394	0.364	0.397	0.37	0.428
c	0.03					0.028	0.034		
c(reach)	0.05							0.041	0.053
c(match)	0.02							0.018	0.025

Notes: Results of the Monte Carlo experiments. Estimates from the simulated method of moments, with moments described in section 3.5. Lower and Upper bounds constructed from 100 Monte Carlo experiments. Exogenous consideration sets are drawn from $\mathcal{U}(3,100)$, while the endogenous consideration sets is the search cost is given by $c(n) = c$.

[22]The identification problem is relatively simple as the direction of the search is assumed known.

Comparisons between the true parameters and the estimated coefficients show that parameters are similar in all cases. The parameters are also precisely estimated, confirming our theoretical identification results. Yet, it should be clear that the estimated model does not correspond to the DGP, parameters are less acurately and precisely estimated.

3.4 Estimation

Given beliefs and the identification results, we can estimate preference and search cost parameters. There are two complications. Given the number of choices, it is impossible to compute all potential consideration sets. In addition, as consideration sets are endogenous, constructing the probability requires conditioning on possible consideration sets and the order in which schools are sampled. As a consequence, we use simulation methods. That is, we simulate forward the optimal stopping problem for every student in our sample and construct ROLs. Simulations allow us to deal with complications related to the number of choices and the stochastic components of the utility and the search process.

We estimate the parameters using the Simulated Method of Moments. That is, we match the empirical characteristics of student ranked choices. Formally, let Θ denote the set of parameters to be estimated. The criterion function is given by:

$$\mathcal{L}(\Theta) = -\frac{1}{2}\big(\widehat{m} - m(\Theta)\big)^T \widehat{W}^{-1}\big(\widehat{m} - m(\Theta)\big). \tag{24}$$

where $\widehat{m}$ is a set of empirical moments, and $\widehat{W}$ is the weighting matrix.[23]

The model is simulated using the following algorithm, which is initialized with an initial set of parameters Θ^0:

1. Given the number of discrete individual belief groups $\mathcal{G} = 10$ and school belief groups $\mathcal{H} = 5$, use K-means to determine the weights of each class (ω^t and ω^q) and the corresponding distributional parameters μ^t,σ^t,μ^q and σ^q
2. Given ω^t, draw a class for each student, and then draw a test score belief $\{\mathcal{B}_i^t\}_{i\in\mathcal{I}}$ for all students.
3. Given ω^q, draw a set of belief for all schools $\{\mathcal{B}_{is}^q\}_{i\in\mathcal{I},s\in\mathcal{S}}$
4. Calculate the matrix of utility and admission chances.
5. Draw an initial consideration set.
6. Simulate forward the search problem and derive simulated portfolios.

[23] We use a diagonal weighting matrix, with the elements set equal to the inverse of the diagonal variance-covariance matrix of the empirical moments. Since we have discrete dependents, any approximation of the gradient vector will be sensitive to the chosen step size. Therefore, we calculate the derivative by approximating the function by a low-order polynomial function as we vary each parameter locally.

7. Given portfolios, solve for the matching algorithm, derive placement outcomes for all students, and construct the set of moments.

8. Repeat steps 2 to 7 R times to integrate out the stochastic component in the utility function, and update the criterion function.

The directed search model requires additional details. We group schools into 3 categories which is on line with Avery et al. (2014). In order to simplify this process, we define:

$$\mathcal{J} = \begin{cases} \text{back-up} & \text{if} \quad \mathcal{B}_i^t \in \left(0, (1-\varsigma)\mathcal{B}_{is}^q\right) \\ \text{match} & \text{if} \quad \mathcal{B}_i^t \in \left((1-\varsigma)\mathcal{B}_{is}^q, (1+\varsigma)\mathcal{B}_{is}^q\right) \\ \text{reach} & \text{if} \quad \mathcal{B}_i^t \in \left((1+\varsigma)\mathcal{B}_{is}^q, 1\right). \end{cases} \tag{25}$$

While preference estimates are sensitive to the choice of ς, we find that when ς is between 0.2 and 0.4, estimates are relatively stable. Further, we set ς to 0.3, which is the value that best fits the data.

3.5 Moments

Our final specification of the utility function consists of 11 parameters which capture key choice characteristics, and interactions between individual and choice characteristics. To identify these parameters, we construct empirical analogs that capture the identification content provided by the data. We use three sets of moments:

First, we construct the average characteristics of schools' observable by ranked choice

$$\frac{1}{n}\sum_n Z_{is_i^k}. \tag{26}$$

where s_i^k is the simulated choice of individual i at ranked choice k. These moments consist of the key characteristics of the simulated portfolios (Historic, Boarding, Religiousity, Academic tracks as well as the average school quality) for each ranked choice $k = 1, \ldots, 6$. This first set of moments capture the general features of ranked choices.

Second, we construct a similar set of moments conditioning on some individual characteristics.

$$\frac{1}{n}\sum_n Z_{is_i^k}|X_i. \tag{27}$$

The conditioning variables are gender, and test score $Q1$ (lowest 25% test score) and $Q4$ (highest 25% test score) students. This second set of moments allows us to capture differences related to gender and academic ability in the ranked choices.

Finally, we use a third set of moments that captures the mismatch in the sample. Consistent with the search literature, we match the vacancy rate, and the share of administrative assignment. Then, we use the share of missing choices for each ranked choice as well as the correlation between past and realized cutoffs. Our Monte Carlo study indicates that these moments are crucial to pin down the search cost parameter. We have a total of 267 moments.

4 Results

This section presents our estimation results. First, we discuss the estimates of the model of belief formation. Then, preference estimates are presented. Search cost parameters and the implied consideration sets are then described. Finally, the fit of the model is discussed.

4.1 Beliefs

We present estimates from the model of belief formation. As explained before, we use K-means to group students into ten groups and schools into five groups.

Table 6: Beliefs about test score.

Groups	Beliefs		Weights	Realized test score	
	μ^t	σ^t	ω^t	mean	sd
			Female		
1	0.481	0.323	0.014	0.309	0.037
2	0.568	0.306	0.094	0.421	0.031
3	0.616	0.287	0.225	0.513	0.025
4	0.660	0.260	0.330	0.594	0.024
5	0.720	0.231	0.274	0.680	0.027
			Male		
1	0.484	0.307	0.062	0.388	0.044
2	0.561	0.287	0.287	0.522	0.038
3	0.643	0.259	0.485	0.639	0.037
4	0.775	0.195	0.166	0.754	0.038
5	0.819	0.173	0.064	0.776	0.033

Notes: K-means estimates for the distribution of test score beliefs.

Table 6 shows the parameters (scale and location) as well as the weights of the five distributions for female and male students, which are then associated with the realized distribution of test scores. Not surprisingly, students with higher realized test scores have higher test score beliefs. However, the scale parameters are large enough that the distribution of beliefs overlaps for many groups. The estimation captures the essential aspects of the data: i) students with similar realized test scores have very different beliefs, ii) students with similar beliefs have different realized test scores, and iii) beliefs about test scores may not be precise, but higher realized test score students tend to have higher beliefs.

Table 7 then considers beliefs about cutoffs. Almost half of the choices (48.5%) are perceived to be highly selective, while a little less than 20% of the choices belong to the least selective group.[24] Ex-post heterogenous schools are clustered together in relatively homogenous belief groups. In addition, and contrary to test scores, realized higher selectivity schools are not systematically associated with higher beliefs.

Table 7: Beliefs about cutoffs.

Groups	Beliefs		Weights	Realized cutoffs	
	μ^q	σ^q	ω^q	mean	sd
1	0.082	0.028	0.123	0.550	0.117
2	0.227	0.033	0.060	0.560	0.123
3	0.400	0.044	0.152	0.565	0.125
4	0.605	0.043	0.180	0.581	0.128
5	0.846	0.098	0.485	0.600	0.134

Notes: K-means estimates for the distribution of belief about cutoffs.

Consequently, our main finding is that the two sources of uncertainty have different implications. Students may have accurate beliefs about their academic standings, but there is a sizeable amount of randomness in the BECE performance, which leads to significant noise in test score beliefs. In contrast, when it comes to schools, students do not have accurate beliefs about school selectivity, which may come from the size of the choice set.

4.2 Preferences

We consider preference parameters governing the utility of students. Table 8 presents eight different specifications that correspond to models of full, exogenous, and endogenous considerations.

Column (1) corresponds to the full consideration. Recall that this is a model in which students choose between schools based on observed characteristics, stochastic components, and distance. Analysis of this model is helpful because it provides a benchmark to compare our estimates to the rest of the literature, which is set in full information. The second model adds heterogeneity in consideration but independently from individual characteristics. Columns (2), (3), (4), and (5) present four specifications where the consideration sets are exogenously drawn from uniform distributions $\mathcal{U}(3,10)$, $\mathcal{U}(3,100)$, $\mathcal{U}(3,1000)$, and $\mathcal{U}(3,S)$.[25] The third model considers the

[24] These patterns are unlikely to be driven by selection into the choice set as students submit choices from different types of schools.

