Essays on Matching and Market Design

by

Ming Jiang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in The University of Michigan
2015

Doctoral Committee:

Professor Yan Chen, Chair
Assistant Professor Yinghua He
Assistant Professor Erin Krupka
Associate Professor Stephen Leider

© Ming Jiang 2015

All Rights Reserved

For my family

ACKNOWLEDGEMENTS

First of all, I am deeply indebted to my advisor, Professor Yan Chen, who has been the greatest mentor in both my academic and personal life. She is a great advisor, an exemplary scholar, and a role model. Working, writing, and discussing ideas with her is a great joy.

I am also grateful to my committee members, Professor Erin Krupka, Professor Stephen Leider and Professor Yinghua He. I enjoyed working on my first ever academic paper with Erin and Steve, who really helped me get started as a researcher. I thank Yinghua for going beyond his duty to help me improve my papers with great patience.

I thank my coauthor, Professor Onur Kesten, who is among the smartest person I know. He also provides great support for me on the job market. I thank Professor Paul J. Healy from the economics department at the Ohio State University, for lending me use of the Ohio State economics lab when I needed it the most. I also thank my undergraduate advisor, Professor Xiangdong Qin for keeping providing me advice long after I graduated.

I thank all those who gave great comments on various chapters of this dissertation, including Yusufcan Masatlioglu, Yesim Orhun, Tanya Rosenblat, Neslihan Uler, Tayfun Sönmez, Doug Bernheim, Fuhito Kojima, Sera Linardi, Yusufcan Masatlioglu, Alvin Roth, Marie Claire Villeval, Alistair Wilson, and participants at the Economic Science Association conferences and various seminars. Financial support from Rackham Graduate School at the University of Michigan, and from National

Science Foundation is grateful acknowledged.

I would like to thank my colleagues, friends, and supporting staff at the University of Michigan School of Information. Finally, I would like to thank my wife, Jingchao Li. It is impossible for me to accomplish this without your support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Essays on Matching and Market Design

by

Ming Jiang

Chair: Yan Chen

Using a combination of experimental, theoretical, computational and empirical methods, my dissertation studies matching and market design with applications to education policy including school choice and college admissions. I tackle three problems: the effect of standardized tests on matching mechanisms (Chapter 2); experimental evidence of matching in a large market (Chapter 3); and quasi-experimental evidence of the theoretical properties of matching mechanisms (Chapter 4).

In Chapter 2, I investigate the matching of college admissions, where students' admission priorities and colleges' preferences over students are misaligned, due the imperfect measure of student aptitudes by standardized entrance tests. I show that in this case any matching mechanism that is stable with regard to priority is not stable with regard to preference. The resulting instability leads to market unraveling. However, a manipulable mechanism, combined with limited information about priorities, may succeed in mending this market failure. A laboratory experiment confirms this theoretical prediction.

In Chapter 3, we study the role market size plays in school choice. We evaluate the performance of the Boston and the Deferred Acceptance (DA) mechanism in laboratory with different market sizes. The results show that increasing the market size from 4 to 40 students per match increases participant truth-telling under the DA but decreases it under the Boston mechanism, leading to a decrease in efficiency but no change in the large stability advantage of the DA over the Boston mechanism. Furthermore, increasing the scale to 4,000 students per match has no effect on either individual behavior or mechanism performance. Our results indicate that "large market" in practice is smaller than in theory.

In Chapter 4, we evaluate the Immediate Acceptance (Boston) mechanism and the parallel mechanism in college admissions, both in the laboratory and with naturally-occurring data. Through both channels, we find that the more emphasis a mechanism put on the first choice, the more likely students rely on their rankings in the test to manipulate their reported preferences, which confirms the theoretical predictions. Although in the laboratory, the parallel mechanism proves to be more stable, we do not observe economically significant difference in stability in the field.

CHAPTER I

Introduction

Matching and market design is a growing field in economics. Matching problems involve pairing members of one group of agents with one or more members of another disjointed group of agents. Examples include college admissions, matching medical school graduates to hospitals, assigning students to public schools, assigning students to on-campus housing and overseas trips, facilitating pairwise kidney exchanges, and matching cadets to army branches. My dissertation focuses on the application of matching and market design theory to education policy, particularly the school choice and college admissions problems. I investigate three questions that have not been answered in the literature: What is the effect of standardized tests on matching stability? How does matching in a large market work in an experimental setting? What is the quasi-experimental support for the theoretical differences in matching mechanisms? The next three chapters address each of the three problems using different combinations of theoretical, experimental, computational, and empirical methods.

In Chapter 2, I study matching problems where priorities and preferences are misaligned. In the case of test-based college admissions, students are matched based on their test scores in standardized tests, which determine their priorities. However, the tests, due to measurement errors, are noisy realizations of students' aptitudes, which capture colleges' preferences. I investigate two questions: Why can the Deferred Acceptance mechanism produce unstable matchings and cause market unraveling? How can the Boston mechanism outperform the Deferred Acceptance mechanism in mending such market failure? These research questions are motivated by the real world college admissions matching in China.

To answer the question of *why*, I formally model the game of college admissions with standardized tests. I show that in centralized college admissions, a stable matching mechanism that respects test scores produces unstable matching outcomes in terms of student aptitude. This directly predicts market unraveling, as most cases of market unraveling are caused by perceived instability in matching outcomes.

I then design and conduct a laboratory experiment to answer the question of *how*. I combine the two mechanisms - the Boston mechanism and the Deferred Acceptance (DA) mechanism - with two timing conditions of the submission of students' rank-ordered lists of colleges. The environment is designed as a hybrid of one-sided and two-sided markets, where there is a centralized, one-sided matching stage, and preceding it is a decentralized, two-sided stage where colleges can send early admission offers to students. The experimental results confirm the theoretical predictions, and show that the Boston mechanism combined with pre-exam submission performs best in terms of reducing market unraveling, yet it achieves that with little sacrifice on the stability of eventual matching outcomes.

In Chapter 3, coauthored with Yan Chen, Onur Kesten, Stephane Robin, and Min Zhu, I complement the active and growing literature in matching theory on large market properties within the context of student assignment, using a combination of experiments and computational methods. There is a growing body of theoretical literature showing that as the size of the market becomes larger, many properties of matching mechanisms may change. In terms of policy implication, because laboratory experiments are used as wind tunnels to test mechanisms that are being used or will be used in the real world, and the market size in the real world is usually much larger than in the lab, it is necessary to see if findings in the lab can be generalized to the field.

Creating a large market in the laboratory is challenging because of obvious physical and financial constraints. To test the effect of market size on behavior and matching outcomes, we conduct two experiments, studying two mechanisms, Boston and DA. In one experiment, we compare two markets of size 4 and 40, by replicating the smaller market tenfold. To allow random rematching, each large session has 80 subjects making decisions simultaneously. In the other experiment, instead of pairing human subjects with each other, we pair each of them with either 39 or 3999 computer agents who behave as if they are human subjects. We do so by letting them play the strategies of human subjects in the first experiment. We find that when the size of the market goes from 4 to 40, more subjects report their true preferences under DA, which has truth-telling as the dominant strategy, and more subjects manipulate their preferences under Boston, which itself is a manipulable mechanism. DA also remains stable under large market, and Boston becomes more unstable. However further increasing the size of the market from 40 to 4000 has no significant effect. This result shows that what is considered "large market" in the theoretical literature, might actually be much smaller in application.

In Chapter 4, coauthored with Yan Chen and Onur Kesten, I study how different matching mechanisms in college admissions affect students' strategies and subsequent matching outcomes, using both a naturally occurring experiment and a laboratory experiment. The natural experiment comes from a policy change in China's Sichuan Province between the year 2008 and 2009. Since when students participate in college admissions is mostly determined by birth, while the mechanism change happens much later, it eliminates selection bias to a large extent. Moreover, the mechanism change only affects a portion of the students, therefore it allows us the use of difference-in-differences estimators. The dataset we obtain for a city in that province contains students' test scores, students' reported preferences, as well as the colleges that admitted them for the two years. We find that when the mechanism changes from the sequential mechanism to the parallel mechanism, (1) students list more colleges in their rank-ordered list; (2) students list better colleges as their first choices; (3) students are less likely to list local colleges as their first choices; and (4)fewer students get admitted to their reported first choices.

To further study the effect of mechanism change in a well controlled environment where we can induce students' true preferences, we design a laboratory experiment fashioned after the real world matching market and its mechanisms. Specifically, each of these mechanisms either emphasizes first choice, emphasizes first two choices, or is strategy-proof. The results from the laboratory corroborate most of the findings in the field. Further, we find that the more emphasis a mechanism puts on the first choice, the less likely students will report their true preferences. Both the results from the laboratory experiment and from the field confirm the theoretical predictions.

CHAPTER II

When Do Stable Matching Mechanisms Fail? The Role of Standardized Tests in College Admissions

2.1 Introduction

College admissions mechanisms affect the career choices and labor market outcomes of many young people around the world. They belong to a broader class of matching problems that involve pairing members of one group of agents with one or more members of another disjointed group of agents. Other examples include college admissions (Gale and Shapley, 1962; Roth, 1985; Balinski and Sönmez, 1999), matching medical school graduates to hospitals (Roth, 1984, 1986), assigning students to public schools (Abdulkadiroğlu and Sönmez, 2003; Abdulkadiroğlu et al., 2005a,b), assigning students to on-campus housing and overseas trips (Abdulkadiroğlu and Sönmez, 1999; Chen and Sönmez, 2002; Featherstone, 2013), facilitating pairwise kidney exchanges (Roth et al., 2005), and matching cadets to army branches (Sönmez and Switzer, 2013). *Matching mechanisms* are algorithms used to accomplish assignments in these cases.

For such matching problems, stable and strategy-proof matching mechanisms such as the Deferred Acceptance mechanism[1] (Gale and Shapley, 1962) are widely praised as a superior alternative to priority matching mechanisms, such as the Boston mechanism, which is manipulable and may thus lead to unstable outcomes (Abdulkadiroğlu et al., 2005b; Chen and Sönmez, 2006; Ergin and Sönmez, 2006). This is because if a matching mechanism is stable, it produces matching outcomes in which no pair of agents would both prefer to be matched to each other than to their current partners. If a mechanism is strategy-proof, then participants do not have incentives to lie about their true preferences. A significant issue with unstable matchings is market *unraveling* (Roth and Xing, 1994): the transaction date is pushed further and further ahead, often circumventing the centralized matching process. For example, in the case of matching new physicians to hospitals, when the job market unravels, hospitals sign employment contracts with medical school students sometimes years ahead of their actual graduation date. Market unraveling also manifests itself through early admissions in college admissions (Avery et al., 2009). Roth (1991) shows that if a mechanism frequently yields unstable matchings, even if such market is centralized, it still causes market unraveling, and is no better than the decentralized market it tries to replace. Such unraveling is believed to be inefficient (Fréchette et al., 2007), and reduces market mobility (Niederle and Roth, 2003). In addition to the cost of instability, Ergin and Sönmez (2006) and Chen and Sönmez (2006) also suggest that strategic manipulation in the Boston mechanism leads to efficiency loss both in theory and in the laboratory.

Stability, however, does not ensure that market unraveling would not happen. One instance from two-sided matching comes from Sönmez (1999), which states that stability does not prevent manipulation through pre-arranged match. Even for centralized college admissions, which is a one-sided matching market, market unraveling stills occurs under stable mechanisms. In 2008, the government of Shanghai switched

[1] Note however, the Deferred Acceptance mechanism is strategy-proof for one side of the matching market. It is not so for the other side of the market.

its college admissions mechanism from a variant of the Boston mechanism called the "sequential mechanism" to a variant of the Deferred Acceptance mechanism called the "parallel mechanism," which is more stable and less manipulable (Chen and Kesten, 2014)[2]. Yet, after the policy change, the top universities in Shanghai increased the number of their early admission offers from 893 to 1353[3], an almost 50% increase, which all happened months before the actual college entrance exam. Consequently, more than half of the students those universities admitted were through early admissions, circumventing the centralized matching process. Since a stable matching mechanism would predict less market unraveling, why did this happen? [4]

To explain why this happened, first we need to note that college admissions in China are organized as a centralized matching market, where a central authority administers an annual standardized test, and places students into colleges based on their test scores[5], with the aim of the test being to measure students' aptitudes. This standardized test-based college admissions problem differs from the school choice problem (Abdulkadiroğlu and Sönmez, 2003) in that although colleges, like public schools, follow priorities based on test scores, colleges themselves actually have strong preferences over students that may be different from priorities. It is understandable that colleges prefer better students with higher academic abilities (aptitudes) as better students are more likely to improve college reputations and may bring more future donations. It also differs from decentralized college admissions systems like those in the United States (Gale and Shapley, 1962; Roth, 1985), because colleges can not express their preferences over students during the centralized matching process,

[2]I will henceforth simply refer to them as the Boston mechanism and the Deferred Acceptance mechanism respectively.

[3]http://edu.qq.com/a/20081208/000066.htm. Retrieved on 8/25/2014.

[4]Although the change in mechanism is the only major policy change that year that I am aware of, and students are randomly selected into high school cohorts by birth, there may be other factors that can possibly influence colleges' early admission decisions. In this paper, I try to establish the connect between mechanisms and unraveling (in the form of early admissions) by controlling for other factors, both theoretically and experimentally.

[5]To my best knowledge of the author, the following countries also use centralized college admissions based on standardized test: South Korea, Japan, and Turkey.

since the matching is carried out by a computer program. Colleges' preferences and students' priorities are perfectly aligned only if standardized tests yield scores that do not distort the relative standings of students' aptitudes. In practice, however, this is unlikely, because even if a test is unbiased, that is, the *expected* test scores always respect the relative standings of aptitudes, all tests have *measurement errors*, defined as the lack of "consistency with which the [test] results place students in the same relative position if the test is given repeatedly" (Bloom et al., 1981). Therefore, if a test is taken only once, such as in centralized college admissions, the ranking of test scores may not reflect the ranking of aptitudes.

To see why measurement errors in tests can cause unstable matchings, let us look at an example. Suppose Students A and B are two top students in a college admission market with many students, and Colleges 1 and 2 are the two top colleges, and each has one available seat. Student A is better than B, and College 1 is better than 2. All students prefer better colleges, and colleges better students. Although Student A is better than Student B, she still has 30% of chance to score lower in the test than B, caused by measurement errors in the test. Under the Deferred Acceptance mechanism, since revealing one's true preference is the dominant strategy, both students list College 1 as their first choices, and College 2 second choices. Since priorities are given based on the test scores, therefore, there is a 30% chance that Student A goes to College 2 and Student B goes to College 1 [6]. In this case, both Student A and College 1 prefer each other to what they are currently matched with. So College 1 has an incentive to circumvent the centralized matching, and admits Student A early; and Student A certainly has the incentive to accept such offer. What was it about the mechanism that was used prior to the policy change in Shanghai that it did not have as serious an issue of market unraveling?

The answer to the question above may lie in one policy detail: instead of submit-

[6] This, and the example in the next paragraph assume that no other student can score higher than both Students A and B under any circumstance.

ting their rank-ordered lists of colleges after knowing their test scores and rankings, students report their preferences before taking the exam, therefore the only thing they know is the relative standings of their academic abilities. Let us go back to the previous example. Now when Student B submits his list of colleges without knowing his realized test score, he knows that he has a 70% chance to score lower than A. Since under the Boston mechanism, emphasis is put on the first choice, and the cost of not getting into one's first choice is very high, Student B may not want to take the risk and rank College 1 first, and therefore he will rank College 2 as his first choice. In this case, Student A can go to College 1, and Student B to College 2, with certainty. The matching is therefore stable for the top two students and colleges. So in short, the manipulability of this mechanism, combined with limited information of the priorities (test scores) of the matching, actually helps better students go to better colleges, regardless of their test performances.

Previous literature has also shown, in different aspects, that the Boston mechanism could outperform the Deferred Acceptance mechanism in some aspects under some conditions. In terms of market efficiency, the Boston mechanism is shown to be ex ante more efficient under incomplete information than the Deferred Acceptance mechanism both theoretically (Abdulkadiroğlu et al., 2011) and experimentally (Featherstone and Niederle, 2014). In terms of strategy, contrary to the Deferred Acceptance mechanism, the Boston mechanism is immune to manipulation by schools through misreporting enrollment capacity (Kesten, 2011). In addition to the Boston mechanism, other manipulable mechanisms can also take the advantage of their manipulability and perform better than truth mechanisms. For example, manipulable mechanisms that encourage reporting indifferences can outperform strategy-proof Random Serial Dictatorship mechanism in terms of efficiency (Fragiadakis and Troyan, 2013).

In addition to the properties of mechanisms, the context of this study - centralized college admissions in China - has also been explored in the literature. Chiu and

Weng (2009) show that popular schools will use early commitment to "lock-in" good students. Zhong et al. (2004), and Lien et al. (2012) show that the Boston mechanism combined with preference submission before knowing the test scores has an efficiency advantage in centralized admissions. Wu and Zhong (2014) provide empirical support by using data from a top university in China and find that students admitted through the above mechanism have better academic performance in college.

In this paper, motivated by a real world anomaly, I answer two questions: why the Deferred Acceptance mechanism can produce unstable matchings and cause market unraveling; and how the Boston mechanism can outperform the Deferred Acceptance mechanism in mending such market failure. To answer the question of *why*, I formally model the game of college admissions with standardized tests, and show that without the possibility of early admissions (a form of market unraveling), a stable matching mechanism that respects test scores produces unstable matching outcomes in terms of student aptitude. This directly predicts market unraveling, as most cases of market unraveling are caused by perceived instability in matching outcomes. I then design and conduct a laboratory experiment to answer the question of *how*. I combine the two mechanisms (the Boston mechanism and the Deferred Acceptance mechanism) with two timing conditions of the submission of students' rank-ordered lists of colleges. The environment is designed as a semi-two-sided market, where there is a centralized, one-sided matching stage, and preceding it is a decentralized, two-sided stage where colleges can send early admission offers to students. The experimental results confirm the theoretical predictions, and show that the Boston mechanism combined with pre-exam submission performs best in terms of reducing market unraveling, yet it achieves that with little sacrifice on the stability of eventual matching outcomes.

The remainder of this paper is organized in the following way: Section 2.2 formally models the game of college admissions with standardized test as well as introducing the two mechanisms. Section 2.3 presents theoretical results. Section 2.4 describe the design of the experiment. Section 2.5 presents experimental results. Section 4.8 concludes.

2.2 The Environment

2.2.1 The Model

I outline a matching economy with a continuum of students to be assigned to a finite number of colleges.

The set of students is denoted by S, with any student s whose type is represented by his aptitude $a \in \mathbb{R}$. The total mass of students is normalized to 1, and the distribution of their aptitudes follows a density function $f(\cdot)$. Each student also has a test score t after taking the entrance test. A *standardized test* determines the relationship between the test score and the aptitude: $t = a + \eta$, where $\eta \in \mathbb{R}$ has a distribution with density function $g(\cdot)$, and a and η are independent. It captures the measurement error of the test, and is individually determined for each student. I assume that η has mean 0. That is, a test is unbiased.

The set of colleges is denoted by $C = \{c_1, ..., c_m\}, m \geq 2$. Each college c_i has capacity to enroll a mass of q_{c_i} students. Without loss of generality, I assume that the mass of total quota equals to the mass of students.

A matching μ is an allocation of college slots to students such that the mass of students assigned to any college does not exceed its quota. A matching mechanism accomplishes this matching using students' reported rank-ordered lists (ROL) of colleges, and students' *priorities*. In college admissions, students' priorities at every college during the admissions are determined by their test scores: s has a higher priority than s' if and only if $t > t'$. I also assume that all colleges have the same preferences over students, which are determined by students' aptitudes.

A student also has ranking in aptitude, r_a and ranking in test score, r_t, measured by the mass of students with higher aptitudes and test scores, respectively.

I also make the following assumptions: the distributions of aptitude and measurement error of the test (captured by $f(\cdot)$ and $g(\cdot)$) are common knowledge to all students; every student knows his ranking of aptitude and test score after the exam; and all students have the same preferences over colleges. The assumption that students know their is justified by the fact that after many testings done in high schools, students have good ideas about where they stand within each school, and they are able to extrapolate that knowledge to other schools based on the relative qualities of different schools. The last assumption is based on the nature of college admissions: good colleges are sought after by almost all students and students' preferences are highly correlated.

Next, since both test score-based priorities and aptitude-based preferences coexist, I define two sets of stability criteria.

A matching μ is *stable with respect to test score* if there is no student-college pair (c, s) such that student s prefers college c to the college he is assigned to, and college c has students assigned to it who are ranked lower in test scores than student s.

A matching μ is *stable with respect to aptitude* if there is no student-college pair (c, s) such that student s prefers college c to the college he is assigned to, and college c has students assigned to it who have lower aptitudes than student s.

2.2.2 The Mechanisms

In this section I describe the algorithm of the two mechanisms I study in this paper.

The first mechanism I study is the Boston mechanism (Boston). The procedure of this mechanism is listed below.

There is a priority ordering of students. In the case of centralized college admissions, it is determined by the ranks of students' test scores.

Step 1: The first choices of the students are considered. For each college, consider the students who have listed it as their first choice and assign seats of the college to these students one at a time following their priority order until either there is no seat left or there is no student left who has listed it as her first choice.

Step k ($k > 1$): For the students who have been rejected after step k-1, only the kth choices of them are considered. For each college with available seats, consider those students who have listed it as their kth choice and assign the remaining seats to these students one at a time following their priority order until either there is no seat left or there is no student left who has listed it as her kth choice.

The algorithm terminates when there is no rejected student. The Boston mechanism is manipulable (Abdulkadiroğlu et al., 2005b; Ergin and Sönmez, 2006), and its manipulability has also been observed experimentally (Chen and Sönmez, 2006) and empirically (He, 2014); particularly, it puts heavy emphasize on how students list their first choices, since admission at every step is final. Such manipulability therefore often yields matching outcomes that are unstable (Roth, 1991; Chen and Sönmez, 2006).

The second mechanism studied in this paper is the Deferred Acceptance mechanism (DA) (Gale and Shapley, 1962). This mechanism played a key role in school choice reforms in Boston and New York City (Abdulkadiroğlu et al., 2005b,a). The procedure of the DA mechanism is listed as the following algorithm.

For each college, a priority ordering of students is determined by the ranks of students' test scores.

Step 1: The first choices of the students are considered. For each college, consider the students who have listed it as their first choice and temporarily assign seats of the college to these students one at a time following their priority order until either there is no seat left or there is no student left who has listed it as her first choice.

Step k ($k > 1$): Each student who was rejected applies to the next college on his list. Each college then considers the students it has already temporarily accepted along with the new applicants, accept the ones with the highest priorities within their enrollment capacities among those students, and then rejects the rest.

The algorithm terminates when there is no rejected student. The DA mechanism is strategy-proof (Dubins and Freedman, 1981), and it also produces stable matchings that are most favorable to students (Roth, 1982).

2.3 Theoretical Results

In this section, I present the theoretical predictions for the two mechanisms under the test-based college admissions environment. Since the stability and incentive compatibility of the Deferred Acceptance mechanism with deterministic priorities are well established (Dubins and Freedman, 1981; Roth, 1982), I will start with the incentives for truth-telling under DA when students submit their rank-ordered list of colleges (ROL) before they take the exam. In this case, they only know their rankings in aptitude and the distribution of measurement error. It turns out that the strategy-proof property of DA still holds.

Proposition 1 (Roth, 1989)**.** *It is a dominant strategy to reveal one's true preference under DA with pre-exam ROL submission.*

The next proposition is the extension of the standard stability result for the DA mechanism.

Proposition 2. *The DA mechanism with pre-exam ROL submission yields stable matchings with respect to test score under the Bayesian Nash equilibrium in dominant strategies.*

Proof. First, it is straightforward to see that the set of stable matching (with respect to test score) is a singleton, because all students have identical priorities at every college, which are determined only by test scores. Next, by Proposition 1, we see that truth-telling is a dominant strategy under DA pre-exam, therefore truth-telling is the Bayesian Nash equilibrium in dominant strategies. Moreover, it is unique, because every other strategy is dominated by truth-telling. Finally, DA yields stable outcome (w.r.t. test score) when everyone is truth-telling. □

Different from DA, the Boston mechanism encourages strategic manipulation. Specifically, students manipulate their first choices based on their rankings in test scores. I define the following kind of strategic manipulation under the test-based college admissions, which has been observed in the laboratory as well as in the field (Chen et al.).

Definition 1 (Rank bias)**.** *A student s with rank r_t exhibits* rank bias*, if he ranks the least commonly preferred college as his first choice among all the colleges that satisfy that the total quotas of all more commonly preferred colleges does not exceed his ranking in test score (r_t). That is, a student lists college c_i as his first choice if $\sum_{k=1}^{i-1} q_{c_k} < r_t$, $\sum_{k=1}^{i} q_{c_k} \geq r_t$.*

In other words, it means a student will rank a college whose rank in quality, quota considered, corresponds to the student's rank in test score. The definition of rank bias is closely related to the concept of district bias in the school choice literature (Chen and Sönmez, 2006). The following result shows that it is a Nash Equilibrium strategy under Boston with post-exam ROL submission.

Proposition 3. *Rank bias is a Nash equilibrium strategy under both the Boston and the DA mechanisms with post-exam ROL submission.*

Proof. See Appendix 2.7.1 □

From the above results, we can see that the relative standings of aptitude are totally ignored during the matching. For DA, it is ignored because it is always best to reveal one's true preference, and assignment process depends on priorities that are entirely decided by test scores. It is also ignored under Boston, because as long as students know their test scores, the rankings of test scores are the only thing they need in order to strategize their first choices.

The following theorem then illustrates the consequences of such matching outcomes, when there are a continuum of students whose aptitudes are distributed following $f(\cdot)$, students and colleges have homogeneous preferences over each other, and test scores are noisy realizations of students' aptitudes with measurement error following $g(\cdot)$. It show the conflicting nature of the two stability definitions.

Theorem 1. *In the continuum economy of the test-based college admissions, with a continuum of students whose aptitudes are distributed following the density function* $f(\cdot)$*, any matching outcome that is stable with respect to test score is not stable with respect to aptitude with probability 1. Specifically, the proportion of students each college prefers to be matched with rather than what it is matched with after the centralized admissions is non-zero.*

Proof. See Appendix 2.7.2. □

Theorem 1 provides a non-parametric look at the potential unstable outcomes of the matching. Intuitively, how large the measurement error of a test is affects how unstable the matching outcomes is, with respect to aptitude. To give a concrete example, I conduct a simulation with a set of parameters. In this simulated example, there are 1000 students and five colleges, each with 200 seats. Students' aptitudes are uniformly distributed from 50 to 100, and the measurement error follows a normal distribution with mean 0. All students report their true preferences and are matched using the Deferred Acceptance mechanism. Figure 2.1 reports the number of students with whom colleges would form blocking pairs, after varying the standard deviation of the distribution of measurement error (from 0 to 10). From this graph we can see that the proportion of blocking pairs (out of each college's capacity) increases as the standard deviation of measure error increases.

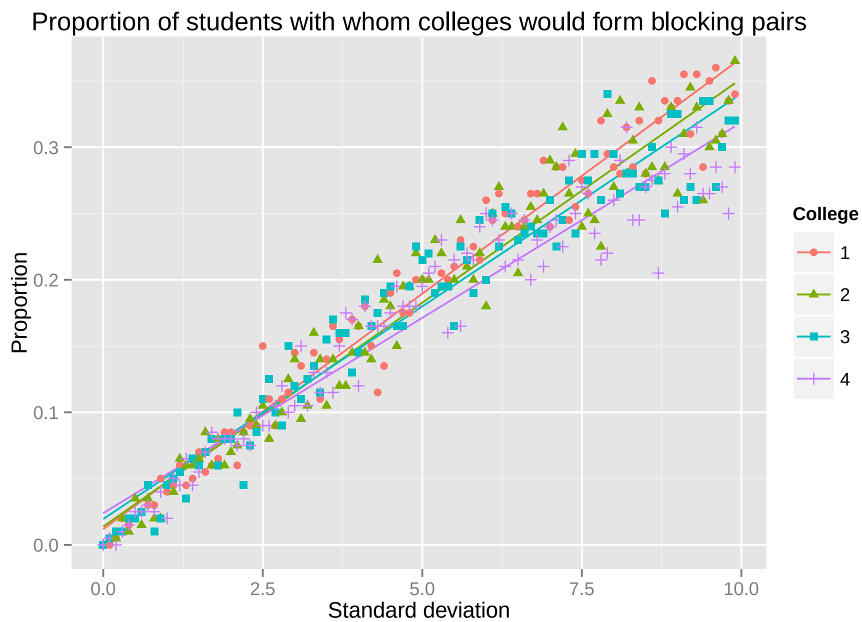


Figure 2.1: Number of blocking pairs increases with measurement error

The next graph (Figure 2.2) goes into what actually happened behind each dot in the previous graph by showing the distribution of aptitudes of students admitted by each college in a single simulation. The simulated matching has a standardized test with standard deviation 8. The distribution curves from left to right are for Colleges 1, 2, 3, 4, and 5, in order. Vertical lines from left to right are 20%, 40%, 60%, and 80% quantiles for students' aptitudes. In a matching that is stable with respect to aptitude under the simulated environment, the top 20% of students should be matched with College 1, the next 20% with College 2, etc. Therefore, for College 1, the area to the right of the 80% quantile line that is a part of the distribution of aptitudes of students admitted by colleges of lower quality is the proportion of students College 1 forms blocking pairs with. The same applies to Colleges 2, 3, and

4. Figure 2.2 illustrates the pattern of unstable mismatch with respect to aptitude through overlapped areas under the aptitude distributions, which are in term caused by the measurement error of the standardized test.

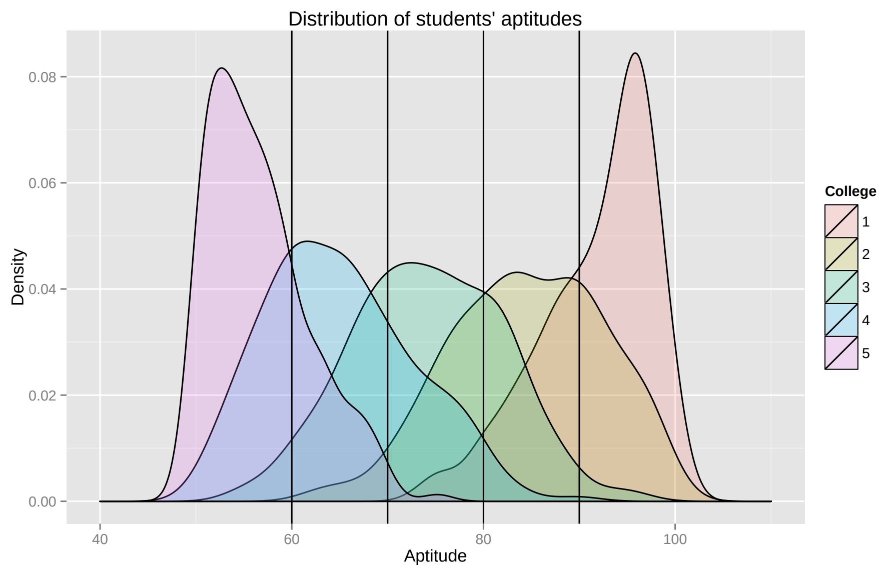


Figure 2.2: Number of blocking pairs increases with measurement error

Since DA under truthful preference revelation, and Boston with post-exam submission under the Nash Equilibrium strategy both yield the same matching outcomes that are stable with respect to test scores, we now have the following corollary.

Corollary 1. *In the test-based college admissions, DA under post-ROL condition, and Boston under post-exam condition yield Nash equilibrium matching outcomes that are not stable with respect to aptitude with probability 1. Moreover, DA under pre-exam condition has a Bayesian Nash equilibrium in dominant strategies that yields an outcome unstable with respect to aptitude with probability 1.*

While the instability with respect to aptitude under the DA mechanism is due to the fact that DA is strategy-proof, and allocation is entirely based on priorities, the reason for instability under the Boston mechanism is different. Under the Boston mechanism, students manipulate based on their priorities. When test scores are revealed after the exam, and ranks are known, aptitude plays no role in decision making. However, when scores are not known, and aptitude and the rank of it are the only clues students can follow to make their decisions, the fact that the Boston mechanism is manipulable may make students reveal more information about their true aptitudes through submitted ROL. To show this, first I extend the definition of rank bias to aptitude.

Definition 2 (Aptitude-based rank bias)**.** *A student s with rank r_a exhibits* aptitude-based rank bias*, if he ranks the least commonly preferred college as his first choice among all the colleges that satisfy that the total quotas of all more commonly preferred colleges does not exceed his ranking in aptitude (r_a). That is, a student lists college c_i as his first choice if $\sum_{k=1}^{i-1} q_{c_k} < r_a$, $\sum_{k=1}^{i} q_{c_k} \geq r_t$.*

Proposition 4. *When students play the strategy of aptitude-based rank bias, the Boston mechanism with pre-exam ROL is not stable with respect to test scores, but is stable with respect to aptitude.*

Proof. Similar to students playing score-based rank bias under Boston with post-exam ROL, playing aptitude-based rank bias yields an matching outcome that is assortative in aptitude, which also means that it is stable with respect to aptitude. The instability result follows from Theorem 1, where we replace test scores with aptitudes, and vice versa. □

The above proposition is only true under the best case scenario, however. The actual strategies students may employ under this mechanism-timing combination depend on their cardinal utilities as well as their risk preferences, therefore generalizing the equilibrium strategies is challenging. In the following section, the matching market in the experiment design serves as an example, showing that in some cases, the Boston mechanism with pre-exam submission indeed outperforms all other combinations in terms of stability w.r.t. aptitude.

