Matching in the Large: An Experimental Study[*]

Yan Chen   Ming Jiang   Onur Kesten   Stéphane Robin   Min Zhu

March 15, 2016

Abstract

Market size has been predicted to play an influential role in a broad class of economic environments. We study the performance of the Boston and the Deferred Acceptance (DA) mechanism in a laboratory where we increase the market size. Our 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. We then further increase the market size to 4,000 by introducing robots. When humans play truthful robots (without strategic uncertainty), we find that scale has no effect on best response behavior. However, when humans play empirical robots (with strategic uncertainty), scale increases best responses under both mechanisms, which is likely due to the increase in the precision of subjects' beliefs about others' strategies.

Keywords: matching, school choice, experiment, scale

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 China and Turkey 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., 2012) 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, 2013; 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 mechanisms 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

---

[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.

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). While much of the theoretical literature on market size focuses on two-sided matching and how properties of mechanisms change with size, as the first experimental study to understand the behavioral effects of market size, we have chosen the one-sided school choice setting where the theoretical predictions about the properties of the two mechanisms we study remain unchanged with market size. In this sense, our approach can be viewed as a behavioral complement to existing literature, rather than a direct test of it.

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 truthful or empirical 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. We then further increase the market size to 4,000. When humans play truthful robots (without strategic uncertainty), we find that scale has no effect on best response behavior. However, when humans play empirical robots (with strategic uncertainty), scale increases best responses under both mechanisms, which is likely due to the increase in the precision of subjects' beliefs about others' strategies. Finally, we find that the matching outcomes under the DA mechanism remain stable as market size increases. Overall, our results provide additional support for the replacement of the Boston mechanism with the DA mechanism in practice.

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. 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.

Note that there are two kinds of large market modeling approaches in the literature. In the context of school choice, one approach holds the number of schools constant while letting the size of each school increase, and the other holds the size of each school constant while the number of schools increases. We adopt the former setting for our experiment as we believe this modeling choice fits well with many school districts in the United States as well as the college admissions setting in China.

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

---

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

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 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), peer information sharing in networks Ding and Schotter (2015b) and intergenerational advice Ding and Schotter (2015a) 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 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

---

[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 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 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

Chen and Sönmez (2002), who study house allocations when the number of students increases from 12 to 60 per 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. To our knowledge, this is the first experiment which systematically investigates the scale effect in matching.

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 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

---

[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.

(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.

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 ($m = 4,000$) matching market. 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 strategic uncertainty with regard to other participant strategies.

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 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 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. 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. The experiment is implemented under complete information. That is, participants know the entire payoff table (Table 2), school capacity and priorities, and their exact lottery number before submitting their rank ordered list.

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.

Table 3: 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

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

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 4: 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		
Block 2	not NE	dominant strategy	$\mu^{S/C}$	$\{\mu^{S/C}, \mu^*\}$
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 4 summarizes the properties of this game.

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.

When human subjects form beliefs about others' strategies, the precision of this belief might vary with market size. The introduction of truthful robots and empirical robots provides crucial variations in the aggregated uncertainty that human subjects face, and thus in their beliefs.

There are three advantages associated with the human-vs-robots design. The first advantage is 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.

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 5: Features of Experimental Sessions

Mechanisms	Composition	Match size	Robot strategies	#Sbj. $\times$ # sessions	Total# of sbj.
Boston	All-human	4	n/a	12×4	48
		40	n/a	80×2	160
	Human-robot	40	Empirical	20×2	40
			Truthful	20×2	40
		4000	Empirical	20×2	40
			Truthful	20×2	40
DA	All-human	4	n/a	12×4	48
		40	n/a	80×2	160
	Human-robot	40	Empirical	20×2	40
			Truthful	20×2	40
		4000	Empirical	20×2	40
			Truthful	20×2	40

Table 5 summarizes the features of the different experimental sessions. For 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] The English translations of the experimental instructions can be found in Appendix A.

5   Results

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

5.1   Truth-telling and district school bias

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.[7]

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

Hypothesis 2 (Truth-telling: scale effect)**.** Scale has no effect on the proportion of truth-telling under either the Boston mechanism or the DA mechanism.

---

[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.

[7]As a robustness check, we use the complete rank ordered list for truth-telling for both mechanisms and find that the mechanism and scale effects remain albeit with different effect size. See Appendix B.1 for the corresponding regression analysis.
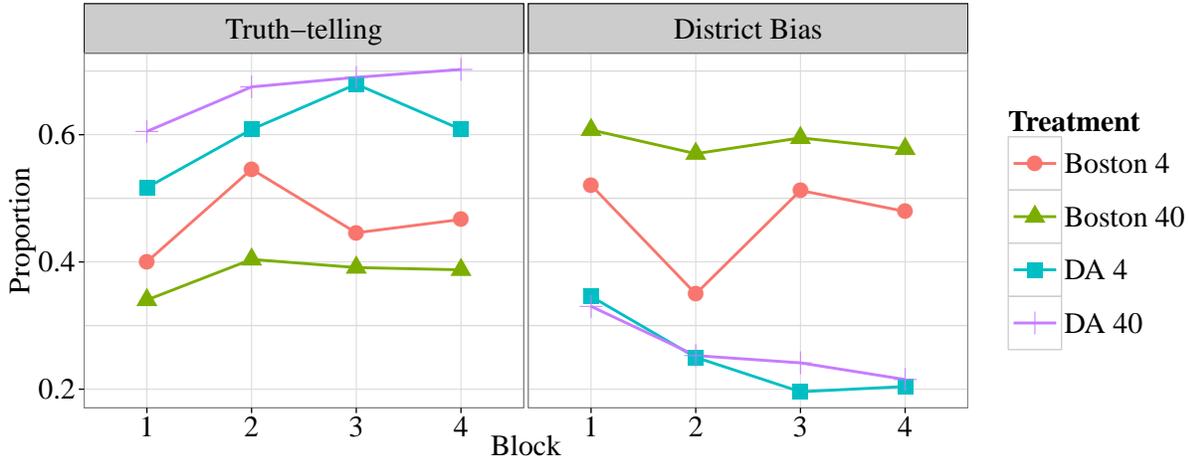
Figure 1: Truth-telling and district school bias in all-human treatments

We first examine our all-human sessions in treatments with $m = 4$, and 40. Figure 1 presents the proportion of truth-telling (left panel) and district school bias (right panel) by block for each of the four all-human treatments. The ranking of mechanisms in the 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 6 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), Period, LotteryPosition, and interaction terms. In each specification, standard errors are clustered at the session level. In this as well as all subsequent probit regressions, we report average marginal effect.[8] We summarize the results below.

Result 1 (Truth-telling: mechanism effect, all human). *The proportion of truth-telling is 14-percentage (29-percentage) points 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 6, we see that the coefficients of DA as well as DA $\times$ Scale40 are positive and significant, indicating that truth-telling under DA in the 4-student (40-student) matches is 14-percentage (29-percentage) points higher compared to that in the corresponding sessions under the Boston mechanism.

---

[8]The marginal effect for an interaction term, e.g., DA $\times$ Scale40, is the difference between the average marginal effect (AME) of the dummy DA at Scale40 $= 1$ and at Scale40 $= 0$; the standard errors are calculated using the Delta method (Ai and Norton, 2003).

Table 6: Truth-telling and district school bias: probit regressions

	Truth-telling		District School Bias	
All-human treatments	(1)	(2)	(3)	(4)
Scale40	-0.084***	-0.084***	0.122***	0.122***
	(0.025)	(0.025)	(0.037)	(0.037)
DA	0.139***	0.139***	-0.217***	-0.216***
	(0.032)	(0.032)	(0.046)	(0.046)
DA × Scale40	0.149***	0.149***	-0.111***	-0.113***
	(0.042)	(0.042)	(0.048)	(0.048)
Period		0.004***		-0.004**
		(0.001)		(0.002)
LotteryPosition		-0.114***		0.121***
		(0.003)		(0.005)
LotteryPosition × DA		0.013**		-0.024***
		(0.006)		(0.008)
LotteryPosition × Scale40		0.021		-0.019
		(0.011)		(0.012)
No. of observations	8320	8320	8320	8320
Human-vs-empirical-robots	(1)	(2)	(3)	(4)
Scale4K	-0.019	-0.018	0.014	0.013
	(0.070)	(0.070)	(0.074)	(0.074)
DA	0.164**	0.164**	-0.242***	-0.243***
	(0.067)	(0.067)	(0.066)	(0.066)
DA × Scale4K	0.062	0.062	-0.066	-0.066
	(0.091)	(0.091)	(0.090)	(0.090)
Period		0.003*		-0.003*
		(0.002)		(0.002)
LotteryPosition		-0.105***		0.105***
		(0.013)		(0.012)
LotteryPosition × DA		0.049***		-0.056***
		(0.018)		(0.018)
LotteryPosition × Scale4K		-0.013		0.006
		(0.018)		(0.018)
No. of human observations	3200	3200	3200	3200

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

By Result 1, we reject the null in favor of Hypothesis 1 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 2 (Truth-telling: scale effect $4 \rightarrow 40$). *The proportion of truth-telling under the Boston (DA) mechanism significantly decreases (weakly increases) when the number of students per match increases from 4 to 40.*

Support. In specification (1) of the upper panel of Table 6, the coefficient of Scale40 is negative and significant (-0.084, $p < 0.01$), indicating that truth-telling in the 40-student sessions under Boston decreases by 8-percentage points compared to that in the 4-student sessions. In comparison, the coefficient of Scale40 + DA × Scale40 is positive and marginally significant (0.065, $p < 0.10$), indicating that truth-telling in the 40-student matches under DA weakly increases by 7-percentage points compared to that in the 4-student matches.

By Result 2, we reject Hypothesis 2 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 3 (District school bias: mechanism and scale effects). *The proportion of DSB under the Boston mechanism is 22-percentage (33-percentage) points 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-percentage (12-percentage) points when the size of the match increases from 4 to 40.*

Support. In specification (3) in the upper panel of Table 6, the coefficient of DA is -0.217 ($p < 0.01$), that of DA × Scale40 is -0.111 ($p < 0.01$), that of Scale40 is 0.122 ($p < 0.01$), that of DA + DA × Scale40 is -0.328 ($p < 0.01$), and that of Scale40 + DA × Scale40 is 0.011 ($p > 0.10$).

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-percentage points per period ($p < 0.01$, specifications (2) and (4)). We also find that, under the Boston mechanism, a one-position increase in the lottery position decreases the likelihood of truth-telling by 11.4-percentage points ($p < 0.01$, specification (2)), but increases the likelihood of DSB by 12.1-percentage points ($p < 0.01$, 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 can be largely explained by individual best response under different market size, which is analyzed in Sections 5.2 and 5.3.