[25] For example, under $\mathcal{U}(3,10)$ consideration sets are drawn from a discrete uniform distribution with boundary parameters 3 and 10. We attempted to estimate the parameters of the uniform distribution, but the optimization of

sequential search model, where individuals endogenously build consideration sets. Compared to the previous model, the size of the consideration sets depends on individual beliefs and luck during the search process. Columns (6) and (7) present specifications with sequential search under random and directed search. Under both models, the search cost is set to c.

real and integer parameters turned out to be complicated. As a consequence, we calibrate those parameters. We tested lower bounds from 1 to 10 and upper bounds ranging from 10 to the number of schools S.

Table 8: Estimates of Preferences and Search Cost.

	Full	Exogenous Consideration				Endogenous Consideration	
	Consideration	$\mathcal{U}(3,10)$	$\mathcal{U}(3,100)$	$\mathcal{U}(3,1000)$	$\mathcal{U}(3,S)$	Random	Directed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Female			
Historic	0.772***	-2.251***	0.085	0.038	0.037	0.048	0.035
Boarding	-0.196***	1.157***	0.888***	1.005***	1.007***	0.823***	0.867***
Religious	-1.316***	-1.282***	0.204	0.217	0.343**	0.287*	0.317**
Single-Sex	-4.42***	0.409	0.38	0.444	0.453	0.412	0.424
School Quality	2.34***	2.495***	0.855***	0.998***	1.005***	0.861***	0.886***
Gen. Science	-0.429	-2***	-0.062	-0.085	0.042	-0.122	-0.092
Business	-1.429***	-0.144	-0.585***	-0.644***	-0.64***	-0.533***	-0.522***
Gen. Arts	0.692***	0.044**	0.619***	0.666***	0.892***	0.743***	0.757***
Home Econ	0.23	-3.059***	1.313***	1.331***	1.445***	1.217***	1.238***
V. Arts	-1.206***	-2.963***	0.361	0.449*	0.32	0.262	0.272
Agri.	-3.108***	0.37*	-0.405	-0.419	-0.538	-0.368	-0.371
σ_ξ	1.577***	1.45***	0.775***	0.885***	0.881***	0.884***	0.905***
σ_ϵ	1.246***	1.216***	0.438***	0.45***	0.44***	0.439***	0.443***
c						0.018**	
c(reach)							0.03**
c(match)							0.017**
				Male			
Historic	1.552***	-1.12***	0.584***	0.657***	0.576***	0.671***	0.676***
Boarding	-0.435	-1.107***	1.302***	1.235***	1.321***	1.323***	1.367***
Religious	-0.645***	0.873***	0.408***	0.549***	0.428***	0.556***	0.586***
Single-Sex	-4.55***	0.567	0.252	0.336	0.418	0.441	0.462
School Quality	1.125***	2.002***	0.456***	0.454***	0.527***	0.554***	0.577***
Gen. Science	-4.252***	0.687*	0.697*	0.765**	0.644*	0.666*	0.698*
Business	-4.465***	0.89**	0.754*	0.779*	0.722*	0.882**	0.901**
Gen. Arts	-0.634***	0.123	0.506***	0.551***	0.444***	0.567***	0.575***
Home Econ	-4.814***	1.115***	-1.574***	-1.767***	-1.56***	-1.756***	-1.729***
V. Arts	-2.753***	0.782***	0.24	0.449*	0.336	0.451*	0.457*
Agri.	-2.411***	1.116***	0.129	0.165	0.101	0.117	0.129
σ_ξ	2.246***	1.906***	0.275	0.439**	0.332*	0.319*	0.33*
σ_ϵ	2.249***	2.223***	0.741***	0.998***	0.925***	0.991***	1.018***
c						0.025**	
c(reach)							0.031**
c(match)							0.011**

Notes: Estimation results for four models and eight specifications. The models are estimated by gender.

We estimate the model flexibly using gender-specific parameters.[26]

In Appendix B, we present the fit of four models: full consideration, exogenous consideration $\mathcal{U}(3, 100)$, endogenous consideration, and directed search. Surprisingly, all these models fit the data reasonably well with the choice attributes (boarding, historical, religious, school quality), but less so for academic tracks. Overall, the directed search provides the best fit to the data, and in most instances, there is almost no difference between the data and the model.[27]

Finally, we should note that several models, namely the directed search, the endogenous and exogenous consideration with $\mathcal{U}(3, 1000)$ lead to almost identical estimates for most preference parameters. Because of these reasons, our comments focus on the parameters of the directed search model.

The coefficients for observed characteristics are generally consistent with the demographic patterns reported in descriptive evidence in Appendix A. On average, students prefer boarding schools. Religious and mixed-sex schools appear to impact utility positively, although the coefficients are not always significant. School quality is arguably an essential determinant of individual choices. Academic tracks such as visual arts and agriculture do not significantly affect workers' utility.

However, there are some differences across gender groups. For example, female students display weaker preferences for historic schools than their male counterparts - which may be related to the competitive nature of these schools. In addition, there is a stark divide in preferences for academic tracks. For example, female students prefer home economics programs, while male students are more attracted to business and general sciences.

Table 8 reports estimated parameters for the distributions of unobserved preferences. The results show essential heterogeneity in unobserved tastes. The estimated coefficients on the idiosyncratic taste σ_ϵ are larger for male than female students, suggesting that male students are more likely to select a school on the basis of individual-school match effects. On the contrary, the relative importance of school-specific matches and idiosyncratic taste σ_ε is reversed. Female students are more likely to select a school based on school-specific matches than male counterparts.

4.3 Search cost and Consideration Sets

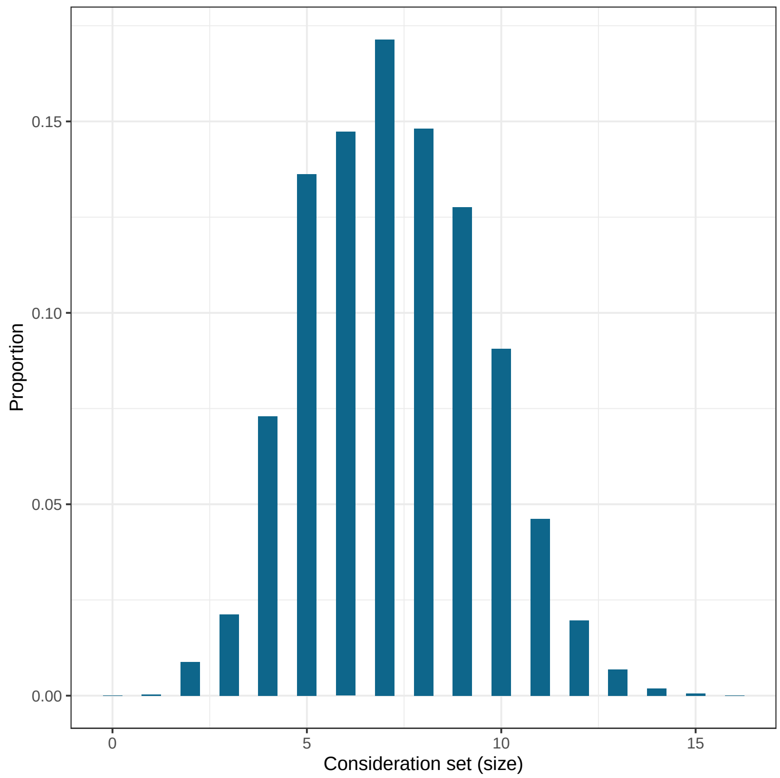
Finally, we quantify the role of search cost. Table 8 displays estimates for the search cost parameters. As the quantitative value of search costs does not have a clear structural interpretation,

[26] Given the complexity of the models, the criteria function displays non-convexities. We find that low dimensional parameters are easier to handle, and as such, we estimate models separately by gender.

[27] The ability of the various models to provide a good fit to the data creates some confoundedness between specification and identification issues. Although all these models are identified in practice, they are not encompassing and may use different tools to rationalize the same observation. For example, consider the observation that more than 25% of the sample submit less than six choices. The full consideration model can rationalize these patterns only through the participation condition, pushing downward estimates for school quality. As such, the full consideration model implies negative utility for most schools. Especially for male students, choices are almost exclusively driven by stochastic shocks. As models with consideration sets allow students not to consider all options, negative utilities are not required.

we simulate choices for all students multiple times and report the implied distribution of consideration in Figure 1. The distribution of consideration sets is identified from information on truncated lists and identifying restrictions.

Figure 1: Distribution of consideration sets



The share of students who consider one to five schools is observed in the data, and normality assumptions on the error terms imply that the distribution of consideration sets is approximately normal. Intuitively, given the observed consideration sets, the model derives a distribution of consideration sets under i) a consideration set formation technology, and ii) individual heterogeneity. The model implies that students consider between 1 to 26 choices. The median student considers seven choices compared to 2,089 choices. This relatively small consideration set speaks to the potential for misallocation but raises important questions about why students do not search more. The model provides two potential explanations. First, the gains to search are relatively small given the number of choices - even within our directed search model, the probability of drawing a highly desirable school is small. Second, biased beliefs may lead to a bias in assessing the value of a ROL.[28]

[28]While any notion of efficient search would be complex in this setting, we can perform a simple simulation where we let students have accurate beliefs about admission chances and see how the distribution of consideration changes. We find that considerations for lower ability students would substantially increase - the median student would consider up to 18 choices.