One of the most important consequences of matchings that are not stable w.r.t. aptitude (i.e. preference) is market unraveling. This phenomenon has been observed in many real world matching markets where matching mechanisms produce unstable outcomes. Examples include matching of couples in the National Resident Matching Program (Roth, 1984), sorority rush at American universities (Mongell and Roth, 1991), and new physicians matching in Newcastle, Edinburgh, Birmingham, UK (Roth, 1991). In the next section, agents of both sides are allowed to circumvent the centralized matching, which captures the unraveling of the market.

2.4 Experimental Design

The previous section outlines the theoretical stability results under different combinations of mechanisms and submission timings. The equilibrium under the Boston mechanism with pre-exam submission, however, is difficult to derive for a general case. In this section, I design a specific, discrete market, where equilibrium strategies can be reasonably derived for the Boston mechanism, and test the associated theoretical predictions using a laboratory experiment.

I implement a 2(timing) by 2(mechanism) factorial design to investigate the performance of the two mechanisms under two different timing conditions regarding when ROLs are submitted (pre-exam and post-exam). The four treatments are henceforth referred to as *Boston-pre*, *Boston-post*, *DA-pre*, and *DA-post*.

2.4.1 The Environment

The game environment consists of three students, $\{1, 2, 3\}$, and three colleges, $\{A, B, C\}$. Each college has exactly one slot, and each student can occupy at most one slot. Three students are ranked by their aptitudes, $1 \succ_c 2 \succ_c 3$, and three colleges their qualities, $A \succ_s B \succ_s C$. Students' and colleges' preferences over each other are described by utilities: $\{30, 20, 10\}$, which is the experimental points they earn given the students or colleges they are matched with.

In the theoretical derivation, I do not explicitly model the early admissions stage. In the experiment, colleges are allowed to make early offers, which captures the consequences of market instability. Therefore, the game in this experiment consists of four stages: Early Admission Stage, Exam Stage, ROL-submission Stage, and Centralized Matching Stage. Under the pre-exam timing condition, the game follows the following order: EarlyAdmission-ROL-Exam-Matching, while under the post-exam condition, the game follows EarlyAdmission-Exam-ROL-Matching.

Before students take the exam and submit their ROLs, colleges are given the opportunities to send binding early admission offers to students by paying a fee of 3 utility points. The fee represents the cost associated with the consequences of market unraveling [7]. The choice of 3 utility points (15% of the average utility) is based on the consideration that some colleges have the incentive to pay the fee to participate in early admissions under some mechanisms, while under others it is too costly. After paying the fee, each college can send exactly one offer to any student. Students who receive the offers can choose to accept or decline the offers. Students and colleges that have reached the early admission agreements are removed from subsequent centralized matching market. The design of the early admissions stage is similar to Kagel and Roth (2000).

During the Exam Stage, the computer simulates the exam by giving each student a test score. The measurement error of the test is characterized by the probability parameter p as well as the spread from the mean. Each student can perform normally with probability p, or either overperform or underperform each with probability $\frac{1-p}{2}$. Students know their types (aptitudes) and distribution of test outcomes of all other students. The detailed testing outcome for the students are shown in Table 2.1. In

[7] This can be thought of as the cost associated with market unraveling, for example, the cost of screening students individually; the cost of students not studying for the rest of their high school since they know they can go to colleges for sure; etc.

this experiment I choose $p = 0.5$.

Table 2.1: Exam outcomes for each student, $p = 0.5$

Student	1	2	3	Prob.
Underperform	12	7	2	0.25
Normal	20	15	10	0.5
Overperform	28	23	18	0.25

Either after or before taking the exam depending on the timing condition, students who have not been admitted early by the colleges submit their ROLs of all three colleges, regardless whether the slot in some colleges are taken or not. If a college has reached an agreement and admitted a student, that college is removed from each student's ROL and does not occupy a position in students' ROLs. For example, Student 1 lists $A-B-C$ and College A has already admitted Student 2, then College A will simply be removed before the algorithm starts and College B becomes Student 1's first choice in the subsequent matching. This design is chosen to let college players think about, through repeated play, what would have happened if they choose to or not to participate in early admissions.

One concern we may have, with respect to still allowing students to rank all three colleges even though one or more colleges have admitted students early, is that students' strategies may alter colleges' early admissions decisions. However, I will now show that as long as a student believes that a college, who should participate in early admissions, does not do so with any positive probability, his best response is to play the strategy as if no college participates in early admissions.

Under the post-exam condition, Student 3 is the only one who will make decisions after early admissions. Suppose he believes College A will not participate with probability $p > 0$ and College B with probability $q > 0$. Then he believes he plays the matching game with three colleges with probability pq, with Colleges B and C with probability $(1-p)q$, and with Colleges A and C with probability $(1-q)p$. Under DA, there are multiple best response strategies for the second and third scenario toward which Student 3 is indifferent. However, the only common best response for all three scenario is the strategy of truth-telling, which is the strategy he will play. The same goes for the Boston mechanism, where the only common best response strategy is to play rank bias.

Under the pre-exam condition, Students 1 and 3 will both list three colleges. Suppose they believe College B will not participate with probability $p > 0$. Then they believe they play the matching game with three colleges with probability p, and with Colleges A and C with probability $1 - p$. It is easy to see that truth-telling is always the best response strategy for Student 1 under both mechanisms. For Student 3, listing $B - A - C$ is the only common best response for both scenarios. Therefore, both Students 1 and 3 will play the strategies under the game where there are colleges.

To predict what will happen in this game environment, first we only look at the equilibrium strategies when there is centralized matching only, without the possibility of early admissions. Since under the DA mechanism under both timing conditions, truth-telling is the dominant strategy, and under Boston-post, rank-bias is the equilibrium strategy once the test scores are revealed[8], for all the three mechanisms, matching outcomes are the same for any given realization of test scores. Table 2.2 summarizes the probability of ranking realizations of test scores and corresponding equilibrium outcomes. Most of the outcomes have blocking pairs and are unstable with respect to aptitude.

[8]Note that the last-ranked student is indifferent among all strategies.

Table 2.2: Possible ranking realizations after the exam

Realized rankings	DA and Post-exam Boston NE	Probability	# Blocking pairs
1-2-3	1-2-3	0.406	0
2-1-3	2-1-3	0.266	1
1-3-2	1-3-2	0.266	1
3-1-2	3-1-2	0.016	2
3-2-1	3-2-1	0.031	3
2-3-1	2-3-1	0.016	2

Given the distribution of matching outcomes, College A will have lower risk-neutral expected utility (26.25) compared to the utility from its match under stable matching (30). It will have sufficient incentive to pay the 3-points fee and send an early admission offer to Student 1, and Student 1 will accept because accepting first-order stochastically dominates not doing so. Conditional on the strategies of College A and Student 1, College B will now have its own incentive to send an early admission offer to Student 2, and Student 2 will accept as well. Only College C will be left now and it admits Student 3. The outcome is now stable, but at a cost of 6 points out of 60. This is summarized in Table 2.3 and 2.4.

Table 2.3: Colleges' incentives to make their move before the centralized matching begins

College	Exp. Payoff post-exam	Payoff Stable Matching w.r.t. Aptitude	Incentive
A	26.25	30	3.75
B	20	20	0
C	13.75	10	0
Total	60	60	

Table 2.4: Colleges' incentives to make their move before the centralized matching begins, conditional on the equilibrium strategy of College A

College	Exp. Payoff post-exam	Payoff Stable Matching w.r.t. Aptitude	Incentive
B	16.875	20	3.125
C	13.125	10	0
Total	30	30	

Boston-pre however is a different story. Since students do not know the eventual realization of their test outcomes, they use strategies that maximize their expected utility. Again we only look at the equilibrium strategies for the centralized matching first. Under this game environment, students will play the following Bayesian Nash equilibrium strategies: Student 1 submits: $A - B - C$; Student B: $B - A - C$; and Student C: $B - A - C$. Note that Student 1 and 2 exhibit aptitude-based rank-bias strategies [9] This Bayesian Nash equilibrium is unique and robust to risk attitude and parameter choices, as simulation using different combinations of parameters shows that equilibrium strategies do not change for $0.4 \leq p \leq 0.6$ and for risk-seeking and very risk-averse players [10].

Table 2.5: Colleges' incentives to make their move before the centralized matching begins

College	Exp. Payoff pre-exam	Payoff Stable Matching w.r.t. Aptitude	Incentive
A	30	30	0
B	16.875	20	3.125
C	13.125	10	0
Total	60	60	

Table 2.6: Colleges' incentives to make their move before the centralized matching begins, conditional on the equilibrium strategy of College B

College	Exp. Util. post-exam	Payoff Stable Matching w.r.t. Aptitude	Incentive
A	28.75	30	1.25
C	11.25	10	0
Total	30	40	

Given such strategies, College B will now have a strong incentive to send an early admission offer to Student 2, because Student 2 will accept it and Student 1 will not. Conditional on College B's play, College A will not have a strong enough incentive to

[9] Aptitude-based rank-bias happens to coincide with truth-telling for Student 1.

[10] This BNE does not change for risk parameter $0.3 \leq \gamma \leq 1.1$ with functional form $U(w) = \frac{w^\gamma}{\gamma}$

send an offer to Student 1. Consequently, Student 1 goes to College A with probability 93.75%, and Student 3 6.25%. This is summarized in Table 2.5 and 2.6.

2.4.2 Experiment Procedure

Each session consists of 15 participants in 3 groups with 5 students in each group. Three participants of them play the role of the students, and two participants play the role of either College 1 or College 2. Since in equilibrium, College 3 never sends any offer, to reduce strategic uncertainty and control the beliefs of other players, it is played by the computer using the equilibrium strategy (”doing nothing”), which is common knowledge explained to the subjects in the instruction.

The experiment repeats for 20 rounds to facilitate learning. Each participant is assigned the role of either a college or a student at the beginning. Each round they are randomly rematched with different group members. Which college or student a participant plays is randomly decided each round. That is, a participant may play as different colleges or students, but he or she also plays as either a college or a student throughout the experiment.

During the early admission stage, I use the strategy method to elicit students' entire strategy profiles. Specifically, before students receive offers from colleges, they will be asked the following questions: ”If you receive offers from both Colleges 1 and 2, what would you do? 1. Accept College 1's offer; 2. Accept College 2's offer; 3. Do not accept any offer.”, ”If you receive offer only from College 1, what would you do? 1. Accept the offer; 2. Do not accept the offer.”, and ”If you receive offer only from College 2, what would you do? 1. Accept the offer; 2. Do not accept the offer.”.

At the end of the experiment, I elicit subjects' risk attitude using the lottery game from Holt and Laury (2002). The features of the experiment is summarized in Table 2.7.

The experiment was conducted in summer 2014 at the Experimental Economics

Table 2.7: Design features of the experiment

	Pre-exam
Boston	Students and colleges submit early admissions decisions Students submit ROLs Test scores are revealed Matches are calculated using the Boston mechanism Risk preference elicitation at the end
DA	Students and colleges submit early admissions decisions Students submit ROLs Test scores are revealed Matches are calculated using the DA mechanism Risk preference elicitation at the end
	Post-exam
Boston	Students and colleges submit early admissions decisions Test scores are revealed Students submit ROLs Matches are calculated using the Boston mechanism Risk preference elicitation at the end
DA	Students and colleges submit early admissions decisions Test scores are revealed Students submit ROLs Matches are calculated using the DA mechanism Risk preference elicitation at the end

Laboratory at the Ohio State University and the Behavioral and Experimental Economics Laboratory at the University of Michigan School of Information. There are 180 subjects across 12 sessions, with 3 sessions for each treatment. No one subject participated in more than one session. The average payment is $19.1, including a $5 show-up fee. Each session lasts about 90 minutes.

2.5 Experimental Results

Since the environment is two-sided, I report the behavior of colleges and students separately. In this section, I first report the decisions by colleges. Then I report the decisions by students. Finally, I report the matching outcomes, including matching stability with respect to both test scores and aptitude. Throughout this section, the general null hypothesis is that there is no difference in behaviors or matching outcomes.

2.5.1 Decisions by Colleges and Efficiency of the Market

First, I examine the proportion of colleges who participate in the early admission process. Recall that as long as a college sends an offer to a student, the college pays a fixed amount of fee, regardless of whether that student accepts the offer or not eventually. Since without early admissions, the sum of payoff is the same for all matching outcomes, the proportion of colleges who send early admission offers therefore serves as the proxy for the efficiency of the market.

Figure 2.3 shows the overall proportion of colleges who participate in early admissions as well as the proportions for College A and B separately. Theorem 1 and equilibrium for the game environment lead to the first hypothesis

Hypothesis 1 (Participation in early admissions)**.** There are significantly fewer colleges who participate in early admissions under the Boston mechanism with pre-exam

ROL submission than under the Boston mechanism with post-exam submission or the Deferred Acceptance mechanism with either of the two timing conditions.

Result 1 (Participation in early admissions)**.** *The Boston mechanism with pre-exam submission performs best in reducing participation in early admissions. Specifically, the four treatments have the following order in occurrences of market unraveling (early admissions): Boston-pre¡DA-post¡DA-pre¡Boston-post.*

Support. Two-sided proportion test shows that Boston-pre has a significantly lower participation rate than the second best, DA-post (46.7% vs 58.1%, $p = 0.002$). Similarly, DA-post has a lower rate than DA-pre (58.1% vs 66.7%, $p = 0.017$), and DA-pre has a lower rate than Boston-post (66.7% vs 76.9%, $p = 0.002$). Additional probit regressions with clustered standard errors confirm this finding, shown in Table 2.15 in the appendix.

By Result 1, we reject the null in favor of Hypothesis 1. This finding confirms the theoretical prediction that the Boston mechanism, combined with pre-exam ROL submission, performs best in reducing market unraveling, while the Deferred Acceptance mechanism can relatively increase the level of such unraveling.

I further break down the differences by colleges, and find that the difference between Boston-pre and other three mechanisms becomes even larger for the top college, College A (46.1% vs 70.6% (second best), $p < 0.001$), and diminishes for the mid-tier college, College B (47.2% vs 40.6%, $p = 0.203$). This is consistent with the equilibrium prediction of the experimental environment, where it predicts that College B under both mechanisms and both timing conditions would have incentives to participate in early admissions, while only under Boston-pre it predicts that College A would not have enough incentive to do so.

In addition, Table 2.8 shows which students colleges send early admission offers to. Not surprisingly, the majority of College A's early admission offers are sent to

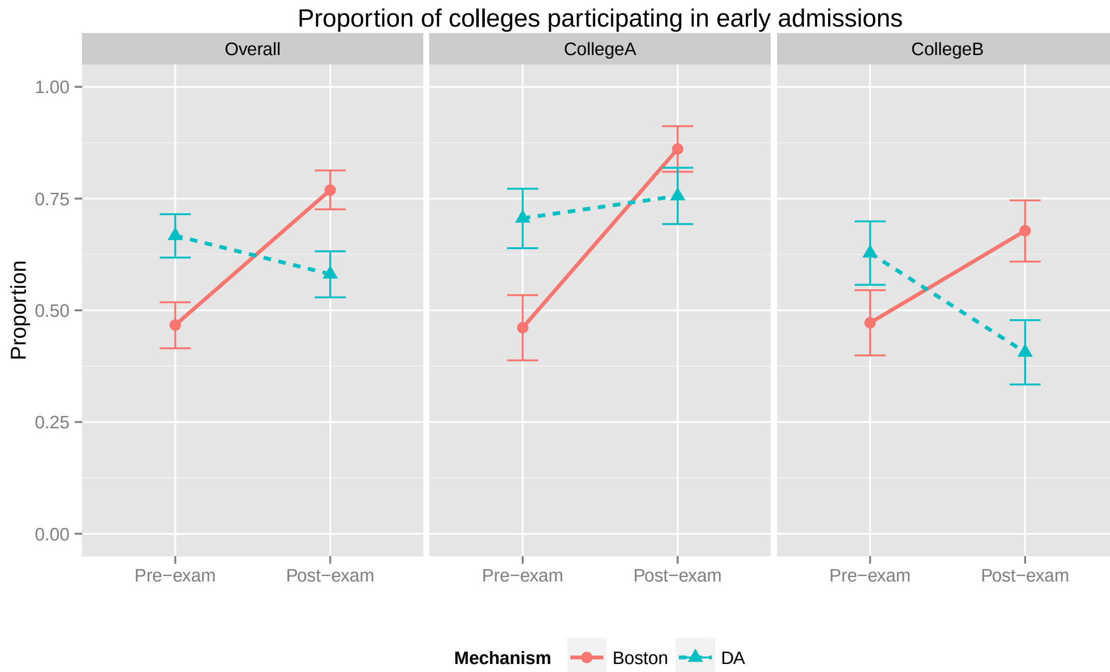


Figure 2.3: Proportion of colleges who send out early admission offers (error bars show 95% confidence intervals)

Student 1, and College Bs send more offers to Student 2 than Student 1. However, contrary to the equilibrium prediction which suggests that College A's will not send out early admission offers, in the experiment I find that there are about 43.9% of College A's nevertheless send out offers to Student 1. These decisions can in part be explained by players' risk attitude, since even under Boston-pre, College A still has 6.25% chance of admitting the worst student. Therefore, sending out offers early ensures College A of a match with a better student (given the equilibrium strategy of Student 1). Regression confirms this conjecture (see Table 2.14 in the appendix): the more risk averse a College A player is, the more likely he or she will send out an early admission offer[11] (probit regression, marginal effect 11.4%, $p = 0.023$, standard errors clustered at the session level). Figure 2.4 looks at the relationship between risk attitude, measured by the switching points in the lottery game, and the likelihood of

[11] 23.3% of subjects who switch multiple times are excluded.

participating in early admissions.

Table 2.8: Percentage of students to whom offers are sent by colleges

	Pre-exam			
	Boston		DA	
	College A	College B	College A	College B
S1 (%)	43.9	16.1	69.4	21.1
S2 (%)	2.2	24.4	0.6	39.4
S3 (%)	0.0	6.7	0.6	2.2
None (%)	53.9	52.8	29.4	37.2
	Post-exam			
S1 (%)	85.0	12.2	75.0	11.7
S2 (%)	1.1	55.0	0.6	27.8
S3 (%)	0.0	0.6	0.0	1.1
None (%)	13.9	32.2	24.4	59.4

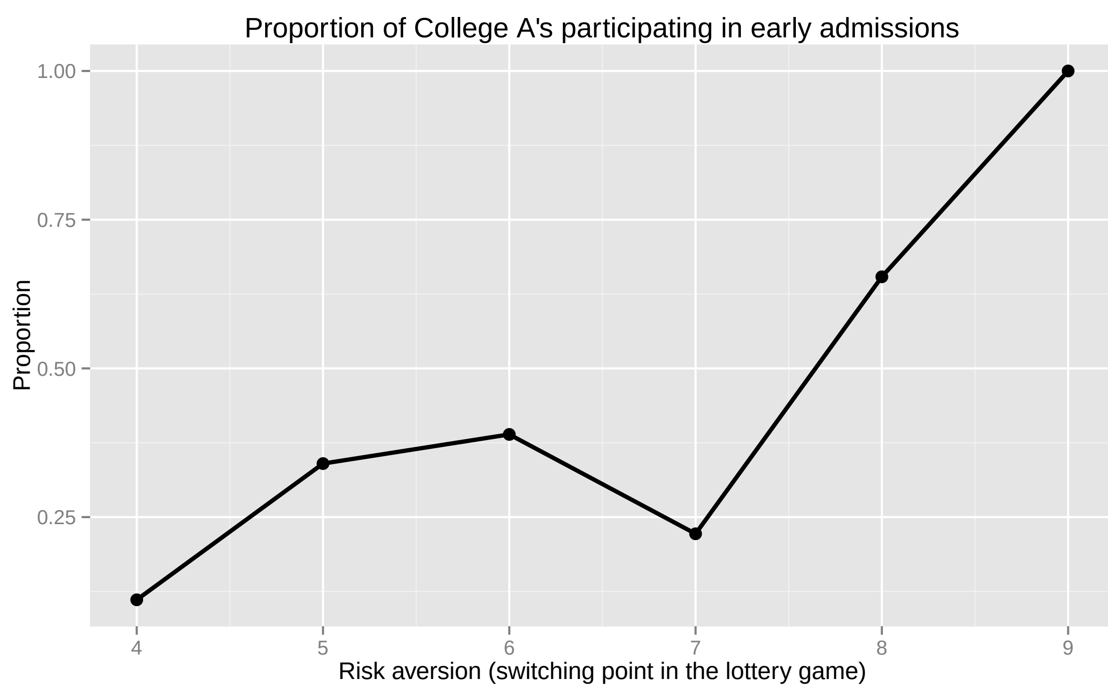


Figure 2.4: Relationship between risk attitude and likelihood of participating in early admissions

2.5.2 Students' Strategies

In this section, I look at the decisions by students who form the other side of the market.

First I report students' decisions during the early admissions stage. Tables 2.9, 2.10 and 2.11 summarize the proportion of decisions when students receive both offers, the offer from only College A, and the offer from only College B, respectively. Consistent with the equilibrium prediction, an overwhelming majority of the students choose the offer from College A over the offer from College B. However, note that there are significant numbers of Student 1s in all treatments that are also willing to accept the only offer from College B. This could be because they simply choose randomly, since assuming Colleges play equilibrium strategies, by accepting the offer from College A, students are either certain to get into the best college (Boston-post/DA-pre/DA-post), or they will not receive the offer from College B (Boston-pre), and therefore this decision becomes irrelevant. Also, under Boston-post, there is a sizable proportion of Student B (proportion test against 0%, $p < 0.01$) who choose College B over College A (Table 2.9), even though that is a strictly dominated strategy. This is equivalent to rank bias, except that this happens during the early admissions. Such bias is also unjustified.

Next, let us turn to the strategies for the ROL-submission during the centralized matching stage. The following hypothesis comes from the theoretical prediction of the truth-telling property of the DA and Boston mechanism. Figure 2.5 reports the proportion of truthful preference revelation among students who are not admitted early and therefore participate in the main admissions (note that students who are admitted early do not make these decisions). We can see clear difference in the proportions of truth-telling between Boston and DA.

Hypothesis 2 (Truthful preference revelation)**.** More students will reveal their true preferences over colleges under the Deferred Acceptance mechanism than under the

Table 2.9: Students' decisions when receiving both offers

	Pre-exam					
	Boston			DA		
	S1	S2	S3	S1	S2	S3
Accept A (%)	98.89	88.89	84.44	98.89	87.78	86.67
Accept B (%)	0.56	8.89	10.56	0.56	12.22	10.56
Reject (%)	0.56	2.22	5.00	0.56	0.00	2.78
	Post-exam					
Accept A (%)	92.2	77.8	75.6	99.4	91.1	81.7
Accept B (%)	3.9	20.0	20.0	0.0	8.9	16.1
Reject (%)	3.9	2.2	4.4	0.6	0.0	2.2

Table 2.10: Students' decisions when receiving offer from College A only

	Pre-exam					
	Boston			DA		
	S1	S2	S3	S1	S2	S3
Accept (%)	99.4	96.7	90.6	98.9	98.3	95.6
Reject (%)	0.6	3.3	9.4	1.1	1.7	4.4
	Post-exam					
Accept (%)	96.1	92.2	87.8	99.4	96.7	93.3
Reject (%)	3.9	7.8	12.2	0.6	3.3	6.7

Table 2.11: Students' decisions when receiving offer from College B only

	Pre-exam					
	Boston			DA		
	S1	S2	S3	S1	S2	S3
Accept (%)	41.1	93.3	95.6	47.8	78.9	96.7
Reject (%)	58.9	6.7	4.4	52.2	21.1	3.3
	Post-exam					
Accept (%)	46.7	83.3	92.2	32.2	75.6	97.8
Reject (%)	53.3	16.7	7.8	67.8	24.4	2.2

Boston mechanism.

Result 2 (Truthful preference revelation)**.** *Conditional on not being admitted early,*

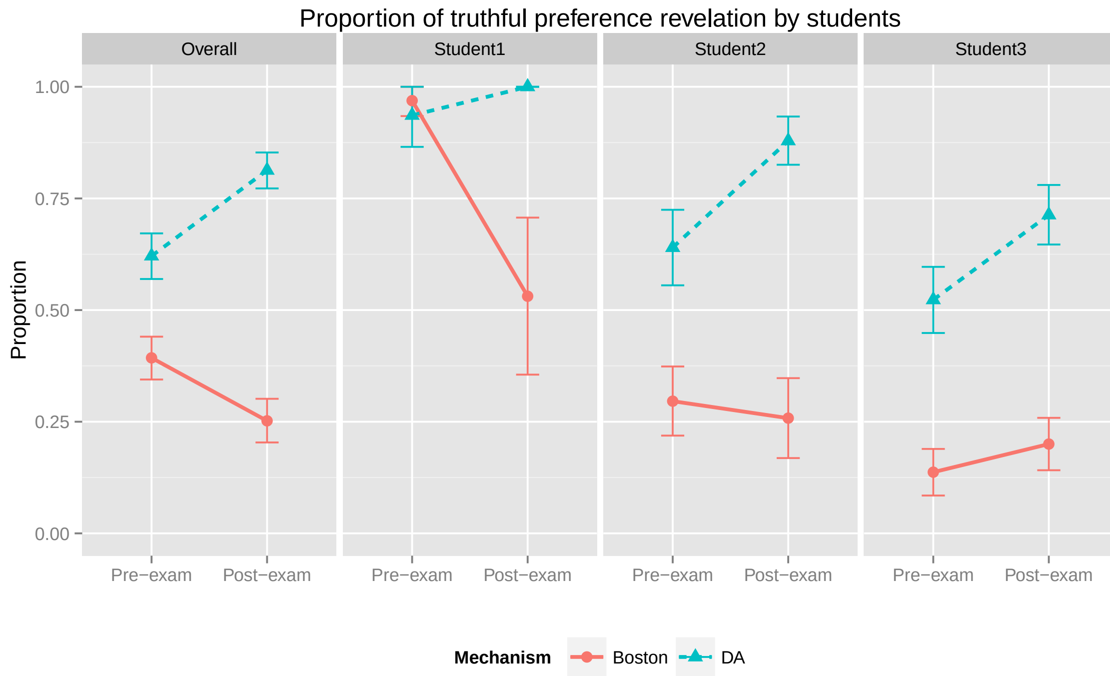


Figure 2.5: Proportion of truthful preference revelation by students (error bars show 95% confidence intervals)

significantly more students report their true preferences of colleges under DA than under Boston.

Support. Two-sided proportion test shows that Boston-pre has significantly lower truth-telling rate than DA-pre (39.3% vs 62.1%, $p < 0.001$). Similarly, Boston-post has lower rate than DA-post (25.5% vs 81.3%, $p < 0.001$). Additional probit regressions with clustered standard errors confirm this finding, shown in Table 2.16 in the appendix.

The above result rejects the null hypothesis that there is no significant difference in truth-telling between the two mechanisms, in favor of Hypothesis 13. In addition to the mechanism effect on truth-telling, I also observe the effect of timing. Significantly more students reveal their true preferences under DA-post than under DA-pre (proportion test, 81.3% vs 62.1%, $p < 0.001$). Further investigation shows that this is mostly attributed to the misrepresentations by Student 2 and Student 3.

On the other hand, the difference in truth-telling under Boston is mostly attributed to the top ranked student, Student 1. Many fewer Student 1s reveal their true preferences under Boston-pre compared to each of the other three treatments. This is not surprising because under Boston, when a student knows that he no longer ranks at the top after the test, the best strategy is not to list the best college as his first choice, but rather the college that corresponds to his ranking instead.

In addition to whether to reveal preferences truthfully, the strategy of rank bias also predicts that what a student actually lists as the first choice is determined by his or her rank in test scores, especially under Boston since it is manipulable. I summarize the prediction in the following hypothesis. Table 2.12 reports four ordered-profit specifications testing the effect of rank in test score on Student 2's first choice. Student 2 is the most interesting target of this task, because most Student 1s are admitted early, and Student 3 is indifferent among multiple strategies.

Hypothesis 3 (ROL strategies)**.** More students will rely on their rankings of test scores when submitting ROLs under the Boston mechanism with post-exam submission, while more students will rely on their rankings of aptitude under the Boston mechanism with pre-exam submission.

Result 3 (ROL strategies)**.** *Students exhibit rank bias under Boston with post-exam ROL submission, while they do not with pre-exam ROL submission.*

Support. Specifications (1) to (4) in Table 2.12 represent the effect of ranks in different treatments. Under the pre-exam timing conditions ((1) and (3)), rank in test score has no significant effect. Under Boston-post, having a higher rank in the test leads to a significantly higher chance of listing a better college as first choice. Surprisingly, we also observe rank-bias under DA-post, although the effect is smaller than under Boston-post.

Table 2.12: Order probit: Rank of test score on Student 2's first choice

Dep. Var.	First Choice College			
	(1)	(2)	(3)	(4)
	BOS-pre	BOS-post	DA-pre	DA-post
Rank at FirstChoice= 1	0.021	-0.252***	0.092	-0.148***
	(0.051)	(0.043)	(0.061)	(0.036)
Rank at FirstChoice= 2	-0.019	0.105**	-0.071*	0.098***
	(0.048)	(0.052)	(0.040)	(0.031)
Rank at FirstChoice= 3	-0.001	0.147***	-0.021	0.051*
	(0.003)	(0.054)	(0.027)	(0.028)
Observations	135	93	125	141

Notes: Standard errors in parentheses are clustered at the session level. Reporting marginal effects for different outcomes. *** p¡0.01, ** p¡0.05, * p¡0.1

By Result 19 we reject the null in favor of Hypothesis 14. In addition to what is stated in the hypothesis, I also observe that submission timing even has an effect under DA as well, which is consistent with the district bias findings in experimental school choice literature (e.g. Chen and Sönmez (2006); Chen and Kesten (2014)), where students exhibit district school (high-priority) bias under DA.

I do not separately report the regressions for aptitude-based rank bias, because (1) for Student 1s, it is captured by their truth-telling behavior; (2) for Student 2s, non-significant coefficients in score-based rank bias mean that the first choices are not dependent on the ranks in the test; and (3) Student 3s are indifferent among all three colleges for their first choices under conditions other than Boston-pre, therefore the results are very noisy.

2.5.3 Matching Outcomes

After looking at the decisions by both sides of the market, I now analyze the matching outcomes. The prediction for the proportion of students and colleges who are matched early is summarized in the following hypothesis. Figure 2.6 shows the proportion of colleges who admit students early, while Figure 2.7 shows the proportion of students who are admitted by colleges early. This helps to explain the empirical observation that after the policy change in Shanghai, more students were admitted early by elite colleges.

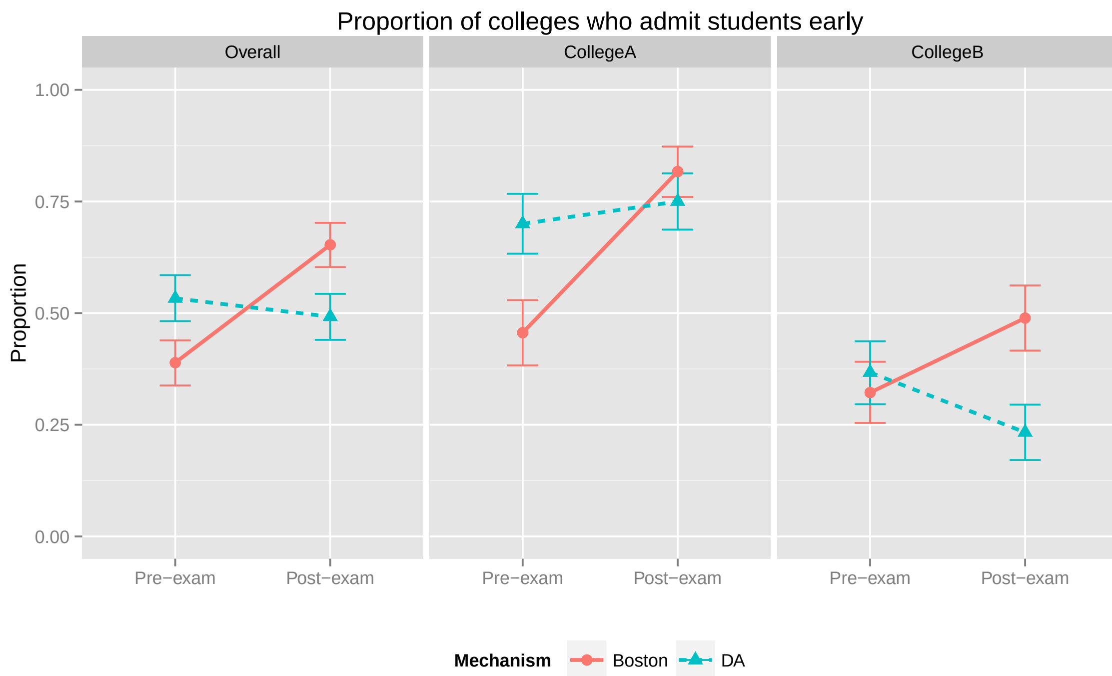


Figure 2.6: Proportion of colleges who admit students early (error bars show 95% confidence intervals)

Hypothesis 4 (Early matchings)**.** Fewer students and fewer colleges will be matched through early admissions under the Boston mechanism with pre-exam ROL submission than under the Boston mechanism with post-exam submission or the Deferred Acceptance mechanism with either of the two timing conditions.