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 17 in Appendix B.2 presents p-values computed from Fisher's exact 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 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 6 presents the results of four probit regressions investigating the scale and mechanism effects in the human-vs-empirical-robot treatments. The dependent variable is again Truth-telling in (1) and (2), and District School Bias in (3) and (4). The independent variables are similar to those in the upper panel, except that the scale variable is 4K, with the omitted category being $m = 40$. In each specification, standard errors are clustered at the individual subject level, as each subject is an independent observation. We summarize the results below.

Result 4 (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' lot-*
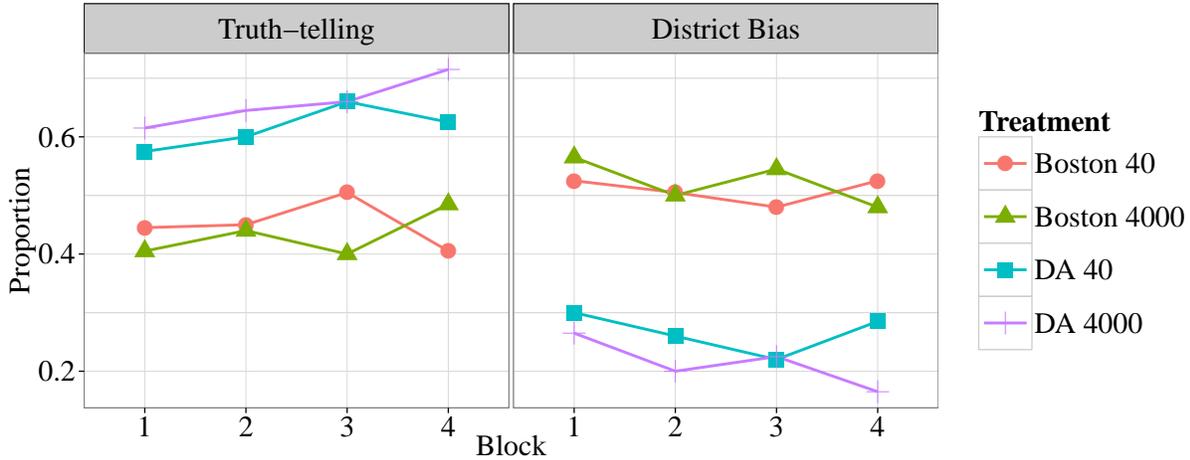
Figure 2: Truth-telling and district school bias in human-vs-empirical-robot treatments

tery positions, i.e., a one-position increase in lottery position decreases the likelihood of truth-telling by 10-percentage (6-percentage) points under the Boston (DA) mechanisms.

Support. In specification (1) of Table 6 (lower panel), the coefficients of DA (0.164, $p < 0.05$) and DA + DA × 4K (0.226, $p < 0.01$) are both positive and significant. The effect of the lottery position under the Boston mechanism, indicated by the coefficient of LotteryPosition, is -0.105 ($p < 0.01$), while the effect under the DA mechanism is indicated by LotteryPosition + LotteryPosition × DA (0.049, $p < 0.01$) in specification (2).

By Result 4, we again reject the null in favor of Hypothesis 1 in our human-vs-empirical-robots treatments. However, we note that the coefficient of Scale4K (Scale4K + DA × Scale4K) is negative (positive) but insignificant, indicating a lack of scale effect. This leads to our next result.

Result 5 (Truth-telling: scale effect $40 \rightarrow 4K$). *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 6 (lower panel), the coefficients of Scale4K and Scale4K + DA × Scale4K are both insignificant ($p > 0.10$), 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 5, we fail to reject Hypothesis 2 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 6, we see that the proportion of district school bias under the DA mechanism is 23-percentage points lower 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 3. To understand the mechanism and scale effects on truthtelling, we analyze human subjects' best response behavior in each treatment in the next two subsections.

5.2 Best response: Human-vs-truthful-robots treatments

In this section, we investigate mechanism and scale effects on best response behavior in the human-vs-truthful-robots treatments, where subjects are told that the robots report their preferences truthfully. Therefore, this environment presents no strategic uncertainty regarding others' behavior.

Under the DA mechanism, when human players are matched with truthful robots, all strategies yield the same outcome. Therefore, there is no single best response. However, significantly more first choices are truthful (48.94%) compared to the random choice benchmark of 25% (χ^2 goodness-of-fit test against multinomial distribution, $p < 0.01$).

In comparison, under the Boston mechanism, best responses vary with student type and lottery positions. Table 7 presents the best response strategies for each type in each lottery block under the Boston mechanism.

Table 7: Best Responses in the Human-vs-Truthful-Robots Treatment under the Boston Mechanism

Block (lottery position)	Type 1	Type 2	Type 3	Type 4
Block 1 (1-2-3-4)	T1	DSB	T1	DSB
Block 2 (2-3-4-1)	T12 or DSB	T1	T1	DSB
Block 3 (3-4-1-2)	T1	DSB	T1	DSB
Block 4 (4-1-2-3)	T1	DSB	T12 or DSB	T1

Note: T1 = Truthful first choice; T12 = Truthful first two choices; DSB = District school bias

Based on our best response analysis in Table 7, we define four types of strategies, which collectively characterize 98% of observed strategies, with the remaining two-percent strictly dominated by one of the four strategies. We first decompose truth-telling into *best-response* and *naïve truth-telling* (Ergin and Sönmez, 2006). The former reveals one's true preference (T1 or T12) when it is a best response to do so, whereas the latter reveals one's true preference when district school bias is the best response. Likewise, we categorize district school bias into *justified* and *unjustified DSB*. The former refers to district school bias when it is a best response to do so, whereas the latter refers to district school bias when truth telling is a better response. Table 8 summarizes the proportion of best responses as well as the proportion of strategies that could be classified as each of the four

types listed above. The left panel of Figure 3 presents the same information by period.

Table 8: Breakdown of Strategies in the Boston Truthful Robot Treatments.

	Medium Scale	Large Scale
Best-response truth-telling	0.465	0.505
Justified DSB	0.419	0.416
Total best response	0.884	0.921
Naïve truth-telling	0.023	0.031
Unjustified DSB	0.079	0.031
Total non-best response	0.102	0.063

While the proportion of best responses is high, it might not be an accurate representation of player types. For example, a subject using justified DSB in one period should not be viewed as a best-response type if she continues to exhibit DSB in a later block when it is no longer a best response, such as a Type-4 student moving from blocks 3 to 4. Consequently, we define a strategy in a given period as a *consistent best response* if a subject always best responds in all subsequent periods. Although this definition does not allow for any mistake, it enables us to identify subjects who best respond in a consistent manner across different environments.
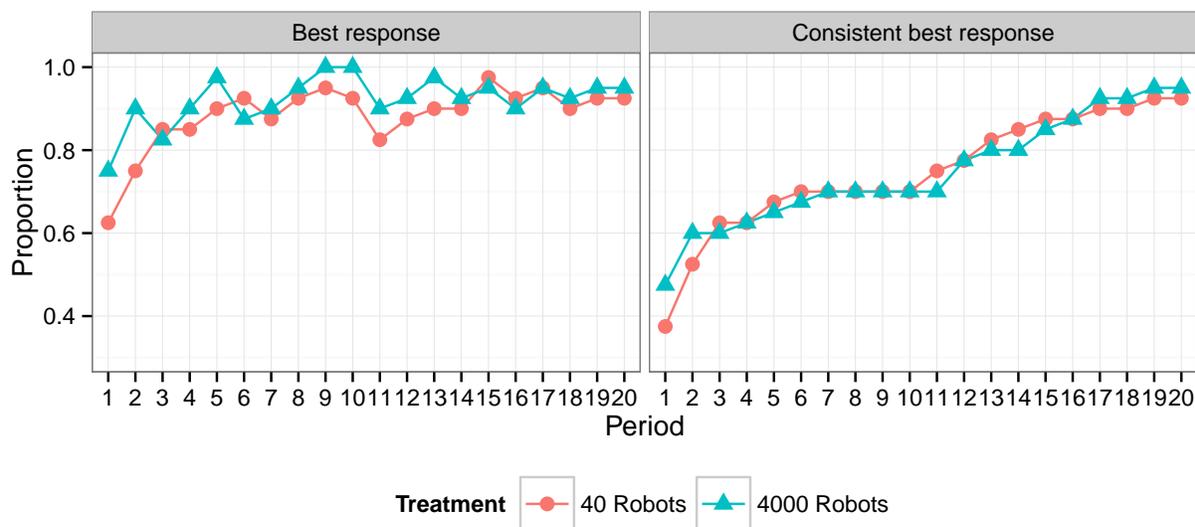

Figure 3: Proportion of best response (left panel) and consistent best response (right panel) over time under the Boston mechanism

The right panel in Figure 3 presents the proportion of consistent best responses over time in the two human-vs-truthful-robots treatments under the Boston mechanism. While the proportion

of consistent best responses is non-decreasing by definition, we observe a sharp increase in the first three periods and continued growth until the last period, indicating that subjects learn to best respond consistently over time in an environment with no strategic uncertainty.

Table 9: Best responses in human-vs-truthful-robots treatments: Probit regressions

Dependent Variable	Best response		Consistent best response	
	(1)	(2)	(3)	(4)
Scale4K	0.038	0.037	0.003	0.003
	(0.035)	(0.036)	(0.076)	(0.076)
Period		0.006***		0.022***
		(0.002)		(0.003)
No. of observations	1600	1600	1600	1600

Notes: Standard errors in parentheses are clustered at the individual levels; coefficients are average marginal effects. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Result 6 (Best response without strategic uncertainty). *Under the Boston mechanism, when humans play with truthful robots, they learn to best respond over time. Furthermore, we do not observe a scale effect on best response from $m = 40$ to $m = 4,000$.*

Support. Table 9 reports four probit specifications, with best response (specifications 1 and 2) and consistent best response (3 and 4) as the dependent variable. Independent variables include Scale4K and period. In none of the specifications is Scale4K significant, indicating a lack of scale effect between the two treatments. In comparison, the period variable is positive and significant (0.006 in (2) and 0.022 in (4), $p < 0.01$).

Result 6 indicates that when there is no strategic uncertainty in the environment, player beliefs are accurate. They learn to best respond over time, and there is no scale effect.