5 Counterfactual Simulations

In this section, we analyze the efficiency content of our model. Given individual preferences and technological constraints on vacancies, we investigate whether a social planner could achieve a better allocation. Then, we build on this result to explore ways to improve the allocation of students. The rest of this section is based on the specification with a fixed search cost.

5.1 Quantifying inefficiencies

Inefficiencies stem from the existence of search friction and uncertainty.[29] Inefficiency is reflected in the existence of missing choices, vacancies, administrative assignments, and potential mismatches among the matched students.

We quantify the particular importance of search frictions and uncertainty for welfare. We propose a simple utilitarian welfare function $\sum_i U_i^*$ that aggregates individual utilities, where U_i^* is the utility individual i derives from the school she was assigned. Individuals who do not gain admission into any of their choices are randomly assigned to a local school.

We consider three settings. Frictional Application is our benchmark case. In our benchmark, students are subject to uncertainty and frictions. Then, the problem of the Constrained Planner (CP) is analyzed. Under CP, the planner maximizes total welfare subject to preference and technology constraints. Specifically, the planner collects preferences and assigns students based on the number of vacancies at each school. Since matching is centralized, we assume that the planner can mitigate the effect of frictions and uncertainty.

Finally, we evaluate welfare sequentially, relaxing the sources of inefficiencies: search frictions, test score uncertainty, and aggregate uncertainty. The effects of these changes need to account for the possibility that cutoffs may change endogenously for students as there are no uncertainty or search frictions. As a consequence, we solve for the Bayesian Nash Equilibrium. That is, given an initial set of cutoffs denoted by q^0, and preferences estimated in our setting summarized by u, we solve for the following algorithm.

1 Individuals select the rank-order lists that maximize expected utility.

2 Given submitted lists, students get admitted to schools, and the realized matching determines the new distribution of cutoffs.

3 Repeat until cutoffs converge.

The solution concept is similar to Walters (2018), with the difficulty of accounting for belief bias. In practice, starting from a distribution of test score and cutoffs beliefs, we let students use past realizations to update these beliefs. Table 9 reports the results.

[29] Another source of inefficiency comes from the discrepancy between private and social values of search, which arises because of the standard "overcrowding" among students: when an extra person lists a school, it reduces the availability of vacancies for other students. This is an externality that students are not likely to internalize. See Abdulkadiroglu et al. (2015) for a theoretical analysis of this problem in school choice. However, we do not address this. Also, the fact that students can submit only six choices induces additional efficiency concerns.

We compute the aggregate utility under each of these scenarios using estimates from the directed search model and compare them to the efficient allocation.[30]

Table 9: Efficiency

		No Search Frictions			With Search Frictions		
	Benchmark	No Unc.	AU only	TU only	No Unc.	AU only	TU only
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Welfare	0.34	1	0.852	0.795	0.726	0.547	0.481
Decomposition of welfare by student test score							
Students Q4	0.45	0.33	0.52	0.39	0.40	0.53	0.48
Students Q3	0.28	0.29	0.24	0.32	0.31	0.22	0.26
Students Q2	0.22	0.26	0.15	0.18	0.19	0.18	0.16
Students Q1	0.05	0.12	0.09	0.11	0.10	0.07	0.10

Notes: Welfare under different scenarios and welfare decomposition by student ability. No Unc = No uncertainty; AU only = aggregate uncertainty only; TU only = test score uncertainty only.

Column (1) presents the welfare effect under our benchmark, including search frictions, aggregate uncertainty, and test score uncertainty. Column (2) corresponds to the normalized welfare under the efficient allocation. Columns (3) and (4) quantify the welfare loss that can be attributed to coordination frictions, and test score uncertainty when there are no search frictions. Column (5) quantifies the welfare loss that can be attributed to search frictions. Finally, Columns (6) and (7) characterize the welfare loss that can be attributed to coordination frictions, and test score uncertainty when there are search frictions.

We show that the constrained planner achieves approximately three times more welfare than our benchmark. More precisely, the welfare under the benchmark accounts for 34% of the efficient allocation. These estimates are relatively robust across the various models. That is, we obtain respectively 36% and 41% under the sequential random search and exogenous consideration models.

As we measure welfare in terms of willingness-to-travel, our results indicate that the cost of boarding could be divided by three in the economy. This magnitude of the gap between the two allocations suggests that administrative assignment alone cannot explain the difference. Indeed, we find a substantial mismatch, as more than 60% of students gain admission into a higher utility choice.

Efficiency is not the only problem that comes from search frictions and uncertainty, the allocation of resources is more unequal. That is, the highest ability students receive 33% of total welfare

[30] As it is not obvious how to treat the search cost when interpreting these results, we do not consider it. As a consequence, our results may be viewed as a lower bound on the welfare gain.

under the efficient allocation to be compared to 45% under the benchmark. As a consequence, school choice as implemented in Ghana, may not be a progressive policy.

Finally, we quantify the relative importance of these sources of inefficiencies in shaping welfare. Further computations show that 37.4% of the welfare loss can be attributed to the inability of students to gather information about alternatives, while test score uncertainty accounts for 20.5% of the welfare loss and aggregate uncertainty for 14.8%. Of these inefficiencies, test score uncertainty is likely the most straightforward problem to solve. In the next section, we consider an alternative design that alleviates these problems.

5.2 Choice Paradigm and Welfare

School choice is based on the premise that students (or parents) know best about which school to attend. The standard paradigm in choice theory, *the more options, the better*, reinforces the notion that expanding the horizon of choices beyond an assigned neighborhood, for example, improves welfare. This is the rationale that motivates a national placement system in Ghana. However, students may be worse off when school decisions are made without the full examination of all available options. Furthermore, as the number of choices increases, it becomes almost impossible for decision-makers to consider all choices. One extreme case of restricting choice is the efficient allocation, where the planner has full information on the preferences of all students. In this section, we analyze whether realistically restricting choice could be welfare-improving.

We propose a different mechanism, where instead of asking for specific schools, the planner collects preferences over the limited number of school attributes. As noted before, informational constraints, uncertainty, and the inability of students to construct ROLs under interdependent admission chances are the primary sources of inefficiencies. However, it is easy to see that an elicitation mechanism that would reveal preferences over school attributes will likely alleviate all these concerns. We refer to this mechanism as the Ranked-Attributes-List (RAL).

In theory, RAL retains the key properties of the serial dictatorship algorithm, as students are asked to resolve the trade-off between different school attributes using a ranking of which characteristics to consider first. For example, an example of RAL could be: i) selectivity: D9, ii) District: Accra, iii) Boarding, iv) Religious, and v) general science. Given this list, the planner assigns the student to a school belonging to the 10% most selective schools in the Accra District, with boarding facilities, in a religious institution, and academic track general science. The planner fulfilled as many choice characteristics as possible focusing on the highest-ranked factors. If, for example, there is a choice $S1$ that fulfills requirements i) and ii) and another $S2$ that satisfies i), iii) and iv), the student is assigned to $S1$. In an extreme case, where there is a choice $S1$ that fulfills requirement i) and another $S2$ that satisfies ii), iii) and iv), the student is assigned to $S1$. In principle, it is possible to weigh differently ranked attributes lists; or adopt a mix of attribute and school choices - these extensions are left for future research.

As we have only preference estimates, it is not trivial to implement this policy in our counterfactual simulations. We proceed as follows. First, we calculate the utility of students for all choices but focus on the 200 preferred choices. Then, we perform a utility decomposition to isolate the effect of each school attribute. As we consider only discrete co-variates, the reported school characteristics can be discretized. This is mostly relevant for capturing the effect of school quality and distance. We use indicators for the deciles of school selectivity and each pair of home and school districts. Then, we calculate the partial R_2 for each of these characteristics and derive an RAL based on their respective importance. There are two reasons why all students do not have the same RAL. First, we calculate utilities accounting for the stochastic components (ξ_j and ϵ_{ij}), as such, different draws may affect the relative ranking of school attributes. Second, as we discretize distance, the effect of location is heterogeneous across individuals. We repeat these calculations multiple times to integrate out the stochastic components. Table 10 reports the average welfare.

Table 10: Restricting Choices and Welfare

	Benchmark	RAL
Utilities	1	0.876
Vacancies	0.325	0.09

Notes: Welfare under a ranked attributes list (RAL) mechanism.

We find that school choice under RAL generates substantial welfare gains - recovering most of the welfare loss. RAL sidesteps the effect of uncertainty and search frictions on the application behavior and achieves approximately 87.6% of the efficient allocation with a minimal change to the current system.

6 Discussion and conclusions

This paper develops and estimates a model to understand individual preferences for schools in a large matching market. We introduce imperfect information in the standard school application problem. Search allows students to learn about the characteristics of schools and build consideration sets.

