The Instability of Matching with Overconfident Agents: Laboratory and Field Investigations

Siqi Pan*

November 20, 2017

Abstract

Many centralized college admissions markets allocate seats to students based on their performance on a standardized exam. A single exam's measurement error can cause the exam-based priorities to deviate from colleges' aptitude-based preferences. Previous literature proposes correcting such an error by combining pre-exam preference submission with a Boston algorithm. In a laboratory experiment, I find that pre-exam preference submission is skewed by overconfidence, which leads to even more severe and more varied welfare distortions than the exam's measurement error alone. Moreover, the mechanism exhibits a tendency to reward overconfidence and punish underconfidence, thus serving as a gender penalty for women. I also analyze field data from China and reach similar conclusions.

1 Introduction

In admissions, colleges usually prefer students with certain qualities or aptitudes that are not readily observable. Students are therefore evaluated by noisy signals such as SAT scores and high-school transcripts. Some countries—including China, South Korea, Turkey, Russia, and Greece—simply use a standardized exam as an evaluation system, which means that every year around the world, more than 13 million students' college admissions are determined by their performance on a single exam.[1]

*Department of Economics, The University of Melbourne, VIC 3010, Australia (e-mail: siqi.pan@unimelb.edu.au). I am deeply indebted to my advisor Paul J. Healy for countless conversations and encouragement while advising me on this project. I would like to thank John Kagel, Lucas Coffman, Katie Baldiga Coffman, Huanxing Yang, Yaron Azrieli, Hal Arkes, Clayton Featherstone, Yan Chen, John Hatfield, SangMok Lee, Marek Pycia, Alex Gotthard-Real, Ritesh Jain, Anthony Bradfield, seminar participants at the Ohio State University, and audiences at the 2015 North American Economic Science Association Conference for helpful comments and inspiration. I am grateful to Ming Jiang, Hengli Zhang, Ming Pan, and Qiquan Xu for their help in the data collection process. This project was supported by JMCB Grants for Graduate Student Research in Economics. All remaining errors are mine.

[1] Table 9 in Appendix C provides the statistic for each representative country.

Since the exam is conducted only once a year, it greatly simplifies the admissions process and is thus especially favorable for large markets. However, a single exam always entails measurement error, and a student who underperforms on the exam may lose her placement at a preferred college to someone with a lower aptitude. Previous literature (discussed in more detail below) proposes a hopeful hypothesis that such an error in the evaluation system could be corrected with proper market design. Under the proposed mechanism, students apply to colleges or submit their preferences before taking the exam. If we assume that students have perfect knowledge about their relative aptitudes, this mechanism triggers a self-sorting process, with students who have higher (lower) aptitudes targeting more (less) preferred colleges. However, given the considerable evidence regarding self-evaluation biases such as overconfidence, I argue that in practice, students may not be able to sort themselves perfectly. As a result, the proposed solution may introduce a behavioral error into the admissions procedure that is larger than the measurement error from the exam alone. This paper presents a tradeoff between these two errors that may emerge in real markets in the presence of overconfident agents.

Formally, the above issue can be described as a college admissions problem, which is a matching problem that involves pairing members of one group of agents (students) with members of another group of agents (colleges). A centralized procedure, called a matching mechanism, is adopted to solve the problem based on students' submitted preferences for colleges along with their priority ordering at each college. The priority ordering, which is solely determined by the score ranking on a single standardized exam, serves as a noisy proxy for colleges' true preferences regarding students.

Suppose that all colleges prefer students with higher aptitudes and that students have correlated preferences for colleges. Then a matching mechanism is considered socially desirable or "fair" if it matches more preferred colleges with students who have higher aptitudes, not just with those who have higher realized scores. To capture the notion of such fairness, a market outcome is said to be *stable with regard to aptitude* ("aptitude-stable") if it eliminates all cases in which a student with a higher aptitude is not assigned to a preferred college but instead another student with a lower aptitude is assigned to that college.[2]

Previous studies such as Jiang (2014), Lien, Zheng, and Zhong (2015, 2017), and Wu and Zhong (2014) propose a hopeful hypothesis that aptitude-stability is more likely to be achieved by a "pre-exam Boston mechanism" (henceforth "PreExam-BOS"), which combines a Boston matching algorithm with pre-exam preference submissions—that is, students are asked

[2]A formal definition that considers the college's unoccupied seats is given in Section 2.1. Aptitude-stability, defined according to the true preferences of the market participants, is simply called "stability" in the standard matching literature (Gale and Shapley, 1962). Here I emphasize the word "aptitude" because in the current setting, we can also define stability according to the exam-based priorities, which may deviate from colleges' aptitude-based preferences due to a single exam's measurement error (see Section 2.1 for more details).

to submit their preferences before taking the exam.[3] The argument goes as follows. A Boston algorithm is not strategy-proof; instead of truthful revelation, a student should indicate her first choice with a "safer" college at which she is more likely to win a seat. With pre-exam preference submissions, each student should employ a self-sorting strategy based on her ex-ante expected exam performance, which perfectly reflects her aptitude. Then the seats at more desirable colleges will be "reserved" for students with higher aptitudes, since students with lower aptitudes will sort themselves into lower-ranked colleges.