Result 4 (Early matchings)**.** *Fewer colleges admit students early under Boston-pre, and fewer students are admitted early by colleges under Boston-pre, than under each of the other three treatments.*

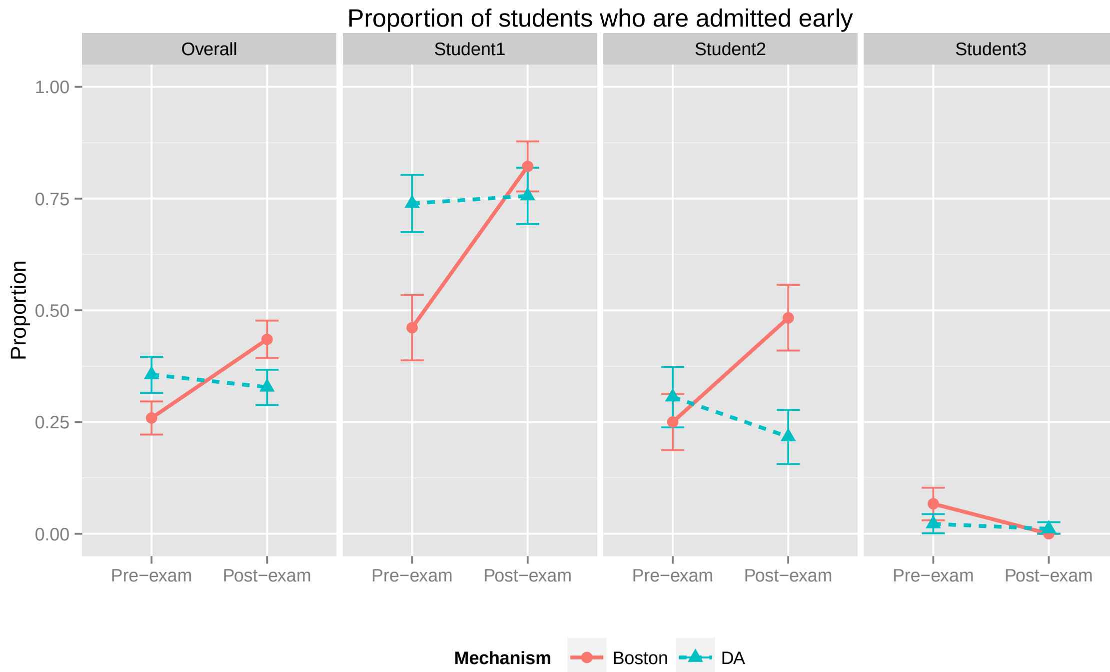


Figure 2.7: Proportion of students who are admitted early (error bars show 95% confidence intervals)

Support. For colleges, two-sided proportion test shows that Boston-pre has significantly lower early admission rate than the second best, DA-post (38.9% vs 49.2%, $p = 0.005$). DA-post is not significantly different from DA-pre (49.2% vs 53.3%, $p = 263$), and DA-pre has lower rate than Boston-post (53.3% vs 65.3%, $p = 0.001$). For students, two-sided proportion test shows that Boston-pre has significantly lower early admission rate than the second best, DA-post (25.9% vs 32.8%, $p = 0.013$). Similarly, DA-post is not significantly different from DA-pre (32.8% vs 35.6%, $p = 0.336$), and DA-pre has lower rate than Boston-post (35.6% vs 43.5%, $p = 0.007$). Additional probit regressions with clustered standard errors confirm this finding, shown in Table 2.17 and Table 2.18 in the appendix.

By Result 4, we reject the null in favor of Hypothesis 4. The above result can be further broken down to different types of colleges and students. Generally, College

As admit more students early than College Bs. Consistent with prediction, College As admit far fewer students early under Boston-pre than three other mechanisms. However, such difference does not exist for College Bs. Instead, College Bs admit most students under Boston-post, and fewest under DA-post.

On the students' side, the proportion of students who are admitted early is directly related to the student quality of each type, and almost no Student 3 is admitted early. Pattern similar to colleges' early admission outcomes can be observed for students as well: far fewer Student 1s are admitted early under Boston-pre than three other mechanisms, and more Student 2s are admitted early under Boston-post.

Next, I look at the matching outcomes in terms of the two criteria of stability: stability with respect to test score, and stability with respect to aptitude. Note that in the experimental game environment, through early admissions, the Boston mechanism with post-exam submission and the Deferred Acceptance mechanism under both timing conditions completely eliminate instability with respect to aptitude, there is still a small chance (6.25%) of a matching outcome being unstable under the Boston mechanism with pre-exam submission. The opposite is true for stability w.r.t. test score. Stability predictions are summarized in the following two hypotheses. Note that the predictions are the opposite of what would happen in the absence of early admissions. Figure 2.8 shows the proportion of matching outcomes that satisfy each of the two stability notions.

Hypothesis 5 (Stability w.r.t. test score)**.** When early admissions are allowed, the stability w.r.t. test score for the four treatments has the following order: Boston-pre>Boston-post=DA-pre=DA-post.

Result 5 (Stability w.r.t. test score)**.** *DA is more stable than Boston with respect to test score under both timing conditions. There is no difference within mechanisms between the two timing conditions.*

Support. Two-sided proportion test shows that Boston-pre has significantly lower proportion of stable matchings w.r.t. test score than DA-pre (48.9% vs 61.1%, $p = 0.020$). Similarly, Boston-post has lower proportion than DA-post (50.0% vs 67.8%, $p < 0.001$). There is no timing effect ($p = 0.186$ and $p = 0.833$ for Boston and DA respectively). Additional probit regressions with clustered standard errors confirm this finding, shown in Specification (1) in Table 2.19 in the appendix.

Result 5 fails to reject the null hypothesis that there is no difference in stability (w.r.t. test score) within in Boston between pre and post timing conditions, when early admissions are present. It also contradicts the prediction that (1) Boston-pre is more stable than DA-pre and (2) there is no difference across mechanisms under the post timing condition. These contradictions to theoretical prediction can be explained by colleges' failure to employ Nash equilibrium early admission strategies. Note that under DA, if colleges do not participate in early admissions, truth-telling by students leads to matching outcomes that are stable w.r.t. test scores. Table 2.8 indeed shows that significant proportion of college players deviating from the equilibrium strategies.

Next we explore stability with respect to aptitude.

Hypothesis 6 (Stability w.r.t. aptitude)**.** When early admissions are allowed, the stability w.r.t. aptitude for the four treatments has the following order: Boston-pre<Boston-post=DA-pre=DA-post.

Result 6 (Stability w.r.t. aptitude)**.** *Boston-post is more stable with respect to aptitude than Boston-pre. There is no difference between Boston-pre and DA-pre. There is also no difference between the two treatments under DA.*

Support. Two-sided proportion test shows that Boston-pre has significantly lower proportion of stable matchings w.r.t. aptitude than Boston-post (62.8% vs 73.9%, $p = 0.024$). However, there is no significant difference between Boston-pre and DA-pre (62.8% vs 66.7%, $p = 0.440$), Boston-post and DA-post (73.9% vs 67.8%, $p = 0.202$),

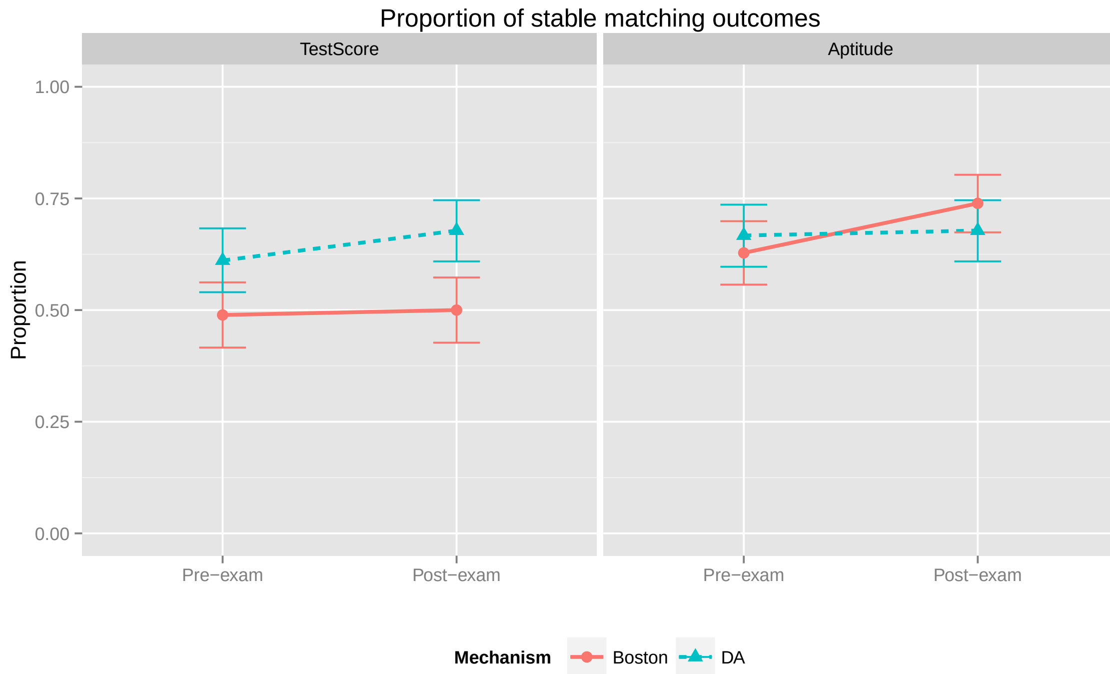


Figure 2.8: Proportion of stable matchings w.r.t. test score (left) and w.r.t. aptitude (right) (error bars show 95% confidence intervals)

and DA-pre and DA-post (66.7% vs 67.8%, $p = 0.822$). Additional probit regressions with clustered standard errors confirm this finding, shown in Specification (2) in Table 2.19 in the appendix.

From the above result, although we can reject the null that there is no timing effect under Boston, there is no significant mechanism difference under the pre condition. For the rest of this section, I explore why this is the case. Note that in order to reach predicted matching outcomes, both colleges and students need to play Nash equilibrium strategies. However, from previous finding we see that this is not the case. Therefore, I look at how the stable matchings (w.r.t. aptitude) are achieved, and why some matching outcomes are unstable (w.r.t. aptitude).

First I investigate how stable matchings are achieved. Figure 2.9 presents the proportion of stable matchings (w.r.t. aptitude) that are achieved without any early admissions under the four mechanism-timing conditions I study.

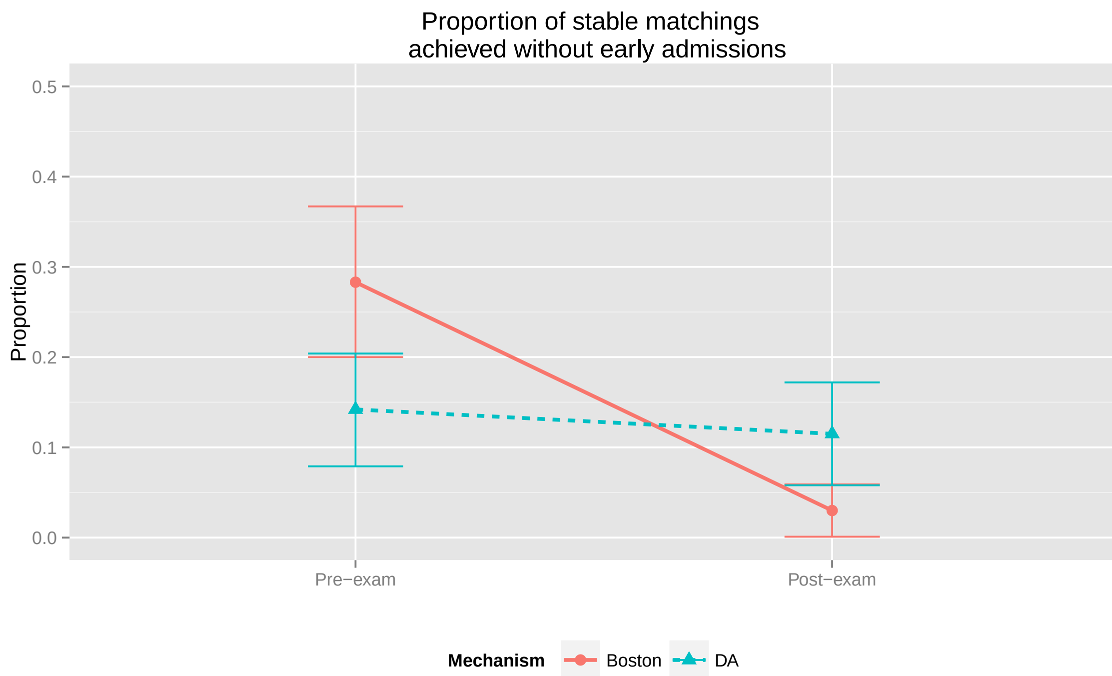


Figure 2.9: Proportion of stable matchings achieved without early admissions (error bars show 95% confidence intervals)

Result 7. *More stable matchings w.r.t. aptitude are achieved without going through any early admissions under Boston-pre.*

Support. Two-sided proportion test shows that Boston-pre has significantly higher proportion of stable matchings (w.r.t. aptitude) achieved without early admissions, compared to the second best, DA-pre (28.3% vs 14.2%, $p = 0.008$). There is no significant difference in proportion between DA-pre and DA-post (14.2% vs 11.5%, $p = 0.531$). Boston-post has the lowest proportion (3.0% vs 11.5%, $p = 0.008$). Additional probit regressions with clustered standard errors confirm this finding, shown in Specification (3) in Table 2.19 in the appendix.

The above result is encouraging: we were worried that, per equilibrium prediction, Boston-pre trades more early admissions for more unstable matching outcomes w.r.t.

aptitude. Now we see that the difference in stability is so small, especially when compared to the reduction in market unraveling, that the benefits outweigh the costs.

One may argue that by deliberately allowing early admissions as a policy choice, even though it is costly, it can solve the problem of aptitude mismatch ex post. The above result, however, shows that simply through the design of the centralized matching, similar outcomes can be achieved, for a much lower cost. Of course, if the top concern is stability w.r.t. test score, regardless of timing conditions, DA is still the preferred choice; then it is probably even better to simply restrict the use of early admissions.

Table 2.13: Causes of outcomes that are unstable w.r.t. aptitude

	Pre-exam		Post-exam	
	Boston	DA	Boston	DA
Measurement error with three students participating	46.3%	25%	29.8%	29.3%
Measurement error with two students participating	22.4%	51.7%	57.4%	62.1%
Suboptimal early admission decisions	31.3%	23.3%	12.8%	8.5%
Total cases	67	60	47	58

Finally, I decompose the matching outcomes that are unstable w.r.t. aptitude. Table 2.13 presents the classification of causes of unstable matchings. I identify three categories of causes: (1)no student participates in early admissions (thus all three participate in the centralized matching), and measurement error of the test causes unstable matchings; (2)one student participates in early admissions (thus two participate in the centralized matching), and measurement error of the test causes unstable matchings; and (3)colleges send “wrong” offers to students, therefore early admission itself directly leads to unstable matchings. This table shows that Boston-pre has the highest occurrence of unstable matchings caused by the measurement error in the test with all students participating in the centralized matching. However, it has the lowest occurrence of unstable matchings when one student is admitted through early admissions. Additionally, Boston-pre again has the highest number of cases of unstable matchings caused by colleges admitting early worse students than the students they could have got under the equilibrium. Many of such cases can be avoided by not participating in the early admission at all.

2.6 Conclusion

Using standardized tests to evaluate students and select them into higher levels of education has become a hot topic, both in the United States and around the world. One of the main problems with standardized tests, however, is the associated measurement errors. Therefore, the misalignment of priorities and preferences in the matching process is almost inevitable. Although stable matching mechanisms such as the Deferred Acceptance mechanism are generally considered superior to unstable alternatives, especially in reducing market unraveling, this may not be the case under such case. In this paper, motivated by the college admissions matching in Shanghai, I try to explain why a transition to an apparently more stable matching mechanism increases rather than decreases the level of unraveling in this market.

I find that under the context of college admissions, when students' priorities in the matching are determined by test scores from a standardized test, which is a noisy realization of students' aptitudes, any matching mechanism that produces outcomes that are stable with respect to test score, produces outcomes that are not stable with respect to aptitude. If colleges only participate in the centralized admission process, then they may prefer students who are matched with other colleges, and these students also prefer to be matched with these colleges rather than the colleges they are currently matched with. This gives colleges incentives to circumvent the centralized matching, and try to admit these students in advance. Since it is the dominant strategy to reveal one's true preference over colleges under the Deferred Acceptance mechanism, the matching absolutely respects test scores, with no regard to a student's actual aptitude. Therefore, this mechanism will cause market unraveling through early admissions by colleges. On the contrary, the Boston mechanism is manipulable, and the information of students' rankings, either of test scores or of aptitudes, is used by students to strategize their rank-ordered lists, therefore we can take the advantage such manipulability by revealing the set of information that is desirable (in this case, the rankings of aptitudes rather than the rankings of test scores).

To test the theoretical predictions, I design a laboratory matching market, where colleges can send out early admission offers to students ahead of the centralized matching. Yet, they can not actively participate in the matching itself once it starts. Experimental results show that, consistent with theoretical predictions as well as real world evidence, the Deferred Acceptance mechanism causes more market unraveling as more colleges send out early admission offers, circumventing the centralized matching. However, when students are required to submit their rank-ordered lists before they take the exam, when they only know their rankings in aptitudes, the Boston mechanism significantly reduces the occurrence of early offers. Furthermore, the Boston mechanism with pre-exam submission does not yield more unstable matching outcomes with respect to students' aptitudes compared to other mechanisms, even though fewer students and colleges go through early admissions. Even for the stable matching outcomes, significantly more of them are achieved without anyone participating in early admissions. Thus, I conclude that the Boston mechanism under pre-exam submission condition performs at least as well as all other mechanisms and timing conditions in stability with respect to aptitude, and it achieves that with much less sacrifice in market efficiency caused by unraveling.

This paper, from another perspective, shows that the Boston mechanism can actually outperform the Deferred Acceptance mechanism, and thus may to some extend explain why such mechanism is still widely used in many places where centralized matching is carried out. Given that standardized tests are still one of the most popular ways to evaluate students and select them into higher levels of education, policy makers may need to be very careful when deciding to switch from the Boston mechanism to the Deferred Acceptance mechanism, especially if their goal is to organize such markets in a centralized and efficient way.

2.7 Appendix

2.7.1 Proof of Proposition 3

Proof. Given the definition of rank bias, the mass of students who list college c_j as their first choices using rank bias strategy equals to q_{c_j}. This also implies that all students get into their first choice colleges.

No student has the incentive to deviate from that strategy. Students who are ranked top q_{c_1} will not deviate because they will certainly be admitted by the top college. Given that, students who are ranked between $q_{c_1}+1$ and q_{c_2} cannot profitably deviate as well: if they rank a college that is below the second best college, they are strictly worse off; if they rank a college that is above the second best college, their chance of getting into that is 0. The same reasoning can be applied to all students down the rank. Finally for the bottom ranked students, they are indifferent between playing rank bias or ranking any college as their first choice. □

2.7.2 Proof of Theorem 1

Proof. The proof consists of following steps.

First, I show that the admissions results are assortative. That is, top q_{c_1} students in the test are admitted into college c_1, next q_{c_2} students scores are admitted into college c_2, etc. This is evident given the truth-telling property of the DA mechanism and the homogeneous preferences of students.

Denoting the density function of test scores as $h(\cdot)$, we partition students' test scores into m parts with cut-off points $p_{t_1}, ..., p_{t_{m-1}}$, such that

$$\int_{p_{t_{i-1}}}^{p_{t_i}} h(x)dx = \frac{q_{c_i}}{\sum q_c}$$

. From the above result we know that a student with test score $t, p_{t_{i-1}} \leq t < p_{t_i}$ is admitted to college c_i. Note that for colleges are the boundary (best school and worst school), we replace cut-off points with $-\infty$ and ∞

We do the same to aptitude so that we have cut-off points $p_{a_1}, ..., p_{a_{m-1}}$. A student with score-aptitude pair (t, a) admitted by a non-boundary college c_i will form a block pair with a least one better college if his aptitude satisfies $a > q_{a_i}$. The probability this will happen is given by: $Prob(a > q_{a_i} | q_{t_{i-1}} \leq t < q_{t_i})$, which is expanded to

$$Prob(a > p_{a_i} | p_{t_{i-1}} \leq t < p_{t_i}) = \frac{Prob(a > p_{a_i}, p_{t_{i-1}} \leq t < p_{t_i})}{Prob(p_{t_{i-1}} \leq t < p_{t_i})}$$

We know that $t = a + \eta$, where a has a p.d.f. f and η has a p.d.f. g. We need to calculate the joint p.d.f. of $(a, a+\eta)$. Let $X = a, Y = a + \eta$. Solving the linear equation, we have $a = X, \eta = Y - X$. We calculate the determinant of the Jacobian matrix:

$$\begin{vmatrix} \frac{da}{dX} & \frac{da}{dy} \\ \frac{d\eta}{dX} & \frac{d\eta}{dy} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1$$

So the joint distribution of $(a, a+\eta)$ is given by $f_{X,Y}(x, y) = f(x)g(y - x)$. Go back to the probability equation, we have:

$$Prob(a > p_{a_i} | p_{t_{i-1}} \leq t < p_{t_i}) = \frac{\int_{p_{a_i}}^{\infty} \int_{p_{t_{i-1}}}^{p_{t_i}} f(x)g(y-x)dydx}{\int_{p_{t_{i-1}}}^{p_{t_i}} \int_{-\infty}^{\infty} f(x)g(z-x)dxdz} > 0$$

given the non-trivial support of g. □

2.7.3 Additional Regression Tables

Table 2.14: Probit regression: risk preference on early admission decision

Dep. Var.	Early Admission
	(1)
HL switch point	0.114**
	(0.050)
Observations	138

Notes: Standard errors in parentheses are clustered at the session level; reporting marginal effects. *** p¡0.01, ** p¡0.05, * p¡0.1

Table 2.15: Probit regression: treatment effects on early admission decision

Dep. Var.	Early Admission		
	(1)	(2)	(3)
Post	0.299***	0.421***	0.421***
	(0.079)	(0.076)	(0.076)
DA	0.187	0.229**	0.229**
	(0.118)	(0.115)	(0.115)
Post×DA	-0.381***	-0.373***	-0.373***
	(0.135)	(0.132)	(0.132)
CollegeB		0.015	0.015
		(0.028)	(0.028)
Post×CollegeB		-0.240***	-0.240***
		(0.057)	(0.057)
DA×CollegeB		-0.095*	-0.095*
		(0.053)	(0.053)
Period			0.001
			(0.002)
Observations	1,440	1,440	1,440

Notes: Standard errors in parentheses are clustered at the session level; reporting marginal effects. *** p¡0.01, ** p¡0.05, * p¡0.1

Table 2.16: Probit regression: treatment effects on truth-telling

Dep. Var.	Truth-telling		
	(1)	(2)	(3)
Post	-0.134***	-0.345***	-0.338***
	(0.020)	(0.103)	(0.104)
DA	0.197***	0.151	0.159
	(0.026)	(0.132)	(0.128)
Post×DA	0.331***	0.190***	0.186***
	(0.059)	(0.065)	(0.064)
Student2		-0.609***	-0.606***
		(0.051)	(0.052)
Post×Student2		0.349***	0.346***
		(0.102)	(0.101)
DA×Student2		0.130	0.123
		(0.111)	(0.108)
Student3		-0.715***	-0.714***
		(0.061)	(0.059)
Post×Student3		0.348***	0.343***
		(0.117)	(0.116)
DA×Student3		0.112	0.106
		(0.129)	(0.126)
Period			0.004*
			(0.002)
Observations	1,416	1,416	1,416

Notes: Standard errors in parentheses are clustered at the session level; reporting marginal effects. *** p¡0.01, ** p¡0.05, * p¡0.1

Table 2.17: Probit regression: treatment effects on early matchings for colleges

Dep. Var.	Matched Early		
	(1)	(2)	(3)
Post	0.261***	0.347***	0.346***
	(0.068)	(0.064)	(0.064)
DA	0.142	0.216*	0.216*
	(0.117)	(0.114)	(0.114)
Post×DA	-0.302**	-0.291**	-0.290**
	(0.134)	(0.133)	(0.133)
College2		-0.125***	-0.125***
		(0.012)	(0.012)
Post×College2		-0.193***	-0.193***
		(0.019)	(0.019)
DA×College2		-0.171***	-0.171***
		(0.018)	(0.018)
Period			0.006**
			(0.002)
Observations	1,440	1,440	1,440

Notes: Standard errors in parentheses are clustered at the session level; reporting marginal effects. *** p¡0.01, ** p¡0.05, * p¡0.1

Table 2.18: Probit regression: treatment effects on early matchings for students

Dep. Var.	Matched Early		
	(1)	(2)	(3)
Post	0.175***	0.227***	0.227***
	(0.046)	(0.041)	(0.041)
DA	0.100	0.161**	0.160**
	(0.081)	(0.078)	(0.078)
Post×DA	-0.202**	-0.196**	-0.195**
	(0.093)	(0.092)	(0.092)
Student2		-0.141***	-0.141***
		(0.014)	(0.014)
Post×Student2		-0.086***	-0.086***
		(0.028)	(0.028)
DA×Student2		-0.132***	-0.133***
		(0.024)	(0.025)
Student3		-0.381***	-0.381***
		(0.078)	(0.078)
Post×Student3		-0.364**	-0.365**
		(0.148)	(0.148)
DA×Student3		-0.193	-0.193
		(0.118)	(0.118)
Period			0.004**
			(0.002)
Observations	2,160	2,160	2,160

Notes: Standard errors in parentheses are clustered at the session level; reporting marginal effects. *** p¡0.01, ** p¡0.05, * p¡0.1

Table 2.19: Probit regression: treatment effects on stability

Dep. Var.	Stable w.r.t. test score	Stable w.r.t. aptitude	Stable w/o early admission
	(1)	(2)	(3)
Post	0.011	0.112*	-0.262***
	(0.056)	(0.058)	(0.079)
DA	0.119**	0.037	-0.100
	(0.053)	(0.072)	(0.087)
Post×DA	0.058	-0.101	0.236**
	(0.090)	(0.085)	(0.100)
Observations	720	720	488

Notes: Standard errors in parentheses are clustered at the session level; reporting marginal effects. *** p¡0.01, ** p¡0.05, * p¡0.1. The dependent variable for specification (1) is whether a matching outcome is stable w.r.t. test scores; for specification (2) is whether a matching outcome is stable w.r.t. aptitude; and for specification (3) is whether a stable matching outcome is not achieved through early admissions.

CHAPTER III

Matching in the Large: An Experimental Study

3.1 Introduction

Market size has been predicted to play an influential role in a broad class of economic environments. In a large market, existing impossibility results about incentives, welfare, and stability may be overturned, while other existing possibility results may be sharpened to unique solutions. This theoretical phenomenon becomes particularly important for practical market design cases, such as combinatorial auctions, school choice, labor market clearinghouses, course allocation, and kidney exchange, where the market size can range from hundreds to millions. For example, the National Resident Matching Program matches roughly 30,000 doctors and hospitals per year. In another setting, the centralized college admissions systems in Turkey and China match millions of students respectively. Within the United States, school assignments in New York City match nearly 100,000 students per year.

Motivated by practical concerns, there has been a surge of interest in the theoretical study of large matching markets within the last decade by either investigating the asymptotic properties of finite discrete markets or modeling either or both sides of the market as a continuum mass of agents. Within this literature, a significant number of papers have examined the question of whether good incentive and stability properties under the Gale-Shapley deferred acceptance mechanism (Gale and Shapley, 1962) hold as the market size grows, as conjectured in Roth and Peranson (1999). One strand of this literature has shown that, under this mechanism, partner incentives for preference misrepresentations in marriage problems (Immorlica and Mahdian, 2005), college incentives for capacity and preference misrepresentations in college admissions problems (Kojima and Pathak, 2009), and school incentives to disrespect quality improvements in school choice problems (Hatfield et al., 2011) vanish with the market size. Similarly, another line of research has shown that, in a large market satisfying certain regularity assumptions, this mechanism always produces a stable matching in a discrete two-sided matching model, allowing for complementarities (in the form of couples) in the context of the entry-level labor market for U.S. doctors (Kojima et al., 2013; Ashlagi et al., 2011) as well as in continuum models of many-to-one and many-to-many matching (Azevedo and Leshno, 2011; Azevedo and Hatfield, 2012; Che et al., 2013).

In designing practical markets, institutions have relied on economic theory, computation, and controlled laboratory experiments (Roth, 2002). In the school choice reforms in New York City (Abdulkadiroğlu et al., 2005a) and Boston (Abdulkadiroğlu et al., 2005b), for example, matching theorists were directly involved to influence the adoption of the Gale-Shapley deferred acceptance mechanism (Gale and Shapley, 1962; Abdulkadiroğlu and Sönmez, 2003). This represents an improvement over the Boston mechanism, which is severely vulnerable to strategic manipulation in the laboratory (Chen and Sönmez, 2006). Experimental data helped to persuade the Boston public-school authorities to switch from the Boston mechanism to the student-optimal deferred acceptance mechanism in 2005 (Abdulkadiroğlu et al., 2005b).

Such institutional redesigns rely on computational as well as laboratory experiments to provide the first data on a theoretically superior mechanism for which field data is not yet available and to compare the performance of different mechanisms at a level of detail that cannot be obtained from field data. Yet another reason for experimentally studying large market dynamics can be attributed to limitations of theoretical analysis. While the theoretical literature characterizes the performance of the canonical matching mechanisms in the large, it is often silent about the question of how large is "large."[1] To answer this question, empirical, simulation, and experimental work might shed some light.

Matching experiments in particular have been used to study mechanism stability as well as related unraveling issues under different types of clearinghouses (Kagel and Roth, 2000; Niederle and Roth, 2003), performance evaluation of different assignment methods in complete information environments (Chen and Sönmez, 2004; Chen and Kesten, 2014) and the effect of informational and institutional constraints on mechanism performance (Pais and Pintér, 2008; Calsamiglia et al., 2010). Despite the theoretical interest in understanding effects of the market size, attention has not been transferred to the experimental setting for several reasons. First, while laboratory experiments are often compared to a wind tunnel for evaluating new institutions, the scale of a laboratory experiment is often small compared to the corresponding real-world implementations. For example, the largest school choice experiments have 36 subjects per match (Chen and Sönmez, 2006; Calsamiglia et al., 2010), a far cry from the hundreds of thousands of students in Beijing and Shanghai assigned to various high schools each year (He, 2014).

Given the practical importance of understanding how market size impacts mechanism performance, our goal in this paper is to experimentally complement the active and growing literature in matching theory on large market properties within the context of student assignment. To bridge the gap between laboratory experiments and real-world implementations, we compare the performance of two school choice mecha-

[1] A notable exception is Rustichini et al. (1994) who show in the context of double auctions six traders of each type are sufficient to obtain efficiency to within one percent. In the context of assignment problems, for a fixed set of object types and a given agent's utility function, Kojima and Manea (2010) calculate the number of copies of each object type needed in order for the probablistic serial mechanism to become strategy-proof.

nisms in the laboratory when market size increases. Studying large matching markets in the laboratory is of interest for two main reasons. First, large markets often have different theoretical properties than their smaller counterparts. Second, a large market may impact participant behavior due to the complexity of thought it brings to decision making.

We focus our experiment on school choice as it is a widely-debated education policy across the world (Hoxby, 2003; He, 2014), affecting the education experiences and labor market outcomes for millions of students each year. To contribute insight to this debate, we compare and contrast the large market characteristics of the Gale-Shapley deferred acceptance (DA) mechanism to the Boston mechanism. Unlike the DA, which is known in this context to be strategy-proof regardless of market size, the Boston mechanism has been shown to be prone to strategic play in both small and large markets (Kojima and Pathak, 2009; Azevedo and Budish, 2013).

Creating a large market in the laboratory is challenging because of both physical and financial constraints. To address this challenge, we first vary the scale of our all-human sessions from 4 to 40 students per match. To enable random rematching, the latter treatment requires 80-subject sessions. We then let human subjects play against robots whose strategies are drawn from empirical human strategies from the 40-subject matches, and further increase the scale from 40 to 4,000 students per match. In doing so, we find that when the number of students increases from 4 to 40, the proportion of truth-telling significantly increases under the DA but decreases under Boston. These results point to nuanced scale effects under each mechanism, though in opposite directions. Our finding of a scale effect under the DA mechanism is unexpected, whereas the increased manipulation under Boston with an increase in scale leads to an even stronger finding than the theoretical view that Boston is likely to remain manipulable in a large market. Further we find that increasing the scale from 40 to 4,000 has no additional effect on the performance of either mechanism.

Finally, we find that the matching outcomes under the DA mechanism remain stable as market size increases. Overall, our results indicate that "large" might be much smaller than predicted in the theoretical literature and provide additional support for the replacement of the Boston mechanism with the DA mechanism in practice.

3.2 Literature Review

Incentives for truthful preference revelation in large markets have been studied in matching theory as well as other economic contexts. For example, Roberts and Postlewaite (1976) have shown that the Walrasian mechanism is approximately strategy-proof in a large exchange economy.[2] In auction theory, Gresik and Satterthwaite (1989), Rustichini et al. (1994), Pesendorfer and Swinkels (2000) and Cripps and Swinkels (2006), and Fudenberg et al. (2007) all show that strategic misreporting vanishes in double auctions in large markets under various informational assumptions. Swinkels (2001) shows a similar result for uniform-price and pay-as-bid auctions.