Using additional survey data on beliefs, we isolate the effect of consideration sets, preferences and beliefs. We show that there is substantial test score uncertainty, and students have biased beliefs about cutoffs. Our analysis of welfare shows that only 34% of the efficient allocation is realized. Further computations show that 37.4% of the welfare loss can be attributed to the inability of students to gather information about alternatives, while test score uncertainty accounts for 20.5% of the welfare loss and aggregate uncertainty for 14.8%.

Our findings raise new questions about school choice in large matching markets. Given the number of choices, students may not be aware of all schooling opportunities. The size of the choice set imposes several challenges on low ability students. The first is related to the lack of social ties that allow students to easily collect information about schools. Second, liquidity constraints are likely to affect those students more severely. Finally, the expected gain from search is relatively small for lower ability students, which decreases their value of search.[31]

The methods used in this paper provide several avenues for future research. Although our analysis focuses on key features of the education system in Ghana, the potential behavioral implications can be extended to many countries, as well large school districts in the US. One key extension would be to augment these types of administrative datasets with surveys on search behavior to get a better understanding of the nature of search as in Arteaga et al. (2021). These extensions are left for future work.

[31] Unfortunately, our data do not allow us to quantify the respective importance of these channels.

References

Abaluck, J. and Adams, A. (2017). What do consumers consider before they choose? identification from asymmetric demand responses. Working Paper 23566, National Bureau of Economic Research.

Abdulkadiroglu, A., Agarwal, N., and Pathak, P. A. (2017). The welfare effects of coordinated assignment: Evidence from the new york city high school match. *American Economic Review*, 107(12):3635–89.

Abdulkadiroglu, A., Che, Y.-K., and Yasuda, Y. (2015). Expanding "Choice" in School Choice. *American Economic Journal: Microeconomics*, 7(1):1–42.

Abdulkadiroglu, A., Pathak, P. A., and Roth, A. E. (2005). The new york city high school match. *American Economic Review*, 95(2):364–367.

Abdulkadiroglu, A. and Sonmez, T. (2003). School Choice: A Mechanism Design Approach. *American Economic Review*, 93(3):729–747.

Agarwal, N. and Somaini, P. (2018). Demand analysis using strategic reports: An application to a school choice mechanism. *Econometrica*, 86(2):391–444.

Agarwal, N. and Somaini, P. J. (2022). Demand analysis under latent choice constraints. Working Paper 29993.

Ajayi, K. F., Friedman, W. H., and Lucas, A. M. (2020). When Information is Not Enough: Evidence from a Centralized School Choice System. NBER Working Papers 27887, National Bureau of Economic Research, Inc.

Arteaga, F., Kapor, A. J., Neilson, C. A., and Zimmerman, S. D. (2021). Smart matching platforms and heterogeneous beliefs in centralized school choice. Working Paper 28946.

Avery, C., Howell, J. S., and Page, L. (2014). A review of the role of counseling, coaching, and mentoring in students? postsecondary outcomes. *College Board Research Brief*.

Azevedo, E. M. and Leshno, J. D. (2016). A Supply and Demand Framework for Two-Sided Matching Markets. *Journal of Political Economy*, 124(5):1235–1268.

Barseghyan, L., Coughlin, M., Molinari, F., and Teitelbaum, J. C. (2019). Heterogeneous Choice Sets and Preferences. Papers 1907.02337, arXiv.org.

Calsamiglia, C., Fu, C., and Güell, M. (2020). Structural estimation of a model of school choices: The boston mechanism versus its alternatives. *Journal of Political Economy*, 128(2):642–680.

Caplin, A. and Dean, M. (2011). Search, choice, and revealed preference. *Theoretical Economics*, 6(1).

Carvalho, J. R., Magnac, T., and Xiong, Q. (2019). College choice, selection, and allocation mechanisms: A structural empirical analysis. *Quantitative Economics*, 10(3):1233–1277.

Dinerstein, M., Einav, L., Levin, J., and Sundaresan, N. (2018). Consumer price search and platform design in internet commerce. *American Economic Review*, 108(7):1820–59.

Duflo, E., Dupas, P., and Kremer, M. (2021). The impact of free secondary education: Experimental evidence from Ghana. NBER Working Papers 28937, National Bureau of Economic Research, Inc.

Fack, G., Grenet, J., and He, Y. (2019). Beyond truth-telling: Preference estimation with centralized school choice and college admissions. *American Economic Review*, 109(4):1486–1529.

Gabaix, X., Laibson, D., Moloche, G., and Weinberg, S. (2006). Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, 96(4):1043–1068.

Goeree, M. S. (2008). Limited information and advertising in the u.s. personal computer industry. *Econometrica*, 76(5):1017–1074.

Haeringer, G. and Klijn, F. (2009). Constrained School Choice. *Journal of Economic Theory*, 144(5):1921–1947.

Heiss, F., McFadden, D., Winter, J., Wuppermann, A., and Zhou, B. (2016). Inattention and Switching Costs as Sources of Inertia in Medicare Part D. NBER Working Papers 22765, National Bureau of Economic Research, Inc.

Honka, E., Hortaçsu, A., and Vitorino, M. A. (2017). Advertising, consumer awareness, and choice: evidence from the u.s. banking industry. *The RAND Journal of Economics*, 48(3):611–646.

Jolivet, G. and Turon, H. (2014). Consumer Search Costs and Preferences on the Internet. IZA Discussion Papers 8643, Institute of Labor Economics (IZA).

Kapor, A. (2016). Distributional effects of race blind affirmative action. In *Working Paper, Yale University*.

Kapor, A., Neilson, C. A., and Zimmerman, S. D. (2018). Heterogeneous beliefs and school choice mechanisms. Working Paper 25096, National Bureau of Economic Research.

Luflade, M. (2017). The value of information in centralized school choice systems (job market paper).

Manzini, P. and Mariotti, M. (2014). Stochastic choice and consideration sets. *Econometrica*, 82(3):1153–1176.

Masatlioglu, Y. and Nakajima, D. (2013). Choice by iterative search. *Theoretical Economics*, 8(3).

Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. (2012). Revealed attention. *American Economic Review*, 102(5):2183–2205.

Shorrer, R. I. (2019). Simultaneous search: Beyond independent successes.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, 69(3):213–225.

Walters, C. R. (2018). The demand for effective charter schools. *Journal of Political Economy*, 126(6):2179–2223.

Appendix

A Motivation

Our data comes from Ghana, where the national school system consists of six years of primary school, three years of junior high school (JHS), and three years of senior high school (SHS). In contrast to most higher-income nations, high-school completion (SHS) is the final qualification for almost 80% of students (Duflo et al., 2021). Starting in 2005, students completing junior high school apply for admission to senior high school through a centralized application system. One may wonder about the rationale for organizing a nationwide education system for teenagers. As we show later, top academic programs are located in few regions; a national education system gives a pathway to elite schools for students in rural locations.[32]

Students apply to specific academic programs within a school and can submit a ranked list of up to six choices. Available programs include agriculture, business, general arts, general science, home economics, visual arts, and several occupational programs offered by technical or vocational institutes.[33]

After submitting their rank-order lists, students take a standardized Basic Education Certification Exam (BECE). The application system then allocates students to schools based on a serial dictatorship where priorities are determined by the BECE score.

Students who are unassigned at the end of the algorithm are administratively assigned to a nearby program with remaining vacancies. Our data, which provides individual choices along with BECE scores as well as admission outcomes, consists of final year (grade 9) students from the universe of junior high schools (grades 6-9) in 2008.

In principle, students are not able to switch schools and academic tracks once admitted through the centralized system.[34]

We checked the consistency between applications and admission outcomes, and we find a 95% matching rate. The inconsistency comes from students who submit a truncated ROL. It is possible that the admission office administratively assigned all students with missing choices. Because of this inconsistency, in the rest of the paper, we use the matching algorithm to generate admission outcomes, and vacancies.[35]

In the remainder of this section, we study in detail individuals' application behavior as well as admission outcomes, revealing some regularities that will guide our modeling strategy.

[32]Student placement before the centralized system was based on yearly regional selection meetings.

[33]For exposition simplicity, we use the term school or program to refer to a bundle school$\times$program. When strictly referring to a school, we use the term senior high school (SHS).

[34]However, there are some exceptional cases where some students manage to get admitted into a different program or school, likely through illegal means. As we do not have any data on these changes, we abstract from them.

[35]In the original data, a larger share of high selectivity schools have vacancies due to this problem.

A.1 Students and Schools.

This section reports the underlying statistics behind our data. We describe the characteristics of students before considering schools.

The full sample consists of 340,823 students, among which, 160,869 students (47%) passed the qualifying exam and are therefore considered for the matching. We consider only qualified students, as individual test score is not available for unqualified students.

Table A reports the basic characteristics of students. Over half (58.2 %) are male, and the average age is 16.6. More than 47% of students live in the Ashanti and Accra (capital) regions. Student performance, measured by the BECE exam ranges from 158 to 469 points out of a possible 600, which is re-scaled to take values between 0 and 1.[36] As such, students have very heterogeneous chances of gaining admission to any given program. Table A also reports that younger male students are over-represented among higher test score students. Similarly, high test score students are over-represented in the Accra and Ashanti regions. In the absence of information on family background, we proxy for it using measures of academic success at the junior high school level. JHS quality measures the average test score of students from a JHS, while JHS pass rate is the share of students who obtain a passing grade at the BECE. As expected, higher test score students are associated with higher quality JHSs.