The preceding argument hinges on the assumption that students are able to correctly sort themselves before taking the exam, which requires that each student have perfect knowledge of the relative standing of her aptitude among all students in the market. Such knowledge is then "reported" to the market designer through the way each student misrepresents her preferences under PreExam-BOS. However, considerable evidence has established the existence and heterogeneity of biases, such as overconfidence, in self-evaluation. In other words, while the proposed solution theoretically diminishes the effect of the measurement error of a single exam, it might introduce a behavioral error due to self-evaluation biases. As I show in this paper, these biases have important consequences for aptitude-stability.

Using a college admissions model, I first offer theoretical predictions regarding the market outcome under different matching mechanisms. Following previous literature, I mainly compare PreExam-BOS to a "post-score Serial Dictatorship mechanism" (henceforth "PostScore-SD"), which combines a Serial Dictatorship algorithm with post-score preference submission—that is, students submit their preferences after seeing the exam results.[4] In contrast to PreExam-BOS, PostScore-SD is strategy-proof. With every student truthfully revealing her preferences, the matching outcome under PostScore-SD only depends on exam-based priorities and is thus distorted from aptitude-stability by the exam's measurement error. On the other hand, as a dominant strategy under PostScore-SD, truth-telling is not affected by a student's self-evaluation biases. Hence, compared to PreExam-BOS, PostScore-SD is more vulnerable to the noise from the single-exam system but less influenced by students' over- or underconfidence. Which mechanism will create smaller distortions from aptitude-stability depends on the relative magnitudes of these two effects.

Since we cannot exogenously vary the choice of mechanisms in the field, I conduct a lab experiment to investigate students' strategic behaviors and the market outcomes under different mechanisms. In an experimental market, each subject, playing the role of a student, is asked to take an exam, guess the exam results, and submit her preferences regarding simulated colleges to a matching algorithm. The exam is designed as a real-effort task in the lab, and a

[3] The procedure of a Boston algorithm is described in Section 2.2.1.

[4] I also discuss a third timing of preference submission named "halfway," under which students submit preferences after the exam but before the revelation of exam results. Thus, a total of six mechanisms are considered: two algorithms (BOS and SD) combined with three timings of preference submission (pre-exam, halfway, and post-score). The procedure of a Serial Dictatorship algorithm is described in Section 2.2.2.

subject's aptitude is evaluated as her average performance on multiple exams. As treatments, different matching algorithms and timings of preference submission are adopted.

The experimental data confirm that a majority of students report their preferences truthfully under PostScore-SD, while under PreExam-BOS, their strategies are significantly skewed by over- or underconfidence. Thus, neither mechanism fully achieves aptitude-stability on the market level. To measure how much a student's welfare is distorted from an aptitude-stable matching, I compare the desirability of her aptitude-stably matched college to that of her assignment under each treatment. The result shows that PreExam-BOS creates more severe, noisier distortions from aptitude-stability than PostScore-SD. This is because not only are fewer students assigned to their aptitude-stably matched colleges, but the magnitudes of such welfare distortions are also more spread out among students.[5] In other words, under PreExam-BOS, some students receive a large advantage while some others are considerably hurt, and neither the gains nor the losses can be justified by the students' aptitudes.

These results can be explained by three observations: (i) on average, students exhibit overconfidence; (ii) there is significant heterogeneity in their levels of overconfidence; (iii) students make more heterogeneous strategic choices under PreExam-BOS, as opposed to the highly aligned truth-telling behavior under PostScore-SD. In particular, under PreExam-BOS, subjects tend to choose more aggressive or optimistic strategies than self-sorting based on their guessed exam performances. Thus, in the setting of this study, PreExam-BOS introduces more noise into the admissions procedure through both self-evaluation biases and strategic behaviors. Regarding the welfare effect of overconfidence, I find that PreExam-BOS tends to reward those who are overconfident and punish those who are underconfident. Since women in the lab sample tend to exhibit less overconfidence than the men in the sample, PreExam-BOS serves on average as a gender penalty for women.