In indivisible good allocation problems large market arguments have been used to support particular market design approaches. In kidney exchange problems, Roth et al. (2005, 2007) show that conducting only small size kidney exchanges is sufficient to achieve full efficiency when the number of incompatible patient-donor pairs is large. When multiple hospitals are involved in an organized exchange program, Ashlagi and Roth (2011) show that it becomes individually rational for each hospital to participate in the joint exchange program (as opposed to conducting exchanges internally) when the population of hospitals and patients grows large. Finally, in the object assignment context, Kojima and Manea (2010) show that the probabilistic serial mechanism of Bogomolnaia and Moulin (2001) becomes exactly strategy-proof in a sufficiently large finite market.

[2] Specifically, they show that when equilibrium prices vary continuously with reports, truthtelling is approximately ex-post optimal.

Relevant to our experimental design, Azevedo and Budish (2013) propose the concept of strategy-proofness in the large (SP-L). They examine the manipulability properties of well known mechanisms as the market size increases and provide a unified view of the large market approaches to incentive issues. In the school choice context, using two different large-market models, Kojima and Pathak (2009) and Azevedo and Budish (2013) both conclude that the Boston mechanism remains manipulable even in a large market. For finite markets Pathak and Sönmez (2013) introduce a metric for ranking different mechanisms according to their manipulability. Interestingly, our experimental results indicate the possibility of varying degrees of manipulability of the same mechanism depending on the scale of the market.

It has been well-documented that the DA mechanism is manipulable in a two-sided matching market (Dubins and Freedman, 1981; Roth, 1982). Some empirical studies and simulations results support the large market predictions of the DA and Boston mechanisms. For example, Roth and Peranson (1999) analyze the NRMP data by conducting simulations on randomly generated simple markets and show that when there is a bound on the length of preferences acceptable to one side of agents, the set of stable matchings becomes small as the market grows.[3] They observe that, of the more than 20,000 applicants and 3,000-4,000 programs in their study, less than one percent could benefit from truncating preference lists or capacities. During the transition from the previous NRMP algorithm to the new algorithm, less than one percent of applicants and programs are found to be affected by such change, which is comparably small and unsystematic.[4] By contrast, the DA is strategy-proof in the

[3]Recently, Azevedo and Leshno (2014) develop a continuum model of the college admissions problem to present a more appealing and general property about the set of stable matchings: that is, a generic continuum economy has a unique stable matching, to which a sequence of the set of stable matchings in large discrete economies converges. Such "core convergence" and uniqueness properties of set of stable matchings have an important implication on other asymptotic properties of the DA mechanism. In particular, Azevedo and Leshno (2014) and Azevedo and Leshno (2011) investigate the DA with single tie-breaking and show that the mechanism is robust to aggregate randomness: whatever tie-breaking results, students will be assigned to almost the same schools under the DA mechanism in a large market.

[4]In the school choice context, Abdulkadiroğlu et al. (2009) also find small changes in matching

school choice context where only students are assumed to be strategic agents. Thus the non-manipulability of DA is robust to market size. Using data from the Boston school district prior to the reform, Abdulkadiroğlu et al. (2006) find empirical evidence for both strategically sophisticated and naive play under the Boston mechanism. Based on these theoretical and empirical results, we expect to see misrepresentation of preferences in the Boston mechanism regardless of market size. We expect no scale effect on the proportion of truth-telling under either mechanism.

With the development of matching theory, a growing number of laboratory experiments have tested mechanism performance as well as participant behavior under different incentives. In one study with 36 students per match, Chen and Sönmez (2006) observe that the proportion of preference manipulation under the Boston mechanism is significantly higher than either the DA or the Top Trading Cycles (TTC) mechanisms. Subsequent studies have examined the impact of different information conditions (Pais and Pintér, 2008), a limit on the number of schools in the rank order list (Calsamiglia et al., 2010), participant risk attitude and preference intensities (Klijn et al., 2013) on participant behavior in school choice experiments. Featherstone and Niederle (2014) observe that the Boston mechanism achieves higher efficiency than the DA mechanism when preferences are private information and when school priorities involve ties which are broken randomly.

Table 3.1 summarizes the design features of several representative experimental studies of school choice. In addition to the mechanisms examined in each study, we document the number of students per match. While these experiments use various market sizes, none studies the scale effect. To our knowledge, the only other matching experiment which studies the scale effect is that of Chen and Sönmez (2002), who study house allocations when the number of students increases from 12 to 60 per

outcomes when investigating New York City school choice preference data. In several different runs of the DA algorithm, using different lottery outcomes to break the ties in priorities, the aggregate statistics of the match do not vary much.

Table 3.1: Representative Experimental Studies of School Choice

Representative Studies	Mechanisms	# per match
Chen and Sönmez (2006)	Boston, DA, TTC	36
Pais and Pintér (2008)	Boston, DA, TTC	5
Calsamiglia et al. (2010)	Boston, DA, TTC	36
Featherstone and Niederle (2014)	Boston, DA	5
Klijn et al. (2013)	Boston, DA	3
Chen and Kesten (2014)	Boston, DA, Chinese Parallel	4, 6
This paper	Boston, DA	4, 40, 4,000

match. They find that the change in scale has no significant effect on the proportion of truth-telling or participation rate under either TTC or Random Serial Dictatorship with Squatting Rights.

3.3 The matching problem and two mechanisms

A school choice problem (Abdulkadiroğlu and Sönmez, 2003) is comprised of a set of students, each of whom is to be assigned a seat at one school from a set of schools. Each school has a number of available seats called the quota for that school.[5] For each school, there is a strict priority order for all students, and each student has strict preferences over all schools.

Within this context, a *matching* μ is a list of assignments such that each student is assigned to one school and the number of students assigned to a particular school does not exceed the quota of that school. A matching μ is *Pareto efficient* if there is no other matching which makes all students at least as well off and at least one student better off.

The *college admissions problem* (Gale and Shapley, 1962) is closely related to the school choice problem. By contrast though, in the college admissions problem, schools have preferences over students, whereas in a school choice problem, schools

[5] We assume that there are enough seats for all the students, an assumption often met in practice. However, the model is easily modified to accommodate outside options and a shortage of seats.

are objects to be consumed. A key concept in the school choice problem is *stability*, i.e., there is no unmatched student-school pair (i, s) such that student i prefers school s to his assignment, and either school s has not filled its quota or student i has a higher priority than at least one student j who is enrolled in s. In the latter case, we say that student i *justifiably envies* student j for school s. A (school choice) mechanism is a systematic procedure that chooses a matching for each problem. A mechanism is Pareto efficient (stable) if it always selects Pareto efficient (stable) matchings. A mechanism φ is *strategy-proof* if it is a dominant strategy for each student to truthfully report her preferences.

In our study, we focus on two school choice mechanisms. Our first mechanism, the Boston mechanism, is the most common school choice mechanism observed in practice. Its outcome can be calculated via the following algorithm:

Step 1: For each school, consider only those students who have listed it as their first choice. Those students among them with the highest priority for that school are assigned that school up to its quota.

Step k, $k \geq 2$*: Consider the remaining students who are unassigned and the schools that have not filled their quota. For each such school, consider only those students who have listed it as their k-th choice. In this group, those students with the highest priority for that school are assigned that school up to its remaining quota.*

The algorithm terminates when there are no students left to assign. Importantly, note that the assignments in each step are final. Based on this feature, an important critique of the Boston mechanism is that it gives students strong incentives for gaming through misreported preferences. Because a student who has high priority for a school may lose her priority advantage for that school if she does not list it as her first choice, the Boston mechanism forces students to make hard and potentially costly choices, which leads to a high-stakes game among participants with different levels of strategic sophistication (see e.g., Abdulkadiroğlu and Sönmez (2003); Ergin and Sönmez (2006); Chen and Sönmez (2006); Pathak and Sönmez (2008); He (2014)).

In addition to the Boston mechanism, we consider the student-optimal stable mechanism (Gale and Shapley, 1962), which has played a central role in the school choice reforms in Boston and New York City (Abdulkadiroğlu et al., 2005b,a) and, more recently, in Paris. Its outcome can be calculated via the following *deferred acceptance (DA) algorithm*:

Step 1: Each student applies to her favorite school. Each school tentatively retains those applicants who have the highest priority at that school. The remaining applicants are rejected.

Step k, $k \geq 2$: Each student rejected from a school at step $k-1$ applies to his next choice school. Each school then tentatively retains those applicants who have the highest priority among the new applicants as well as those tentatively retained at an earlier step. The remaining applicants are rejected.

The algorithm terminates when each student is tentatively retained at some school. Note that, in the DA, assignments in each step are temporary, until the last step. The DA has several desirable theoretical properties, most notably in terms of incentives and stability. First, the DA is strategy-proof (Roth, 1982; Dubins and Freedman, 1981). Furthermore, it produces the stable matching that is most favorable to each student. Although its outcome is not necessarily Pareto efficient, it is constrained efficient among the stable mechanisms.

3.4 Experimental Design

We design our experiment to compare the performance of the Boston and the DA mechanisms in a small scale ($m = 4$ per match), a medium scale ($m = 40$), and a large scale matching market ($m = 4,000$). We adapt our economic environment from the four-school treatment in Chen and Kesten (2014) to capture the key aspects of the school choice problem under complete information.

To study the impact of scale on mechanism performance, our experiment replicates the $m = 4$ per match economy to $m = 40$ and $m = 4000$. We use all-human sessions for the small and medium scale treatments. To make large scale matching market possible in the laboratory, we let human subjects play with computerized agents (robots) programmed to follow empirical strategies previously used by our human subjects under similar conditions. To check for any behavioral differences when human subjects play with other humans versus "empirical" robots, we also design an "empirical" human-robot treatment under the medium scale condition. Lastly, we use a human-vs-truthful-robot design, where all robots always reveal their true preferences, to study subject behavior when there is no uncertainty in opponent strategies and when the human best responses are easy to calculate.

3.4.1 Economic Environment

In our experiment, there are four schools, $s \in \{A, B, C, D\}$. Each school has 1, 10 or 1000 slots, corresponding to the scale of the matching market $m \in \{4, 40, 4000\}$, respectively. There are four types of students, $i \in \{1, 2, 3, 4\}$, with 1, 10 or 1000 of each type, again corresponding to the scale of the matching market.

The payoffs for each type are presented in Table 3.2. The square brackets, [ ], indicate the district school of the student, where she has higher priority than non-district applicants. Payoffs range from 16 points for the most preferred school to 5 points for the least-preferred school. Each student resides in the district of her second preferred school.

Table 3.2: Payoff Table

	A	B	C	D
Payoff to Type 1	[11]	7	5	16
Payoff to Type 2	5	[11]	7	16
Payoff to Type 3	7	16	[11]	5
Payoff to Type 4	5	16	7	[11]

The game preserves the properties of the four-school design in Chen and Kesten (2014): (1) no one lives in her top or bottom choices; (2) the first choice accommodation index, i.e., the proportion of first choices an environment can accommodate, is 1/2, with a fair amount of competition; and (3) the average efficiency under truth-telling is 81% for DA, reflecting the trade-off between stability and efficiency.

For our all-human small-scale (medium-scale) sessions, there are 12 (80) human subjects of four different types in each session. Subjects are randomly assigned to one of the four types at the beginning of the session and keep their type throughout the experiment. At the beginning of each subsequent round, they are randomly re-matched into three (two) groups of 4 (40) in each small-scale (medium-scale) session. Each 4-participant (40-participant) group has 1 (10) student(s) of each type.

For our human-robot medium-scale (large-scale) sessions, each human subject is paired with 39 (3999) robots. Including the human subject, there are 10 (1000) of each type per match. While a human-robot experimental session can start with any number of human subjects, we control the size of each session to between 19 and 21 human subjects.

In all treatments, we run the experiment for 20 periods to facilitate learning. To investigate whether participant strategies are conditional on their priority, we change the priority queue for each student type every five periods, as indicated in Table 3.10 in Appendix B. The priority lottery within each type is randomly drawn at the beginning of every five-period block and remains fixed for the block. A smaller lottery number indicates a higher priority.

Given this design, we compute the stable outcomes for the school choice game. Using the fact that all stable outcomes are contained between the student-optimal (μ^S) and school-optimal DA outcome (μ^C) (Roth and Sotomayor, 1990), μ^S and μ^C coincide in our environment. The unique stable outcome is that each student is assigned to his or her district school. Therefore, while the scale of the market increases, the number of stable outcomes remains as one.

$$\mu^{S/C} = \mu^S = \mu^C = \begin{pmatrix} 1 & 2 & 3 & 4 \\ A & B & C & D \end{pmatrix}; \mu^* = \begin{pmatrix} 1 & 2 & 3 & 4 \\ A & D & C & B \end{pmatrix}.$$

Furthermore, in the Boston mechanism, as the Nash Equilibrium outcomes equal to the stable outcomes under complete information (Ergin and Sönmez, 2006), there is a unique Nash equilibrium outcome $\mu^{S/C}$ in our experiment design, in which all students get into their district schools. By contrast, the DA mechanism has one more Nash equilibrium outcome, μ^*.

More generally, under our replication method, the number of stable outcomes and the school each type of student gets into do not change with the scale of the market.

Table 3.3: Truth-telling and Nash Equilibrium Outcomes in the Four-School Game

	Truthful Preference Revelation		Nash Equilibrium Outcomes	
	Boston	DA	Boston	DA
Block 1	not NE	dominant strategy	$\mu^{S/C}$	$\{\mu^{S/C},\mu^*\}$
Block 2	not NE	dominant strategy		
Block 3	not NE	dominant strategy		
Block 4	not NE	dominant strategy		

In addition, we also look at subjects' incentives to reveal their true preferences in large markets. Here, we find that truth-telling is not a Nash equilibrium strategy under the Boston mechanism for any of the priority queues. Table 3.3 summarizes the properties of this game.

3.4.2 Human-vs-Robot Design

A unique design feature in our study is the pairing of human subjects with computerized agents (”robots”) to create a large matching market in the laboratory setting. Previous studies have included robots in their experimental design. In auction experiments, robots follow the dominant strategy in multi-unit Vickrey and English auctions (Kagel and Levin, 2001), Vickrey and iBEA package auctions (Chen and Takeuchi, 2010) and a single-unit Vickrey auction (Davis et al., 2010). In Chen and Takeuchi (2010), robots follow a random bidding strategy. Lastly, in a VCM public goods game, robots follow a pre-determined set of actions to eliminate potential other-regarding behavior (Ferraro and Vossler, 2010).

In our experiment, we design two kinds of robots. Our “empirical robots” use strategies previously used by human subjects under similar conditions in the medium scale sessions, i.e., human subjects of the same type, in the same period, with a corresponding priority lottery number. For example, in the medium scale (40 participants per match) sessions, a robot of Type 2 with priority lottery number 15 in period 12, will randomly pick one out of the two choices human subjects of Type 2 with priority lottery number 15 made at period 12. Likewise, in the large-scale (4000 participants per match) sessions, 100 robots with priority lotteries 1401 to 1500, will randomly pick one of the two choices human subjects of Type 2 with priority lottery number 15 made at period 12, etc. Our human subjects in the human-robot sessions know how their robots counterparts’ strategies are drawn and used. To our knowledge, our empirical robot design is new to the experimental literature.

Our second type of robot is our “truthful robot,” who always ranks schools truthfully, regardless of its priority, a dominant strategy under the DA mechanism, but a “naïve” one under the Boston mechanism (Ergin and Sönmez, 2006). Again, our human subjects know the robot strategies.

There are three advantages associated with the human-vs-robots design. The first advantage is, of course, the scale. In an all-human experiment, the number of subjects in a group is limited by the capacity of the lab, whereas in a human-vs-robots design, the scale is only limited by the processing power of the computers. The second advantage is to reduce the strategic uncertainty faced by human subjects when they play robots with well defined strategies, which enables the experimenter to study human subjects responses toward opponents of different levels of strategic sophistication. The third advantage is related to statistical independence. Since there is no interaction among human subjects, each human subject is an independent observation.

3.4.3 Experimental Procedure

In each experimental session, each subject is randomly assigned an ID number and seated in front of the corresponding terminal in the laboratory. The experimenter reads the instructions aloud. Subjects are given the opportunities to ask questions, which are answered in public. We check subjects' understanding of the instructions by asking them to answer incentivized review questions at their own pace. After everyone finishes the review questions, the experimenter distributes the answers and goes over the answers in public. Afterwards, subjects go through 20 periods of the school choice experiment. In each period, each subject is asked to submit a full ranking of schools. Robots also submit the rankings of schools under certain strategies in the human-robot sessions. After all rankings are submitted, the server allocates the schools and informs each subject of his allocated school and respective payoff. At the end of the 20 periods, each subject fills out a demographics and strategy survey on the computer, and is then paid in private. Each session lasts approximately 90 minutes, of which 30 minutes are devoted to instruction. The experiment is programmed in z-Tree (Fischbacher, 2007) and Python.

Table 3.4 summarizes the features of the different experimental sessions. For

Table 3.4: Features of Experimental Sessions

Mechanisms	Composition	Match size	Robot strategies	#Sbj. × # sessions
Boston	All-human	4	n/a	12 × 4
		40	n/a	80 × 2
	Human-robot	40	Empirical	20 × 2
			Truthful	20 × 2
		4000	Empirical	20× 2
			Truthful	20× 2
DA	All-human	4	n/a	12 × 4
		40	n/a	80 × 2
	Human-robot	40	Empirical	20 × 2
			Truthful	20 × 2
		4000	Empirical	20 × 2
			Truthful	20 × 2

each mechanism, we conduct four independent sessions for the all-human small-scale treatments, two independent sessions for the all-human intermediate-scale treatments, and two independent sessions for each human-robot intermediate-scale and large-scale treatment, respectively. All sessions are conducted in Chinese at the Experiment Economics Laboratory and the Finance Simulation Laboratory at Beijing Normal University between June 2012 and May 2013. The subjects are students from Beijing Normal University and the Beijing University of Posts and Telecommunications. No subject participates more than once. This gives us a total of 12 independent sessions for the all-human treatments and 320 independent observations for the human-robot treatments. In total, 736 subjects participated in the experiment.

The exchange rate is 5 experiment points for 1 RMB for all sessions. Each subject also receives a participation fee of 5 RMB. The average earning (including participation fee) is 63.8 RMB.[6]

[6]The average wage of part-time work for university students in Beijing was around 30 RMB per hour. The exchange rate at the time of the experiment was around $1 = 6 RMB.

3.5 Results

In this section, we present results for our experiment, first for the effect of scale on individual behavior and then for the effect of scale on mechanism performance.

3.5.1 Individual Behavior

We first examine the extent to which individuals reveal their preferences truthfully, as well as any patterns in preference manipulation for the two mechanisms when the scale changes. In particular, when there is misrepresentation, we look at a common behavior: listing one's district school as one's first choice.

For the Boston mechanism, we define truth-telling as reporting an entire ranking as identical to one's true preference ranking. For the DA, however, we define truth-telling as reporting a ranking that is identical to the true preference ranking from the first choice to one's district school, as the remaining rankings are irrelevant under DA.

Based on Azevedo and Budish (2013), we expect that participants will misrepresent their preferences under the Boston mechanism regardless of the size of the market, whereas they will reveal their preferences truthfully under the DA mechanism regardless of the size of the market. Therefore, we formulate the following hypotheses on truth-telling:

Hypothesis 7 (Truth-telling: mechanism effect)**.** The proportion of truth-telling under the DA is greater than that under the Boston mechanism, regardless of scale.

Hypothesis 8 (Truth-telling: scale effect)**.** Scale has no effect on the proportion of truth-telling under either the Boston mechanism or the DA mechanism.

We first examine our all-human sessions in treatments with $m = 4$, and 40. Figure 3.1 presents the proportion of truth-telling (left panel) and district school bias (right panel) for each of the four all-human treatments. The ranking of mechanisms in the

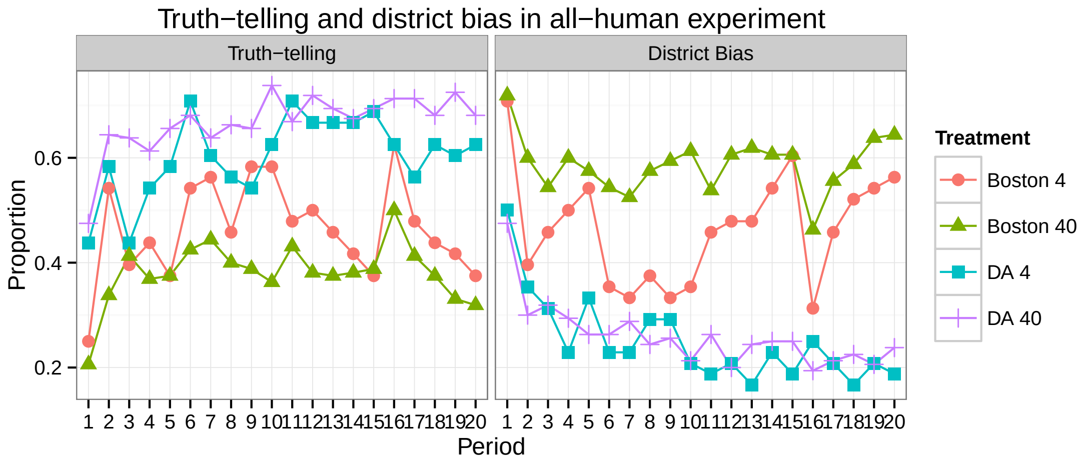


Figure 3.1: Truth-telling and district school bias in all-human treatments

truth-telling graph follows the order of DA-40 > DA-4 > BOS-4 > BOS-40, whereas the proportion of district school bias follows the reverse order.

Table 3.5 presents the results from four probit regressions investigating the scale and mechanism effects in the all-human (upper panel) and human-vs-empirical-robot treatments (lower panel). The dependent variable is truth-telling in specifications (1) and (2), and District School Bias in specifications (3) and (4). In the upper panel, the independent variables include (omitted variables in parentheses): Scale40 (Scale4), DA (Boston), DA × Scale40, Period, LotteryPosition, LotteryPosition × DA, LotteryPosition × Scale40. In each specification, standard errors are clustered at the session level. We summarize the results below.

Result 8 (Truth-telling: mechanism effect, all human)**.** *The proportion of truth-telling is 13% (28%) higher under the DA mechanism than under the Boston mechanism in the 4(40)-student environment.*

Support. In specification (1) of the upper panel of Table 3.5, we see that the coefficients of DA as well as DA × Scale40 are positive and significant, indicating that truth-telling under DA in the 4-student (40-student) matches is 13% (28%) higher

Table 3.5: Truth-telling and district school bias: probit regressions

	Truth-telling		District School Bias	
All-human treatments	(1)	(2)	(3)	(4)
Scale40	-0.081***	-0.122***	0.110***	0.166***
	(0.024)	(0.033)	(0.033)	(0.041)
DA	0.132***	0.111***	-0.211***	-0.205***
	(0.031)	(0.025)	(0.046)	(0.046)
DA × Scale40	0.146***	0.143***	-0.098**	-0.094*
	(0.041)	(0.041)	(0.049)	(0.049)
Period		0.004***		-0.004**
		(0.001)		(0.002)
LotteryPosition		-0.125***		0.127***
		(0.010)		(0.012)
LotteryPosition× DA		0.009		-0.003
		(0.006)		(0.009)
LotteryPosition × Scale40		0.017		-0.022*
		(0.012)		(0.013)
No. of observations	8,320	8,320	8,320	8,320
Human-vs-e-robots treatments	(1)	(2)	(3)	(4)
Scale4K	-0.018	0.020	0.012	-0.011
	(0.069)	(0.083)	(0.066)	(0.080)
DA	0.160**	0.040	-0.229***	-0.128
	(0.064)	(0.082)	(0.061)	(0.078)
DA × Scale4K	0.063	0.064	-0.072	-0.073
	(0.091)	(0.091)	(0.088)	(0.088)
Period		0.003*		-0.003*
		(0.002)		(0.002)
LotteryPosition		-0.097***		0.090***
		(0.017)		(0.015)
LotteryPosition × DA		0.047**		-0.039**
		(0.020)		(0.019)
LotteryPosition × Scale4K		-0.016		0.009
		(0.020)		(0.019)
No. of Human Observations	3,200	3,200	3,200	3,200

Notes: Standard errors in parentheses are clustered at the session (individual) level for the all-human (human-vs-empirical-robot) treatments; coefficients are marginal effects. *** p ¡ 0.01, ** p ¡ 0.05, * p ¡ 0.1.

compared to that in the corresponding sessions under the Boston mechanism.

By Result 8, we reject the null in favor of Hypothesis 7 under both $m = 4$ and $m = 40$. Our mechanism effect under $m = 4$ is consistent with the identical treatments in Chen and Kesten (2014), as well as with prior experimental studies of school choice in comparable markets. We also note that the gap between the two mechanisms more than doubles when we increase the scale by a factor of 10. This leads to our next result.

Result 9 (Truth-telling: scale effect $4 \to 40$)**.** *The proportion of truth-telling under the DA (Boston) mechanism significantly increases (decreases) when the number of students per match increases from 4 to 40.*

Support. In specification (1) of the upper panel of Table 3.5, the coefficient of DA $\times$ Scale40 is positive and significant, indicating that truth-telling in the 40-student matches under DA increases by 7% compared to that in the 4-student matches. In comparison, the coefficient of Scale40 is negative and significant, indicating that truth-telling in the 40-student sessions under Boston decreases by 8% compared to that in the 4-student sessions.

By Result 9, we reject Hypothesis 8 that scale has no effect on truth-telling under either mechanism. Indeed, we find a significant and sizeable scale effect for both mechanisms. Specifically, when the number of students per match increases from 4 to 40, participants exhibit less manipulation of the DA but more manipulation of the Boston mechanism. Consequently, we conclude that scale magnifies the performance gap between the two mechanisms.

We next explore the patterns of manipulation behind our scale effect. As documented in several experiments (Chen and Sönmez, 2006; Calsamiglia et al., 2010), the most prevalent form of manipulation in school choice experiments is the district school bias, where a student gives her district school a higher ranking than its true ranking.

In particular, when a district school is ranked as the top choice, a participant is guaranteed to be assigned to this school under the Boston mechanism. In our experiment, we find significant scale and mechanism effects in this type of manipulation.

Result 10 (District school bias: mechanism and scale effects)**.** *The proportion of DSB under the Boston mechanism is 21% (30%) higher than that under the DA mechanism in the 4-student (40-student) treatment. Under the DA (Boston) mechanism, the proportion of district school bias increases by 1% (11%) when the size of the match increases from 4 to 40.*

Support. In specification (3) in the upper panel of Table 3.5, the coefficient of DA is -0.21 ($p < 0.01$), that of DA $\times$ Scale40 is -0.098 ($p < 0.05$), and that of Scale40 is 0.11 ($p < 0.01$).

Furthermore, we observe a significant, albeit moderate, learning effect across periods, i.e., the proportion of truth-telling (DSB) increases (decreases) by about 0.4% per period ($p < 0.01$, specifications (2) and (4)). We also find that, under the Boston mechanism, a one point increase in the lottery position decreases the likelihood of truth-telling by 12.5% ($p < 0.01$, specification (2)), but increases the likelihood of DSB by 12.7% ($p < 0.01$, specification (4)), indicating a tendency to seek secure allocations as one's priority deteriorates. In comparison, such effect under the DA mechanism is smaller ($p < 0.01$, specifications (2) and (4)).

Our finding of a significant scale effect ($4 \rightarrow 40$) under both mechanisms is consistent with individual players adopting secure strategies as the size of the market increases. Under the Boston mechanism, district school bias guarantees that a player is matched with her second choice, regardless of what others do. Similarly, truth-telling is a weakly dominant strategy under the DA mechanism. Thus, regardless of what others do, a player cannot be worse off adopting this strategy. However, the unstable Nash equilibrium outcome under the DA requires half of the players to coordinate their manipulation. Doing so becomes increasingly difficult when the scale increases. Therefore, when the market size increases by a factor of 10, we observe increased adoption of the dominant strategy under the DA mechanism and increased district school bias under the Boston mechanism.

Next, we investigate the scale effect when the number of students per match increases from 40 to 4,000. As we transition from our all-human to human-vs-empirical-robot sessions, we note that our human subjects have the same behavioral responses in the all-human $m = 40$ treatments as in the human-vs-39-empirical-robots treatments. Table 3.12 in the Appendix presents p-values computed from Kolmogorov-Smirnov tests, comparing the probability distributions of submitted preference rankings between the all-human 40-student and human-versus-39-empirical-robots sessions for each mechanism period by period. These results show no statistically significant difference in participant behaviors in the respective treatments. These results indicate that humans do not respond differently when playing against robots whose strategies are drawn from the same human population.

In our human-vs-empirical-robot treatments, each human participant plays against either 39 or 3,999 robots whose strategies are randomly drawn from the all-human 40-student treatments. Figure 3.2 presents the proportion of truth-telling (left panel) and district school bias (right panel) for each of the human-vs-empirical-robot treatments. The ranking of treatments in the truth-telling graph follows the order of DA-4000 $>$ DA-40 $>$ BOS-40 $\sim$ BOS-4000, whereas the proportion of district school bias follows the reversed order.

The lower panel of Table 3.5 presents the results of four probit regressions investigating the scale and mechanism effects in the human-vs-empirical-robot treatments. The dependent variable is Truth-telling in specifications (1) and (2), and District School Bias in (3) and (4). The independent variables include (omitted variables in parentheses): Scale4K (Scale40), DA (Boston), DA $\times$ Scale4K, Period, LotteryPosition, LotteryPosition $\times$ DA, LotteryPosition $\times$ Scale4K. In each specification,

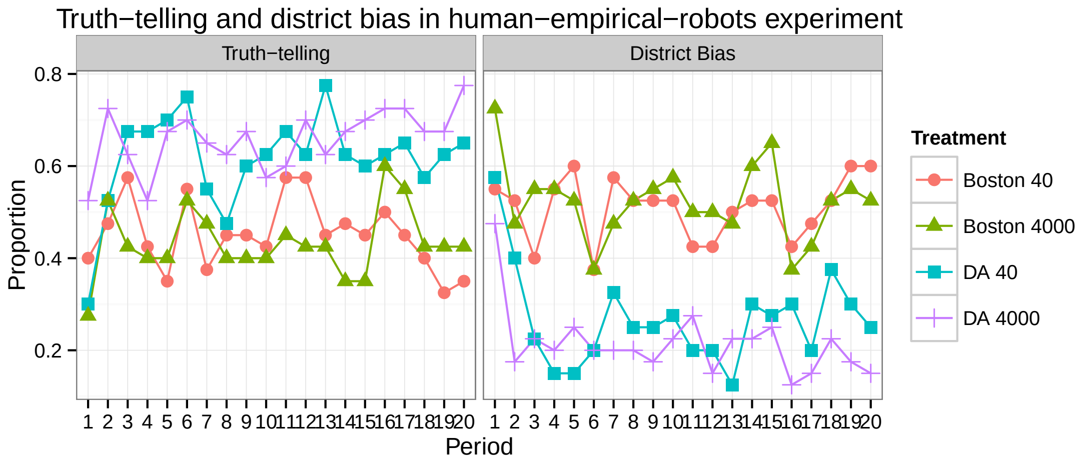


Figure 3.2: Truth-telling and district school bias in human-vs-empirical-robot treatments

standard errors are clustered at the individual subject level as each subject is an independent observation. We summarize the results below.

Result 11 (Truth-telling: mechanism effects, human-vs-empirical-robots)**.** *The proportion of truth-telling is significantly higher under the DA than under the Boston mechanism in the human-vs-empirical-robots treatments. Furthermore, this mechanism effect works through the subjects' lottery positions, i.e., a one-position increase in lottery position decreases the likelihood of truth-telling by 10% (5%) under the Boston (DA) mechanisms.*

Support. In specification (1) of Table 3.5 (lower panel), the coefficient of DA is positive and significant, indicating that the proportion of truth-telling under the DA in the 40-student sessions is about 16% higher compared to that in the corresponding sessions under the Boston mechanism. The effect of the lottery position under the Boston mechanism, indicated by the coefficient of LotteryPosition, is -0.097 ($p < 0.01$), while the effect under the DA mechanism is indicated by the sum of the coefficients of LotteryPosition and LotteryPosition × DA (0.047, $p < 0.05$) in specification (2).

By Result 11, we again reject the null in favor of Hypothesis 7 in our human-vs-empirical-robots treatments. However, we note that the coefficient of Scale4K (DA $\times$ Scale4K) is negative (positive) but insignificant, indicating a lack of scale effect. This leads to our next result.

Result 12 (Truth-telling: scale effect $40 \to 4K$)**.** *The proportion of truth-telling under the DA (Boston) mechanism increases (decreases) when the size of the match increases from 40 to 4,000; however, this effect is statistically insignificant.*

Support. In specification (1) of Table 3.5 (lower panel), the coefficient of Scale4K (DA $\times$ Scale4K) is negative (positive) but insignificant, indicating that the proportion of truth-telling in the 40-student sessions under Boston (DA) is not different from that in the 4000-student sessions.

By Result 12, we fail to reject Hypothesis 8 when the scale increases from 40 to 4,000 students per match in our human-vs-empirical-robots treatments.

Finally, looking at specification (3) in the lower panel of Table 3.5, we see that the proportion of district school bias under the DA mechanism is 23% less than that under the Boston mechanism when the size is 40 ($p < 0.01$), comparable to the magnitude of bias in the all-human 40-student per match treatments in Result 10. Additionally, we have analyzed human subjects' likelihood of best response in each treatment, and find both a mechanism (more best responses under the DA mechanism) and a scale effect (more best responses with increased scale). This analysis is relegated to Appendix C due to space constraints.