Table A: Students Characteristics

		Students test score (quartiles)			
Characteristics	All	0.337 − 0.537	0.537 − 0.603	0.603 − 0.691	0.69 − 1
Age	16.662	17.26	16.959	16.552	15.851
Male	0.582	0.573	0.583	0.592	0.581
JHS quality	0.621	0.553	0.584	0.626	0.721
JHS pass rate	0.682	0.496	0.616	0.739	0.882
Regions					
Ashanti	0.228	0.144	0.217	0.272	0.283
Accra	0.243	0.15	0.185	0.241	0.401
Central	0.083	0.111	0.093	0.078	0.05
Eastern	0.101	0.119	0.112	0.097	0.074
Volta	0.066	0.091	0.076	0.062	0.034
Western	0.093	0.114	0.104	0.088	0.063
Number of Observations	160869	41079	40168	40019	39602

Notes: The table shows the average characteristics of junior high schools students who qualify for senior high school placement. Average characteristics are computed for the full sample, and by interval of student test score. For concision, the largest six of the ten regions are reported.

[36] There are 281 distinct values of the BECE score. Nevertheless, ties in the matching process are rare because of the size of the choice set. For example, the mode of the test score distribution is 262, with 1,324 students. Among the first choice of this group, there are 654 distinct options. When ties occur during the estimation, we break them randomly.

Then, we consider the other side of the market, which consists of senior high schools. There is a total of 641 schools, and some offer as many as 33 programs.[37] Altogether, there are 2,230 school-programs.[38] Table B reports the characteristics of all schools, regardless of students' choices.

Table B: Schools/program Characteristics

		School cutoffs (quartiles)			
Characteristics	All	0.337 – 0.458	0.458 – 0.512	0.512 – 0.612	0.612 – 0.923
Boarding	0.562	0.300	0.412	0.657	0.879
Historic	0.083	0.006	0.010	0.038	0.280
Religious	0.213	0.147	0.167	0.253	0.282
Size	78.434	78.919	72.313	78.820	83.655
Gender					
Boys Only	0.042	0.002	0.006	0.025	0.136
Girls Only	0.063	0.021	0.016	0.044	0.172
Coed	0.895	0.977	0.979	0.931	0.692
Programs					
Agriculture	0.146	0.187	0.179	0.134	0.084
Business	0.172	0.185	0.163	0.165	0.176
General Arts	0.212	0.209	0.175	0.215	0.249
General Science	0.118	0.049	0.111	0.107	0.203
Home Economics	0.174	0.172	0.198	0.197	0.130
Technical	0.064	0.091	0.093	0.050	0.023
Visual Arts	0.113	0.108	0.080	0.132	0.134

Notes: The table shows the average characteristics of all schools/programs. The size variable is defined as the number of vacancies reported by the school. The gender category reports the gender exclusivity of the school. Average characteristics are computed for the full sample and by interval of school selectivity measured by realized cutoffs in 2008. For concision, all the technical programs have been grouped into one category.

There is substantial variation across schooling options. Over half of the choices (56.2%) offer boarding facilities. The presence of boarding facilities implies that students may gain admission everywhere in the country. The British colonial administration established the historic schools (8.3%) before Ghana gained independence in 1957, and a little more than 21.3% of the programs were offered in schools with a religious affiliation. The average program has the capacity to admit 78.4 students, with a range from 10 to 120. While co-education has been generalized over the years, 10.5% of schools are still single-sex, with three-fifths of them being girls-only programs. A substantial share of the single-sex schools are also religious and were established

[37] This includes traditional high school, and both technical and vocational training institutions.

[38] Our attempts at reducing the dimensionality of the problem have failed as there are no systematic matching patterns between individuals and schools. As such, restrictions on the set of schools or individuals considered may alter the matching outcomes, and limit the scope of any counterfactual analysis.

pre-independence.

Finally, general arts is the most commonly offered program, accounting for approximately 21.2% of available academic tracks. Technical, which includes all vocational education programs represents 6.4% of the choices. We now consider the same characteristics by school selectivity. In this setting, selectivity is based on observed cutoffs in 2007.[39] There is a strong correlation between school quality and the indicators for boarding, pre-independence, and coed status. That is, a vast majority of high-selectivity schools offer boarding facilities (87.9%), over a quarter of them date back to the colonial era, and single-sex schools are over-represented among them. We also note that although there is not a monotonic relationship between school quality and size, more selective schools appear to offer more seats. Finally, exploring the level of selectivity by programs shows a consistent pattern: choices in general arts and general sciences are the most over-represented among high selectivity options. On the contrary, agriculture and technical programs are the least selective.

A.2 Choices.

The matching mechanism is based on serial dictatorship, which is strategy-proof when individuals are allowed to submit an unlimited number of choices. However, in our case, ROLs are truncated at six (6) choices, which prompts individuals to be strategic. Determining which subset of schools to submit is a complicated problem. Students must find the right balance between sought-after schools, which are likely to be selective while insuring themselves against the risk of un-assignment. While there is no simple strategy to construct a portfolio, the literature has provided some results about the properties of the optimal portfolio.

Proposition 1 ***Haeringer and Klijn (2009)****. Let $N_p = \mathcal{S} = \|U_n > 0\|$ be the set of alternatives with positive utilities. Then, the optimal strategy consists of choosing N among N_p, and ranks them according to the true preference ordering.*

Proposition 1 illustrates a simple property: while finding the optimal portfolio may not be obvious, the ordering within the portfolio is. Specifically, not ranking choices according to true preferences conveys the risk of getting assigned to a less preferred option.

[39] We have the same results when using realized cutoffs in 2008.

Table C: Characteristics of the ranked choices

	Choices					
	1	2	3	4	5	6
Historic	0.250	0.147	0.100	0.069	0.026	0.018
Boarding	0.869	0.812	0.754	0.671	0.604	0.592
Coed	0.749	0.873	0.921	0.949	0.965	0.978
Cutoff	0.677	0.635	0.603	0.570	0.526	0.516
Distance	34.012	32.747	30.457	26.980	30.953	31.890
Home District	0.358	0.336	0.342	0.382	0.220	0.205
Programs						
Agriculture	0.058	0.072	0.082	0.095	0.085	0.085
Business	0.198	0.215	0.205	0.191	0.196	0.188
General Arts	0.406	0.397	0.400	0.398	0.430	0.438
General Science	0.137	0.102	0.088	0.078	0.083	0.074
Home Economics	0.098	0.104	0.110	0.116	0.106	0.113
Technical	0.038	0.038	0.038	0.039	0.036	0.032
Visual Arts	0.066	0.072	0.078	0.082	0.065	0.070

Notes: The table shows the average characteristics of all schools/programs by ranked choices for 6 choices. Average distance (in miles) is evaluated between the centroid of the junior high school and senior high school districts using GPS coordinates.

In addition to the ordering of choices reflecting true preferences, students are often advised to diversify their rank-order lists, including selective schools as well as safer options. Using, this simple framework, we describe individual choices in detail.

Table C presents descriptive statistics on students' ranked choices. We report the characteristics of each ranked option to determine whether there exists a consistent pattern across choices. As mentioned before, students were allowed to list six choices in 2008. We find that students are more likely to list a school that was established pre-independence as their first choice. That is, 25.0% of first choices are historic schools, to be compared to 1.8% for the sixth choice. A similar pattern is observed for schools with boarding facilities, and single sex status.

Then, we examine the distance between a student's junior high school and selected senior high schools. We do not have exact coordinates for school locations, so we measure the distance between centroids of the 110 administrative districts in the country. Ghana's school choice system is truly national, and some students apply to schools as far as 450 miles away (roughly the distance from London to Geneva). At the same time, a substantial share of students apply only in their home region. Students' first choice is, on average, 34.01 miles away from their junior high schools, and their second choice programs are 1.27 miles closer to them. Their third and fourth-ranked choices are 30.46 and 26.98 miles away, but their last two choices are further away at a distance of 30.95 and 31.89 miles on average. Interestingly, individuals' first choices are more likely to be to located in their home district compared to their sixth choices. The arguments against a simple geography-based matching are better illustrated by the share of students who apply to schools in their home district. That is, 36% of first choices are to schools in their home district to be compared to 20% for the last choice. As such, it is unlikely that students are applying to their neighborhood schools as back-up options.