5.3 Best response: All-human and human-vs-empirical-robots experiment

In this section, we apply the same best response analysis to the all-human and human-vs-empirical-robots treatments. This time, however, players do not know others' strategies when they decide theirs, thus these treatments entail more strategic uncertainty compared to the truthful robot treatments. To identify whether a human player best responds, we implement counterfactual simulations using a strategy of either truth-telling or district school bias. For the all-human treatments, for each observation, we control the strategies of all other players in the group, calculate her payoffs under truth-telling and district bias respectively, and compare them with the actual earning: if

the actual earning is greater than or equal to the higher payoff between truth-telling and DSB, we categorize the human strategy as a best response. For the human-vs-empirical-robot treatments, we again recreate 2000 groups for each human observation, calculate the expected payoff under truth-telling and DSB respectively, and compare them with the player's actual earning: if the actual earning is greater than or equal to the highest average simulated earnings between truth-telling and DSB, we again categorize the human strategy as a best response.
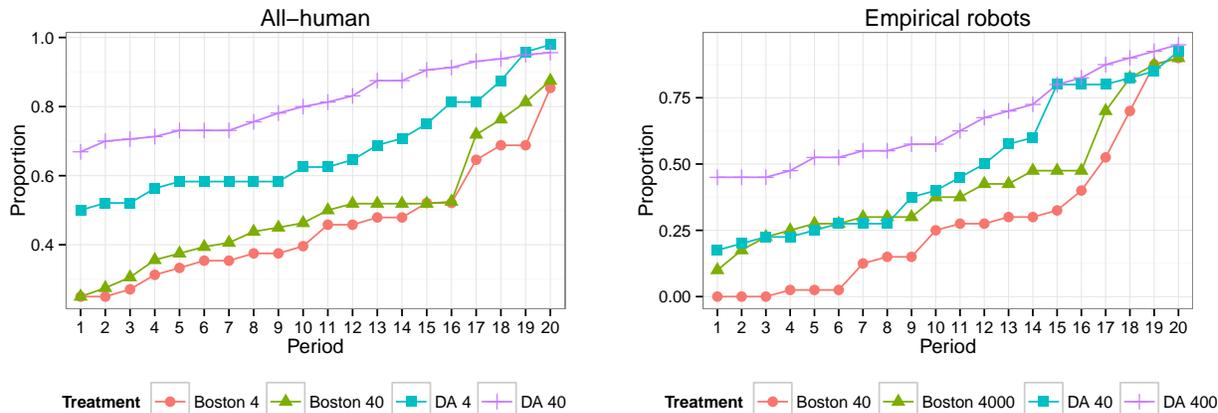
Figure 4: Proportion of consistent best responses in all-human (left) and human-vs-empirical-robots (right) treatments

Figure 4 presents the proportion of consistent best responses in the all-human (left panel) and human-empirical-robots (right panel) treatments. Similar to Result 6, we see that human subjects learn to best respond over time. Further, we see a separation between the two mechanisms and among the different scales. We next investigate these effects using regression analysis.

Table 10 presents six probit specifications for the all-human (1-3) and human-vs-empirical-robots (4-6) treatments. The dependent variables are best response in specifications (1) and (4), naï¿œve truth-telling in (2) and (5), and unjustified DSB in (3) and (6). The independent variables (omitted) include: Scale, DA (Boston), DA × Scale, and Period. Note that the variable Scale (omitted) refers to Scale40 (Scale4) in (1) - (3), and Scale4K (Scale40) in (4) - (6). Likewise, Table 11 presents four probit specifications investigating the mechanism and scale effects on the likelihood of adopting consistent best response. We summarize the results below.

Result 7 (Best response with strategic uncertainty: mechanism effect). *In all-human and human-vs-empirical robot treatments, significantly more subjects best respond and consistently best respond under the DA than under the Boston mechanism.*

Support. In specifications (1) and (4) of Table 10, the coefficient of DA is 0.137 ($p < 0.01$)

Table 10: Strategies in all-human and human-vs-empirical-robots treatments: Probit regressions

Treatment	All-human treatments			Human-vs-empirical-robots		
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variables	Best Response	Naï¿œve Truth-telling	Unjustified DSB	Best Response	Naï¿œve Truth-telling	Unjustified DSB
Scale	0.055**	-0.029***	-0.010	0.102***	-0.075***	-0.024
	(0.028)	(0.008)	(0.023)	(0.025)	(0.022)	(0.016)
DA	0.137***		-0.048**	0.122***		-0.028*
	(0.026)		(0.022)	(0.025)		(0.170)
DA × Scale	-0.036		0.004	-0.046		0.003
	(0.030)		(0.023)	(0.033)		(0.018)
Period	0.002***	-0.001	-0.001*	0.006***	-0.005***	-0.003***
	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	8,320	4160	8320	3,200	1600	8320
Scale + DA × Scale	0.019		-0.007	0.056***		-0.021**
	(0.013)		(0.004)	(0.021)		(0.008)
DA + DA × Scale	0.102***		-0.044***	0.076***		-0.026***
	(0.016)		(0.009)	(0.020)		(0.007)

Notes: Standard errors in parentheses are clustered at the session level for specification (1), (2) and (3), at subject level for specification (4), (5) and (6). *** p < 0.01, ** p < 0.05, * p < 0.1. Scale refers to Scale40 in (1) - (3), and Scale4K in (4) - (6).

Table 11: Probability of consistent best response: Probit regressions

Treatment	All-human treatments		Human-vs-empirical-robots	
	(1)	(2)	(3)	(4)
Scale	0.046*	0.046*	0.145**	0.153**
	(0.025)	(0.025)	(0.067)	(0.067)
DA	0.222***	0.222***	0.209***	0.215***
	(0.072)	(0.073)	(0.064)	(0.065)
DA × Scale	0.094	0.093	0.021	0.010
	(0.088)	(0.088)	(0.103)	(0.102)
Period		0.022***		0.035***
		(0.001)		(0.002)
Observations	8320	8320	3200	3200
Scale + DA × Scale	0.140*	0.139	0.166**	0.163**
	(0.085)	(0.085)	(0.078)	(0.076)
DA + DA × Scale	0.316***	0.315***	0.23***	0.225***
	(0.051)	(0.050)	(0.080)	(0.078)

Notes: Standard errors in parentheses are clustered at the session level for specification (1) and (2), at subject level for specification (3) and (4). *** p < 0.01, ** p < 0.05, * p < 0.1. Scale refers to Scale40 in (1) and (2), and Scale4K in (3) and (4).

and 0.122 ($p < 0.01$), indicating sizeable mechanism effects under the small and medium scale, respectively. Likewise, the coefficient of DA + DA × Scale is 0.102 ($p < 0.01$) and 0.076 ($p < 0.01$), indicating sizeable mechanism effects under the medium and large scale, respectively. Table 11 presents larger mechanism effects on consistent best response across small (0.222, $p < 0.01$), medium (0.315 for all-human, 0.215 for human-vs-empirical-robots, $p < 0.01$) and large scales (0.225, $p < 0.01$).

Result 7 provides further support that strategy-proof mechanism is "easier" to understand in practice. Since the DA mechanism is strategy-proof whereas the Boston mechanism is not, best response is easier to use under the DA mechanism as naïœve truth-telling is considered a best response. In contrast, under the Boston mechanism, students have to learn to abandon naïœve truth-telling.

Result 8 (Best response with strategic uncertainty: scale effect)**.** *In all-human and human-vs-empirical robot treatments, the likelihood of best response increases with scale under both the Boston and DA mechanisms.*

Support. In specifications (1) and (4) of Table 10, the coefficient of Scale is 0.055 ($p < 0.05$) and 0.102 ($p < 0.01$), indicating an increase of best response under the Boston mechanism from small to medium, and from medium to large scale, respectively. Similarly, the coefficient of Scale + DA × Scale is 0.056 ($p < 0.01$) in (4), indicating an increase of best response under the DA mechanism from medium to large scale. Table 11 presents similar scale effects on consistent best response from medium to large scale (columns 3 and 4) under the Boston (0.153, $p < 0.05$) and the DA (0.163, $p < 0.05$) mechanisms, respectively.

When looking at non-best response strategies (Table 9, we find that an increase in scale significantly reduces naïœve truth-telling under the Boston mechanism (-0.029, $p < 0.001$ in (2); -0.075, $p < 0.001$ in (5)), as well as unjustified district school bias under the DA mechanism (-0.021, $p < 0.05$ in (6)).

In sum, our best response analysis provides deeper insights into the scale effect on individual behavior. Absent of strategic uncertainty, we find that scale has no effect on best response behavior under the Boston mechanism, as predicted by theory. With strategic uncertainty, however, scale increases best responses under both mechanisms, which is likely due to the fact that the precision of subjects' beliefs about others' strategies grows with market size.

5.4 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, indi-

vidual 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 ($m = 40, 4000$).

5.4.1 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 3 (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 7 and 8 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 12 reports the results from four probit specifications for students' true (specifications 1 and 2) and reported first choices (specifications 3 and 4) for both the all-human (upper panel) and the human-vs-empirical-robots treatments (lower panel). The independent variables for the upper panel (omitted) include: Scale40 (Scale4), DA (Boston), Period, LotteryPosition, and interaction terms. 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 9 (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 under the DA mechanism.*

Support. In the upper panel of Table 12, the coefficients of DA are negative and significant ($p < 0.01$ in each of the four specifications), indicating a 8.9-percentage (29.4-percentage) points 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 + DA $\times$ Scale40 are also negative and significant in all four specifications ($p < 0.01$).

26

Table 12: First choice accommodation: Probit regressions

	True First Choice		Reported First Choice	
All-human treatments	(1)	(2)	(3)	(4)
Scale40	-0.069***	-0.068***	0.057*	0.056*
	(0.018)	(0.017)	(0.029)	(0.029)
DA	-0.089***	-0.088***	-0.294***	-0.293***
	(0.023)	(0.023)	(0.036)	(0.036)
DA × Scale40	0.010	0.009	-0.106***	-0.107***
	(0.023)	(0.023)	(0.040)	(0.040)
Period		2.950		-0.005***
		(0.001)		(0.001)
LotteryPosition		-0.053***		0.075***
		(0.003)		(0.006)
LotteryPosition × DA		0.037***		0.004
		(0.006)		(0.010)
LotteryPosition × Scale40		0.020***		0.005
		(0.006)		(0.015)
No. of observations	8320	8320	8320	8320
Human-vs-e-robots treatments	(1)	(2)	(3)	(4)
Scale4K	0.011	0.011	0.026	0.026
	(0.029)	(0.029)	(0.054)	(0.054)
DA	-0.083***	-0.083***	-0.302***	-0.302***
	(0.020)	(0.020)	(0.055)	(0.055)
DA × Scale4K	-0.019	-0.017	-0.087	-0.088
	(0.031)	(0.030)	(0.073)	(0.073)
Period		-0.001		-0.005***
		(0.001)		(0.002)
LotteryPosition		-0.050***		0.060***
		(0.010)		(0.014)
LotteryPosition × DA		0.030***		-0.029
		(0.012)		(0.020)
LotteryPosition × Scale4K		-0.001		0.004
		(0.012)		(0.020)
No. of human observations	3200	3200	3200	3200

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

In comparison, from the lower panel of Table 12, we see that the coefficients of DA are also negative and significant ($p < 0.01$ in each of the four specifications), indicating a 8.3-percentage (30.2-percentage) points reduction in receiving one's true (reported) first choice under the DA in the $m = 40$ treatments in specification 1 (3). The coefficients of DA + DA $\times$ Scale4K are also negative and significant in all four specifications ($p < 0.01$).