3.5.2 Aggregate Performance

In this section, we examine the scale and mechanism effects on measures of aggregate performance: the proportion of students admitted by both their reported and true first choice schools, individual rationality, efficiency and stability. For each measure, we first compare the all-human treatments ($m = 4, 40$), and then compare the human-vs-empirical-robots treatments ($m = 40, 4000$).

3.5.2.1 First Choice Accommodation and Individual Rationality

We first look at the first-choice accommodation rate, differentiating between the proportion of students admitted by their true versus reported first choice schools.

Based on Ergin and Sönmez (2006), we expect that a higher proportion of participants will receive their *true* first choices under the DA mechanism compared to the Boston mechanism. In contrast, we expect that a larger proportion will receive their *reported* top choices under the Boston mechanism (Chen and Kesten, 2014).

Hypothesis 9 (First-choice accommodation: mechanism effect)**.** If subjects play Nash equilibrium strategies, a higher proportion will receive (a) their true first choices under the DA mechanism, and (b) their reported first choices under the Boston mechanism.

Figures 3.4 and 3.5 in Appendix B present the first-choice accommodation rate for the all-human and the human-vs-empirical-robots treatments, respectively, comparing the proportion of subjects receiving their reported (top panel) and true first choices (bottom panel). We see that a greater proportion receive their reported top choice under the Boston mechanism, but the gap between the two mechanisms is much smaller when examining subjects' true top choices.

Table 3.6 reports the results from four probit specifications for students' true (1-2) and reported first choices (3-4) for both the all-human (upper panel) and the human-vs-empirical-robots treatments (lower panel). The independent variables for the upper panel include: Scale40 (Scale4), DA (Boston), DA $\times$ Scale40, Period, LotteryPosition, LotteryPosition $\times$ DA, and LotteryPosition $\times$ Scale40. The lower panel reports similar specifications except for Scale4K (Scale40). In the all-human sessions, we cluster the standard errors at the session level, whereas in the human-vs-empirical-robots treatment, we cluster the standard errors at the individual level as each human subject interacts with only robots.

Result 13 (First-choice accommodation: mechanism effect)**.** *Regardless of scale, both the reported and the true first-choice accommodation rates are significantly higher under the Boston than the DA mechanism.*

Support. In the upper panel of Table 3.6, the coefficients of DA are negative and significant ($p < 0.01$ in each of the four specifications), indicating a 5.8% (26.6%) reduction in receiving one's true (reported) first choice under the DA in the $m = 4$ treatments in specification 1 (3). Furthermore, the coefficients of DA $\times$ Scale40 are negative and significant ($p < 0.05$ in specifications (1) and (2), $p < 0.01$ in specifications (3) and (4)), indicating a further reduction of 4.4% (10.4%) of students receiving their true (reported) first choice under the DA in the $m = 40$ treatments.

In comparison, from the lower panel of Table 3.6, we see that the coefficients of DA are also negative and significant ($p < 0.01$ in each of the six specifications), indicating a 8.7% (27.9%) reduction in receiving one's true (reported) first choice under the DA in the $m = 40$ treatments in specification 1 (3). However, the coefficients of DA $\times$ Scale4K are negative but insignificant ($p > 0.10$), indicating no further reduction in receiving one's first choice under $m = 4,000$ treatments.

By Result 13, we reject the null in favor of Hypothesis 9 (b), but fail to reject the null in favor of Hypothesis 9 (a). Our findings confirm the theoretical prediction that the Boston mechanism places heavy weight on how students report their first choices. In our study, the true first-choice accommodation rate is lower under the DA because the dominant strategy equilibrium assigns everyone to his second choice.

Result 14 (First-choice accommodation: scale effect)**.** *The increase in scale from* $4 \rightarrow 40$ *significantly decreases (increases) the true (reported) first-choice accommo-*

Table 3.6: First choice accommodation: Probit regressions

Dependent Variable:	True First Choice		Reported First Choice	
All-Human Treatments	(1)	(2)	(3)	(4)
Scale40	-0.042***	-0.035**	0.055**	0.030
	(0.010)	(0.014)	(0.028)	(0.036)
DA	-0.058***	-0.067***	-0.266***	-0.279***
	(0.016)	(0.019)	(0.033)	(0.039)
DA × Scale40	-0.044**	-0.044**	-0.104***	-0.107***
	(0.020)	(0.018)	(0.039)	(0.039)
Period		0.000		-0.005***
		(0.001)		(0.001)
LotteryPosition		-0.034***		0.066***
		(0.004)		(0.013)
Lottery × DA		0.005		0.006
		(0.007)		(0.011)
Lottery × Scale40		-0.003		0.010
		(0.005)		(0.015)
Number of observations	8,320	8,320	8,320	8,320
Human-vs-e-Robots	(1)	(2)	(3)	(4)
Scale4K	0.007	0.010	0.025	0.010
	(0.019)	(0.029)	(0.052)	(0.066)
DA	-0.087***	-0.057*	-0.279***	-0.214***
	(0.021)	(0.032)	(0.050)	(0.067)
DA × Scale4K	-0.025	-0.023	-0.089	-0.090
	(0.030)	(0.030)	(0.073)	(0.073)
Period		-0.001		-0.005***
		(0.001)		(0.002)
LotteryPosition		-0.031***		0.054***
		(0.008)		(0.016)
Lottery × DA		-0.017		-0.026
		(0.012)		(0.020)
Lottery × Scale4K		-0.002		0.006
		(0.010)		(0.020)
Number of human obs.	3,200	3,200	3,200	3,200

Notes: Standard errors in parentheses are clustered at the session (individual) level for the all-human (human-vs-empirical-robots) treatments; reporting marginal effects. *** p ¡ 0.01, ** p ¡ 0.05, * p ¡ 0.1

dation rate by 4% (5%) under the Boston mechanism, but significantly decreases the rate by 8% (5%) under the DA mechanism. By contrast, the increase in scale from $40 \rightarrow 4,000$ *has no significant effect on either first-choice accommodation rate.*

Support. In the upper panel of Table 3.6, the coefficients of Scale40 are -0.04 ($p < 0.01$ in specification (1), $p < 0.05$ in (2)), and 0.055 ($p < 0.05$) in specification (3). Furthermore, the coefficients of DA $\times$ Scale40 are -0.044 ($p < 0.05$) in specifications (1) and (2), and -0.104 ($p < 0.01$) in specification (3) and (4). In comparison, in the lower panel of Table 3.6, none of the coefficients of Scale4K or DA $\times$ Scale4K is significant ($p > 0.10$).

Result 14 is consistent with the scale effect on truth-telling (Results 9 and 12). Increased district school bias under the Boston mechanism under the medium scale leads to a significantly higher (lower) proportion receiving their reported (true) first choice schools. By contrast, increased truth-telling under the DA leads to an increased proportion receiving their second choice, which is the dominant strategy equilibrium.

We next examine any scale and mechanism effect on individual rationality, i.e., the proportion of students placed at a school which is at least as good as her district school. Since students are guaranteed a seat at their district schools under the DA mechanism by playing the truth-telling strategy, which is not the case under the Boston mechanism, we have the following hypothesis.

Hypothesis 10 (Individual rationality)**.** More students will receive individually rational allocations under the DA than under the Boston mechanism regardless of scale.

Table 3.7 presents four probit specifications for the all-human (1-2) and human-vs-empirical-robots treatments (3-4), whereas Figure 3.6 in Appendix B presents the proportion of individual rational allocations in each treatment.

Result 15 (Individual rationality: mechanism and scale effects)**.** *Regardless of scale, the proportion of individually rational allocations is significantly higher under the DA*

Table 3.7: Individual rationality in all-human and empirical robots treatments: Probit regressions

Dependent Variable:	Individual Rationality			
Treatments:	All-human		Human-vs-empirical-robots	
	(1)	(2)	(3)	(4)
Scale	0.027***	0.048***	-0.015	-0.008
	(0.009)	(0.018)	(0.018)	(0.027)
DA	0.072***	0.105***	0.048***	0.094***
	(0.016)	(0.026)	(0.018)	(0.028)
DA × Scale	-0.007	-0.007	0.004	0.005
	(0.023)	(0.022)	(0.025)	(0.025)
Period		0.001**		0.003***
		(0.000)		(0.001)
Lottery		0.029***		0.025***
		(0.005)		(0.007)
Lottery × DA		-0.015*		-0.020**
		(0.008)		(0.008)
Lottery × Scale		-0.009		-0.003
		(0.006)		(0.008)
Observations	8,320	8,320	3,200	3,200

Notes: Standard errors in parentheses are clustered at the session (individual) level for the all-human (human-vs-empirical-robots) treatments; reporting marginal effects. *** p ¡ 0.01, ** p ¡ 0.05, * p ¡ 0.1

than the Boston mechanism. The increase in scale from $4 \to 40$ *significantly increases the proportion of individually rational allocations under both mechanisms.*

Support. In Table 3.7, the coefficients of DA are 0.072 and 0.048 for all-human and human-vs-empirical robots treatments respectively ($p < 0.01$ in all specifications). In addition, increasing the scale in all-human treatment from 4 to 40 also increase the proportion of students receiving individually rational allocations for both mechanisms ($p < 0.01$).

3.5.2.2 Efficiency and Stability

To study the effect of scale on mechanism efficiency, we define payoff-based efficiency in our all-human treatments as the sum of the individual payoffs in each match in each period, normalized according to the following formula:

$$\text{Payoff-based Efficiency} = \frac{\text{Actual sum of payoffs} - \text{Minimum sum of payoffs}}{\text{Maximum sum of payoffs} - \text{Minimum sum of payoffs}}.$$

In our environment, the minimum sum of payoffs is 240 for the 40-student environment, and 24,000 for the 4,000-student environment. Likewise, the maximum sum of payoffs is 540 for the 40-student environment, and 54,000 for the 4,000-student environment.

In the human-vs-empirical-robots treatments, the equivalence of the payoff-based efficiency is the human players' expected payoffs. Specifically, for each human subject in each period, we recombine the subject with 39 (or 3999) group members in the same period, of the same type, and with the same (or corresponding) lottery number, taken from the all-human session. While there are 2^{39} (2^{3999}) possible recombinations, to reduce computation, we randomly generate 2000 groups for each subject. We then estimate the expected payoff for each human subject by averaging her payoffs over the 2000 group recombinations.

Tables 3.8 and 3.9 report the results of our OLS regression analysis of efficiency and justified envy in the all-human and human-vs-empirical-robots treatments, respectively. Specifications (1) and (2) in Table 3.8 indicate that increasing the scale from 4 to 40 significantly reduces payoff-based efficiency under the Boston mechanism (-0.014, $p < 0.01$), but has no further significant effect on the DA. Similarly, specifications (1) and (2) in Table 3.9 indicate that further increasing the scale from 40 to 4000 has no effect on the efficiency of either mechanism.

Result 16 (Efficiency: scale effect)**.** *From* $4 \to 40$*, payoff-based efficiency under the Boston (DA) mechanism decreases by 1.4% (3.3%). Further increasing the scale to 4000 has no additional effect on the efficiency of either mechanism.*

Support. In specifications (1) and (2) of Table 3.8, the coefficients for both Scale40

Table 3.8: Payoff-based efficiency and justified envy in all-human treatments: OLS regressions

Dependent Variable:	Payoff-based Efficiency		Justified Envy	
	(1)	(2)	(3)	(4)
Scale40	-0.014***	-0.014***	-0.006	-0.006
	(0.004)	(0.004)	(0.022)	(0.022)
DA	0.004	0.004	-0.076***	-0.076***
	(0.011)	(0.011)	(0.021)	(0.021)
DA × Scale40	-0.019	-0.019	-0.024	-0.024
	(0.012)	(0.012)	(0.028)	(0.028)
Period		0.000		0.000
		(0.001)		(0.001)
Constant	0.690***	0.688***	0.140***	0.141***
	(0.003)	(0.012)	(0.018)	(0.024)
Observations	640	640	640	640
R-squared	0.018	0.019	0.096	0.096

Notes: Standard errors in parentheses are clustered at the session level. *** p ¡ 0.01, ** p ¡ 0.05, * p ¡ 0.1

(-0.014, $p < 0.01$) and Scale40 + DA × Scale40 (-0.033, $p = 0.0147$) are negative and significant. In comparison, none of the coefficients for the scale variables is significant in specifications (1) and (2) in Table 3.9.

The small but significant decrease in efficiency is likely due to the increased manipulation of preferences under the Boston mechanism, as the scale increases from 4 to 40. While the efficiency comparisons between the DA and the Boston mechanism depend on the environment, in our environment, we do not observe a mechanism effect.

Finally, we investigate the mechanism and scale effect on stability. Empirically, we evaluate mechanism stability by calculating the proportion of students in each group who exhibit justified envy toward at least one other student. As the DA mechanism is stable while the Boston mechanism is not, we expect that the proportion of students exhibiting justified envy will be lower under the DA at any given scale.

Table 3.9: Efficiency and justified envy in human-vs-empirical-robots treatments: OLS regressions

Dependent Variable:	Expected Payoff		Justified Envy	
	(1)	(2)	(3)	(4)
Scale4K	-0.052	0.004	0.012	-0.001
	(0.224)	(0.346)	(0.023)	(0.039)
DA	-0.277*	-0.231	-0.063***	-0.135***
	(0.162)	(0.279)	(0.019)	(0.035)
DA × Scale4K	0.055	0.055	-0.012	-0.012
	(0.240)	(0.240)	(0.027)	(0.027)
Period		-0.002		-0.004***
		(0.008)		(0.001)
Lottery		-0.048		-0.037***
		(0.064)		(0.009)
Lottery × DA		-0.019		0.029***
		(0.071)		(0.009)
Lottery × LargeScale		-0.022		0.005
		(0.071)		(0.009)
Constant	11.172***	11.312***	0.116***	0.251***
	(0.154)	(0.295)	(0.017)	(0.038)
Observations	3,200	3,200	3,200	3,200
R-squared	0.005	0.007	0.017	0.037

Notes: Standard errors clustered at individual level, shown in parentheses. *** p ¡ 0.01, ** p ¡ 0.05, * p ¡ 0.1

Hypothesis 11 (Stability: mechanism effect)**.** The proportion of students exhibiting justified envy is lower under the DA than under the Boston mechanism at any given scale.

Theoretical work on the scale effect implies that, in general, the set of stable matchings decreases for both mechanisms when the scale increases. However, since we have a unique stable outcome in our environment, we do not expect any scale effect on stability.

Hypothesis 12 (Stability: scale effect)**.** Scale has no effect on the proportion of students exhibiting justified envy under either the Boston or the DA mechanism.

Figure 3.3 presents the results for the average proportion of students with justified envy in our all-human treatments (left panel) and the average expected probability of having justified envy in the human-vs-empirical-robots (right panel) treatments. For the human-vs-empirical-robot treatments, we investigate stability through the use of simulations. More specifically, we randomly generate groups of size 40 or 4000 for each period and mechanism by drawing on human subject behavior from the $m = 40$ all-human sessions. For each simulated group, students are assigned to schools based on their reported preferences. We then randomly generate 2000 groups for each human observation, and calculate the probability that a student may exhibit justified envy in the 2000 simulated matchings. The results are summarized below.

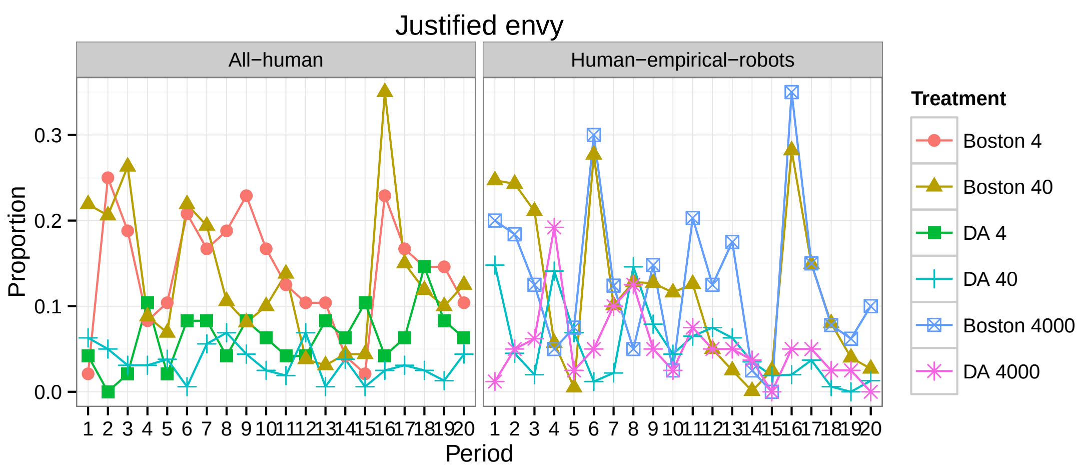


Figure 3.3: Justified envy in all-human (left panel) and human-vs-empirical-robots (right panel) treatments

Result 17 (Stability: mechanism effect)**.** *For each given market size, the proportion of students exhibiting justified envy is significantly lower under the DA than under the Boston mechanism.*

Support. In specifications (3) and (4) of Table 3.8, the coefficients for DA are negative and significant (-0.076, $p < 0.01$), whereas neither of the coefficients for DA $\times$

Scale40 is significant. Likewise, the coefficients for DA are negative and significant (-0.063 and -0.135, $p < 0.01$) in specifications (3) and (4) in Table 3.9, whereas neither of the coefficients for DA $\times$ Scale4K is significant.

By Result 17, we reject the null in favor of Hypothesis 11. We also observe that the none of the scale variables in Tables 3.8 and 3.9 is significant, indicating the absence of any scale effect on stability, which is consistent with the theoretical predictions. Thus, we fail to reject Hypothesis 12.

3.6 Conclusion

In practice, matching mechanisms are implemented across all size markets, from hundreds of students in course allocations, to tens of thousands in school choice, to millions in centralized college admissions markets. However, it is unclear whether market size plays a role in either participant behavior or mechanism performance. Complicating the issue is the fact that most laboratory matching experiments work with only small-scale markets.

In this paper, we use the school choice context to present a laboratory experiment investigating how matching market scale affects individual behavior and the performance of the Boston and the DA mechanisms. Specifically, we investigate the scale effect by varying the number of students per match from 4 to 40, and then from 40 to 4,000. The results of our study reveal a significant scale effect on individual behavior when we increase the scale from 4 to 40. Specifically, subjects become more likely to reveal their preferences truthfully under the DA mechanism but less likely to do so under the Boston mechanism. Thus, we conclude that the well-known preference manipulation gap between the DA and the Boston mechanisms in small markets increases when the size of the market increases by tenfold. Furthermore, we find that this magnification effect is due to an increased likelihood of individual players to adopt secure strategies in larger markets. Under the Boston mechanism, this secure strategy comes through in district school bias, while under the DA mechanism, it comes through in truthful preference revelation. However, increasing the market size to 4,000 students per match has no additional effect on individual behavior.

Examining mechanism performance at the aggregate level, we find that, while the Boston mechanism assigns more students to their first choices, the DA mechanism assigns more students to their top two choices, i.e., individually rational outcomes. Furthermore, the DA mechanism has a large and significant stability advantage over the Boston mechanism, which remains robust across all size markets. A further increase to 4,000-student per match makes no additional change on the aggregate performance of either mechanism.

In addition to our finding regarding participant behavior and mechanism performance, we contribute to the literature by using robots to create our large matching markets in a standard laboratory setting. Specifically, we group human subjects with robots whose strategies are drawn from empirical human strategies obtained from our all-human sessions. Previous experiments with robots endow the robots with either a dominant or random choice strategy. In comparison, we endow them with empirical strategies, providing a novel solution to the problems associated with conducting large group experiments in the lab.

Lastly, our findings provide insight into the ongoing theoretical debate on when market size becomes a factor in matching mechanism performance. Our results indicate that "large" might be much smaller than predicted in the theoretical literature.

Table 3.10: Priority Queue

	Scale	Type 1	Type 2	Type 3	Type 4
Block 1: periods 1-5	4 students	1	2	3	4
	40 students	1~10	11~20	21~30	31~40
	4000 students	1~1000	1001~2000	2001~3000	3001~4000
Block 2: periods 6-10	4 students	4	1	2	3
	40 students	31~40	1~10	11~20	21~30
	4000 students	3001~4000	1~1000	1001~2000	2001~3000
Block 3: periods 11-15	4 students	3	4	1	2
	40 students	21~30	31~40	1~10	11~20
	4000 students	2001~3000	3001~4000	1~1000	1001~2000
Block 4: periods 16-20	4 students	2	3	4	1
	40 students	11~20	21~30	31~40	1~10
	4000 students	1001~2000	2001~3000	3001~4000	1~1000

Table 3.11: Proportion of truthful preference revelation and district school bias

All-human treatments

	Truthful Preference Revelation				District School Bias			
	BOS-4	DA-4	BOS-40	DA-40	BOS-4	DA-4	BOS-40	DA-40
Session1	0.421	0.696	0.353	0.644	0.550	0.183	0.607	0.259
Session2	0.483	0.596	0.409	0.693	0.375	0.271	0.568	0.261
Session3	0.492	0.554			0.496	0.204		
Session4	0.463	0.567			0.442	0.338		
Average	0.465	0.603	0.381	0.668	0.466	0.249	0.588	0.260

Human-vs-empirical-robot treatments

	Truthful Preference Revelation				District School Bias			
	BOS-40	DA-40	BOS-4K	DA-4K	BOS-40	DA-40	BOS-4K	DA-4K
Session1	0.419	0.602	0.498	0.708	0.538	0.283	0.460	0.216
Session2	0.487	0.629	0.368	0.614	0.476	0.247	0.585	0.212
Average	0.453	0.616	0.433	0.661	0.507	0.265	0.523	0.214

3.7 Appendix

3.7.1 Experimental Design

Results - Individual Behavior

We present two additional tables for individual behavior. Table 3.11 presents the proportion of truthful preference revelation and district school bias at the session level.

Table 3.12: K-S test for all-human 40-student and human-vs-39-empirical-robots sessions

Boston-40 (H_a : All human $\neq$ Empirical robot)				DA-40 (H_a : All human $\neq$ Empirical robot)			
Period	P-value	Period	P-value	Period	P-value	Period	P-value
1	0.218	11	0.914	1	0.078	11	0.627
2	0.997	12	0.814	2	0.254	12	0.989
3	0.868	13	0.868	3	0.949	13	0.691
4	0.914	14	0.997	4	0.563	14	0.974
5	0.997	15	0.949	5	0.754	15	0.814
6	0.389	16	0.989	6	1	16	0.691
7	1	17	0.868	7	0.868	17	0.627
8	0.949	18	0.627	8	0.997	18	0.218
9	0.502	19	0.914	9	1	19	0.814
10	0.295	20	0.949	10	0.914	20	0.989

Notes: p-values calculated for H_a : All human $\neq$ Empirical robot, with K-S test at the individual level.

Table 3.12 presents the results of the Kolmogorov-Smirnov tests of the equality of distributions of strategies between the all-human 40-student treatment and the corresponding human-vs-39-empirical-robots treatment, period by period. We do not pool across all periods because of the interdependency of strategies across periods.

Results - Aggregate Performance

Figures 3.4 and 3.5 present the first-choice accommodation rate for the all-human and the human-vs-empirical-robots treatments, respectively, comparing the proportion of subjects receiving their reported (top panel) and true first choices. We see that a greater proportion receive their reported top choice under the Boston mechanism, but the gap between the two mechanisms is much smaller when examining subjects' true top choices.

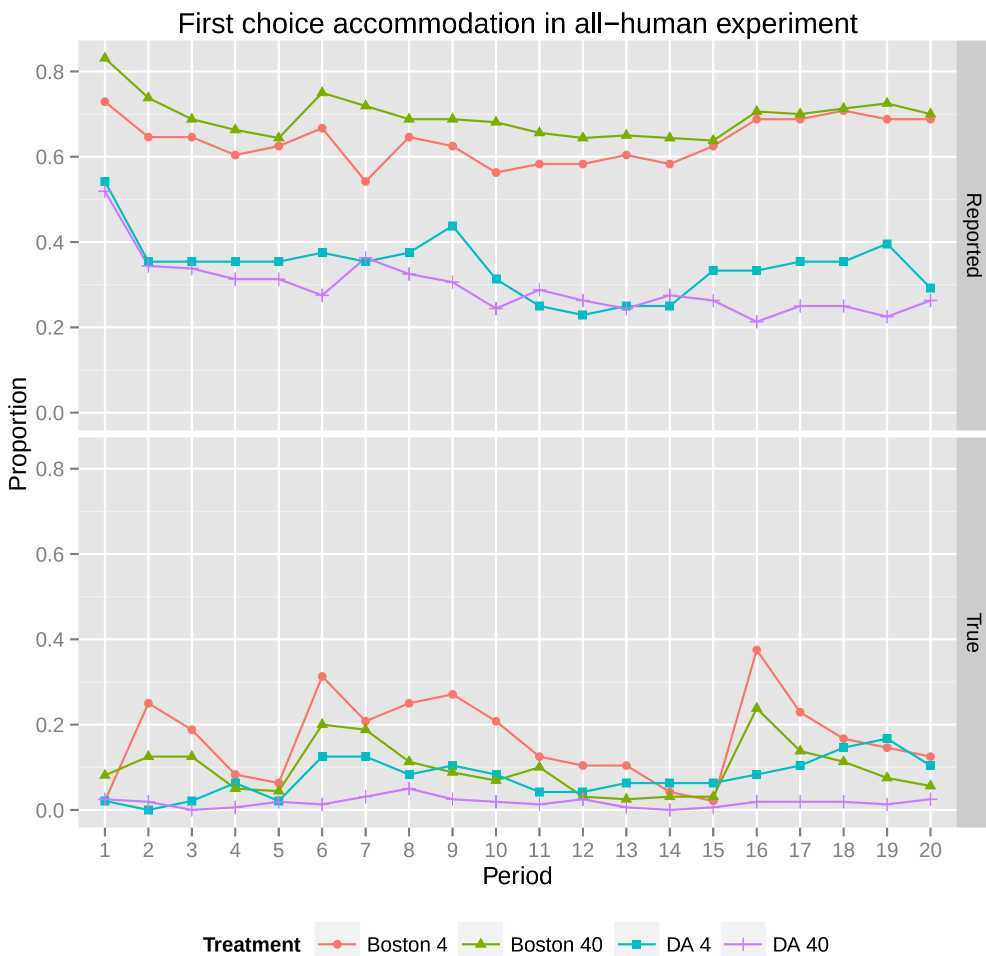


Figure 3.4: First choice accommodation in all-human experiment

Figure 3.6 presents the proportion of individually rational allocations in each

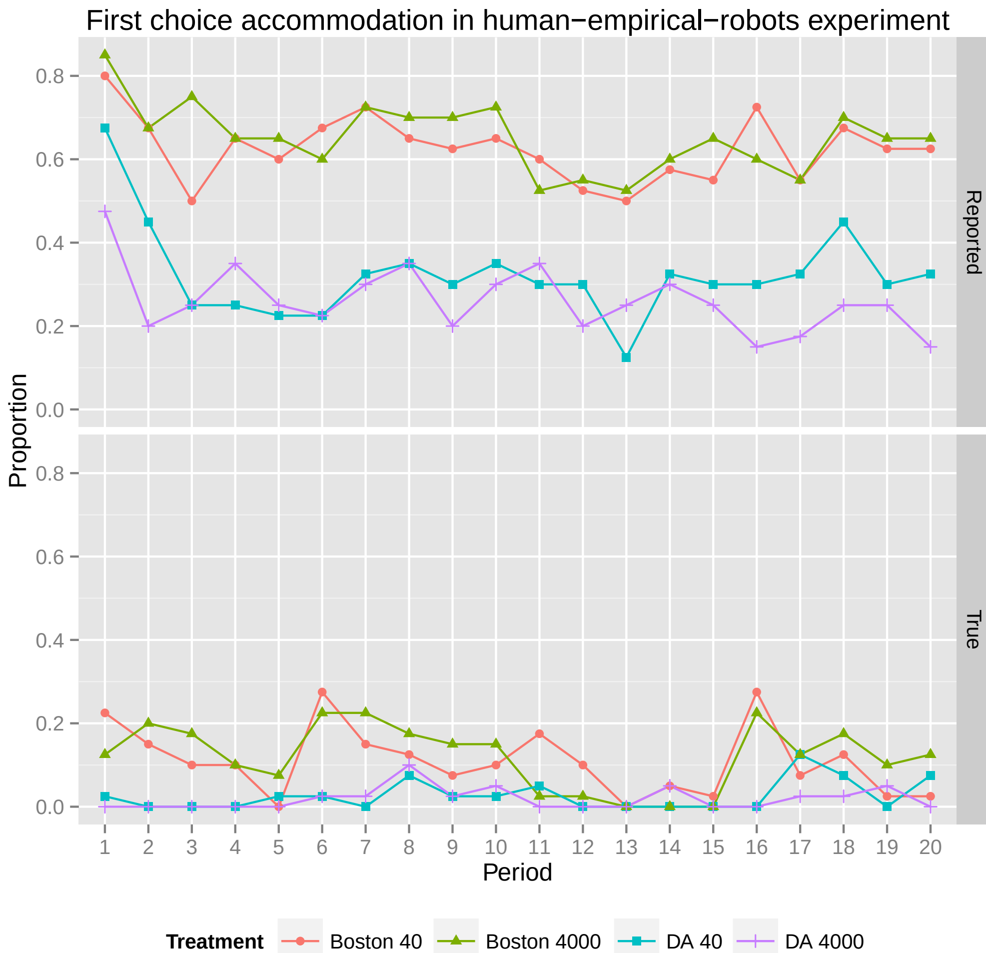
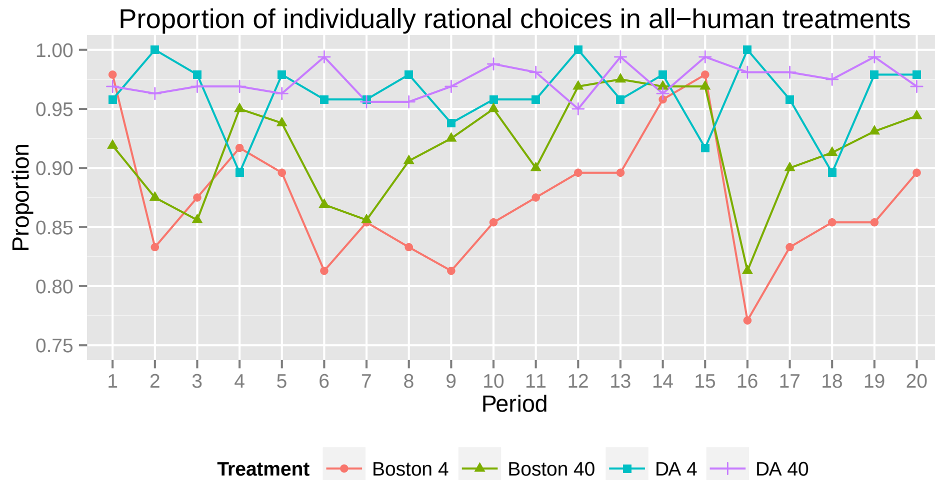


Figure 3.5: First choice accommodation in human-vs-empirical-robots experiment

treatment.