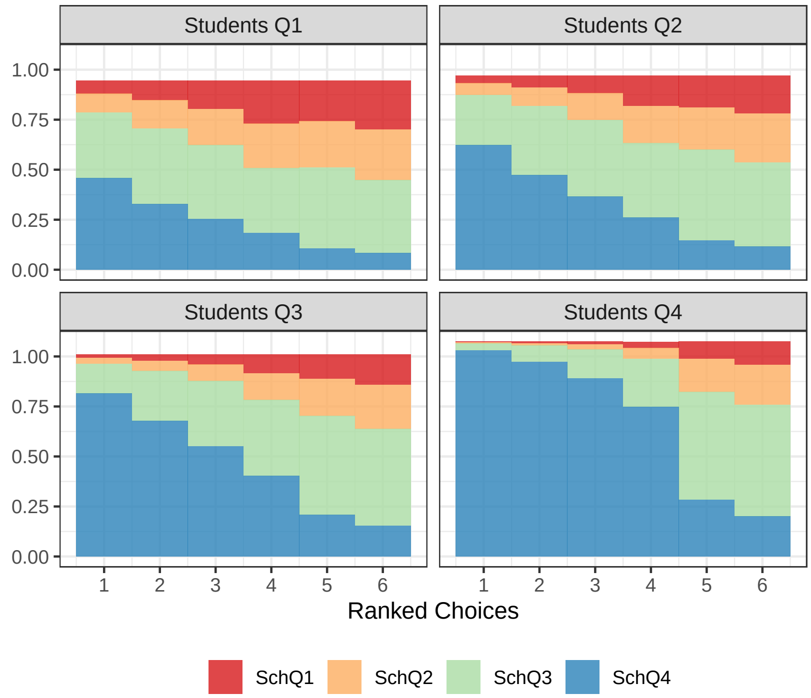
In contrast to preferences for distance, the selectivity of ranked programs decreases monotonically. The cutoff score of a students' first choice is 0.677 but falls to 0.516 for the lowest ranked-choice, which represents a difference of almost 1.5 standard deviations in the BECE score distribution.

Finally, we examine discrete program characteristics and reveal additional characteristics of aggregate choices in Table C. General arts is the most popular program track, with over 40% of students choosing it as their first and sixth choices, which is mostly explained by the large supply of general arts academic tracks. General science has the steepest gradient in choices. 13.7% of students choose a general science program as their first choice to be compared to 7.4% as a sixth choice. The remaining programs are more equally represented across choices, with 19.8% of students choosing a business program, 9.8% choosing home economics, 5.8% choosing agriculture, 6.6% choosing visual arts, and 3.8% choosing technical programs.

A.2.1 Selectivity in Choices.

After reporting the aggregate characteristics in choices, we provide a more in-depth analysis of school selectivity in ranked choices. As explained before, students should find a balance between selective schools and include some safety options. In this section, we discretize school selectivity by equally sized intervals and then report choices by school selectivity in Figure A.

Figure A: Diversification of Schools



Notes: Q1 corresponds to the 25 lowest percentiles. Q4 corresponds to 25 highest percentiles. SchQ1, SchQ2, SchQ3, and SchQ4 refer to schools belonging to the first, second, third, and fourth quantiles of selectivity respectively. StudentsQ1, StudentsQ2, StudentsQ3, and StudentsQ4 refer to students belonging to the first, second, third, and fourth quantiles of the test score distribution.

Half of the first choice applications of the lowest test score students are to the most selective schools compared to less than 10% for the sixth choice. Similar patterns are observed for all ability groups, and approximately 71% of first choices consist of the most selective schools, a ratio that decreases to 16% for sixth choices. However, the sharp decline in the share of top schools over ranked choices is almost entirely compensated by applications to schools belonging to the 25-75% percentiles of the cutoff distribution. More interestingly, almost 50% of sixth choice applications from the lowest ability group are to the two most selective groups of schools. Given these patterns, the risk of administrative assignment is real for many students. As a consequence, *some individuals may not be diversifying as they should*.

A.2.2 Diversification.

While instructive about aggregate patterns, the tables above do not speak to the internal consistency of individual choices. We, therefore, analyze whether individuals target a specific set of characteristics in their application behavior.

Table D: Repeated Characteristics in Choices

	Number of distinct choices					
	1	2	3	4	5	6
Programs	0.149	0.342	0.325	0.157	0.025	0.002
Schools	0.002	0.005	0.025	0.113	0.192	0.662
District	0.047	0.143	0.252	0.284	0.201	0.073
Historic	0.581	0.280	0.102	0.030	0.006	0.000
Boarding*	0.044	0.093	0.162	0.248	0.261	0.179
Religious*	0.314	0.239	0.124	0.041	0.008	0.001
Coed	0.005	0.023	0.060	0.159	0.296	0.457

Notes: Table analyses whether individuals target specific program characteristics in their applications. The first three rows describe the distribution of the number of distinct programs, schools, and districts ranked. The remaining rows report the number of listed options that satisfy each of the row labels. Reading: 14.9% of students apply to a single program track through out their application. $\star$ in front of boarding (resp. religious) indicates that some students do not report any school with boarding facilities (resp. religious affiliation).

Table D investigates whether students apply to choices with the same set of characteristics – such as academic tracks, SHS, districts, and regions. Our intuition is that individuals may target specific characteristics, and in the pursuit of these characteristics, choices may not reflect a thorough trade-off. Yet, we recognize that students' preferences over schools/programs are not unidimensional, as such our goal is simply to rule out the trivial cases. Only 14.9% of individuals apply to a single academic track throughout their entire list, which suggests that the vast majority of individuals do not attach a high value to a single academic track. A larger share of students apply to two and three program tracks (respectively, 34.2% and 32.5%). Regarding senior high schools (SHS), individuals almost exclusively apply to multiple SHS. Finally, choices are scattered geographically; with more than 28.4% of students are applying to four distinct districts.

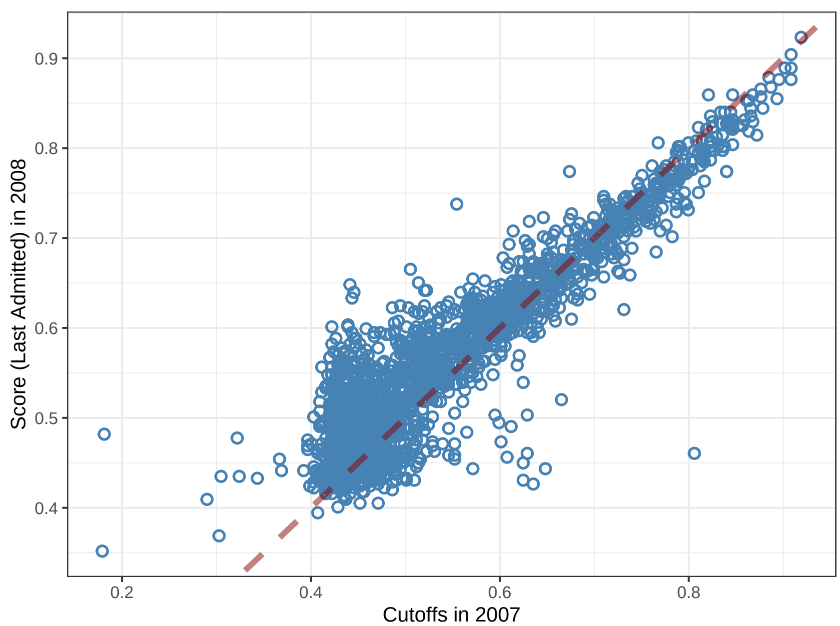
The observation that students do not attach a strong value to program nor specific SHS points to a genuine attempt to construct portfolios of schools that balance ambition and preferences. However, the fact that a limited set of variables cannot characterize choices suggests that the portfolio construction problem may be complex, with potential substitution between multiple choices. The complexity of constructing a portfolio may lead to mistakes.

A.2.3 Uncertainty.

In this section, we introduce the notion of uncertainty. There are two sources of uncertainty in our setting. The first, which we refer to as individual uncertainty, comes from the fact that students apply to schools before taking the exam that determines their ranking in the matching algorithm. The second, which we refer to as aggregate uncertainty, comes from limited information about the preferences of other students.

Figure B reports a strong correlation between cutoffs across years (0.979). However, the correlation drops to 0.762, when we account for schools with at least one opening (by setting the cutoff to zero for schools with open vacancies). Nonetheless, we note that there is more variation in cutoffs for lower selectivity schools. Uncertainty is a key characteristic of the market we analyze. We use data from a field experiment to collect additional information about beliefs.

Figure B: Stability of Cutoffs



Notes: Cutoffs are defined as the test score of the last individual admitted, regardless of capacity.

A.2.4 Survey Data

We use additional data from a field experiment that surveys 12,871 students across 450 junior high schools, which took place in the Ashanti Region in 2016 (Ajayi et al., 2020).[40] Although the survey sample is from a single region, Table E shows that in key dimensions such as gender, age and test score, there is almost no difference between the survey sample and the population. The survey contains detailed information about beliefs (admission chances, aspirations, and expectations) as well as detailed socio-demographic attributes. Specifically, the survey asks all students for a discrete measure of their likely BECE, which allows us to create a mapping between beliefs and realizations.

[40] As the survey and choice data are collected different years, our main assumption is that uncertainty is constant across periods.