By Result 9, we reject the null in favor of Hypothesis 3 (b), but fail to reject the null in favor of Hypothesis 3 (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 10 (First-choice accommodation: scale effect). *The increase in scale from $4 \rightarrow 40$ significantly decreases (increases) the true (reported) first-choice accommodation rate by 7-percentage (6-percentage) points under the Boston mechanism, but significantly decreases the rate by 6-percentage (5-percentage) points 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 12, the coefficients of Scale40 are -0.069 ($p < 0.01$) in specification (1), and 0.057 ($p < 0.10$) in specification (3). Furthermore, the coefficients of Scale40 + DA $\times$ Scale40 are -0.059 ($p < 0.01$) in specifications (1) and (2), and -0.049 ($p < 0.10$) in specifications (3) and (4). In comparison, in the lower panel of Table 12, none of the coefficients of Scale4K or Scale4K + DA $\times$ Scale4K is significant ($p > 0.10$).

Result 10 is consistent with the scale effect on truth-telling (Results 2 and 5). 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 4 (Individual rationality). More students will receive individually rational allocations under the DA than under the Boston mechanism regardless of scale.

Table 13 presents four probit specifications for the all-human (1-2) and human-vs-empirical-robots treatments (3-4), whereas Figure 9 in Appendix B presents the proportion of individual rational allocations in each treatment.

Table 13: 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	0.042***	0.042***	-0.019	-0.019
	(0.015)	(0.015)	(0.023)	(0.023)
DA	0.087***	0.088***	0.044**	0.044**
	(0.018)	(0.018)	(0.017)	(0.017)
DA $\times$ Scale	-0.030	-0.030	0.011	0.012
	(0.020)	(0.020)	(0.027)	(0.027)
Period		0.001**		0.003***
		(0.000)		(0.001)
LotteryPosition		0.032***		0.030***
		(0.006)		(0.008)
LotteryPosition $\times$ DA		-0.028***		-0.027***
		(0.007)		(0.009)
LotteryPosition $\times$ Scale		-0.016***		-0.001
		(0.006)		(0.009)
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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Result 11 (Individual rationality: mechanism and scale effects)**.** *Regardless of scale, the proportion of individually rational allocations is significantly higher under the DA than the Boston mechanism. The increase in scale from* $4 \rightarrow 40$ *significantly increases the proportion of individually rational allocations under both mechanisms.*

Support. In Table 13, the coefficients of DA are 0.087 and 0.044 for all-human and human-vs-empirical robots treatments respectively ($p < 0.01$ in specification (1), and $p < 0.05$ in specification (3)). 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 ($p < 0.01$).

5.4.2   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.

Table 14: 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Tables 14 and 15 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 14 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 15 indicate that further increasing the scale from 40 to 4000 has no effect on the efficiency of either mechanism.

Result 12 (Efficiency: scale effect). *From* $4 \rightarrow 40$, *payoff-based efficiency under the Boston (DA) mechanism decreases by 1.4-percentage (3.3-percentage) points. Further increasing the scale to*

4000 has no additional effect on the efficiency of either mechanism.

Support. In specifications (1) and (2) of Table 14, the coefficients for both Scale40 (-0.014, $p < 0.01$) and Scale40 + DA × Scale40 (-0.033, $p < 0.05$) are negative and significant. In comparison, none of the coefficients for the scale variables is significant in specifications (1) and (2) in Table 15.

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.

Table 15: 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 × Scale4K		-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

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.

Hypothesis 5 (Stability: mechanism effect)**.** The proportion of students exhibiting justified envy is lower under the DA than under the Boston mechanism at any given scale.

Figure 5 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.
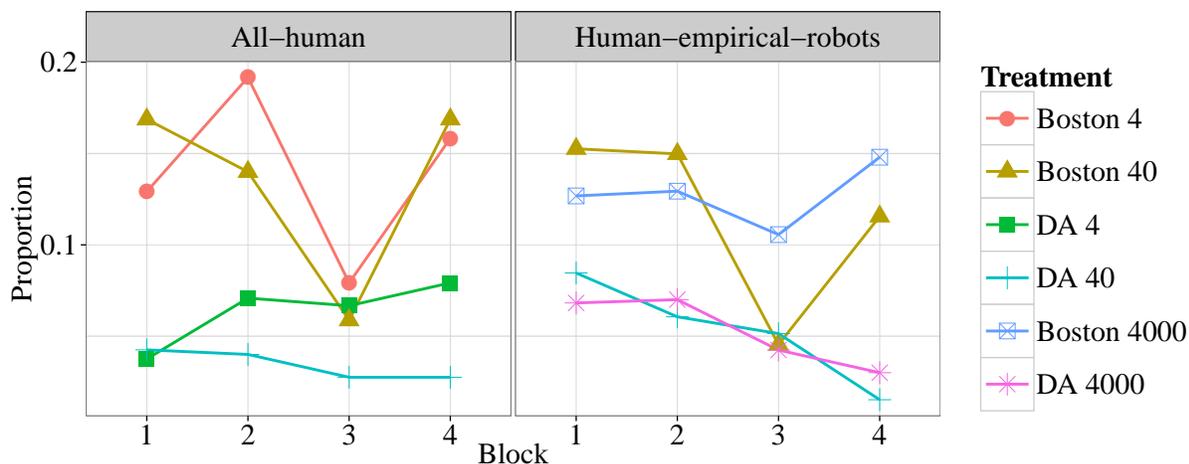
Figure 5: Justified envy in all-human (left panel) and human-vs-empirical-robots (right panel) treatments

Result 13 (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 14, the coefficients for DA and DA + DA × Scale40 are negative and significant ($p < 0.01$). Likewise, the coefficients for DA and DA + DA × Scale4K are negative and significant (-0.063 and -0.135, $p < 0.01$) in specifications (3) and (4) in Table 15.

By Result 13, we reject the null in favor of Hypothesis 5. 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. We also observe that the none of the scale variables in Tables 14 and 15 is significant, indicating the absence of any scale effect on stability, which is consistent with the theoretical predictions.

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. We then further increase the market size to 4,000 by introducing robots. When humans play truthful robots (without strategic uncertainty), we find that scale has no effect on best response behavior. However, when humans play empirical robots (with strategic uncertainty), scale increases best responses under both mechanisms, which is likely due to the increase in the precision of subjects' beliefs about others' strategies.

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.

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. Previous experiments with robots endow the robots with either a dominant or random choice strategy. In comparison, we endow them with either truthful or 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 point to a behavioral factor ignored by the theoretical literature, i.e., scale might increase the precision of participant beliefs about others' strategies when there is strategic uncertainty, and thus increasing the likelihood of best responses.

Subsequent appendices are for online publication only.

A Instructions (Translated from Chinese)

The complete instructions for the Boston Empirical 4000 treatment (i.e., the treatment for the Boston mechanism where each human subject plays with empirical robots to form a group of 4000) are shown here. The instruction for the Boston Empirical 40 treatment are identical except for the number of players, and instructions for the Boston Truthful 4000 treatment are identical except for the subsection "Computer Programs' Strategy ..."; the instructions for the Boston Truthful 40 treatment are identical except for the number of players and the subsection "Computer Programs' Strategy..."; the instructions for the Boston All-human treatment are identical except that the number of players per group is 40 and all the players are human subjects (there is no computer program). Thus, only the different parts of the instructions for the other Boston mechanism treatments are presented here.

The instructions for the DA treatments are identical to their Boston counterparts except for the subsection "The process school admission..." and the worksheet for Review Questions # 1 and #11. Hence, only these two parts for DA Empirical 4000 instruction are provided here; instructions for the other DA mechanism treatments are omitted, but are available from the authors upon request.

A.1 Instructions for the Boston Empirical 4000 Treatment

Instruction - BE4000

This is an experiment in the economics of decision making. In this experiment, we simulate a procedure to allocate students to schools. The procedure, payment rules, and student allocation method are described below. The amount of money you earn will depend upon the decisions you make and on the decisions other participants make. Do not communicate with each other during the experiment. If you have questions at any point during the experiment, raise your hand and the experimenter will help you. At the end of the instructions, you will be asked to provide answers to a series of review questions. Once everyone has finished the review questions, we will go through the answers together.

Procedure

- In this experiment, you are matched with 3999 automated computer programs as a group of 4000, which will make decisions under the same situation as yours. We will describe their strategies later.

- Other than you, there is no other human participant in your group. 4000 participants (including 3999 computer programs and yourself) are classified into 4 different types, depending on

their preferences over schools. Therefore, each group contains 1000 of each of the four different types. Your type will be shown on your computer screen when the experiment starts. Your type remains the same throughout the experiment.

- In this experiment, 4 schools are available for each group. Each school has 1000 slots.

- Your payoff amount depends on the school you are assigned to at the end of each period. Payoff amount for 4 types of students are outlined in the following 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

Your district school is underlined. In other words, you reside in the underlined school district. If a school is your district school, you have the higher priority than other non-district-school students at this school.

- In this experiment, participants are defined as belonging to the following school districts:

  The district school for Type 1 students is school A;

  The district school for Type 2 students is school B;

  The district school for Type 3 students is school C;

  The district school for Type 4 students is school D.

- Your assigned school and your payoff have the following relation:

  If you are Type 1 student:

You will be paid	11	points if you hold a slot of School A at the end of a period.
You will be paid	7	points if you hold a slot of School B at the end of a period.
You will be paid	5	points if you hold a slot of School C at the end of a period.
You will be paid	16	points if you hold a slot of School D at the end of a period.

  The same goes for other types.

- Just like you, computer programs will also make the decision based on the above payoffs table and whether the school is district/non-district school.

- The experiment consists of 20 periods. In each period, you are matched with 3999 computer programs to form a group of 4000; each group contains 1000 of each of the four different types. Your earnings for each period depend on your choices as well as the choices of the 3999 other participants (computer programs) you are matched with.

- Every period, each participant (including computer programs) will rank the schools. Note that you need to rank all 4 schools in order to indicate your preferences.

- After all participants have submitted their rankings, the server will start the school placement in each group and inform each person of his/her admitted school and respective payoff. Note that your allocation in each period is independent of your allocations in the previous periods.