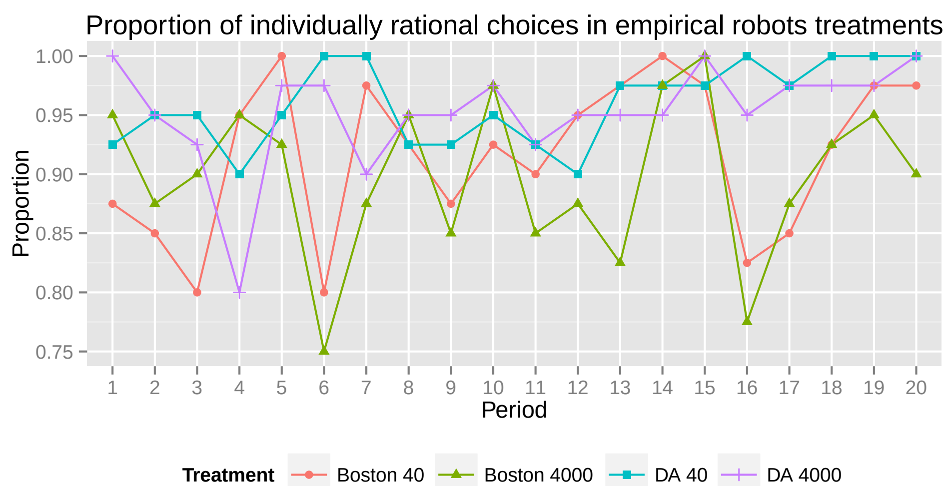


Figure 3.6: Individual rationality in all-human (upper panel) and human-vs-empirical-robots (lower panel) treatments

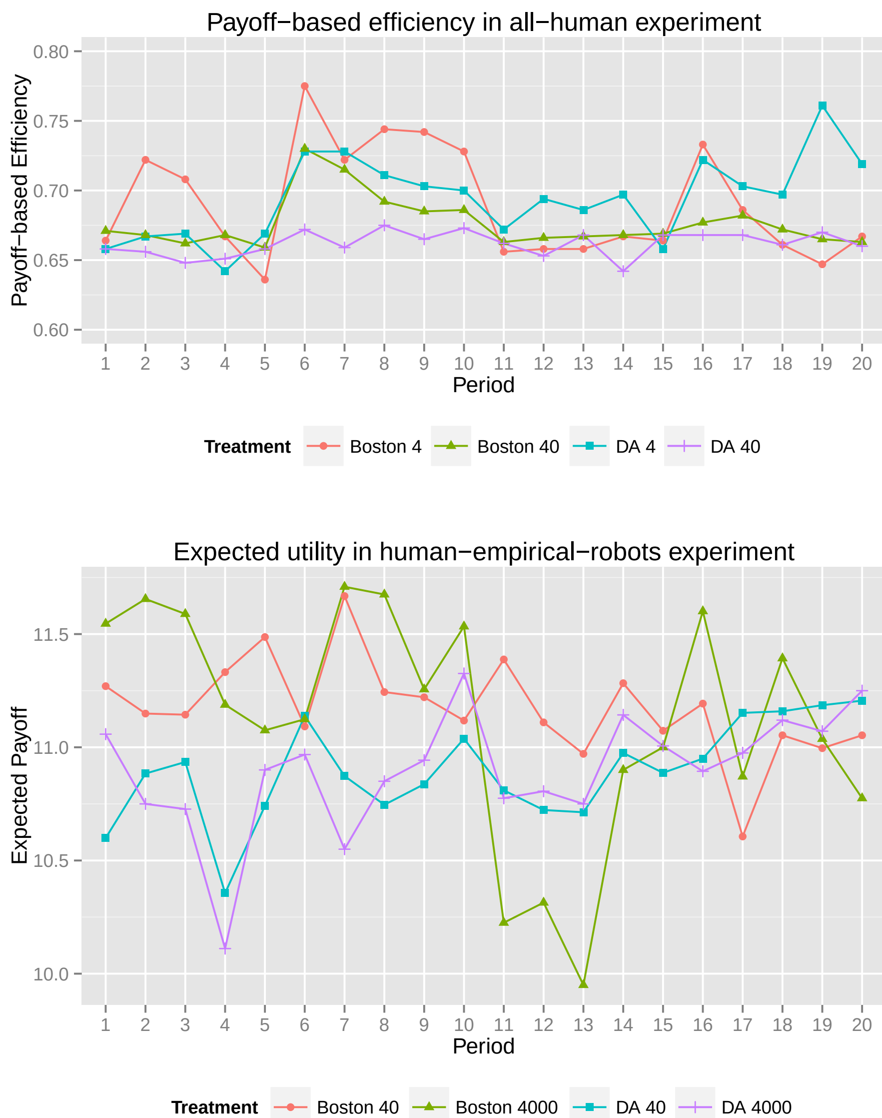


Figure 3.7: Payoff-based efficiency in all-human treatments (upper panel) and expected payoff of human players in human-vs-empirical-robots treatments (lower panel)

CHAPTER IV

Chinese College Admissions Reforms: Experimental and Empirical Evaluations

4.1 Introduction

College admissions mechanisms have significant impact on student education experiences and labor market outcomes in many countries. Each year approximately 10 million high school seniors compete for 6 million seats through a centralized college admissions system in China, making it the largest centralized matching market in the world. In the last fifteen years, many provinces have moved from an Immediate Acceptance mechanism to various versions of a 'parallel' mechanism. Because of the significance of the effect of different mechanisms on admission results, the temporal and regional variations in Chinese college admissions serve an important role in evaluating the performance of various matching mechanisms.

This higher education selection process in countries like China, Japan, Korea, and Turkey are organized in an centralized way, which is different from the college admissions in the United States, which is decentralized in nature. This *centralized* college admissions problem, also called student placement problem (Balinski and Sönmez, 1999), has some unique properties, compared to other matching markets such as school choice. One major differentiator is that students' priorities in college matching are usually determined by their test scores in a standardized college entrance exam, therefore their priorities are identical at every college within a market. Moreover, the prestige of a college is a major concern for all students. Take China for an example: top universities with national prestige are sought after by almost all students. This phenomenon implies that the preferences of students are highly correlated. As a result, college admissions are typically much more competitive than student allocation in a school district. These two factors significantly affect how students strategize how they report their preferences under different mechanisms.

Testing matching theories in this market using experimental and empirical methods both have their advantages and disadvantages. While testing theories using field data has higher external validity, since in the field, only students' *stated* preferences are revealed, without any information regarding students' *true* preferences. It is therefore important to provide experimental insights into the college admissions problem and complement current theory literature, as well as any empirical study, since experiments allow us to induce true preferences and to evaluate the performance of college admissions mechanisms more accurately. However, laboratory experiment alone could not capture the large scale and high stakes nature of the real world college admissions.

Therefore, in this paper, we study how different matching mechanisms in centralized college admissions affect students' strategies and matching outcomes. We study this using both a laboratory experiment and a natural experiment. The natural experiment comes from a change in matching mechanisms in China's Sichuan Province between the year 2008 and 2009. Since when students participate in college admissions is mostly determined by birth, while the mechanism change happened much later, students do not self-select into different mechanisms, which eliminates selection bias to a large extent. Moreover, the mechanism change only affected a portion of the students, therefore it allows us the use of difference-in-differences estimators. Both the results from the laboratory experiment and from the field confirm some of the theoretical predictions of the mechanisms we study.

The rest of this paper is organized in the following way: Section 4.2 provides an overview of the related literature. Section 4.3 outlines the theoretical properties of the mechanisms we study; Section 4.4 discusses the design and procedure of this experiment, and its associated theoretical results; Section 4.5 presents the experimental results; Section 4.6 describes the background and the dataset for the natural experiment; Section 4.7 presents the empirical results; finally, Section 4.8 concludes.

4.2 Related Literature

Gale and Shapley (1962) pioneered the study of the matching problems of marriage and college admissions, and proposed the first stable matching mechanism. Since then, matching theory has been applied to different domains, including school choice (Abdulkadiroğlu and Sönmez, 2003; Abdulkadiroğlu et al., 2005a,b; Chen and Sönmez, 2006), hospital matching residents program (Roth, 1984, 1990; Kagel and Roth, 2000; Niederle and Roth, 2003), house allocation (Abdulkadiroğlu and Sönmez, 1998; Abdulkadiroğlu and Sönmez, 1999; Chen and Sönmez, 2002), and kidney exchange (Roth et al., 2005). A success story in matching research is the transition of the New York and Boston public school systems (Abdulkadiroğlu et al., 2005a,b) from the Immediate Acceptance mechanism (also called the Boston mechanism) (Abdulkadiroğlu and Sönmez, 2003), to the Deferred Acceptance mechanism (Gale and Shapley, 1962).

One strand of matching literature that is closely related to the centralized college admissions problem is the school choice problem, and it has been extensively studied using laboratory experiments. The experiment by Chen and Sönmez (2004) praises the Deferred Acceptance mechanism for both its stability and efficiency. Pais and Pintér (2008) find in the lab that the Top Trading Cycles mechanism is more efficient and less vulnerable to manipulation than the Immediate Acceptance mechanism and the Deferred Acceptance mechanism in the school choice scenario, and is also less sensitive to the amount of information that participants hold. Featherstone and Niederle (2014) show that incomplete information on the students' side changes the efficiency and truthfulness of mechanisms. Calsamiglia et al. (2010) find that constraining students' ability to reveal their preferences leads to more manipulation and lower efficiency. In the same paper, Chen and Kesten (2014) also experimentally study the Chinese Parallel mechanism in the school choice context, and find that while this mechanism is slightly less efficient that the Immediate Acceptance mechanism, it is less manipulable and more stable.

In addition to laboratory experiments, properties of matching mechanisms have also been tested in the field. Mongell and Roth (1991) studies the "preferential bidding system" that matches students to sororities, and find that preference manipulation can prevent an unstable mechanism from unraveling. Braun et al. (2010) study the centralized college admissions in Germany, and find that in that system, truth-telling by failing to understand the mechanism hurts students with good grades. Carvalho and Magnac (2010) and He (2014) take a structural approach to the problem. The former find that the mechanism used to select majors in Brazilian universities are unstable and prone to strategic manipulation. The latter uses the data from middle school admissions in Beijing to study the strategic behavior of parents under the Immediate Acceptance (Boston) mechanism, and suggests that teaching parents to play best response may be more beneficial than switching to the Deferred Acceptance mechanism. More closely related to this paper, Wu and Zhong (2014) compare variations in provincial matching mechanisms in China by looking at students' achievement in colleges after they have been admitted. They find that better students are admitted to a top college when they participate under the Immediate Acceptance mechanism without know their test scores in the entrance exam.

This paper contributes to the college admissions and the broader matching literature by testing a common set of hypotheses using a laboratory experiment and a natural experiment simultaneously.

4.3 The Mechanisms and Their Theoretical Properties

In this section, we introduce the three mechanisms for college admissions and summarize the main theoretical results pertaining to this family of mechanisms. These theoretical results form the basis for the design of our laboratory experiments and our empirical evaluation of these mechanisms using naturally occurring field data.

4.3.1 The college admission problem and three mechanisms

Formally, the college admissions problem (Balinski and Sönmez, 1999) is a tuple (S, C, P_s, P_C), where: (1) a set of students $S = \{s_1, ..., s_n\}$; (2) a set of colleges $C = \{c_1, ..., c_m\}$; (3) a capacity vector $q = (q_{c_1}, ..., q_{c_m})$ where q_{c_i} is the capacity of college c_i ; (4) a list of student preferences $P_S = (P_{s_1}, ..., P_{s_n})$ where P_{s_i} is the strict preference relation of student s_i over colleges including the no-college option; (5) a list of college preferences $P_C = (P_{c_1}, ..., P_{c_m})$ where P_{c_i} is the strict preference relation of college c_i over a set of students, which is determined by students' scores in a centralized college entrance exam, and therefore $P_{c_i} = P_{c_j}$, $\forall i, j \in \{1, ..., m\}$. A matching μ is an allocation of college slots to students such that the number of students assigned to any college does not exceed its quota: $\mu : S \to C \cup \{c_0\}$ such that $|\mu^{-1}(c)| \leq q_c$, for all $c \in C$.

A matching μ is *stable* if there is no student-college pair (c, s) such that student s prefers college c to the college he is assigned to, and college c prefers student s to at least one student who is assigned to it. That is, a matching μ is stable if there is no student s_k and college c_q such that both $c_q P_{s_k} \mu(s_k)$ and $s_k P_{c_q} \mu(c_q)$. A matching is *Pareto efficient*, if every student or college likes the matching at least as well as any other matching, and at least one student or college strictly prefer this matching to any other matchings. Moreover, one matching Pareto dominates another, if every student or college likes the first matching at least as well as the other matching, and at least one student or college strictly prefers it to the other.

A college admissions mechanism, or simply a mechanism, selects a matching for each problem. A mechanism is Pareto efficient (stable) if it always selects Pareto efficient (stable) matchings. In addition, a mechanism is strategy-proof if no student can gain by misrepresenting her preferences. We now describe the three mechanisms that are central to our study.

4.3.1.1 Immediate Acceptance mechanism (IA)

The Immediate Acceptance mechanism (IA), sometimes called the Boston mechanism, is commonly used in the U.S. public school choice. It was the prevalent college admissions mechanism in China in the 1980s and 1990s. The IA mechanism is a special case of the priority matching mechanisms identified by Roth (1991). The outcome of the IA mechanism in centralized college admissions can be calculated via the following algorithm for a given problem:

For each college, a priority ordering of students is determined by the ranking of students' test scores.

Step 1: Each student applies to his first choice. For each college, consider the students who have listed it as their first choice and assign seats of the college to these students one at a time following their priority order until either there is no seat left or there is no student left who has listed it as her first choice.

Step k ($k > 1$): Each student who was rejected in the previous step applies to the next college on his list. For each college with available seats, consider those students who have listed it as their kth choice and assign the remaining seats to these students one at a time following their priority order until either there is no seat left or there is no seat or student left.

The algorithm terminates when no more applications can be rejected or no more seats are left. Importantly, the assignments in each step are final. Based on this feature, an important critique of the IA mechanism is that it gives students strong incentives to misrepresent their preferences. As a student who has high priority for a school may lose it if she does not list it as her first choice, the IA mechanism forces students to make hard and risky strategic choices (see e.g., Abdulkadiroğlu and Sönmez 2003; Ergin and Sönmez 2006; Chen and Sönmez 2006; He 2014).

4.3.1.2 The Deferred Acceptance mechanism (DA)

A second matching mechanism is the Student-Optimal Deferred Acceptance mechanism (Gale and Shapley, 1962), which finds the stable matching that is most favorable to each student. Its outcome can be calculated via the following *Deferred Acceptance (DA) algorithm* for a given problem:

Step 1: Each student applies to his first choice. Colleges temporarily retain their most preferred students within their enrollment capacities, and rejects the rest.

Step k ($k > 1$): Each student who was rejected in the previous step applies to the next college on his list. Each college then considers the students it has already temporarily retained along with the new applicants, retains their most preferred within their enrollment capacities, and rejects the rest.

The algorithm terminates when no more applications can be rejected or no more seats are left.

Note that, in DA, assignments after each step are temporary until the last step. DA has several desirable theoretical properties, most notably in terms of incentives and stability. Under DA, it is a dominant strategy for students to state their true preferences (Roth, 1982; Dubins and Freedman, 1981). Furthermore, it is stable.

Although it is not Pareto efficient, it is the most efficient among the stable college admissions mechanisms.

In practice, the DA has been the leading mechanism for school choice reforms. For example, the DA has been adopted by New York City (in 2003) and Boston public school systems (in 2005), which had suffered from congestion and incentive problems from their previous assignment systems, respectively (Abdulkadiroğlu et al., 2005a,b).

4.3.1.3 The Chinese Parallel Mechanism (PA)

The Chinese parallel mechanism was first implemented in Hunan Tier 0 college admissions in 2001.[1] From 2001 to 2012, variants of the mechanism have been adopted by 28 provinces as the parallel college admissions mechanisms to replace the IA mechanisms (Wu and Zhong, 2014).

Chen and Kesten (2014) present a parametric family of *application-rejection* mechanisms where each member is characterized by a positive number $e \in \{1, 2, \ldots, \infty\}$ of parallel and periodic choices through which the application and rejection process continues before assignments are finalized. As parameter e varies, we go from the Immediate Acceptance mechanism ($e = 1$) to the Chinese parallel mechanisms ($e \in [2, \infty)$), and from those to the DA ($e = \infty$).

We now describe a stylized version of the Chinese parallel mechanisms in its simplest form, with two parallel choices per choice-band ($e = 2$), which is the version we study in our laboratory experiment.

- An application to the first ranked college is sent for each student.
- Throughout the allocation process, a college can hold no more applications than its quota.

  If a college receives more applications than its quota, it retains the students with the highest priority up to its quota and rejects the remaining students.

[1] See Chen and Kesten (2014) for a brief history of the Chinese college admissions reform.

- Whenever a student is rejected from his first-ranked college, his application is sent to his second-ranked college. Whenever a student is rejected from his second-ranked college, he can no longer make an application in this round.
- Throughout each round, whenever a college receives new applications, these applications are considered together with the retained applications for that college. Among the retained and new applications, the ones with the highest priority up to the quota are retained.
- The allocation is finalized every *two* choices. That is, if a student is rejected by his first two choices in the initial round, then he participates in a new round of applications together with other students who have also been rejected from their first two choices, and so on. At the end of each round the assigned students and the slots assigned to them are removed from the system.

The algorithm terminates when no more applications can be rejected or no more seats are left.

4.3.2 Theoretical Properties and Main Hypotheses

Analyzing this parametric family of application-rejection mechanisms, Chen and Kesten (2014) find that, as one moves from one extreme member of this family to the other, the experienced trade-offs are in terms of strategic immunity and stability. At the individual strategy level, they show that, whenever any given member can be manipulated by a student, any member with a smaller e number can also be manipulated but not vice versa (Theorems 1 & 3). This implies that any parallel mechanism ($e \in [2, \infty)$) is less manipulable than the Immediate Acceptance mechanism which it replaces. Furthermore, DA is non-manipulable. We summarize these theoretical characterizations into the following hypotheses.

Hypothesis (Manipulability)**.** The extent of preference manipulation follows the order of DA $<$ PA $<$ IA.

With data from laboratory experiments, we can directly compare the stated preferences with true preferences; with field data, although true preferences are not observable, preference manipulation can exhibit themselves in a variety of ways, such as truncation of the rank-ordered list (**?**), and listing a safe college as one's top choice, either in the form of a less prestigious college or a local college which allocates more quota to local students.

At the outcome level, a parallel mechanism provides the students with a certain sense of "insurance" by allowing them to list their equilibrium assignments under the Immediate Acceptance mechanism as a safety option while listing more desirable options higher up in their preferences, which in turn leads to an outcome at least as good as that of the Immediate Acceptance mechanism for everyone (Proposition 5 in Chen and Kesten 2014). Therefore, we expect:

Hypothesis (Insurance)**.** Students will list more prestigious/more preferred colleges as their first choices under the parallel mechanism, compared to the Immediate Acceptance mechanism.

In terms of choice accommodation, Chen and Kesten (2014) show that the IA mechanism is the most generous in terms of first choice accommodation, whereas DA is the least. In general, a parallel mechanism (e) assigns more students to their top e choices than any other mechanism.

Hypothesis (Choice Accommodation)**.** (a) The Immediate Acceptance mechanism assigns a higher number of students to their reported first choices than does any parallel mechanism. (b) In comparison, the parallel mechanism (e) assigns more students to their top e choices than any other mechanism (e'), where $e \neq e'$.

Chen and Kesten (2014) show that when $e' = ke$ for some $k \in \mathbb{N} \cup \{\infty\}$, any stable equilibrium of the application-rejection mechanism (e) is also a stable equilibrium of the application-rejection mechanism (e') but not vice versa (Theorems 2 & 4). In this sense, they find that every newly adopted parallel mechanism is more stable than the Immediate Acceptance mechanism it replaced.

Hypothesis (Stability)**.** The DA mechanism is more stable than any parallel mechanism, which in turn, is more stable than the Immediate Acceptance mechanism.

Although it is well-known that the dominant strategy equilibrium outcome of DA Pareto dominates any equilibrium outcome of the Immediate Acceptance mechanism (Ergin and Sönmez, 2006), Chen and Kesten (2014) show that there is no clear dominance of DA over a parallel mechanism. Therefore, we are agnostic with regard to the efficiency comparison of the two mechanisms. We next outline the design of a laboratory experiment, which enables us to test these theoretical predictions.

4.4 Experimental Design

We design our experiment to compare the performance of the Immediate Acceptance (IA, $e = 1$), the simplest version of the Chinese Parallel mechanism (PA, $e = 2$) and the Deferred Acceptance (DA, $e = \infty$) mechanisms based on the theoretical characterization of the family of application-rejection mechanisms. Our design captures the following key features of Chinese college admissions: student preferences over colleges are correlated, and college priorities over students are solely determined by their test scores.

4.4.1 The Environment

We design a matching market with 6 colleges and 6 students, similar to the 6-school environment in the Chen and Kesten (2014) experiment. Different from school choice experiments, in our design, colleges have identical priority rankings of students based on student test scores, which are randomly assigned to participants in the game. This design choice is based on the fact that in the centralized college admissions in China, test score is the main determinant of priorities.[2]

We number students in each group as 1, 2, ..., 6, and colleges as A, B, ..., F. For each session, there are 18 subjects, randomly re-matched into 3 groups at the beginning of each period. Each session lasts 20 periods to facilitate learning. At the beginning of each period we randomly assign each subject his rank of test score based on a random number drawn from a uniform distribution on the open interval $(0, 1)$, which in turn determines their admission priorities.

In addition, we construct two types of student preference over colleges (see Table 4.1), which are randomly assigned to an equal number of students. That is, there are 3 students of each type in a group. A student keeps his or her type throughout the experiment. The preference orders of these two types of students are shown in Table 4.1. In this environment, subject preferences are correlated to reflect the competitiveness of college admissions, but imperfectly correlated to reflect preference heterogeneity. Participant payoffs are adopted from Chen and Kesten (2014), with a monetarily salient payoff difference from the most to the least preferred college spanning {16, 13, 11, 9, 7, 5}.

The experiment uses a complete information design, in which subjects know their own score ranking and those of others in their group. For example, Student 1 may know that he is of Type 1, and he also knows the score ranking of the other two Type 1 students and of the three Type 2 students.

In each period, each participant submits a rank ordered list (ROL) of colleges. After all participants have submitted their ROLs, the server assigns students to colleges based on the matching mechanism, and inform each participant of his or her

[2]Other considerations, such as affirmative action for ethnic minorities, are incorporated as a 20-point bump to a student's test score.

Table 4.1: Student types and their payoffs over colleges

Colleges:	A	B	C	D	E	F
Payoff to Type 1	16	13	11	9	7	5
Payoff to Type 2	16	11	13	9	5	7

matching outcome and the corresponding payoff.

Based on the matching market parameters, we now characterize the stable matching outcomes of the college admissions game in our environment under students' true preferences. Since all students have identical priorities in each college, there is only one stable allocation for each type profile realization. We summarize these allocations in Table 4.2.

Table 4.2: Summary of stable matching outcomes for all type/rank realizations

Type profile	Stable matching outcome	Type profile	Stable matching outcome
112212	123456	121212	132456
211212	123456	122112	132456
212112	123456	221112	132456
212211	123456	122211	132456
111222	123465	221211	132456
112122	123465	222111	132456
112221	123465	121122	132465
211122	123465	121221	132465
211221	123465	122121	132465
212121	123465	221121	132465

Notes: For each cell, the profile of students' types is ordered, from left to right, by students' rankings in test scores from the highest to the lowest.

Theorem 1 in Ergin and Sönmez (2006) shows that the set of Nash equilibrium outcomes of this game is equal to the set of stable matchings. Therefore, for each type/rank realization, there is only one Nash equilibrium outcome as well.

We now characterize the Nash equilibrium strategies. First we define a *rank-biased* strategy A_i for student $i \in \{1, ..., 6\}$, where i is the ith-ranked student in her group. For example, student $i = 1$ is ranked first by all colleges. For $i \in \{1, 2, 4, 5\}$, student

i lists his or her ith preferred college, whatever the type is. For $i = 3$, if the type of student i is 1, and the type of student $i-1$ is 2, then student i lists his or her $i-1$th preferred college as the first choice; otherwise he or she lists the ith preferred college as the first choice. For $i = 6$, i can list any arbitrary order of colleges. For choices other than the first one, students arbitrarily arrange the colleges that have not been listed. The following proposition characterizes this strategy as a Nash equilibrium strategy.

Proposition 5 (Rank Bias)**.** *The rank-biased strategy profile, $A = \{A_i\}_i$, is a Nash equilibrium strategy under the Immediate Acceptance mechanism.*

Proof. Listing one's most preferred college as first choice is a dominant strategy for the top-ranked student, regardless of type. Given the top-ranked student's dominant strategy, it is optimal (and weakly optimal under the DA mechanism) for Student 2 to list his second preferred college as the first choice. Given the strategies of Students 1 and 2, and observing the type of Student 2, Student 3 then finds it optimal to list his second preferred college as the first choice when Student 2 is of Type 2, and to list his third preferred college as the first choice when Student 2 is of Type 1. The same argument goes for students ranked 4th to 6th. □

Using the same argument, it is easy to see that such rank-biased strategy is also a Nash equilibrium strategy under the PA and DA mechanisms as well. However, for these two mechanisms, there are other Nash equilibrium strategies, such as the one characterized in Proposition 6 which follows from Proposition 5 in Chen and Kesten (2014).

Proposition 6 (Insurance)**.** *Any strategy listing the first two choices which (a) include the first-choice Nash equilibrium college, and (b) preserves the true order of colleges, is a Nash equilibrium strategy under PA and DA.*

Proof. By Proposition 5 in Chen and Kesten (2014), each student's assignment under strategies that satisfy above two conditions is at least as good as that under the IA mechanism. However, we have already shown that there is only one Nash equilibrium outcome under each type/rank realization, therefore such assignment is already the best outcome and students can not unilaterally improve their assignments by playing other strategies. □

For example, for the second-ranked student of Type 1 ($i = 2$), any first two choices that includes i's second most preferred college (B), and preserves her true preference order of colleges, such as (A, B) or (B, E), is a Nash equilibrium strategy.

Our results about Nash equilibrium strategies also have implications on stable matching outcomes. Although, theoretically, all three mechanisms reach stable outcomes when subjects play Nash equilibrium, in practice the three mechanisms may not be equally stable. Since stable matching outcomes are identical to Nash equilibrium outcomes, to reach such outcomes requires different level of sophistication. Therefore, Propositions 5 and 6 imply different strategic requirements to reach a Nash equilibrium outcome. For the IA mechanism, equilibrium requires carefully planned first choice, whereas the PA mechanism leaves more room for "mistakes" – strict coordination in first choice is not required. Finally, the DA mechanism has the lowest requirement – players can either simply reveal their true preferences, or employ any equilibrium strategies under the IA or the PA mechanisms.

4.4.2 Experimental procedure

In each experimental session, each participant is randomly assigned an ID number and is seated in front of a terminal in the laboratory. The experimenter reads the instructions aloud. Subjects ask questions, which are answered in public. Subjects are then given 10 minutes to read the instructions at their own pace and to finish the review questions in which subjects are paid for each correctly answered question.

After everyone finishes the review questions, the experimenter distributes the answers and goes over the answer in public. Afterwards, participants go through 20 periods of the college admissions game. After the game, we elicit subjects' risk attitude by asking them to filled out an incentivized lottery questionnaire (Holt and Laury, 2002). At the end, they fill out a demographics and strategy survey on the computer. Each participant is paid in private at the end of the experiment. The experiment is programmed in z-Tree (Fischbacher, 2007).

For each mechanism, we conduct four independent sessions, with 18 subjects in each session. In total, we have 12 independent sessions with 216 participants. No one participates in more than one session. Each session consists of 20 periods, and lasts approximately two hours, with the first 20-30 minutes used for instructions.

Our experiments were conducted in May and June, 2011, and June 2013 at the Smith Experimental Economics Research Laboratory at Shanghai Jiao Tong University. Our subjects are students from Shanghai Jiao Tong University. The conversion rate is 1 CNY = 20 points for all treatments. Additionally, each subject receives a participation fee of 5 CNY, and up to 3.5 CNY for answering the Review Questions correctly.[3] The average earning was 84.72 CNY (13.03 USD). Experimental instructions are included in the Appendix.

4.5 Experimental Results

We first examine individual behavior under each of the three mechanisms, paying particular attention to the extent and patterns of their preference manipulation. We then study the performance of the mechanisms at the aggregate level: efficiency, stability, choice accommodation, and the proportion of matching that coincides with Nash equilibrium outcomes.

We introduce some notations when presenting the results. Let $A > B$ denote

[3]The exchange rate at the time of the experiment was approximately 1 USD = 6.5 CNY.

that a measure (e.g. stability) under mechanism A is greater than the corresponding measure under mechanism B at the 5% significance level or less.

4.5.1 Individual behavior

We first examine the extent to which individuals reveal their preferences truthfully, and the pattern of any preference manipulation under each of the three mechanisms. The manipulability hypothesis suggests that the parallel mechanism is less manipulable than the Immediate Acceptance mechanism. Furthermore, under the DA mechanism, truth-telling is a weakly dominant strategy. This leads to our first hypothesis.

Hypothesis 13 (Truth-telling)**.** (a) There will be a higher proportion of truth-telling under the parallel than under the Immediate Acceptance mechanism. (b) Under the DA mechanism, participants will be more likely to reveal their preferences truthfully than under Immediate Acceptance mechanism. (c) Under the DA mechanism, participants will be more likely to reveal their preferences truthfully than under the parallel mechanism.

Figure 4.1 and the upper panel of Table 4.3 report the proportion of truthful revelation of all choices. We report the information averaged over all rounds, where p-values are computed from one-sided permutation tests, treating each session as an independent observation.

Result 18 (Truth-telling)**.** *The proportion of truth-telling under the three mechanisms has the following order: DA > PA > IA.*

Support. Using each session as an independent observation, pairwise comparisons of the proportion of truthful revelation of all choices each yields $p = 0.014$ (one-sided permutation test). By Result 18, we reject the null hypotheses in favor of Hypothesis 13.

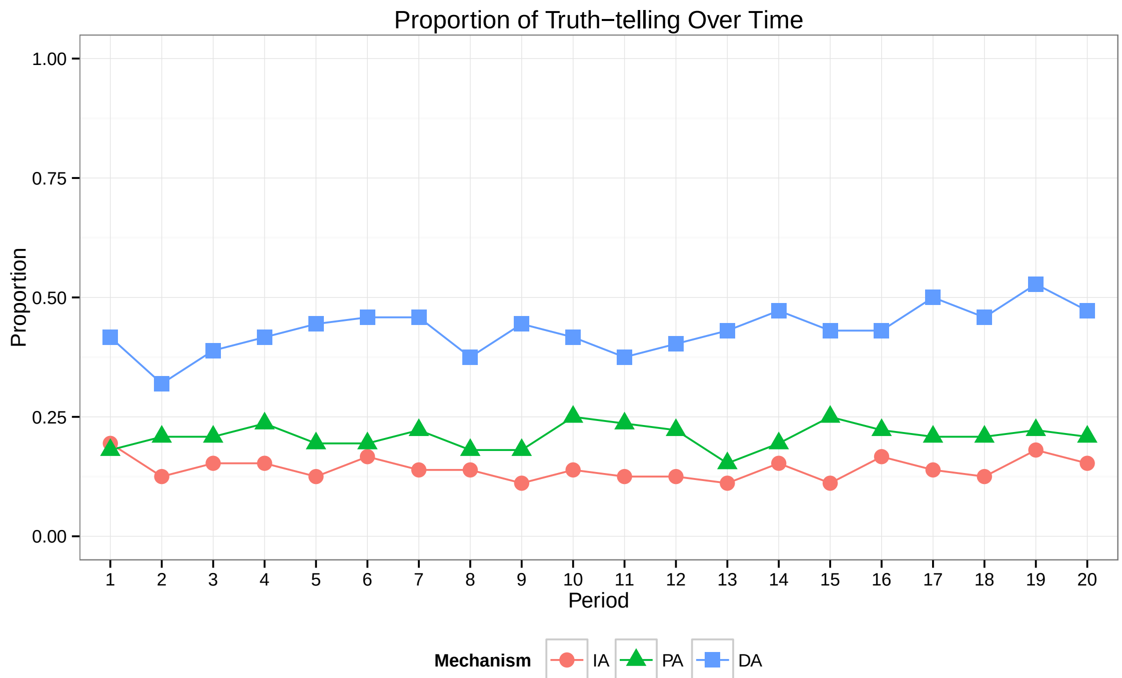


Figure 4.1: Proportion of truth-telling over time.

Table 4.3: Proportion of truth-telling and first choice truth-telling at session level

Proportion of truth-telling					
Session	IA	PA	DA	H_a	p-value
1	0.125	0.203	0.378	IA $<$ PA	0.014
2	0.131	0.203	0.356	PA $<$ DA	0.014
3	0.150	0.206	0.561	IA $<$ DA	0.014
4	0.161	0.225	0.433		
Overall	0.142	0.209	0.432		
Proportion of first choice truth-telling					
Session	IA	PA	DA	H_a	p-value
1	0.178	0.258	0.531	IA $<$ PA	0.014
2	0.178	0.242	0.439	PA $<$ DA	0.014
3	0.172	0.256	0.600	IA $<$ DA	0.014
4	0.192	0.244	0.486		
Overall	0.180	0.250	0.514		

Notes: p-values computed from one-sided permutation tests with each session treated as an independent observation.

Although neither IA nor PA is strategy-proof, IA specifically puts emphasis on students' first choices, while PA does not. Therefore, we pay further attention to the first choice truth-telling behavior. The bottom panel of Table 4.3 reports the proportion of truthful revelation of the first choice. We can see that it has the same order of truth-telling, that is, DA > PA > IA. Statistical tests using one-sided permutation support this finding ($p = 0.014$ for all pairwise comparisons at session level).

The above findings are consistent with the experimental findings of Chen and Sönmez (2006) comparing IA and DA, and Chen and Kesten (2014) comparing all three mechanisms.

In addition to knowing the extent to which students manipulate their strategies, we are also interested in the ways they manipulate. Propositions 5 and 6 imply that under IA, to play Nash requires careful manipulation of one's first choice, and such requirement is relaxed to first two choices under PA. This is further relaxed under DA. Similar to the *district bias* identified in Chen and Sönmez (2006) and Chen and Kesten (2014), we define *rank bias* as the tendency of students to list colleges whose positions in the students' preference rankings are equal or close to these students' rank of scores; i.e., students' position in colleges' priority lists. For example, a 4th ranked student tends to list her 4th preferred college as her first choice.

Hypothesis 14 (Rank bias)**.** The proportion of rank bias under the three mechanisms follows the order of: DA > PA > IA.

Result 19 (Rank bias)**.** *Students are significantly more likely to exhibit rank bias under the IA than under the PA mechanism, and more under the PA than under the DA mechanism.*

Support. Table 4.4 reports probit (1 and 2) and OLS (3 and 4) specifications to examine the extend to which students manipulate their ROLs based on their rankings in the test. Specifications (1) and (2) report that the lower a student's rank of score is, the more likely he will manipulate the reported first choice (marginal effect of going down one rank is -25.3%, $p < 0.01$). The magnitude of manipulation follows the order of: IA > PA > DA, as marginal effects of interaction terms $Ranking \times DA$ and $Ranking \times PA$ are 19.5% ($p < 0.01$) and 7.2% ($p < 0.01$), respectively, and significantly different from each other (F-test $p < 0.01$). Additionally, specifications (3) and (4) report the effect of students' rankings on the first choice colleges' positions' within students' true preference rankings. Similarly, we can see that the lower a student's rank of score is, the worse the college in the preference ranking he will list as the first choice (the effect on the position of first choice colleges of going down one rank is 0.201, $p < 0.01$). The magnitude has the following order as well: IA > PA > DA, as coefficients of interaction terms $Ranking \times DA$ and $Ranking \times PA$ are -0.331 ($p < 0.01$) and -0.045 ($p < 0.01$), respectively, and significantly different from each other (F-test $p < 0.01$). The results are robust after further controlling for risk preference measured by the switch point in the Holt-Laury lottery game[4] and period. Additional tests using ordered probit specifications are in the appendix.