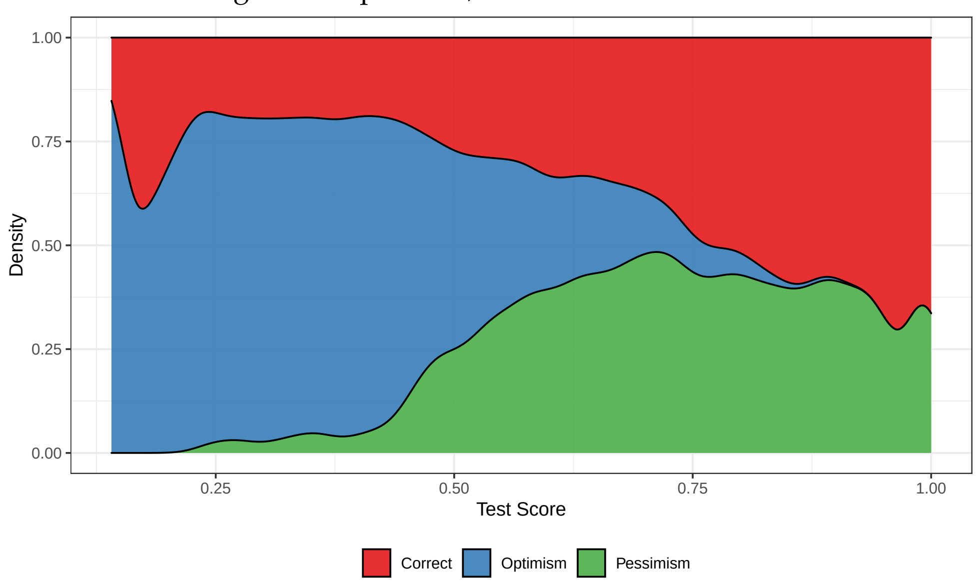
Table E: Coverage

Characteristics	Survey	Population
Female	0.437	0.418
Age		
Mean	15.854	16.662
Median	16.000	16.000
Test score		
Mean	0.583	0.621
Median	0.591	0.603

Notes: Comparison between survey and population. 'Population' is universe of 2008 students at time surveyed. 'Survey' describes surveyed individuals.

Figure C illustrates the differential effect of uncertainty on students. High test score students tend to be pessimistic about their admission chances, while low test score students are generally too optimistic. While pessimism implies biased beliefs, its consequences in terms of admission probabilities are not as severe as optimism.

Figure C: Optimism, Pessimism and Beliefs



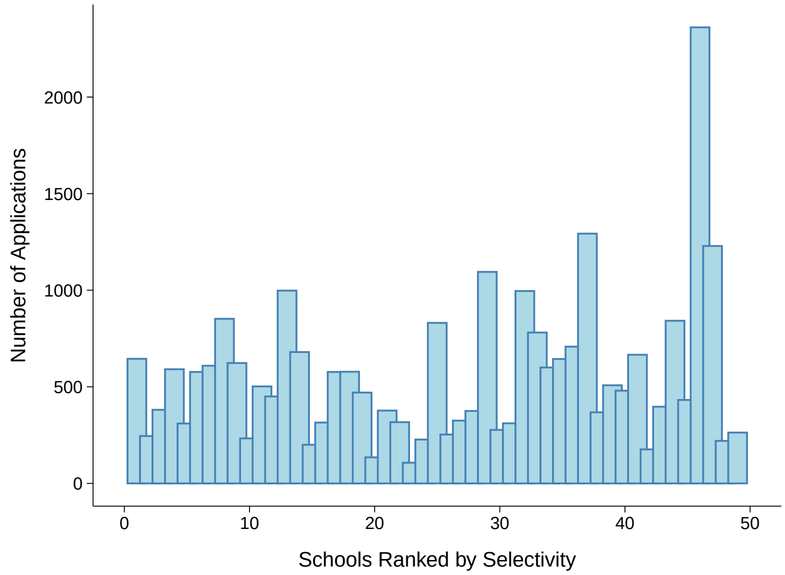
Notes: Figure indicates the differential effect of uncertainty on students based on test scores.

A.2.5 Applications and Selectivity

In this section, we show that seemingly similar schools face very heterogeneous number of applications. Figure D shows the number of first-round applications for the 50 most selective schools.

For example, we find that the 13^{th} most selective school receives three times fewer applications than the 14^{th} despite being located in the same school district and offering the same academic track.

Figure D: Number of applications to the 50 most selective schools



Notes: Total number of applications for the first 50 schools. Schools are ranked by selectivity.

Strategic interactions are likely to play an important role in generating these patterns. Given that programs in our setting are high schools, one should not expect the kind of trade-offs faced by college students between majors. As such it is hard to conceive why only 300 individuals will apply to the second most selective school, whereas the third most selective school receives more than 500 applications.

A.2.6 Vacancies

Finally, we document the existence of vacancies.

First, we explain how we obtain information about vacancies. The government maintained a registry of schools with information about seats and vacancies from 2005 to 2009. However, vacancy outcomes may not be reliable when individuals do not take up with admission outcomes. As a consequence, using applications alone, and information about seats, we compute the serial dictatorship algorithm and retrieve placement outcomes as well as the number of vacancies for each school. Table F shows that only 71% of the schools end-up at capacity.

Not surprisingly, the vacancy rate is decreasing in school selectivity. However, vacancies are not confined to low selectivity schools. That is, only 78% of the 25% most selective schools are at capacity, a ratio that increases to 79.5% when we consider the 5% most selective schools. While the median high selectivity school has one remaining seat, the least selective schools have more vacancies.

Table F: Vacancies

	Full sample	At-capacity	Below-capacity
Share		0.715	0.285
Total Seats	78.434	79.532	75.748
Number of Vacancies			40.938
Historic	0.083	0.113	0.010
Boarding	0.562	0.660	0.319
Programs			
Agriculture	0.146	0.120	0.211
Business	0.172	0.169	0.181
General Arts	0.212	0.223	0.186
General Science	0.118	0.142	0.057
Home Economics	0.174	0.176	0.168
Technical	0.064	0.052	0.094
Visual Arts	0.113	0.118	0.102

Notes: Table reports the occurrence of vacancy at the school level and illustrates the characteristics of vacant and non vacant schools.

The existence of vacancies and administrative assignment suggests a deeper problem, which we posit to be the existence of imperfect information. As such, one side of the market (students) may not be able to collect information about the other side of the market, resulting in violation of the law of one price, and mismatches. Thus, the existence of administrative assignments, vacancies and truncated list is a textbook example of imperfect information.

B Fit

This section presents the fit for the complete set of moments used for the estimation.