- Your total payoff equals the sum of your payoffs in all 20 periods. Your earnings are given in points. At the end of the experiment you will be paid based on the exchange rate:

1 RMB = 5 points.

You will be paid 5 points for answering each Review Question correctly.

In addition, you will be paid 5 RMB for participation. Everyone will be paid in private and you are under no obligation to tell others how much you earn.

Admission Method

- **The priority order for each school depends on two factors:**

  - Whether the student lives within the school district: the school will give higher priority to participants who live within the school district, and give low priority to participants who do not live within the school district.

  - Random lottery number: the server will generate random lottery number at the beginning of every 5 periods. Each student (human or computer program) gets a unique lottery number in the range 1~4000. The smaller the lottery number is, higher the priority a student has among the district students or non-district students.

The lottery numbers for each type in each period are shown below:

	Periods 1-5	Periods 6-10	Periods11-15	Periods 16-20
Type 1	1~1000	3001~4000	2001~3000	1001~2000
Type 2	1001~2000	1~1000	3001~4000	2001~3000
Type 3	2001~3000	1001~2000	1~1000	3001~4000
Type 4	3001~4000	2001~3000	1001~2000	1~1000

For example, if you are student of Type1, in the periods 1~5, you will be given a random number in the range 1~1000 at the beginning of 1st period, and the number is fixed for the rest 4 periods; in the periods 6~10, you will be given a random number in the range 3001~4000 at the beginning of 6th period; the rest can be done in the same manner.

- **The process school admission is described by the following method:**

Step1: a. An application to the first ranked school is sent for each participant.
     b. Each school accepts the student with highest priority in that school. These students and their assignments are removed from the system.The remaining applications for each respective school are rejected.

Step2: a. The rejected applications are sent to his/her second choice.
     b. If a school is still vacant, then it accepts the student with the highest priority and rejects the remaining applications.

Step3: a. The application of each participant who is rejected by his/her top two choices is sent to his/her third choice.
     b. If a school is still vacant, then it accepts the student with the highest priority and rejects the remaining applications.

Step4: Each remaining participant is assigned a slot at his/her last choice.

Computer Programs' Strategy

Computer programs will make their decisions based on the following rule:

Before this experiment, there was another similar experiment. The differences between these two experiments are:

- In this experiment, there are 3999 computer programs in a 4000-students group; in the previous experiment, all 40 participants in the group are human participants like you (not computer programs). There are 4 types of students in the previous experiment, which are the same as this experiment. But there are only 10 students of each type in the previous experiment. There are 4 schools, each with 10 slots in the previous experiment, not 1000 slots as this experiment.

- There are 80 participants in that experiment. Two 40-participants groups are randomly formed at the begining of every round. That is, each round each participant is matched with 39 different others.

Everything else, such as the student allocation method is the same in these two experiments.

In this experiment, each period, a computer program will randomly pick one out of the two participants (not computer program) of the same type and with the corresponding lottery numbers in the previous experiment, and rank schools in the same order as her or him.

The correspondence of lottery numbers between the two experiments: the rule of generating lottery numbers in the previous experiment is similar as this experiment, which is tabulated below.

	Periods 1-5	Periods 6-10	Periods 11-15	Periods 16-20
Type 1	1~10	31~40	21~30	11~20
Type 2	11~20	1~10	31~40	21~30
Type 3	21~30	11~20	1~10	31~40
Type 4	31~40	21~30	11~20	1~10

In each period, students of the same type, with lottery number 1 correspond to the computer programs of the same type with the lottery numbers 1~100; students with lottery number 2 correspond to the computer programs of the same type with the lottery numbers 101~200; and so forth.

For example:

A computer program is a student of Type 1; in the 2nd round, its lottery number is 150, which corresponds to the lottery number 2 in the previous experiment. Meanwhile in the previous experiment, two participants of Type 1 in that same round, with lottery number 2, ranked schools in the order of D, B, C, A and A, B, C, D, respectively. Then the computer program will rank schools in the order of D, B, C, A with probability 50%, and A, B, C, D with probability 50%. Other computer programs of Type 1, with lottery numbers 101~200 will also rank schools based on such rule.

All computer programs in each period will follow this rule.

An example:

We will go through a simple example to illustrate how the admission method works. This example has fewer students than the actual decisions you will make. You will be asked to work out the allocation of this example for Review Question 1. There are 6 blank spaces to fill in on the computer screen. You will be given 5 points for entering each correct answer in each space.

Students and Schools: in this example, there are 6 students, 1-6, and 3 schools, A, B and C.

| Student ID number: 1,2,3,4,5,6 | Schools: A, B, C |

School A is the district school of student 1 and 2; school B is the district school of student 3 and 4; school C is the district school of student 5 and 6.

Ranking of schools based on the lottery: Students' lottery numbers are indicated in the table below.

Student	Lottery number
1	1
2	2
3	3
4	4
5	5
6	6

Submitted School Rankings:

Student	1st Choice	2nd Choice	3rd Choice
1	C	A	B
2	C	A	B
3	A	B	C
4	A	B	C
5	C	B	A
6	A	C	B

The assignment method consists of the following steps: Please use this sheet to work out the allocation and enter it into the computer for Review Question 1.

Step 1: Each student applies to his/her first choice. If a school receives more applications than its capacity, then it accepts the applicant with the highest priority and rejects the remaining students.

Student		School		Accept	Reject
3,4,6	$\longrightarrow$	A	$\longrightarrow$		
		B			
1,2,5	$\longrightarrow$	C	$\longrightarrow$		

Step 2: Each student rejected in Step 1 applies to his/her second choice. If a school still has vacancy, it accepts the applicant with the highest priority and rejects the rest. If a school is already full, it rejects all new applications.

Student		School		Accept	Reject
	$\longrightarrow$	A	$\longrightarrow$		
		B			
	$\longrightarrow$	C	$\longrightarrow$		

Step 3: Each student rejected in Step 2 applies to his/her next choice.

Student		School		Accept	Reject
$\longrightarrow$		A	$\longrightarrow$		
		B			
$\longrightarrow$		C	$\longrightarrow$		

The allocation process terminates at the end of step 3.

- Please enter your answer to the computer for Review Question 1. After all have entered their answers, we will distribute an answer sheet and go through the answer together.

- Afterwards, you will be asked to answer another 10 review questions. When everyone is finished with them, we will again go through the answers together.

- Feel free to refer to the experimental instructions before you answer any question.

Answer to Review Question 1: The allocation method consists of the following steps:

Step 1: Each student applies to his/her first choice: students 3, 4 and 6 apply to A, and student 1, 2 and 5 apply to C.

- School A accepts the applications of students 3 and 4, rejects student 6, because 3 and 4 have smaller lottery numbers than 6, and none of them are school district residents of school A.

- School C accepts the applications of student 1 and 5, rejects student 2, because 5 is a district resident, and therefore has the highest priority; between 1 and 2, 1 has a higher priority based on lottery. There is only one vacancy for school C, so it accepts student 1 and rejects student 2.

Student		School		Accept	Reject
3,4,6	$\longrightarrow$	A	$\longrightarrow$	3,4	6
		B			
1,2,5	$\longrightarrow$	C	$\longrightarrow$	1,5	2

Step 2: Each student rejected in Step 1 applies to his/her next choice. Student 2 applies to school A, and student 6 applies to school C. Although they both live in the school district, since school A and C are already full, they are both rejected.

41

Student		School		Accept	Reject
2	$\longrightarrow$	A	$\longrightarrow$	3,4	2
		B			
6	$\longrightarrow$	C	$\longrightarrow$	1,5	6

Step 3: Each student rejected in Step 2 applies to his/her next choice. Both students 2 and 6 apply to School B.

School B accepts the applications of students 2 and 6.

Student		School		Accept	Reject
		A		3,4	
2,6	$\longrightarrow$	B	$\longrightarrow$	2,6	
		C		1,5	

Answers to Review Questions 2-11:

2. How many participants are there in your group each period?
   Answer: 4000

3. How many human participants are there in your group each period?
   Answer: 1

4. True or false: Participant living in a school district has higher priority than applicants for that school who do not live in the school district.
   Answer: True.

5. True or false: The priority among residents of a school district is determined by their lottery number.
   Answer: True.

6. True or false: The priority among non-residents of a school district is determined by their lottery number.
   Answer: True.

7. True or false: The lottery is fixed for the entire 20 periods.
   Answer: False. It is randomly determined each five periods.

8. True or false: A lottery number of 1 means that I have the highest priority among the other non-resident applicants in a school.
   Answer: True.

9. True or false: If you are accepted by a school of your choice, the schools ranked below are irrelevant.
   Answer: True.

10. True or false: If you are not rejected at a step, then you are accepted into that school.
    Answer: True.

11. True or false: The allocation is final at the end of each step.
    Answer: True.

You will have 5 minutes to go over the instructions at your own pace. Feel free to earn as much as you can.

Are there any questions?

A.2   Instructions for the Boston Empirical 40 Treatment

(......)

Procedure

- In this experiment, you are matched with 39 automated computer programs as a group of 40, which will make decisions under the same situation as yours. We will describe their strategies later.

- Other than you, there is no other human participant in your group. 40 participants (including 39 computer programs and yourself) are classified into 4 different types, depending on their preferences over schools. Therefore, each group contains 10 of each of the four different types. Your type will be shown on your computer screen when the experiment starts. Your type remains the same throughout the experiment.

- In this experiment, 4 schools are available for each group. Each school has 10 slots.

   (......)

- The experiment consists of 20 periods. In each period, you matched with 39 computer programs to form a group of 40; each group contains 10 of each of the four different types.

Your earnings for each period depend on your choices as well as the choices of the 39 other participants (computer programs) you are matched with.

(......)

Admission Method

(......)

- Random lottery number: the server will generate random lottery number at the beginning of every 5 periods. Each student (human or computer program) gets a unique lottery number in the range 1~40. The smaller the lottery number is, higher the priority a student has among the district students or non-district students.

The lottery numbers for each type in each period are shown below:

	Periods 1-5	Periods 6-10	Periods11-15	Periods 16-20
Type 1	1~10	31~40	21~30	11~20
Type 2	11~20	1~10	31~40	21~30
Type 3	21~30	11~20	1~10	31~40
Type 4	31~40	21~30	11~20	1~10

For example, if you are student of Type1, in the periods 1~5, you will be given a random number in the range 1~10 at the beginning of 1st period, and the number is fixed for the rest 4 periods; in the periods 6~10, you will be given a random number in the range 31~40 at the beginning of 6th period; the rest can be done in the same manner.

(......)

Computer Programs' Strategy

Computer programs will make their decisions based on the following rule:

Before this experiment, there was another similar experiment. The differences between these two experiments are:

- In this experiment, there are 39 computer programs in a 40-students group; in the previous experiment, all 40 participants in the group are human participants like you (not computer programs).