By Result 19, we reject the null hypothesis in favor of Hypothesis 14. Further, Figure 4.2 illustrates that nearly all students who are ranked first report their true first choices. The proportion of truth-telling diminishes after rank 1 for IA and after rank 2 for PA. While lower than that of the top ranked students, the proportion of truth-telling remains steady and significantly higher than zero under DA, providing further support for the theoretical predictions of Proposition 5 and 6.

Results from the experiment for individual behavior confirm the theoretical predictions: the more emphasis a mechanism puts on the first choice (which has the order of: $IA > PA > DA$), the less likely students will report their true preferences. Specifically, they manipulate according to their ranks in test scores. The following

[4]Subjects who switched multiple times are excluded.

Table 4.4: Rank Bias

Dependent variable:	First choice truth-telling		True position of reported first choice college	
	(1)	(2)	(3)	(4)
DA	-0.190***	-0.211***	0.201***	0.254***
	(0.041)	(0.041)	(0.050)	(0.046)
PA	-0.042	-0.073	-0.167***	-0.146**
	(0.060)	(0.060)	(0.050)	(0.050)
Ranking	-0.253***	-0.263***	0.735***	0.742***
	(0.017)	(0.021)	(0.011)	(0.012)
Ranking × DA	0.195***	0.208***	-0.331***	-0.359***
	(0.018)	(0.022)	(0.035)	(0.033)
Ranking × PA	0.072***	0.091***	-0.045**	-0.053**
	(0.028)	(0.030)	(0.015)	(0.020)
Period		-0.000		0.006
		(0.001)		(0.004)
HL Switch Point		-0.009		0.021
		(0.008)		(0.019)
Constant			0.439***	0.215
			(0.033)	(0.129)
Observations	4,320	3,920	4,320	3,920
R^2			0.609	0.618

Notes: (1) and (2): Probit regression report the marginal effect of ranking on the probability of first-choice truth-telling; (3) and (4): OLS regression reporting the effect of ranking on the true position of the reported first-choice colleges; standard errors in parentheses are clustered at session level. ***: $p < 0.01$; **: $p < 0.05$; *: $p < 0.1$. When controlling for risk attitude, 20 students (400 observations) are dropped due to multiple switching points in the Holt and Laury lottery game.

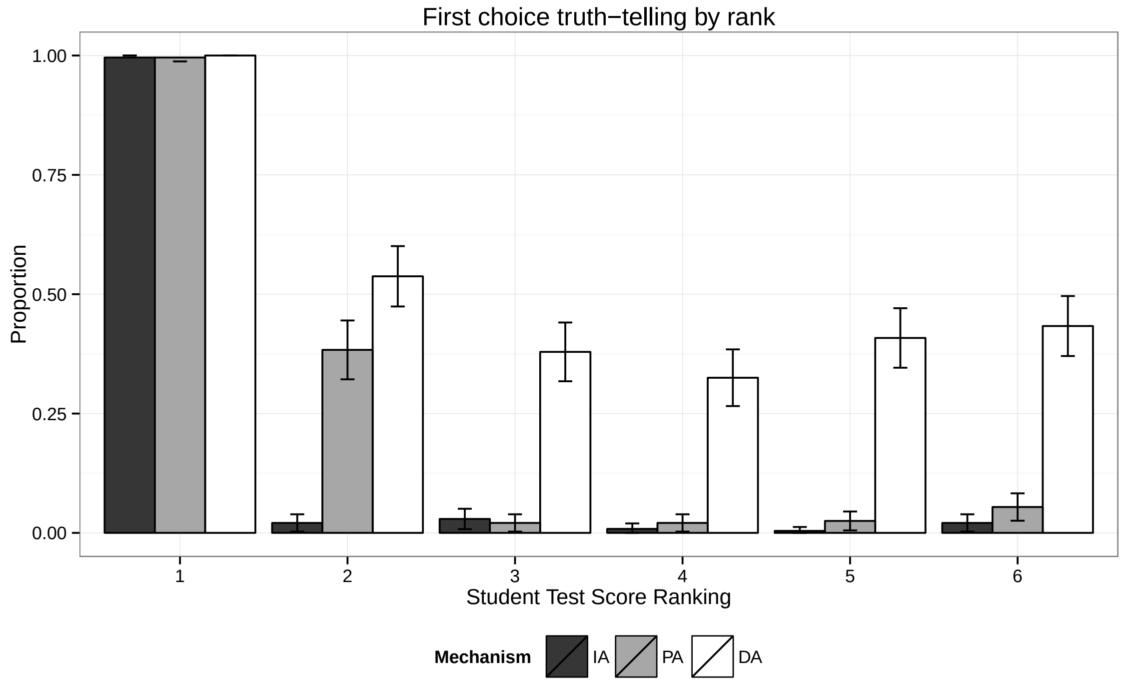


Figure 4.2: First choice truth-telling sorted by student test score ranking

subsection investigates whether the predictions still hold on an aggregate level.

4.5.2 Aggregate Performance

Next we turn to the aggregate performance of the three mechanisms. First we present the stability result. We measure stability at the individual level: whether a student exhibits justified envy toward at least one other student or not. Figure 4.3 illustrates the proportion of students exhibiting justified envy. Similar to the definition in the school choice literature, if a student exhibits justified envy, it means that a college he prefers to his current match admits a student that scores lower in the test. If a student exhibits justified envy, then the matching is unstable.

Figure 4.5 and Table 4.5 report the proportion of students who exhibit justified envy. We report the information averaged over all rounds, where p-values are computed from one-sided permutation tests, treating each session as an independent observation.

The Nash equilibrium matching outcomes under all three mechanisms are stable. However, under Immediate Acceptance and PA, stable outcomes require students play Nash, while under DA playing dominant strategies would also lead to stable outcomes.

Hypothesis 15 (Stability)**.** The matching is more stable under the DA than under the PA mechanism, and is more stable under the PA than under the IA mechanism.

Result 20 (Stability)**.** *The matching stability has the following order: DA > PA > IA.*

Support. Using each session as an independent observation, pairwise comparisons of the proportion of students exhibiting justified envy using one-sided permutation test, IA > PA yields $p = 0.086$, PA > DA $p = 0.043$, and IA > DA $p = 0.014$.

Note that while there is difference in stability in general, the difference disappears in the last five periods. This may be explained by subjects learning to play equilibrium strategies, which lead to stable matching outcomes under all three mechanisms. I will explore this later in this section.

By Result 20, we reject the null hypothesis of no difference in favor of Hypothesis 15.

Next we present the efficiency result. We measure the welfare/efficiency of the mechanisms by summing individual payoffs in a group. We then normalize the payoff using the following formula:

$$\text{Normalized Payoff-based Efficiency} = \frac{\text{Actual sum of payoffs} - \text{Minimum sum of payoffs}}{\text{Maximum sum of payoffs} - \text{Minimum sum of payoffs}}.$$

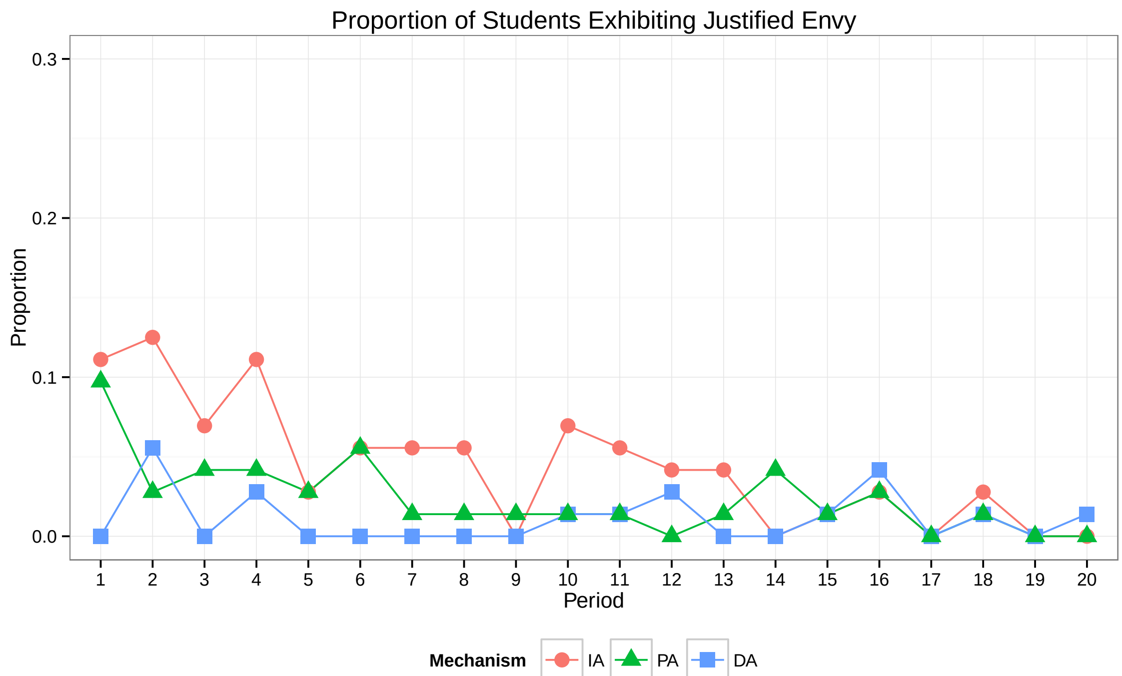


Figure 4.3: Proportion of students exhibiting justified envy.

Table 4.5: Proportion of students exhibiting justified envy at session level.

Proportion of students exhibiting justified envy					
Session	IA	PA	DA	Ha	p-value
1	0.033	0.019	0.006	IA < PA	0.086
2	0.056	0.017	0.006	PA < DA	0.043
3	0.022	0.036	0.019	IA < DA	0.014
4	0.067	0.022	0.014		
Overall	0.044	0.024	0.011		

Notes: p-values computed from one-sided permutation tests with each session treated as an independent observation..

We also calculate the rank-based efficiency: we sum up the rank of matched colleges in students' true preferences for each group. We then normalized it based on the following formula:

$$\text{Normalized Rank-based Efficiency} = \frac{\text{Actual sum of ranks} - \text{Minimum sum of ranks}}{\text{Maximum sum of ranks} - \text{Minimum sum of ranks}}.$$

In our environment, the minimum sum of payoffs is 57, and the maximum is 65; the minimum sum of ranks is 19, and the maximum is 23. Table 4.6 summarizes the normalized efficiency at the session level. We report the information averaged over all rounds and sessions, where p-values are computed from two-sided permutation tests, treating each session as an independent observation. Recall that when students play Nash equilibrium strategies, all three mechanisms lead to the same group payoff. Therefore we do not have systematic efficiency hypothesis.

Result (Efficiency)**.** *There is no significant efficiency difference between any two of the mechanisms.*

Support. In Table 4.6, pairwise comparisons show no significant difference in either payoff-based or rank-based efficiency between IA and DA ($p = 0.400$), IA and PA ($p = 0.457$), or DA and PA ($p = 0.857$).

The above result shows that not only the three mechanisms, while varying greatly in manipulation, are similar in efficiency, it also shows the level of efficiency is high, reaching approximately 80% in payoff-based efficiency for all three mechanisms. The reason why this happens may lie in the strategic sophistication of students, which we explore next, through reported choice accommodation and proportion of Nash equilibrium outcomes.

First we look at the reported choice accommodation rate. The matching under the Immediate Acceptance mechanism highly depends on students' first choices. Therefore, the success of manipulation under the Immediate Acceptance mechanism relies

Table 4.6: Efficiency

Dependent variable:	Prestige of first-choice college		
	(1)	(2)	(3)
Y2009	-0.051***	-0.052***	-0.052***
	(0.003)	(0.003)	(0.004)
Tier1	0.191***	0.200***	0.198***
	(0.005)	(0.004)	(0.005)
Y2009 × Tier1	0.044***	0.046***	0.047***
	(0.010)	(0.011)	(0.012)
Percentile Ranking		0.294***	0.293***
		(0.015)	(0.015)
Science			0.022
			(0.011)
Rural			-0.009
			(0.006)
Female			-0.009
			(0.009)
Constant	0.567***	0.412***	0.408***
	(0.013)	(0.011)	(0.009)
Observations	6,757	6,757	6,757
R^2	0.117	0.221	0.223

Notes: p-values computed from two-sided permutation tests with each session treated as an independent observation.

on how to list one's first choice strategically, as suggested in Proposition 5. While under the Deferred Acceptance mechanism, truth-telling is a weakly dominant strategy. Moreover, Proposition 3 in Chen and Kesten (2014) shows that PA with $e = 2$ assigns a higher number of students to the first two choices. The following hypothesis summarizes this finding.

Hypothesis 16 (Reported choice accommodation)**.** The proportion of students receiving their reported first choice will be the highest under the IA mechanism, followed by the PA mechanism, which is in turn followed by the DA mechanism. The proportion of students receiving their reported first two choices will be the highest under the PA mechanism, followed by the IA, which is in turn followed by the DA mechanism.

Result 21 (Reported choice accommodation)**.** *The reported-first choice accommodation rate has the following order: IA > PA > DA; the reported first two choices accommodation rate has the following order: PA > IA > DA.*

Support. Using each session as an independent observation, pairwise comparisons of the reported first choice accommodation rate using one-sided permutation test each yields $p = 0.014$. Pairwise comparisons of the reported first two choice accommodation rate between PA and IA yields $p = 0.1$, IA and DA $p = 0.014$, and PA and DA $p = 0.014$.

By Result 21, we reject the null hypotheses in favor of Hypothesis 16. Notably, the reported first choice accommodation rate of IA reaches 80%, and the reported first two choices accommodation rate of PA almost reaches 92%.

The upper panel of Table 4.7 reports the reported first choice accommodation rate at the session level, while the lower upper panel reports the reported first-two choice accommodation rate. Figure 4.4 further shows that the only important choice considered by students under IA is the first choice, while both first and second choices are considered important under PA with $e = 2$. DA exhibits less of such a bias.

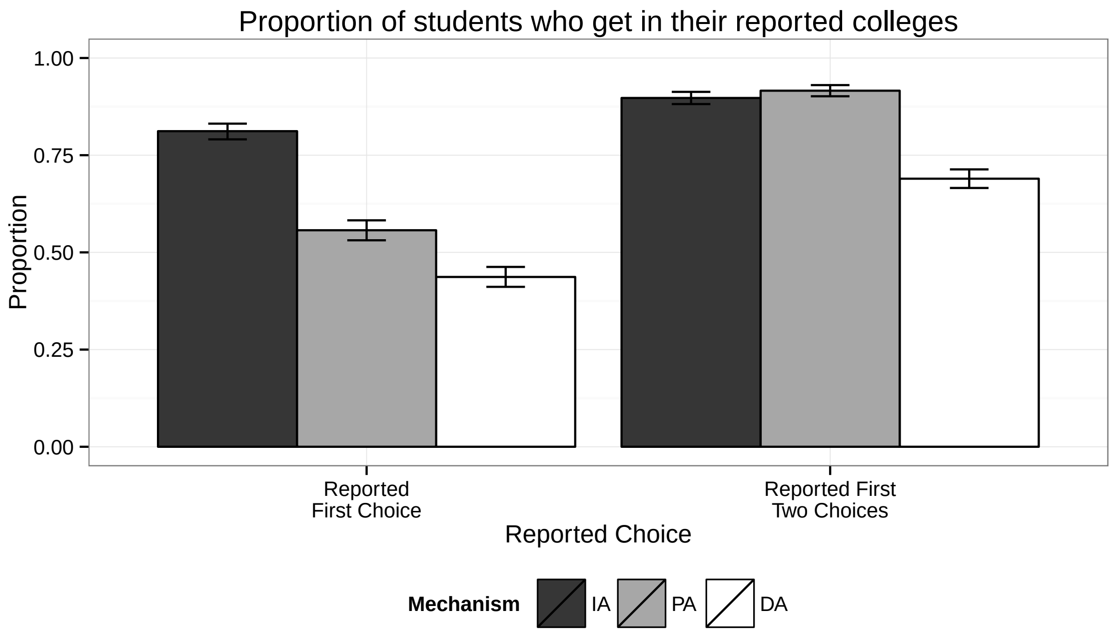


Figure 4.4: Choice accommodation.

Table 4.7: Reported choice accommodation rate at session level

First choice accommodation					
Session	IA	PA	DA	Ha	p-value
1	0.822	0.622	0.394	IA > PA	0.014
2	0.822	0.547	0.511	PA > DA	0.014
3	0.822	0.525	0.375	IA > DA	0.021
4	0.781	0.533	0.467		
Overall	0.812	0.557	0.437		
First two choices accommodation					
Session	IA	PA	DA	Ha	p-value
1	0.922	0.922	0.672	PA > IA	0.1
2	0.900	0.928	0.733	IA > DA	0.014
3	0.892	0.917	0.647	PA > DA	0.014
4	0.875	0.897	0.706		
Overall	0.897	0.916	0.690		

Notes: p-values computed from one-sided permutation tests with each session treated as an independent observation.

Next we connect the choice accommodation rate with matching outcomes. The following hypothesis is closely related to the stability hypothesis, given that stable outcomes coincide with Nash equilibrium outcomes. Moreover, previous study shows that the dominant strategy equilibrium, when coinciding with Nash equilibrium, is more likely to be chosen. This predicts that when the Nash equilibrium outcome is also the result of truth-telling (under DA), it will happen more frequently.

Hypothesis 17 (NE outcome)**.** More groups will achieve Nash equilibrium outcomes under the Deferred Acceptance mechanism that under either of the other two mechanisms.

Result 22 (NE outcome)**.** *The proportion of Nash equilibrium outcomes under DA is significantly higher than under PA and IA, while the proportion is not significantly different between PA and IA.*

Support. Using each session as an independent observation, pairwise comparisons of the proportion of group level stable/NE outcomes report that the stability of DA is significantly higher than both IA and PA ($p = 0.043$, $p = 0.071$ respectively), while there is no significant difference between IA and PA ($p = 0.443$). The result rejects the null hypotheses in favor of the Hypothesis 17.

Table 4.8: Proportion of NE outcomes

Proportion of NE Outcomes					
Session	IA	PA	DA	Ha	p-value
1	0.883	0.933	0.967	IA $<$ PA	0.443
2	0.817	0.833	0.967	PA $<$ DA	0.07
3	0.917	0.917	0.883	IA $<$ DA	0.043
4	0.767	0.600	0.933		
Overall	0.846	0.821	0.938		

Notes: p-values computed from one-sided permutation tests; independent observation at session level.

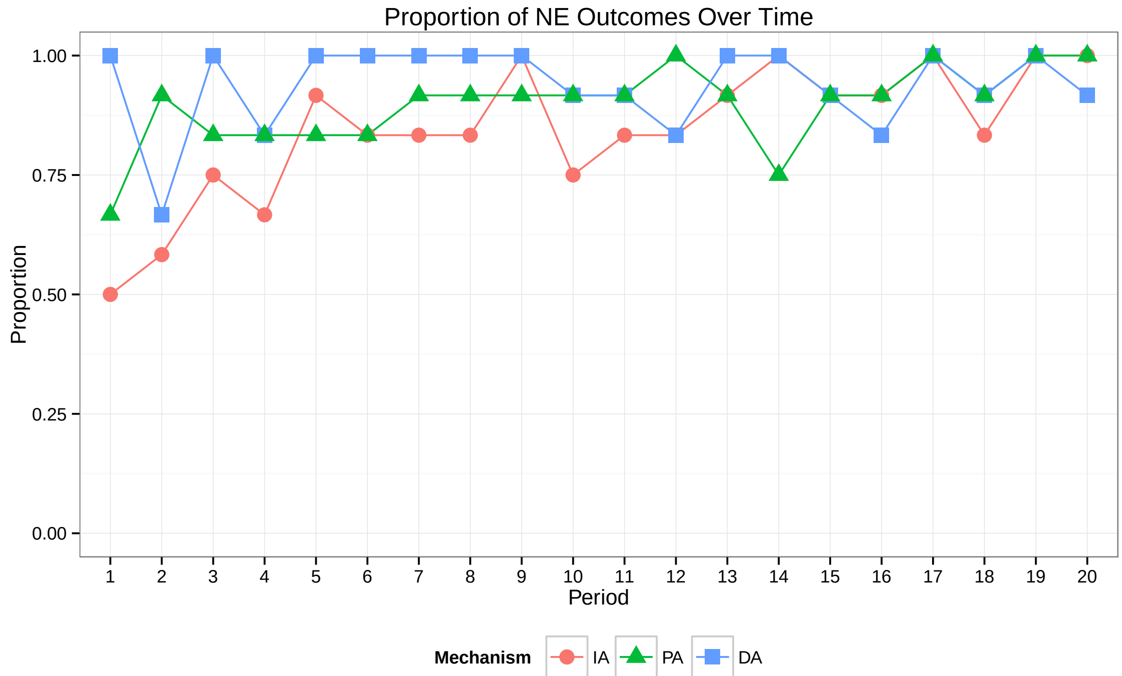


Figure 4.5: Proportion of NE outcomes over time.

High first choice accommodation rate and proportion of Nash equilibrium outcomes from the above results require less conflicted first choices among students, which in turn means subjects implicitly coordinate with each other on the Nash equilibrium strategies through observed types/preferences to achieve highly efficient aggregate outcomes. Such level of strategic coordination requires a certain level of sophistication of our subjects. Based on our post-experiment survey, 99% of our subjects participated in the Chinese college admission systems, and they know what are the best things for them to do. We test this sophistication conjecture by looking at the time students take to make decisions, and have the following result, illustrated in Figure 4.6.

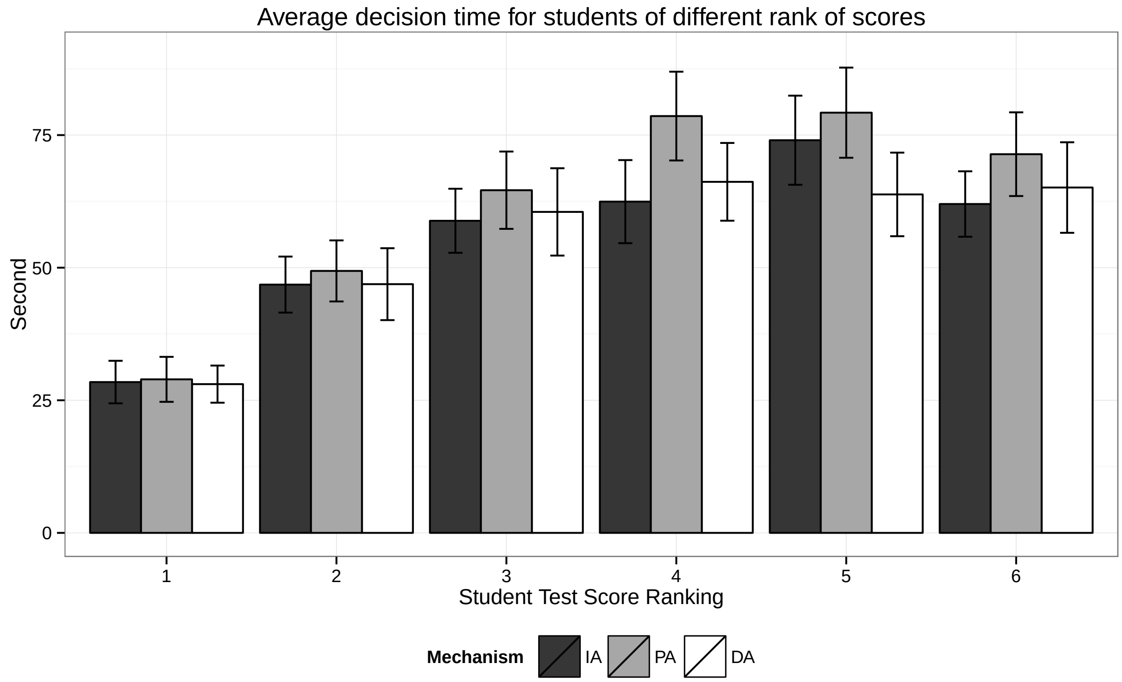


Figure 4.6: Average decision time for students of different rank of scores (in seconds).

Result (Decision time)**.** *Students who are ranked in the middle take significantly longer time to list their college preferences.*

Support. Students who are ranked below first take longer to make decisions than if they are ranked at the top ($p < 0.001$; two-sided permutation test with each period as an independent observation).

4.6 Data and Empirical Methods

We supplement our laboratory experiment with an empirical study of a natural experiment. Our dataset consists of the college admissions data of Santai (sān tái) County in Mianyang (mián yáng) City in the Sichuan Province in southwestern China for the year 2008 and 2009. The county has a population of 1.47 million in 2013. It is

84.67% rural[5], with per capita GDP of USD 2747.5, below the national average of USD 6169.8 [6] [7]. Out dataset includes students' test scores in the National College Entrance Exam, students' reported preferences of colleges (rank-ordered list), the colleges each student eventually got admitted into, and student demographic information.

Chinese colleges are categorized into tiers of decreasing prestige and quality. For example, Tier 1 colleges are generally considered better than Tier 2 colleges, etc. College admissions mechanisms are executed sequentially across tiers. When assignments in the first tier are finalized, the assignment process in the second tier starts, and so on. Our dataset contains all students who participated in the Tier 1 and Tier 2 admissions in 2008 and 2009.

For the period of our dataset, students submitted their rank-ordered lists of colleges after taking the college entrance exams and knowing their test scores as well as their relative standings among all the students in the province. The Provincial College Admissions Office determined whether a student was eligible to participate in the admissions of each tier by setting up an endogenously determined cutoff score, such that the number of students who were above the Tier 1 cutoff was approximately 120% of the total quota of all Tier 1 colleges; the number of students who were above the Tier 2 cutoff was approximately 120% of the total quota of all Tier 1 and Tier 2 colleges; etc.

Additionally, there were two separate matching markets each year for the two academic tracks: humanities and social sciences (humanities henceforth), and science and engineering (science henceforth). Students separate into one of the two tracks in their second year of high school, and take the corresponding set of exams. Likewise, each college has separate quota for each of the two tracks.

Between the college entrance exams of 2008 and 2009, the government of Sichuan

[5] The national rural population is 47.43%.

[6] http://www.stats.gov.cn/tjsj/ndsj/2013/indexeh.htm. Retrieved on 3/30/2015.

[7] http://jy.libai.cn/santai/438823895328358400/20140411/486714.html. Retrieved on 3/30/2015.

province announced that it would change the college admissions mechanism for its Tier 1 admissions from the Immediate Acceptance mechanism to the parallel mechanism. The mechanism for all other tiers remain unchanged. This policy change serves as a natural experiment testing the effect of different matching mechanisms. Since students participate in college admissions during their last year of high school, and the policy change were announced after the previous year's admission was finished, students were selected into different treatment groups by birth.

Even though students are randomly selected into different years by birth, there might still be some changes in trend happening between the two years, such as students' overall preferences for humanities and science programs. We exploit the fact that in the province where we collect data from, only the mechanism for Tier 1 admissions changed where as Tier 2 admissions mechanism remained the same. Therefore, we estimate the following difference-in-differences model:

$$y_i = \beta_0 + \beta_1 \cdot Y2009_i + \beta_2 \cdot Tier1_i + \beta_3 \cdot (Y2009_i \cdot Tier1_i) + \gamma \cdot \boldsymbol{X}_i + \epsilon_i,$$

where y_i is the outcome variable, measuring strategies or matching outcomes for each student. $Y2009_i$ and $Tier1_i$ are dummy variables that equal to 1 for the Year 2009 and Tier 1, respectively. The vector, $\boldsymbol{X}_i$, contains students' individual characteristics, including gender, location (rural or urban), academic track (humanities or science), and ranking of test scores.

Table 4.9 presents the summary of the dataset. Students in different academic tracks and from different demographic backgrounds are similarly distributed for both years and both tiers. Note that there are students who participate in the admissions of both tiers (e.g. not admitted by any Tier 1 colleges who then move down a tier), and who list ROLs for both Tier 1 and Tier 2.[8]

[8]Note that ROL submissions happen after the entrance exam and before the admissions start.

Table 4.9: Summary statistics of the data

	2008			
	Total	Female	Rural	Science
Participated in Tier 1 admission	620	32.3%	80.2%	81.8%
Participated in Tier 2 admission	2443	40.4%	80.8%	70.9%
Participated in both	122	43.4%	81.1%	77.9%
Submitted Tier 2 ROL	2868	39.2%	80.8%	72.4%
Submitted Both	547	34.6%	80.8%	80.4%
	2009			
	Total	Female	Rural	Science
Participated in Tier 1 admission	769	36.3%	79.7%	80.6%
Participated in Tier 2 admission	2735	42.7%	83.1%	73.8%
Participated in both	135	46.7%	81.5%	77.0%
Submitted Tier 2 ROL	3266	41.4%	82.7%	74.8%
Submitted Both	665	37.1%	80.9%	79.1%

4.7 Empirical Results

In this section, we report the results of our empirical analysis of the naturally-occurring data. We first report student strategies, then their matching outcomes, as a result of the change in matching mechanisms for Tier 1 admissions. Students can manipulate their strategies by truncating their rank-ordered lists and by misreporting the ranking of different colleges on their rank-ordered list.

The empirical study of the change in college admission mechanisms provide us with a natural way with high external validity to look at how matching mechanisms affect strategies and matching outcomes. However, one significant limitation is that we do not know students' *true* preferences over colleges. To study strategic manipulation as well as matching stability, we must know students' preferences over colleges. In this section, we approximate such preferences using colleges' prestige, based on the assumption that students' preferences are highly correlated.

First, we look at the length of students' rank-ordered lists (ROL). Table 4.11 presents three OLS specifications. The dependent variable is the length of student ROL, whereas the independent variables (omitted) include Y2009 (Y2008), Tier1

(Tier2), Y2009 × Tier1, Science (Humanities), Rural (Urban), Female (Male), and Percentile Ranking. The variable, Percentile Ranking, is calculated as follows. Since matching is carried out separately by year (2008/2009), tier (1/2), and academic track (humanities/science), we calculate student ranking in each of the 8 markets based on their test scores in the college entrance exams. To correct for different market size, we then normalize student ranking to their percentile ranking in their respective market.

Table 4.10: Summary statistics of the results

	Tier 1		Tier 2	
	2008	2009	2008	2009
First choice prestige	0.76	0.75	0.57	0.52
Likelihood of listing local college(%)	33.8	39.3	66.1	75.4
Length of ROL	3.66	4.46	4.21	4.26
First choice accommodation rate(%)	74.0	50.2	67.3	67.2
First 5 choices accommodation rate(%)	78.1	79.8	77.8	75.9
Stability score	0.89	0.88	0.87	0.88

We find several interesting results. First, the main treatment effect, Y2009 × Tier1, is positive and significant, indicating that the change from the Immediate Acceptance to the parallel mechanism in Tier 1 admissions in 2009 increases the average ROL length in 2009 by 0.75, that is, a student lists approximately one more school for Tier 1 in 2009. Since the length of Tier 2 ROL is stable across the two years, the change in the length of Tier 1 ROL is likely due to the change in the matching mechanism. Second, the coefficient of the Tier1 dummy is negative and significant, indicating that the average length of Tier 1 ROL is shorter than the corresponding Tier 2 ROL in 2008. This can be explained by the fact that there are fewer colleges in Tier 1 than in Tier 2: for 2005 and afterward, there are 175 colleges to choose from for Tier 1 and 433 for Tier 2. Furthermore, we note that students in the science track have shorter ROL, women have longer ROLs, and higher ranked students submit shorter ROLs.

Next we look at mechanism effect on the ranking of colleges as students chose as

Table 4.11: Effects of matching mechanisms on the length of rank-ordered lists (OLS)

Dependent variable:	Length of ROL		
	(1)	(2)	(3)
Y2009	0.049	0.049	0.055
	(0.085)	(0.082)	(0.089)
Tier1	-0.550**	-0.551**	-0.515**
	(0.167)	(0.177)	(0.179)
Y2009 × Tier1	0.751***	0.752***	0.738***
	(0.110)	(0.121)	(0.124)
Percentile Ranking		-0.663***	-0.656***
		(0.141)	(0.139)
Science			-0.301***
			(0.058)
Rural			-0.053
			(0.038)
Female			0.127***
			(0.029)
Constant	4.210***	4.541***	4.749***
	(0.044)	(0.121)	(0.155)
Observations	7,523	7,523	7,523
R^2	0.021	0.045	0.061

Notes: Standard errors in parentheses are clustered at the high school level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

their first choices. Since we do not observe student true preferences over colleges, we use published national college rankings for 2008 and 2009 as proxies for student true preferences. Such approximation is based on the fact that student preferences over colleges are highly correlated: most students prefer higher ranked colleges over lower-ranked ones, as ranking is highly correlated with quality. As college rankings fluctuates locally from year to year, we further divide the colleges into coarser groups of ten based on their rankings. The only exception is that we treat the top two (Peking and Tsinghua) universities as their own group, since they have been consistently ranked as the best two universities in China for many years. Our coarse ranking puts Peking and Tsinghua University as group 1, universities ranked 3–10 as group 2, 11-20 as group 3, etc., with ten colleges in each group. We calculate the coarse ranking of colleges for each tier, and then normalized them to [0,1] to correct for different numbers of colleges in each tier. We call this normalized coarse ranking *prestige*. Our normalization also corrects for the natural correlation between college ranking and tier.