On the other hand, previous experimental studies that do not take overconfidence into account find evidence that PreExam-BOS can outperform PostScore-SD in terms of aptitude-stability.[6] The key difference lies in how subjects obtain information regarding their aptitudes before the exam. In Lien et al. (2015) and Jiang (2014), the exam component is abstracted away from the experiment, and each student is instead simply provided with her score distribution (that is, the distribution from which her score will be drawn) together with the score distributions of all the other students in the market.[7] In contrast, the design in the present

[5] In addition to fairness-related concerns, distortions from aptitude-stability can also lead to overall welfare loss from one of the consequences of unstable matching, namely, a costly rematching process. In college admissions, this usually takes the form of a student repeating her last year of high school and re-entering the market a year later.

[6] In Lien et al. (2015), the advantage of PreExam-BOS in terms of aptitude-stability mainly appears in an additional 10-round learning treatment. Such an advantage is not significant in the original treatment due to subjects' deviation from equilibrium strategies, which is consistent with the aggressive strategic choices observed in the present paper.

[7] Under an additional treatment (the "Quiz" treatment) in Lien et al. (2015), subjects are asked to take a

paper comes closer to the field setting: subjects collect information from feedback provided by multiple practice exams, or "mock tests." As a result, overconfidence severely skews pre-exam preference submission and prevents PreExam-BOS from achieving aptitude-stability.

In an attempt to minimize self-evaluation biases under the setting for this study, I introduce additional treatments in which I help subjects collect information by showing them all of the previous scores for every student in the market as well as all of their average scores. However, in these cases as well, the subjects' overconfidence remains at the same level, and PreExam-BOS continues to be inferior to PostScore-SD. Such a result indicates that a subject's overconfidence is mainly driven by her overoptimistic belief about how much she can improve on the upcoming exam. In other words, the observed biases mainly stem from a source that cannot easily be muted with an increased amount of information.

For a matching market to function well in practice, the choice of a matching mechanism should be tailored to the specific market environment, in this case to the students' self-evaluation biases and the exam's measurement error. Therefore, to investigate Chinese college admissions, I collected field data on students' guessed and realized exam results, multiple mock test results (used to measure academic aptitudes), and demographic information. A simple analysis shows that students exhibit biases in self-evaluation, and the variance of such biases can be larger than the variance of the exam's measurement error. This result is interesting due to the strikingly high level of competition and the high stakes involved in students' self-evaluation and preference submission process. The fact that over- or under-confidence survives even with extreme incentives to have correct self-evaluations indicates the prevalence of these biases.

Using the data collected in the field and the strategic patterns observed in the lab, a simulation is conducted to compare the performance of different matching mechanisms in the specific market of interest. The results exhibit a similar pattern as observed in the lab: compared to PostScore-SD, PreExam-BOS tends to create more severe and more varied distortions from aptitude-stability. This suggests a potential explanation for the recent reforms in China's college admissions policy: despite what is recommended by the previous research, most districts are currently in transition from a mechanism that resembles PreExam-BOS to a mechanism more similar to PostScore-SD.[8]

This study is part of the recent literature on school choice and college admissions problems with a single-exam evaluation system. Wu and Zhong (2014), Jiang (2014), and Lien et al.

short quiz; a subject's relative performance on the quiz determines which role she will play in a group. However, the score distribution and the relative aptitude of each role are predetermined and are directly provided to the subjects.

[8] In Chinese college admissions, when a student's welfare is significantly and negatively distorted from aptitude-stability, a typical consequence is that she rejects the assignment and re-enters the market after a year. A newspaper article in Beijing Youth Daily reports that Beijing's policy reform in 2015 yields a higher admission rate, and the number of such students is reduced by about 20%. This provides an anecdotal evidence that the new mechanism may create less severe distortions from aptitude-stability.

(2016) theoretically compare PreExam-BOS and PostScore-SD and show that PreExam-BOS can outperform PostScore-SD in terms of aptitude-stability. However, as mentioned above, it is assumed that students have perfect knowledge of their relative aptitudes before taking the exam. Such information is directly provided to subjects in the lab experiments conducted by Lien et al. (2015) and Jiang (2014), where they find support for the above theoretical results. In contrast, the current paper tries to relax this assumption by allowing biases in self-evaluation. Wu and Zhong (2014) conduct empirical research using data from a top university in China. They show that students admitted under PreExam-BOS exhibit similar or better college academic performance than those admitted through other mechanisms. This result provides some evidence that PreExam-BOS could match the top college with students of better aptitudes, yet it is silent about middle- or lower-ranked colleges or the overall matching outcome.