- There are 80 participants in that experiment. Two 40-participants groups are randomly formed at the begining of every round. That is, each round each participant is matched with 39 different others.

Everything else, such as the types of students, the payoff for being admitted to each school, the student allocation method, and the arrangement of lottery numbers are the same in these two experiments.

In this experiment, each period, a computer program will randomly pick one out of the two participants (not computer program) of the same type and with the same lottery number in the previous experiment, and rank schools in the same order as her or him.

For example:

A computer program is a student of Type 1: in the 2nd round, its lottery number is 2. Meanwhile in the previous experiment, two participants of Type 1 in that same round, with lottery number 2, ranked schools in the order of D, B, C, A and A, B, C, D, respectively. Then the computer program will rank schools in the order of D, B, C, A with probability 50%, and A, B, C, D with probability 50%.

All computer programs in each period will follow this rule.

(......)

Answers to Review Questions 2-11:

2. How many participants are there in your group each period?
   Answer: 40

(......)

A.3   Instructions for the Boston Truthful 4000 Treatment

(......)

Computer Programs' Strategy

Computer programs will make their decisions based on the following rule:

Computer programs will always rank schools based on their true preference orders.

For example, a computer program is a student of Type 2, therefore its preference order of schools is (D, B, C, A). It then will also rank school (D, B, C, A) on his list.

(......)

A.4   Instructions for the Boston Truthful 40 Treatment

(......)

Procedure

- In this experiment, you are matched with 39 automated computer programs as a group of 40, which will make decisions under the same situation as yours. We will describe their strategies later.

- Other than you, there is no other human participant in your group. 40 participants (including 39 computer programs and yourself) are classified into 4 different types, depending on their preferences over schools. Therefore, each group contains 10 of each of the four different types. Your type will be shown on your computer screen when the experiment starts. Your type remains the same throughout the experiment.

- In this experiment, 4 schools are available for each group. Each school has 10 slots.

  (......)

- The experiment consists of 20 periods. In each period, you matched with 39 computer programs to form a group of 40; each group contains 10 of each of the four different types. Your earnings for each period depend on your choices as well as the choices of the 39 other participants (computer programs) you are matched with.

  (......)

Admission Method

  (......)

- Random lottery number: the server will generate random lottery number at the beginning of every 5 periods. Each student (human or computer program) gets a unique lottery number in the range 1~40. The smaller the lottery number is, higher the priority a student has among the district students or non-district students.

The lottery numbers for each type in each period are shown below:

	Periods 1-5	Periods 6-10	Periods 11-15	Periods 16-20
Type 1	1~10	31~40	21~30	11~20
Type 2	11~20	1~10	31~40	21~30
Type 3	21~30	11~20	1~10	31~40
Type 4	31~40	21~30	11~20	1~10

For example, if you are student of Type1, in the periods 1~5, you will be given a random number in the range 1~10 at the beginning of 1st period, and the number is fixed for the rest 4 periods; in the periods 6~10, you will be given a random number in the range 31~40 at the beginning of 6th period; the rest can be done in the same manner.

(......)

Computer Programs' Strategy

Computer programs will make their decisions based on the following rule:

Computer programs will always rank schools based on their true preference orders.

For example, a computer program is a student of Type 2, therefore its preference order of schools is (D, B, C, A). It then will also rank school (D, B, C, A) on his list.

(......)

Answers to Review Questions 2-11:

2. How many participants are there in your group each period?
   Answer: 40

(......)

A.5 Instructions for the Boston All-human Treatment

(......)

Procedure

- There are 80 participants of 4 different types in this experiment. Your type will be shown on your computer screen after the experiment starts. Your type remains the same throughout the experiment.

- You will be randomly matched with other 39 participants into groups of 40 at the beginning of each period. 4 schools are available for each group. Each school has 10 slots.

   (......)

- The experiment consists of 20 periods. In each period, you are randomly matched with other 39 participants to form a group of 40; each group contains 10 of each of the four different types. Your earnings for each period depend on your choices as well as the choices of the 39 other participants you are matched with.

- Every period, each participant will rank the schools. Note that you need to rank all 4 schools in order to indicate your preferences.

   (......)

Admission Method

   (......)

   - <u>Random lottery number</u>: the server will generate random lottery number at the beginning of every 5 periods. Each student gets a unique lottery number in the range 1~40. The smaller the lottery number is, higher the priority a student has among the district students or non-district students.

The lottery numbers for each type in each period are shown below:

	Periods 1-5	Periods 6-10	Periods11-15	Periods 16-20
Type 1	1~10	31~40	21~30	11~20
Type 2	11~20	1~10	31~40	21~30
Type 3	21~30	11~20	1~10	31~40
Type 4	31~40	21~30	11~20	1~10

For example, if you are student of Type1, in the periods 1~5, you will be given a random number in the range 1~10 at the beginning of 1st period, and the number is fixed for the rest 4 periods; in the periods 6~10, you will be given a random number in the range 31~40 at the beginning of 6th period; the rest can be done in the same manner.

(......)

[There is no subsection "Compter Programs' Strategy..." in the instructions for the Boston All-human Treatment.]

(......)

Answers to Review Questions 2-11:

2. How many participants are there in your group each period?
   Answer: 40
3. True or false: You will be matched with the same 39 participants each period.
   Answer: False. You are randomly re-matched with 39 other participants each period.

(......)

A.6   Instructions for the Deferred Acceptance Empirical 4000 Treatment

(......)

- **The process school admission is described by the following method:**

Step1: Each student applies to his first choice school. For each school, up to the number of applicants that does not exceed its capacity, who have the highest priority for this school, are tentatively accepted. The remaining applicants are rejected.

Step2: Each student rejected from a school at step 1 applies to his next favorite acceptable school. For each school, it pools together the new applicants and the students that have already been tentatively placed, and tentatively accepts those who have the highest priority within its capacity. The remaining applicants are rejected.

... ...

The process continues until there is no student being rejected. Then all tentative acceptances become permanent.

(......)

An example:

(......)

Submitted School Rankings:

Student	1st Choice	2nd Choice	3rd Choice
1	C	B	A
2	C	A	B
3	A	B	C
4	A	B	C
5	C	B	A
6	A	C	B

The assignment method consists of the following steps: Please use this sheet to work out the allocation and enter it into the computer for Review Question 1.

Step 1: Each student applies to his/her first choice. If a school receives more applications than its capacity, then it temporarily holds the applications with the highest priority within its capacity and rejects the remaining students.

Student		School		Hold	Reject
3,4,6	$\longrightarrow$	A	$\longrightarrow$		
		B			
1,2,5	$\longrightarrow$	C	$\longrightarrow$		

Step 2: Each student rejected in Step 1 applies to his/her next choice. When a school receives new applications, these applications are considered together with the application on hold for that

school. Among the new applications and those on hold, the one with the highest priority is on hold within the school's capacity, while the rest are rejected.

Hold	New Applicants		School		Hold	Reject
		$\longrightarrow$	A	$\longrightarrow$		
		$\longrightarrow$	B	$\longrightarrow$		
		$\longrightarrow$	C	$\longrightarrow$		

Step 3: Each student rejected in Step 2 applies to his/her next choice. No one is rejected at this step. All students on hold are accepted.

Hold	New Applicants		School		Accept	Reject
		$\longrightarrow$	A	$\longrightarrow$		
		$\longrightarrow$	B	$\longrightarrow$		
		$\longrightarrow$	C	$\longrightarrow$		

The allocation process terminates at the end of step 3.

- Please enter your answer to the computer for Review Question 1. After all have entered their answers, we will distribute an answer sheet and go through the answer together.

- Afterwards, you will be asked to answer another 10 review questions. When everyone is finished with them, we will again go through the answers together.

- Feel free to refer to the experimental instructions before you answer any question.

Answer to Review Question 1: The allocation method consists of the following steps:

Step 1: Each student applies to his/her first choice: students 3, 4 and 6 apply to A, and student 1, 2 and 5 apply to C.

- School A accepts the applications of students 3 and 4, rejects student 6, because 3 and 4 have smaller lottery numbers than 6, and none of them are school district residents of school A.

- School C accepts the applications of student 1 and 5, rejects student 2, because 5 is a district resident, and therefore has the highest priority; between 1 and 2, 1 has a higher priority based on lottery. There is only one vacancy for school C, so it accepts student 1 and rejects student 2.

Student		School		Hold	Reject
3,4,6	$\longrightarrow$	A	$\longrightarrow$	3,4	6
		B			
1,2,5	$\longrightarrow$	C	$\longrightarrow$	1,5	2

Step 2: Each student rejected in Step 1 applies to his/her next choice. Student 2 applies to school A, and student 6 applies to school B.

- School A considers the application of student 2 together with the applications of students 3 and 4 which were on hold. Student 2 is a resident in that school district, and student 3 has a smaller lottery number, therefore school A holds students 2 and 3's applications, and rejects student 4.

- Similarly, school C holds students 5 and 6's applications, and rejects student 1.

Hold	New Applicants		School		Hold	Reject
3,4	2	$\longrightarrow$	A	$\longrightarrow$	2,3	4
			B			
1,5	6	$\longrightarrow$	C	$\longrightarrow$	5,6	1

Step 3: Each student rejected in Step 2 applies to his/her next choice. Student 1 and 4 both apply to School B in the next choice.

School B holds the applications of student 1 and 4.

Since no student is rejected after this step, all tentative holdings become permanent acceptance.

Hold	New Applicants		School		Accept	Reject
2,3			A		2,3	
	1,4	$\longrightarrow$	B	$\longrightarrow$	1,4	
5,6			C		5,6	

Answers to Review Questions 2-11:

(......)

11. True or false: The allocation is final at the end of each step.
    Answer: False. The allocation at the end of every step is temporary except the final step.

Appendix B. Additional Tables and Figures

This appendix presents additional tables and figures, organized by section.