Table 4.12 reports the effects of mechanism change on the prestige of student reported first choices. The dependent variable is the normalized ranking (from 0 to 1) of the student reported first choice based on the coarsified ranking of colleges in each tier and each year. The independent variables (omitted) again include Y2009 (Y2008), Tier1 (Tier2), Y2009 $\times$ Tier1, Science (Humanities), Rural (Urban), Female (Male), and Percentile Ranking.

We find that the main treatment effect, Y2009 $\times$ Tier1, is negative and significant, indicating that students list more prestigious colleges as their Tier 1 first choices in 2009, with a magnitude between 4-5%. This is consistent with the theoretical prediction of the insurance property of the parallel mechanism compared to the Immediate Acceptance mechanism. We further find that student Tier 2 first choices in 2009 are ranked about 5% lower than the corresponding ranking in 2008 (Y2009: $p < 0.001$).

Table 4.12: Effects of matching mechanisms on the prestige of reported first choices (OLS)

Dependent variable:	Prestige of first-choice college		
	(1)	(2)	(3)
Y2009	-0.051***	-0.052***	-0.052***
	-0.003	-0.003	-0.004
Tier1	0.191***	0.200***	0.198***
	-0.005	-0.004	-0.005
Y2009 × Tier1	0.044***	0.046***	0.047***
	-0.01	-0.011	-0.012
Percentile Ranking		0.294***	0.293***
		-0.015	-0.015
Science			0.022
			-0.011
Rural			-0.009
			-0.006
Female			-0.009
			-0.009
Constant	0.567***	0.412***	0.408***
	-0.013	-0.011	-0.009
Observations	6,757	6,757	6,757
R^2	0.117	0.221	0.223

Notes: Standard errors in parentheses are clustered at the high school level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

It is not clear what drives this effect. Consistent with theory, we find that students with a one-percentile increase in test scores increase the ranking of their first-choice colleges by 0.29%. Lastly, Tier 1 first choices in 2008 ranked between 19-20% higher than Tier 2 first choices, which might reflect risk aversion in the event that students did not get into Tier 1 colleges.

Table 4.13: Effects of matching mechanisms on the likelihood of ranking a local college as first choice: Probit

Dependent variable:	Local college as first choice		
	(1)	(2)	(3)
Y2009	0.094***	0.094***	0.093***
	(0.022)	(0.022)	(0.023)
Tier1	-0.289***	-0.293***	-0.297***
	(0.023)	(0.026)	(0.025)
Y2009 × Tier1	-0.043*	-0.044*	-0.043*
	(0.024)	(0.025)	(0.026)
Percentile Ranking		-0.164***	-0.164***
		(0.011)	(0.012)
Science			0.053**
			(0.025)
Rural			-0.019**
			(0.007)
Female			0.013
			(0.013)
Observations	6,757	6,757	6,757

Notes: Standard errors in parentheses are clustered at the high school level; reporting marginal effects. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

In several laboratory studies of school choice, a safe strategy adopted by many students under the Immediate Acceptance mechanism is the district school bias, where students rank their district school higher than its true ranking since student priority is higher in their district school. A similarly safe strategy in the college admissions context is to list one's local college (a college in the same province) as the first choice. Local colleges are generally considered "safer" choices because they have higher quota for local students.

Table 4.13 presents three probit specifications, each examining the effects of mechanism change on the likelihood that a student adopts a safe strategy, i.e., lists a local college as his or her first choice. The dependent variable is whether a student lists a local college as his or her first choice (by tier). The independent variables (omitted) again include Y2009 (Y2008), Tier1 (Tier2), Y2009 $\times$ Tier1, Science (Humanities), Rural (Urban), Female (Male), and Percentile Ranking.

We have several interesting findings. First, in 2009, students are 9% more likely to list a local college as their first choice in Tier 2 compared to the previous year ($p < 0.01$). Second, the main treatment effect, Y2009 $\times$ Tier1, is negative and marginally significant, indicating a 4% decrease in the likelihood of listing a local college as first choice for Tier 1 college ranking (-0.043, $p < 0.10$). This effect, though weakly significant, is consistent with the theoretical prediction. Third, in 2008, students are 29.7% less likely to list a local college in Tier 1 than Tier 2 college rankings. Additionally, we find that students in the science track (from rural areas) are 5% more (2% less) likely to list a local college as their first choice. Finally, students with a one-percentile increase in test scores decrease their likelihood of listing a local college as their first choice by 0.16%.

We now summarize our treatment effect on student ranking strategies below:

Result 23. *Changing Tier 1 admissions mechanism from the Immediate Acceptance to the parallel mechanism in 2009 leads to a significant increase of the length of the rank ordered list by approximate one more college, a significant 5% increase of the prestige of first-choice colleges, and a marginally-significant 4% decrease in the likelihood of listing a local college as their first choice.*

Our analysis indicates that the replacement of the Immediate Acceptance mechanism with the parallel mechanism in Tier 1 admissions in 2009 results in less manipulation by truncation in the rank ordered list, and more risk taking in the ranking of first choices. These empirical results are consistent with the theoretical prediction that the parallel mechanism is less manipulable than the Immediate Acceptance mechanism, and that it offers insurance to students. The result regarding the prestige of the first choice college also mirrors that of the rank bias result in the laboratory experiment: the lower students are ranked in the test, the less prestigious colleges they will rank as their first choice, and controlling for rankings in the test, the "rank" of first choice colleges are lower under the IA mechanism. Therefore, it provides empirical support for Hypothesis 14.

Next we investigate the effects of mechanism change on matching outcomes. First, we look at the effects of mechanism change on the likelihood that a student is admitted by his reported first choice. Table 4.14 reports three probit specifications, each examining the effects of mechanism change on the likelihood of first choice accommodation (1), with student percentile ranking (2) and demographic (3) controls. The independent variables (omitted) again include Y2009 (Y2008), Tier1 (Tier2), Y2009 $\times$ Tier1, Science (Humanities), Rural (Urban), Female (Male), and Percentile Ranking.

Consistent with our theoretical prediction, we find that the main treatment effect, Y2009 $\times$ Tier1, is negative and significant, indicating that students are 24% less likely to be admitted by their reported top choices in Tier 1 admissions in 2009, whereas the likelihood of being admitted by first-choice colleges in Tier 2 in 2009 does not change compared to the previous year (-0.001, $p > 0.10$). Additionally, students from rural areas are 4% more likely to be admitted into their reported first choices. Finally, students with a one-percentile increase in test scores increase their likelihood of being admitted by their first choice by 0.56% ($p < 0.001$).

Next we investigate the effects of mechanism change on matching stability. In the context of centralized college admissions, we characterize a matching outcome to be more stable if the proportion of students a student has justified envy towards is lower. Student 1 exhibits *justified envy* toward student 2 if student 2 scores lower in

Table 4.14: Effects of matching mechanisms on first choice accommodation: Probit

Dependent variable:	Admitted to first choice		
	(1)	(2)	(3)
Y2009	-0.001	0.001	-0.001
	(0.021)	(0.021)	(0.021)
Tier1	0.071	0.081**	0.077**
	(0.050)	(0.038)	(0.038)
Y2009 × Tier1	-0.230***	-0.242***	-0.239***
	(0.060)	(0.056)	(0.056)
Percentile Ranking		0.558***	0.558***
		(0.022)	(0.020)
Science			0.038***
			(0.010)
Rural			0.039***
			(0.012)
Female			0.004
			(0.007)
Observations	6,567	6,567	6,567

Notes: Standard errors in parentheses are clustered at the high school level; reporting marginal effects. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 4.15: Effects of matching mechanisms on first 5 choices accommodation: Probit

Dependent variable:	Admitted to first 5 choice		
	(1)	(2)	(3)
Y2009	-0.018	-0.013	-0.014
	(0.023)	(0.022)	(0.022)
Tier1	0.003	0.016	0.016
	(0.042)	(0.033)	(0.033)
Y2009 × Tier1	0.037	0.032	0.033
	(0.047)	(0.040)	(0.041)
Percentile Ranking		0.455***	0.455***
		(0.024)	(0.021)
Science			0.000
			(0.007)
Rural			0.040***
			(0.011)
Female			0.000
			(0.009)
Observations	6,567	6,567	6,567

Notes: Standard errors in parentheses are clustered at the high school level; reporting marginal effects. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

the college entrance exams, but is admitted into a higher ranked college. Similar to Table 4.12, we use the coarse ranking of colleges as proxies for student preferences to calculate justified envy. Since matching is carried out separately by track, year and tier, we divide our sample into 8 markets, and calculate a stability score for each student in an appropriate market using the following formula:

$$\text{Stability Score}_i = 1 - \frac{\#\text{ of students } i \text{ has justified envy toward}}{\text{Total number of students in the corresponding market}}$$

Table 4.16: Effects of matching mechanisms on stability: OLS

Dependent variable:	Stability score		
	(1)	(2)	(3)
Y2009	0.014***	0.016***	0.016***
	(0.003)	(0.003)	(0.002)
Tier1	0.021**	0.015**	0.016**
	(0.007)	(0.005)	(0.005)
Y2009 × Tier1	-0.017**	-0.019**	-0.019**
	(0.006)	(0.006)	(0.005)
Percentile Ranking		-0.196***	-0.196***
		(0.006)	(0.005)
Science			-0.007***
			(0.002)
Rural			0.007*
			(0.003)
Female			-0.001
			(0.001)
Constant	0.866***	0.969***	0.969***
	(0.007)	(0.005)	(0.005)
Observations	5,594	5,594	5,594
R^2	0.004	0.190	0.191

Notes: Standard errors in parentheses are clustered at the high school level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 4.16 reports the effects of mechanism change on matching stability. The dependent variable is the stability score, whereas the independent variables (omitted) again include Y2009 (Y2008), Tier1 (Tier2), Y2009 × Tier1, Science (Humanities), Rural (Urban), Female (Male), and Percentile Ranking. We find that the main treatment effect, Y2009 × Tier1, is negative and significant (-0.019, $p < 0.05$), contrary to the theoretical prediction. While the effect is statistically significant, it is economically small. The average stability score in Tiers 1 and 2 in 2009 are not significantly different (0.88 in both years). In addition, we find that Tier 2 matching outcomes in 2009 are 1.4% more stable than those in 2008, whereas in 2008, Tier 2 matching outcomes are 2% more stable than those in Tier 1. We also find that stability scores for students in science track are 0.7% lower, whereas a one percentile increase in test scores decreases the stability score by 0.002. We summarize our matching outcome analysis below:

Result 24. *Changing Tier 1 admissions mechanism from the Immediate Acceptance to the parallel mechanism in 2009 leads to a significant 24% decrease in admission by reported top-choices colleges, and a significant 2% decrease in stability scores.*

The first choice accommodation result is consistent with theoretical predictions. It provides empirical support for Hypothesis 16 in addition to the experiment by showing the similar result: students are indeed more focused on getting into their reported first choices under IA . The stability result however, while small, is contrary to theoretical predictions.

The above results regarding strategies and outcomes come from the difference-in-differences estimates, relying on the fact that the mechanism for Tier 2 does not change across the years. However, one concern about the use of DiD in this context is that different mechanisms in Tier 1 may cause different kinds of students to be left unadmitted after admissions for this tier. Those “left-over” students then continue to participate in Tier 2 admissions. Because the composition of these students may be different, the validity of using Tier 2 in DiD may not hold, because the composition of students who participate also changes as a result of the mechanism change in Tier

1. To address this concern, we exclude the bottom 20% ranked students in Tier 1, and the top 20% ranked students in Tier 2: those who are most likely to be left over and go to the next tier. Appendix 4.9.2 summarizes the results from this robustness check. We find that other than the disappearance of the already weak result of local college bias, all results remain robust.

4.8 Conclusion

College admission is one of the most important policy problems throughout the world, and it has significant social welfare and economic development implication. In China, every year there are around 10 millions of students going through the college admission process, with mechanisms for matching students with colleges all over the place, and virtually no two provinces use exactly the same mechanism. Some admission mechanisms used in some Chinese cities/provinces have some good properties, which have not been thoroughly studied. Previous experimental matching literatures in related areas focus on the school choice problem, which is different from centralized college admissions in practice. Moreover, the empirical evidences of the effect of change in mechanisms are few.

In the first part of the paper, we design a laboratory experiment modeled in the fashion of the centralized college admission clearinghouse in China, recruiting subjects from a Chinese college, who have had high stakes and made decisions in the similar real world game. We compare three mechanisms studied in Chen and Kesten (2014) under the school choice context: the Immediate Acceptance mechanism, the Deferred Acceptance mechanism and the Chinese Parallel mechanism, with increasing emphasis on the reported first choices. Consistent with previous theoretical and experimental results, we find that the more emphasis a mechanism puts on the first choice, the more susceptible to strategic manipulation it is. Not surprisingly, a direct result is that more students get admitted to their reported first choices under IA than

PA, which is followed DA. The particular pattern of manipulation we identify in this paper, the rank bias, also follows such order. However, we find that strategic manipulation does not lead to efficiency loss, with all three mechanisms achieving same, very high level of social welfare. This can be explained by both the designed game environment, as well as students' experience with allows them to coordinate well. Manipulation does come at a cost: since stable matchings are identical to Nash equilibrium outcomes, failure to reach the equilibrium outcomes lead to justified envy. Therefore, although in theory all three mechanisms predict stable matchings when students play Nash equilibrium strategies, the more manipulable a mechanism is, the more unstable matching outcomes it produces.

To supplement the experimental findings, we analyze a natural experiment with the help of difference-in-differences estimators. Although some theoretical properties of matching mechanisms can not be directly tested empirically, due to the fact that we do not know students' true preferences, we still can draw some analogies between the lab and the field. Specifically, because of the particular matching market we study – the college admissions matching – we can make some assumptions to approximate students' preferences by college qualities. Mirroring the rank bias result from the laboratory experiment, we find that when the mechanism changes from the Immediate Acceptance mechanism to the parallel mechanism, students list better colleges as their first choices, and they are less likely to list local, "safe" colleges as their first choices. The first-choice accommodation prediction is also confirmed empirically. We also find that students list more colleges in their rank-ordered list, which is not testable by the lab experiment. However, we do not find economically significant difference in stability.

Our findings provide strong support for the theoretical properties of college admissions mechanisms. The findings stand both in a well controlled environment, as well as in an environment with high external validity, where the scale of the market is large, and the stakes are high. Our results also imply that if more truthful revelation of preferences over colleges by students is a top policy concern, then moving from the Immediate Acceptance mechanism to the parallel mechanism most likely will achieve that goal. On the other hand, if matching stability is more important, the evidence at hand is not very obvious.

4.9 Appendix

4.9.1 Additional regression tables for the laboratory experiment

Table 4.17: Rank Bias - Additional Regressions

Dependent variable:	First choice college	
	(1)	(2)
DA	-0.035	0.062
	(0.078)	(0.067)
SH	-0.293***	-0.251***
	(0.085)	(0.087)
Ranking	0.809***	0.833***
	(0.073)	(0.071)
Ranking *times* DA	-0.298***	-0.344***
	(0.047)	(0.042)
Ranking *times* SH	-0.029	-0.047*
	(0.023)	(0.028)
Period		0.008*
		(0.005)
HL Switch Point		0.026
		(0.025)
Constant cut1	1.473***	1.794***
	(0.081)	(0.185)
Constant cut2	2.621***	2.948***
	(0.235)	(0.278)
Constant cut3	3.105***	3.422***
	(0.287)	(0.314)
Constant cut4	4.186***	4.525***
	(0.410)	(0.421)
Constant cut5	6.211***	6.683***
	(0.532)	(0.510)
Observations	4,320	3,920

Notes: Ordered Probit regression testing the effect of ranking on the true position of the reported first-choice colleges; reporting regression coefficients; standard errors in parentheses are clustered at session level. ***: $p < 0.01$; **: $p < 0.05$; *: $p < 0.1$. When controlling for risk attitude, 20 students (400 observations) are dropped due to multiple switches in the lottery game.

4.9.2 Robustness checks

4.9.2.1 Excluding students at the boundary

Table 4.18: Effects of matching mechanisms on the length of rank-ordered lists (OLS), excluding students in the bottom 20% test scores in Tier 1, and in the top 20% test scores in Tier 2.

Dependent variable:	Length of ROL		
	(1)	(2)	(3)
Y2009	0.113	0.113	0.117
	(0.089)	(0.089)	(0.091)
Tier1	-0.524***	-0.606***	-0.585***
	(0.085)	(0.083)	(0.084)
Y2009 × Tier1	0.613***	0.613***	0.602***
	(0.098)	(0.095)	(0.099)
Percentile Ranking		0.420***	0.423***
		(0.077)	(0.075)
Science			-0.158**
			(0.044)
Rural			-0.073
			(0.043)
Female			0.077
			(0.045)
Constant	4.430***	4.262***	4.403***
	(0.096)	(0.111)	(0.128)
Observations	6,036	6,036	6,036
R^2	0.032	0.042	0.050

Notes: Standard errors in parentheses are clustered at the high school level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Compared to the original result in Table 4.11, the average tier effect remains similar. The coefficient for the interaction becomes smaller, indicating that the difference in differences of the length of ROLs becomes smaller. The sign of the coefficient of percentile ranking changes. This may be caused by the difference in this variable's range after 20% of students are excluded based on their ranking of test scores (either bottom or top). Additionally, the significance of the female dummy is gone.

Table 4.19: Effects of matching mechanisms on the prestige of reported first choices (OLS), excluding students in the bottom 20% test scores in Tier 1, and in the top 20% test scores in Tier 2.

Dependent variable:	Prestige of first-choice college		
	(1)	(2)	(3)
Y2009	-0.052***	-0.053***	-0.053***
	(0.005)	(0.005)	(0.004)
Tier1	0.238***	0.185***	0.184***
	(0.004)	(0.002)	(0.003)
Y2009 × Tier1	0.047***	0.048***	0.048***
	(0.011)	(0.012)	(0.012)
Percentile Ranking		0.312***	0.311***
		(0.011)	(0.011)
Science			-0.002
			(0.010)
Rural			-0.017
			(0.009)
Female			-0.005
			(0.010)
Constant	0.533***	0.401***	0.419***
	(0.012)	(0.008)	(0.012)
Observations	5,301	5,301	5,301
R^2	0.166	0.239	0.240

Notes: Standard errors in parentheses are clustered at the high school level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Compared to Table 4.12, this result remains robust and similar.

Table 4.20: Effects of matching mechanisms on the likelihood of ranking a local college as first choice: Probit, excluding students in the bottom 20% test scores in Tier 1, and in the top 20% test scores in Tier 2.

Dependent variable:	Local college as first choice		
	(1)	(2)	(3)
Y2009	0.097***	0.097***	0.097***
	(0.030)	(0.030)	(0.030)
Tier1	-0.324***	-0.325***	-0.325***
	(0.027)	(0.026)	(0.026)
Y2009 × Tier1	-0.031	-0.031	-0.032
	(0.030)	(0.030)	(0.032)
Percentile Ranking		0.006	0.007
		(0.027)	(0.028)
Science			0.017
			(0.022)
Rural			-0.027**
			(0.013)
Female			0.027*
			(0.016)
Observations	5,301	5,301	5,301

Notes: Standard errors in parentheses are clustered at the high school level; reporting marginal effects. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Compared to Table 4.13, the previous weakly significant coefficient of the interaction is now not significant anymore. The coefficient of percentile ranking is also no longer significant, which may similarly be explained by the change in variable range. Additionally, the science dummy becomes non-significant, while the female dummy becomes weakly significant.

Table 4.21: Effects of matching mechanisms on first choice accommodation: Probit, excluding students in the bottom 20% test scores in Tier 1, and in the top 20% test scores in Tier 2.

Dependent variable:	Admitted to first choice		
	(1)	(2)	(3)
Y2009	-0.011	-0.009	-0.011
	(0.023)	(0.023)	(0.023)
Tier1	0.268***	0.175***	0.170***
	(0.037)	(0.026)	(0.024)
Y2009 × Tier1	-0.295***	-0.303***	-0.298***
	(0.051)	(0.051)	(0.049)
Percentile Ranking		0.517***	0.517***
		(0.029)	(0.027)
Science			0.036***
			(0.013)
Rural			0.047***
			(0.014)
Female			-0.001
			(0.010)
Observations	5,261	5,261	5,261

Notes: Standard errors in parentheses are clustered at the high school level; reporting marginal effects. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Compared to Table 4.14, the most obvious change is the coefficent for the tier effect. This is however not surprising given that we exclude the least likely students to get admitted to first choice for Tier 1, and the most likely for Tier 2.

Table 4.22: Effects of matching mechanisms on stability: OLS, excluding students in the bottom 20% test scores in Tier 1, and in the top 20% test scores in Tier 2.

Dependent variable:	Stability score		
	(1)	(2)	(3)
Y2009	0.017***	0.018***	0.018***
	(0.004)	(0.003)	(0.002)
Tier1	-0.017**	0.024***	0.025***
	(0.006)	(0.005)	(0.005)
Y2009 × Tier1	-0.017**	-0.018**	-0.019**
	(0.006)	(0.006)	(0.006)
Percentile Ranking		-0.233***	-0.233***
		(0.008)	(0.008)
Science			-0.011***
			(0.001)
Rural			0.007*
			(0.003)
Female			-0.002
			(0.003)
Constant	0.881***	0.978***	0.981***
	(0.006)	(0.006)	(0.004)
Observations	4,361	4,361	4,361
R-squared	0.013	0.213	0.215

Notes: Standard errors in parentheses are clustered at the high school level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Compared to Table 4.16, the difference lies in the coefficient of percentile ranking, which becomes larger, and coefficient for Tier 1 when ranking is not controlled. This may also be explained by the difference in matching outcomes caused by the exclusion of either bottom or top ranked students.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abdulkadiroğlu, A., Pathak, P., Roth, A. E., and Sönmez, T. (2006). Changing the Boston school choice mechanism. Working paper w11965, National Bureau of Economic Research.

Abdulkadiroğlu, A., Pathak, P. A., and Roth, A. E. (2009). Strategy-proofness versus efficiency in matching with indifferences: Redesigning the NYC high school match. *American Economic Review*, 99(5):1954–1978.

Abdulkadiroğlu, A. and Sönmez, T. (1998). Random serial dictatorship and the core from random endowments in house allocation problems. *Econometrica*, 66:689–701.

Abdulkadiroğlu and Sönmez, T. (2003). School choice: a mechanism design approach. *American Economic Review*, 93(3):729–747.

Abdulkadiroğlu, A., Che, Y.-K., and Yasuda, Y. (2011). Resolving conflicting preferences in school choice: the boston mechanism reconsidered. *American Economic Review*, 101(1):399–410.

Abdulkadiroğlu, A., Pathak, P. A., and Roth, A. E. (2005a). The New York City high school match. *American Economic Review*, 95(2):364–367.

Abdulkadiroğlu, A., Pathak, P. A., Roth, A. E., and Sönmez, T. (2005b). The Boston public school match. *American Economic Review*, 95(2):368–371.

Abdulkadiroğlu, A. and Sönmez, T. (1999). House allocation with existing tenants. *Journal of Economic Theory*, 88:233–260.

Ashlagi, I., Braverman, M., and Hassidim, A. (2011). Matching with couples revisited. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 335–336. ACM.

Ashlagi, I. and Roth, A. (2011). Individual rationality and participation in large scale, multi-hospital kidney exchange. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 321–322. ACM.

Avery, C., Fairbanks, A., and Zeckhauser, R. J. (2009). *The early admissions game.* Harvard University Press.

Azevedo, E. M. and Budish, E. (2013). Strategyproofness in the large. `http://assets.wharton.upenn.edu/~eazevedo/papers/Azevedo-Budish-SPL.pdf`.

Azevedo, E. M. and Hatfield, J. W. (2012). Complementarity and multidimensional heterogeneity in matching markets. Technical report, mimeo.

Azevedo, E. M. and Leshno, J. D. (2011). Can we make school choice more efficient? Unpublished.

Azevedo, E. M. and Leshno, J. D. (2014). A supply and demand framework for two-sided matching markets. http://assets.wharton.upenn.edu/~eazevedo/papers/Azevedo-Leshno-Supply-and-Demand-Matching.pdf.

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

Bloom, B. S., Madaus, G. F., and Hastings, J. T. (1981). Evaluation to improve learning.

Bogomolnaia, A. and Moulin, H. (2001). A new solution to the random assignment problem. *Journal of Economic Theory*, 100(2):295–328.

Braun, S., Dwenger, N., and Kübler, D. (2010). Telling the truth may not pay off: an empirical study of centralized university admissions in germany. *The BE Journal of Economic Analysis & Policy*, 10(1).

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

Carvalho, J. R. and Magnac, T. (2010). College choice and entry exams. Working paper.

Che, Y.-K., Kim, J., and Kojima, F. (2013). Efficient assignment with interdependent values. Technical report, Working paper.

Chen, Y., Jiang, M., and Kesten, O. Chinese college admissions reforms: Experimental and empirical evaluations. Working in progress.

Chen, Y. and Kesten, O. (2014). Chinese college admissions and school choice reforms: Theory and experiments. http://yanchen.people.si.umich.edu/papers/Chen_Kesten_2014_01.pdf.

Chen, Y. and Sönmez, T. (2002). Improving efficiency of on-campus housing: an experimental study. *American Economic Review*, 92(5):1669–1686.

Chen, Y. and Sönmez, T. (2004). An experimental study of house allocation mechanisms. *Economic Letters*, 83(1):137–140.

Chen, Y. and Sönmez, T. (2006). School choice: an experimental study. *Journal of Economic Theory*, 127(1):202 – 231.

Chen, Y. and Takeuchi, K. (2010). Multi-object auctions with package bidding: An experimental comparison of Vickrey and iBEA. *Games and Economic Behavior*, 68(2):557 – 579.

Chiu, Y. S. and Weng, W. (2009). Endogenous preferential treatment in centralized admissions. *The RAND Journal of Economics*, 40(2):258–282.

Cripps, M. W. and Swinkels, J. M. (2006). Efficiency of large double auctions. *Econometrica*, 74(1):47–92.

Davis, A. M., Katok, E., and Kwasnica, A. M. (2010). Do auctioneers pick optimal reserve prices? *Management Science*, pages 1–16.

Dubins, L. E. and Freedman, D. A. (1981). Machiavelli and the gale-shapley algorithm. *American Mathematical Monthly*, 88:485–494.

Ergin, H. I. and Sönmez, T. (2006). Games of school choice under the boston mechanism. *Journal of Public Economics*, 90(1-2):215 – 237.

Featherstone, C. (2013). Rank effciency: Investigating a widespread ordinal welfare criterion. mimeo.

Featherstone, C. and Niederle, M. (2014). School choice mechanisms under incomplete information: an experimental investigation. Working Paper.

Ferraro, P. J. and Vossler, C. A. (2010). The source and significance of confusion in public goods experiments. *The B.E. Journal of Economic Analysis & Policy*, 10(1):Article 53.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10:171–178. 10.1007/s10683-006-9159-4.

Fragiadakis, D. E. and Troyan, P. (2013). Improving welfare in assignment problems: an experimental investigation. mimeo.

Fréchette, G. R., Roth, A. E., and Ünver, M. U. (2007). Unraveling yields inefficient matchings: evidence from post-season college football bowls. *The RAND Journal of Economics*, 38(4):967–982.

Fudenberg, D., Mobius, M., and Szeidl, A. (2007). Existence of equilibrium in large double auctions. *Journal of Economic theory*, 133(1):550–567.

Gale, D. and Shapley, L. S. (1962). College admissions and the stability of marriage. *American Mathematical Monthly*, 69:9–15.

Gresik, T. A. and Satterthwaite, M. A. (1989). The rate at which a simple market converges to efficiency as the number of traders increases: An asymptotic result for optimal trading mechanisms. *Journal of Economic Theory*, 48(1):304–332.

Hatfield, J. W., Kojima, F., and Narita, Y. (2011). Promoting school competition through school choice: a market design approach. *Unpublished manuscript*.

He, Y. (2014). Gaming the Boston school choice mechanism in Beijing. https://26b054f3-a-62cb3a1a-s-sites.googlegroups.com/site/yinghuahe/research/papers/BM.pdf.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655.

Hoxby, C. M. (2003). School choice and school productivity: Could school choice be a tide that lifts all boats? In Hoxby, C. M., editor, *The Economics of School Choice*. University of Chicago Press, Chicago.

Immorlica, N. and Mahdian, M. (2005). Marriage, honesty, and stability. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, pages 53–62, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Kagel, J. H. and Levin, D. (2001). Behavior in multi-unit demand auctions: Experiments with uniform price and dynamic Vickrey auctions. *Econometrica*, 69(2):413–454.

Kagel, J. H. and Roth, A. E. (2000). The dynamics of reorganization in matching markets: a laboratory experiment motivated by a natural experiment. *Quarterly Journal of Economics*, 115(1):201–235.

Kesten, O. (2011). On two kinds of manipulation for school choice problems. *Economic Theory*, pages 1–17.

Klijn, F., Pais, J., and Vorsatz, M. (2013). Preference intensities and risk aversion in school choice: A laboratory experiment. *Experimental Economics*, 16(1):1–22.

Kojima, F. and Manea, M. (2010). Incentives in the probabilistic serial mechanism. *Journal of Economic Theory*, 145(1):106–123.

Kojima, F. and Pathak, P. A. (2009). Incentives and stability in large two-sided matching markets. *The American Economic Review*, 99(3):608–627.

Kojima, F., Pathak, P. A., and Roth, A. E. (2013). Matching with couples: Stability and incentives in large markets. *The Quarterly Journal of Economics*, 1(48):48.

Lien, J. W., Zheng, J., and Zhong, X. (2012). Preference submission timing in school choice matching: Testing efficiency and fairness in the laboratory. Technical report, Working Paper.

Mongell, S. and Roth, A. E. (1991). Sorority rush as a two-sided matching mechanism. *American Economic Review*, 81:441–464.

Niederle, M. and Roth, A. E. (2003). Unraveling reduces mobility in a labor market: gastroenterology with and without a centralized match. *Journal of Political Economy*, 111(6):1342–1352.

Pais, J. and Pintér, A. (2008). School choice and information: an experimental study on matching mechanisms. *Games and Economic Behavior*, 64(1):303 – 328.

Pais, J. and Pintér, Á. (2008). School choice and information: An experimental study on matching mechanisms. *Games and Economic Behavior*, 64(1):303 – 328.

Pathak, P. A. and Sönmez, T. (2008). Leveling the playing field: sincere and sophisticated players in the boston mechanism. *American Economic Review*, 98:1636–1652(17).

Pathak, P. A. and Sönmez, T. (2013). School admissions reform in chicago and england: Comparing mechanisms by their vulnerability to manipulation. *American Economic Review*, 103(1):80–106.

Pesendorfer, W. and Swinkels, J. M. (2000). Efficiency and information aggregation in auctions. *American Economic Review*, pages 499–525.

Roberts, D. J. and Postlewaite, A. (1976). The incentives for price-taking behavior in large exchange economies. *Econometrica*, 44(1):115–27.

Roth, A. (1990). New physicians:a natural experiment in market organization. *Science*, 250:1524–1528.

Roth, A. E. (1982). The economics of matching: stability and incentives. *Mathematics of Operations Research*, 7:617–628.

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

Roth, A. E. (1985). The college admissions problem is not equivalent to the marriage problem. *Journal of Economic Theory*, 36(2):277–288.

Roth, A. E. (1986). On the allocation of residents to rural hospitals: a general property of two-sided matching markets. *Econometrica*, 54(2):425–427.

Roth, A. E. (1989). Two-sided matching with incomplete information about others' preferences. *Games and Economic Behavior*, 1(2):191–209.

Roth, A. E. (1991). A natural experiment in the organization of entry-level labor markets: regional markets for new physicians and surgeons in the united kingdom. *American Economic Review*, 81:414–440.

Roth, A. E. (2002). The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70:1341–1378.

Roth, A. E. and Peranson, E. (1999). The redesign of the matching market for american physicians: some engineering aspects of economic design. *American Economic Review*, 89(4):748–780.

Roth, A. E., Sönmez, T., and Ünver, M. U. (2007). Efficient kidney exchange: Coincidence of wants in markets with compatibility-based preferences. *The American economic review*, pages 828–851.

Roth, A. E., Sönmez, T., and Utku Ünver, M. (2005). Pairwise kidney exchange. *Journal of Economic Theory*, 125(2):151–188.

Roth, A. E. and Sotomayor, M. A. O. (1990). *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge University Press, Cambridge.

Roth, A. E. and Xing, X. (1994). Jumping the gun: Imperfections and institutions related to the timing of market transactions. *The American Economic Review*, pages 992–1044.

Rustichini, A., Satterthwaite, M. A., and Williams, S. R. (1994). Convergence to efficiency in a simple market with incomplete information. *Econometrica*, 62(5):1041–63.

Sönmez, T. (1999). Can pre-arranged matches be avoided in two-sided matching markets? *Journal of Economic Theory*, 86(1):148–156.

Sönmez, T. and Switzer, T. B. (2013). Matching with (branch-of-choice) contracts at the united states military academy. *Econometrica*, 81(2):451–488.

Swinkels, J. M. (2001). Efficiency of large private value auctions. *Econometrica*, 69(1):37–68.

Wu, B. and Zhong, X. (2014). Matching mechanisms and matching quality: Evidence from a top university in china. *Games and Economic Behavior*, 84(0):196 – 215.

Zhong, X., Cheng, N., and He, Y. (2004). Where have all the flowers gone? A game-theoretic analysis of the National College Entrance Exam preference submission mechanisms. *China Economic Quarterly*, 3(3):763–778.