Figure B.1: Male - Academic Tracks

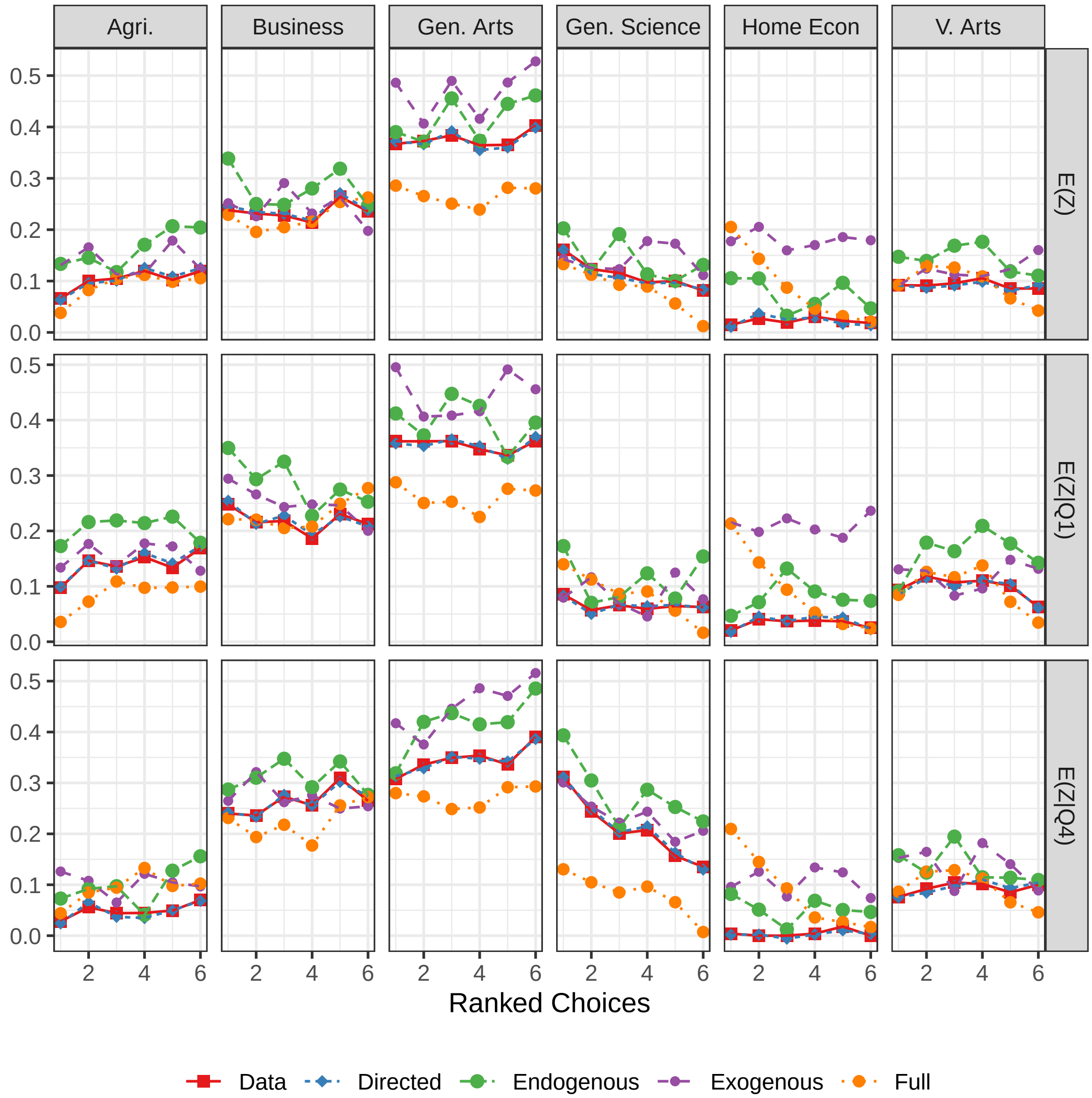


Figure B.2: Male - Characteristics

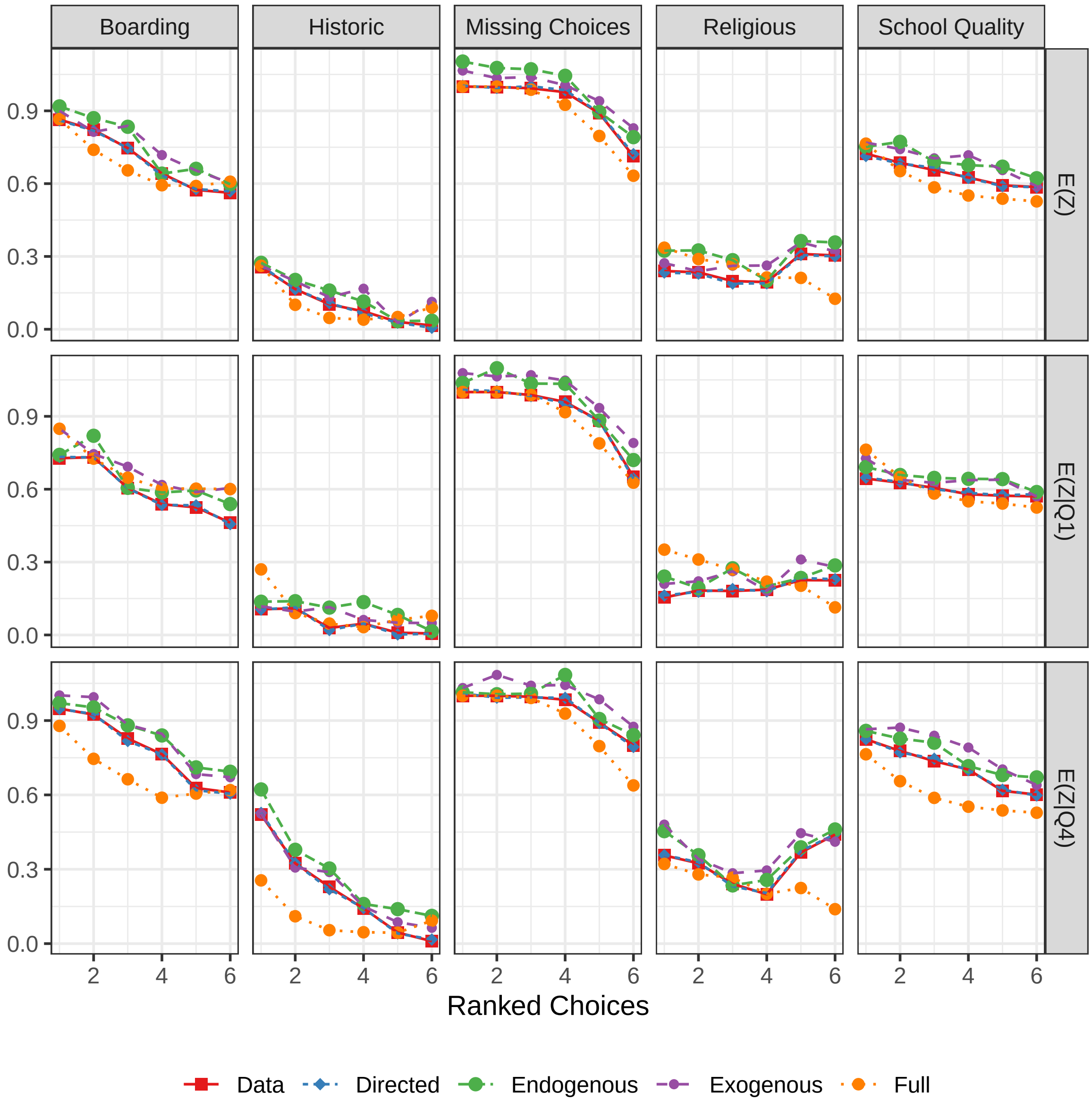


Figure B.3: Female - Academic Tracks

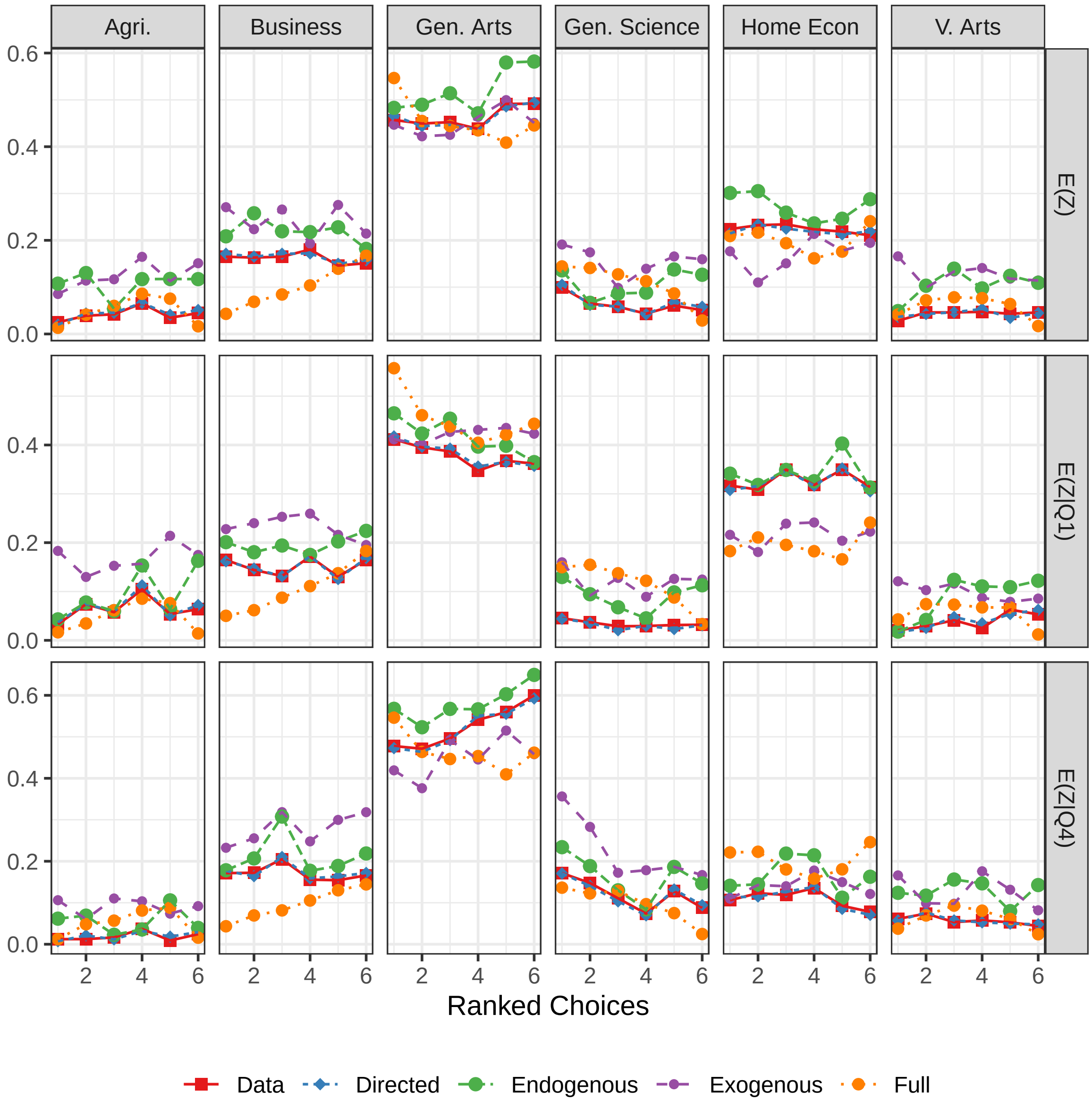


Figure B.4: Female - Characteristics