B.1. Truth-telling Analysis: Robustness Check

For robustness check, we use a consistent definition of truth-telling across the two mechanisms, which requires the entire reported rank-ordered list to be identical to one's true preference ranking. Figure 6 reports the proportion of truth-telling under the new definition. Again, we see a separation of mechanisms where the proportion of truth-telling under the DA is greater than that under the Boston mechanism in both the all-human (left panel) and the human-vs-empirical-robots (right panel) treatments.
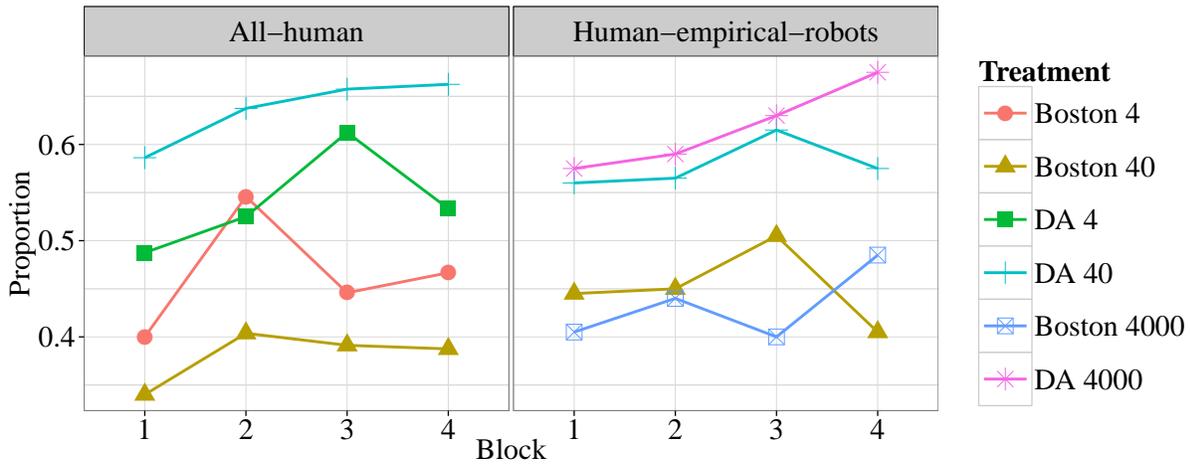
Figure 6: Behavior comparisons based on a consistent definition of truth-telling

Table 16 represents the same probit specifications for truth-telling as those in Table 6, but with the new definition. For the all-human treatments (specifications 1 and 2), we observe significantly more truth-telling under the DA mechanism than under the Boston mechanism, albeit with a smaller difference between the two mechanisms. Likewise, we find a significant scale effect under both mechanisms. For the human-vs-empirical-robots treatments (specifications 3 and 4), we observe a significant mechanism effect and an insignificant scale effect under both mechanisms.

Table 16: Behavior comparisons based on a consistent definition of truth-telling: probit regressions

	All-human treatments		Human-vs-empirical-robots	
	(1)	(2)	(3)	(4)
Scale	-0.084***	-0.084***	-0.019	-0.019
	(0.025)	(0.025)	(0.055)	(0.055)
DA	0.075***	0.075***	0.128***	0.128***
	(0.026)	(0.026)	(0.029)	(0.029)
DA×Scale	0.180***	0.180***	0.058	0.058
	(0.039)	(0.039)	(0.065)	(0.065)
Period		0.003***		0.003
		(0.001)		(0.002)
LotteryPosition		-0.114***		-0.105***
		(0.003)		(0.015)
LotteryPosition×DA		0.015***		0.050***
		(0.005)		(0.017)
LotteryPosition×Scale		0.022**		-0.012
		(0.010)		(0.017)
# 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-robot) treatments; coefficients are marginal effects. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The variable Scale (omitted) refers to Scale40 (Scale4) in specifications (1) and (2), and Scale4K (Scale40) in (3) and (4).

B.2. Comparing all-human and human-vs-39-empirical-robots treatments

Table 17 presents results of Fisher's exact tests of the equality of distributions of strategies between the all-human ($m = 40$, $n = 160$) and the corresponding human-vs-39-empirical-robots ($m = 40$, $n = 40$) treatments under each mechanism, period by period, respectively.[9] As truth-telling and district school bias (DSB) are two main strategies used by the subjects, we also perform proportion tests to compare the proportion of truth-telling as well as district school bias between the two treatments, period by period. We do not pool across all periods because of the interdependency of strategies across periods. With the exception of first-period truth-telling and DSB, we do not find significant difference between the two treatments.

Table 17: P-values for tests comparing all-human 40-student and human-vs-39-empirical-robots sessions

Period	Boston Mechanism			DA Mechanism		
	24 Strategies	Truth	DSB	24 Strategies	Truth	DSB
1	0.307	0.011	0.040	0.021	0.046	0.258
2	0.916	0.106	0.389	0.503	0.166	0.225
3	0.222	0.064	0.104	0.222	0.658	0.247
4	0.754	0.512	0.565	0.564	0.465	0.065
5	0.906	0.770	0.775	0.468	0.600	0.136
6	0.151	0.155	0.056	0.890	0.398	0.414
7	0.136	0.432	0.571	0.262	0.308	0.642
8	0.409	0.565	0.568	0.046	0.029	0.935
9	0.488	0.471	0.431	0.670	0.506	0.935
10	0.497	0.465	0.313	0.341	0.159	0.397
11	0.923	0.103	0.203	0.604	0.940	0.414
12	0.336	0.026	0.038	0.255	0.247	1.000
13	0.227	0.384	0.171	0.543	0.311	0.105
14	0.681	0.279	0.350	0.436	0.549	0.519
15	0.716	0.471	0.350	0.333	0.258	0.746
16	0.274	1.000	0.670	0.531	0.282	0.144
17	0.211	0.667	0.356	0.547	0.440	0.862
18	0.224	0.771	0.475	0.154	0.205	0.052
19	0.142	0.940	0.660	0.109	0.215	0.204
20	0.069	0.706	0.607	0.687	0.706	0.869

Notes: H_a : all-human ($m = 40$) $\neq$ human-vs-39-empirical-robots; Fisher's exact test on strategies distributions; proportion tests on the proportion of truth-telling and district school bias, period by period.

---

[9] We also conduct Kolmogorov-Smirnov tests and χ^2-tests, and find no significant difference between the two treatments.

B.3. Results - Aggregate Performance

Figures 7 and 8 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.
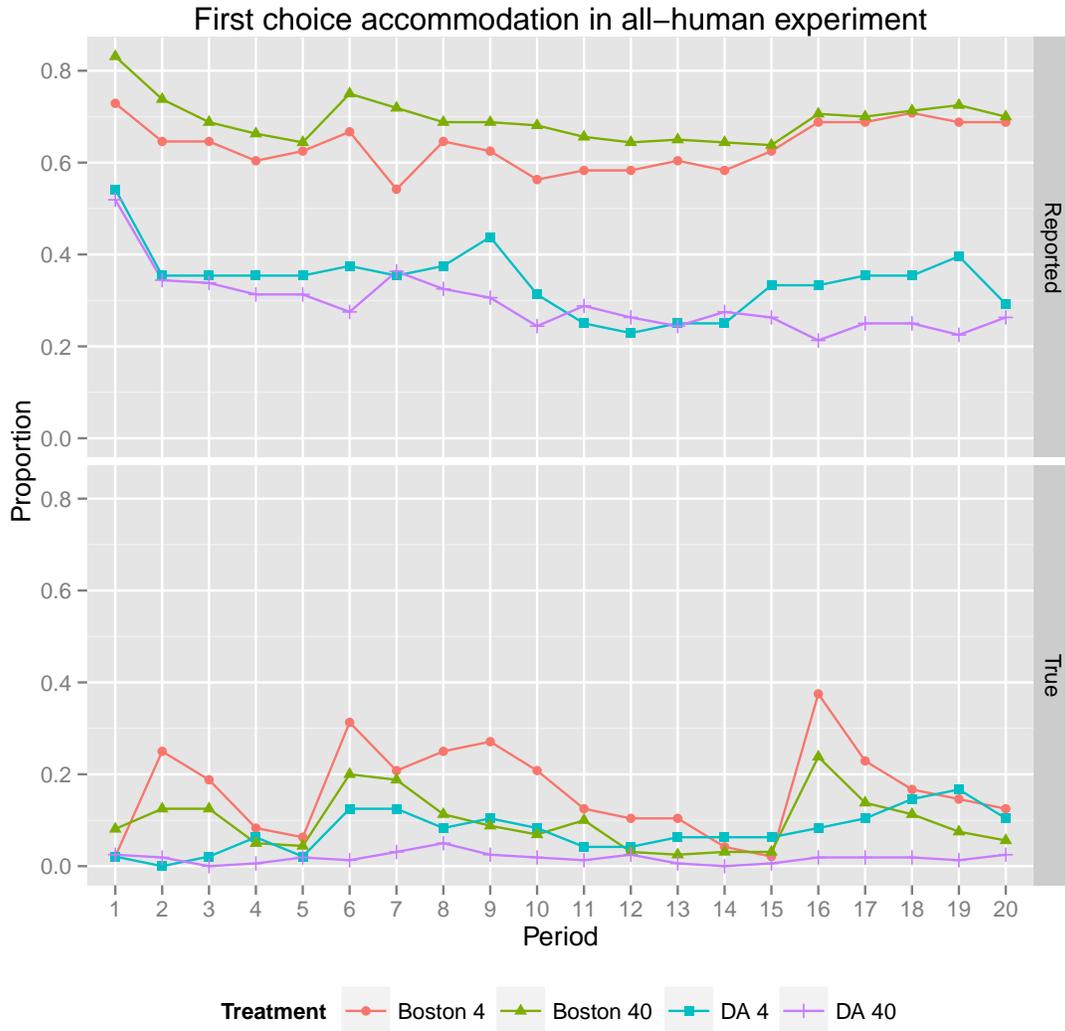

Figure 7: First choice accommodation in all-human experiment

Figure 9 presents the proportion of individually rational allocations in each treatment.
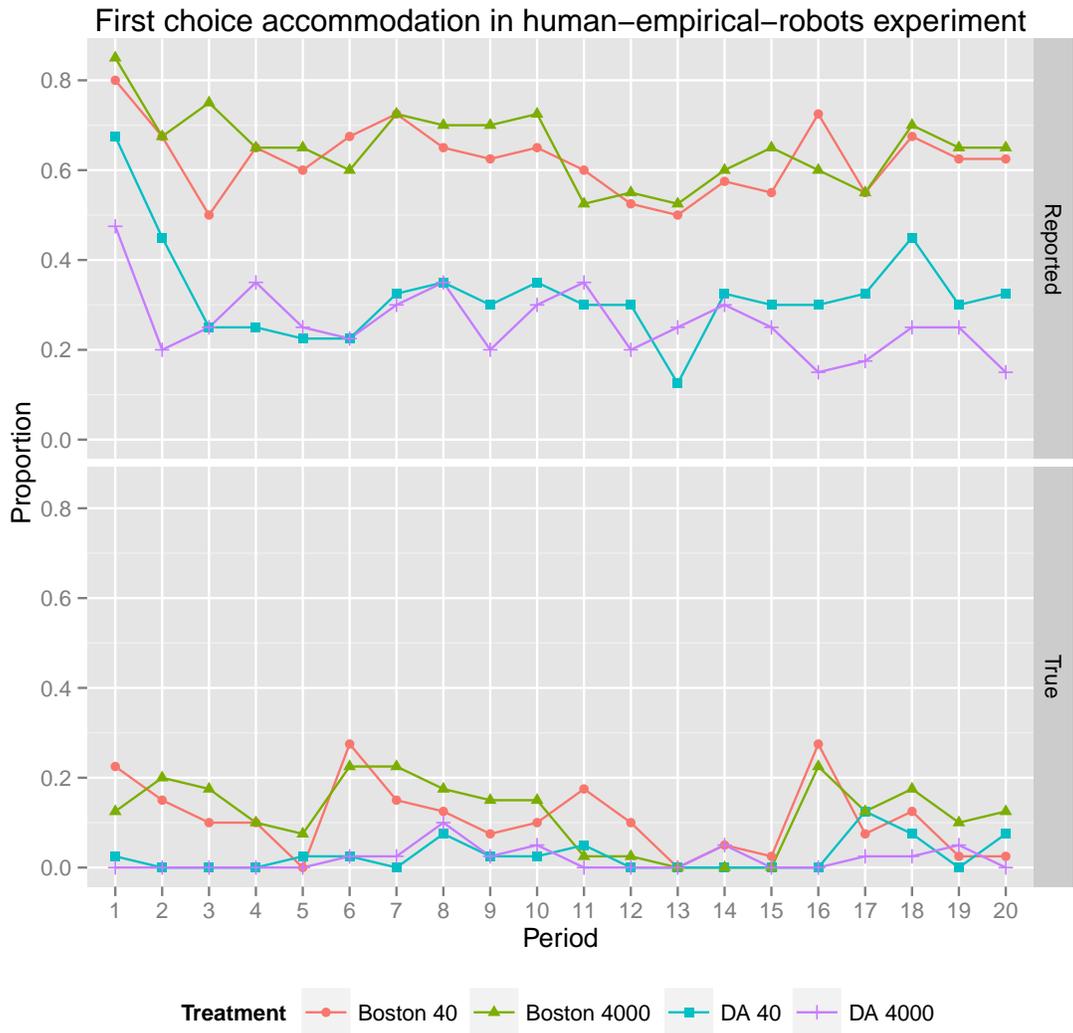
Figure 8: First choice accommodation in human-vs-empirical-robots experiment
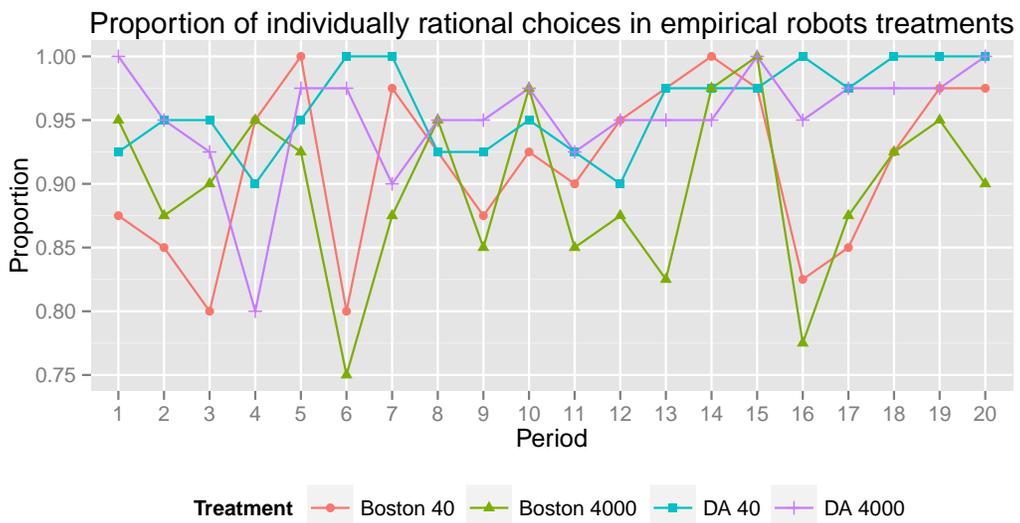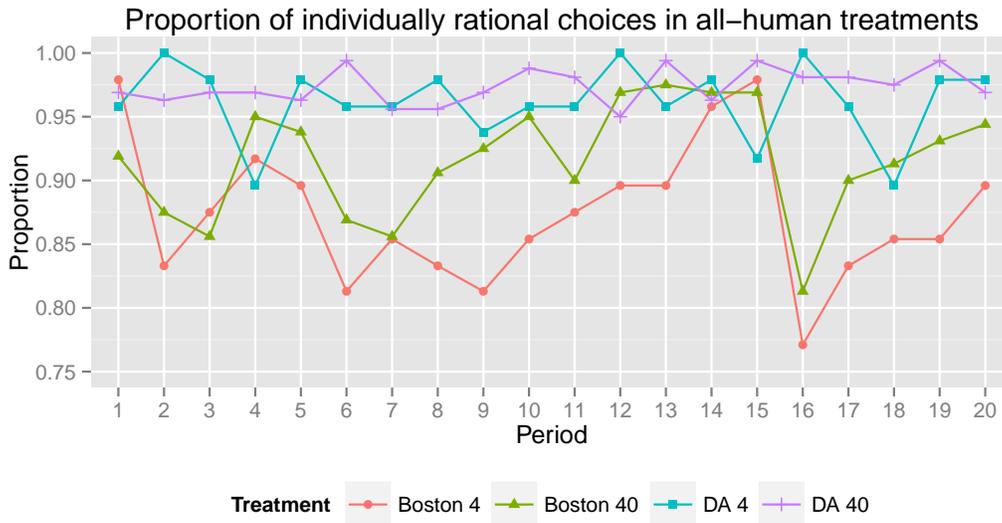
Figure 9: Individual rationality in all-human (upper panel) and human-vs-empirical-robots (lower panel) treatments
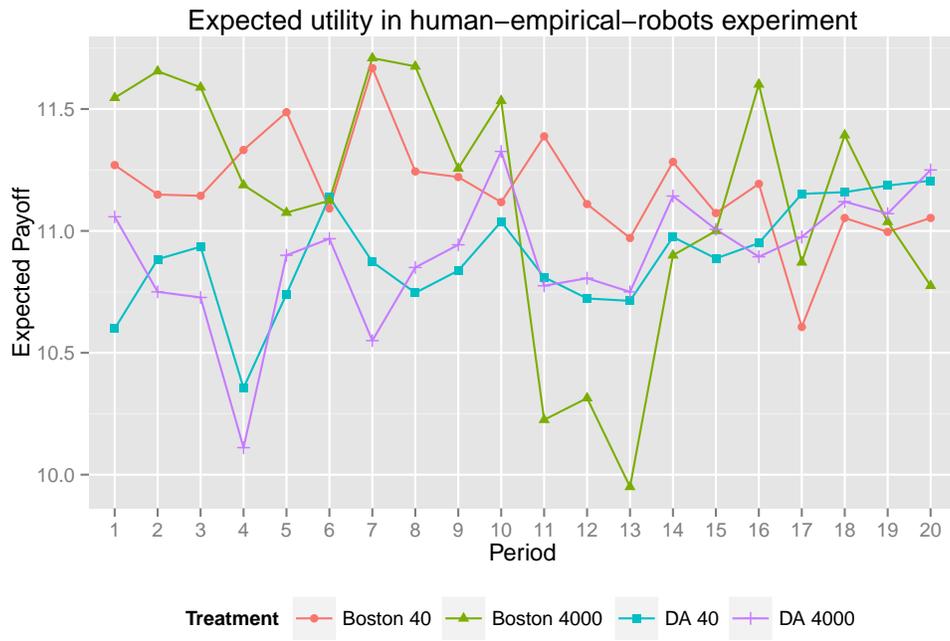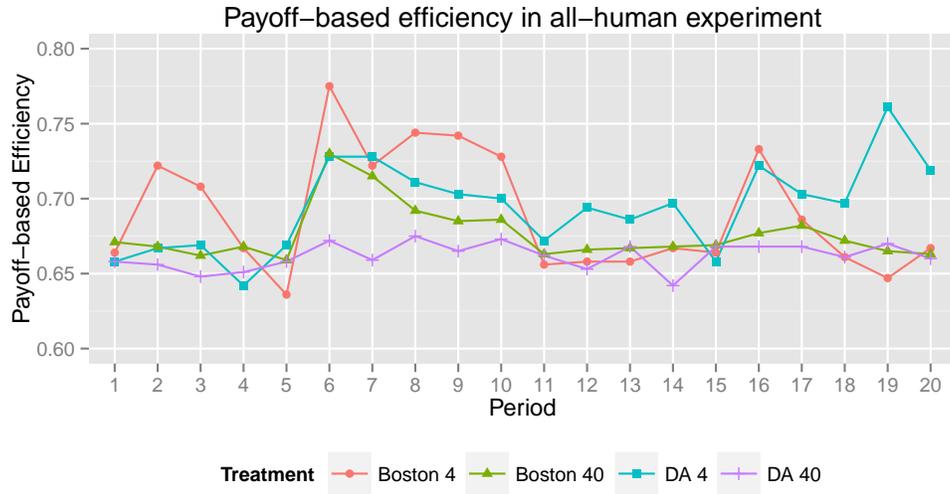
Figure 10: Payoff-based efficiency in all-human treatments (upper panel) and expected payoff of human players in human-vs-empirical-robots treatments (lower panel)

References

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

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. (2005a). The New York City high school match. *American Economic Review*, 95(2):364–367.

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

Ai, C. and Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics letters*, 80(1):123–129.

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.

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. (2013). Complementarity and multidimensional heterogeneity in matching markets. http://assets.wharton.upenn.edu/~eazevedo/papers/Azevedo%20Hatfield%20-%20Existence%20of%20Stable%20Matchings%20with%20Complements.pdf.

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.

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

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

Che, Y.-K., Kim, J., and Kojima, F. (2013). Efficient assignment with interdependent values. http://www.columbia.edu/~yc2271/files/papers/CKK.pdf.

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*, 127:202–231.

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.

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.

Ding, T. and Schotter, A. (2015a). Learning and mechanism design: An experimental test of school matching mechanisms with intergenerational advice. New York University Manuscript.

Ding, T. and Schotter, A. (2015b). Matching and chatting: An experimental study of the impact of network communication on school-matching mechanisms. New York University Manuscript.

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

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

Featherstone, C. and Niederle, M. (2014). Improving on strategy-proof school choice mechanisms: An experimental investigation. http://assets.wharton.upenn.edu/~claytonf/IncInfo14-02-01-2014.pdf.

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.

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. (2012). Promoting school competition through school choice: A market design approach. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1984876.

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.

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.

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. *American Economic Review*, 99:608–627.

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.

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

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

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.

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