1.1 Related Literature

This study contributes to the recent literature on school choice and college admissions problems with a single-exam evaluation system. Lien et al. (2017), Jiang (2014), and Wu and Zhong (2014) theoretically compare PreExam-BOS and PostScore-SD and show that PreExam-BOS can outperform PostScore-SD in terms of aptitude-stability.[9] However, as previously mentioned, these studies assume that students have perfect knowledge of their relative aptitudes before taking the exam. Such information is directly provided to subjects in the lab experiments conducted by Lien et al. (2015) and Jiang (2014), in which they find support for the above theoretical results. In contrast, the present paper tries to relax this assumption by allowing biases in self-evaluation. Wu and Zhong (2014) conduct empirical research using data from a top university in China. They show that students admitted under PreExam-BOS exhibit similar or better college academic performance than those admitted through other mechanisms. This result provides some evidence that PreExam-BOS can match the top college with students who have higher aptitudes, yet it is silent about middle- or lower-ranked colleges as well as the overall matching outcome.

The present paper also contributes to the literature addressing the merits and flaws of the Boston algorithm compared to Serial Dictatorship and other strategy-proof algorithms. In a standard setting without uncertainty or imperfect information, the Boston algorithm is often considered inferior in terms of stability and efficiency (Abdulkadiroğlu and Sönmez, 2003; Chen and Sönmez, 2006; and Ergin and Sönmez, 2006).[10] On the other hand, the manipulability of the Boston algorithm can sometimes improve ex-ante efficiency because it reflects some information that is otherwise unobservable to a market designer (see, for example,

[9] Lien et al. (2017) identify the conditions under which PreExam-BOS can or cannot achieve complete aptitude-stability.

[10] See also Klijn, Pais, and Vorsatz (2013) and Pais and Pintér (2008).

Abdulkadiroğlu, Che, and Yasuda, 2011 and Featherstone and Niederle, 2008).

Finally, the present paper is related to the literature on overconfidence across economics, psychology, and finance.[11] As in the present study, other studies have used overconfidence to explain market failures in various environments, such as excessive business entry and trading volume, corporate investment distortions, and stock market bubbles, to name just a few (see Camerer and Lovallo, 1999; Glaser and Weber, 2007; Malmendier and Tate, 2005; and Odean, 1999). Moreover, it is well established that people exhibit heterogeneous levels of over- or underconfidence, which can be predicted by certain factors, including personality, gender, and cognitive abilities (see Barber and Odean, 2001; Coffman, 2014; Kleitman and Stankov, 2007; Niederle and Vesterlund, 2007; Schaefer, Williams, Goodie, and Campbell, 2004; and Stankov and Crawford, 1996). The present study has also found similar evidence in both lab and field data.

The rest of the paper is organized as follows. In Section 2, I lay out the college admissions model and make theoretical predictions. Section 3 describes the lab experiment and presents experimental results. In Section 4, I introduce the application of Chinese college admissions, describe the field data collection process, and summarize the results of data analysis and simulation. Section 5 concludes the paper.

2 The Model

2.1 A College Admissions Problem

The centralized matching market considered here is a variation of the college admissions problem (Gale and Shapley, 1962): there are a number of students; each of them is to be assigned a seat at one of a number of colleges. Each student has strict preferences over all colleges and each college has strict preferences over all students. There is a maximum capacity at each college, but the total number of seats exceeds the total number of students. The distinguishing feature of this environment is that every college has a priority ordering of all students, which is not necessarily in accord with its preference relation over students: the former is determined by students' performance on a single exam, while the latter is determined by their intrinsic aptitudes. Formally, the college admissions problem consists of:

1. A set of students $I = \{i_1, \ldots, i_n\}$, $n \geq 2$.

2. A set of colleges $C = \{c_1, \ldots, c_m\}$, $m \geq 2$.

3. A capacity vector $q = (q_{c_1}, \ldots, q_{c_m})$.

[11] Moore and Healy (2008) provide a detailed overview of different ways in which the literature has defined overconfidence, i.e., as overestimation, overplacement, or overprecision. They then offer a theory that reconciles these concepts and explains several inconsistencies that can be found in the existing evidence.

4. A list of strict student preferences $P_I = (P_{i_1}, \ldots, P_{i_n})$. The preference relation P_i of student i is a linear order over $C \cup \{i\}$, where cP_ic' means that student i strictly prefers college c to college c' and i denotes remaining unmatched. Students prefer any college to remaining unmatched.

5. A vector of students' aptitudes $a = (a_1, \ldots, a_n)$ and a corresponding vector of aptitude ranks $r_a = (r_{a_1}, \ldots, r_{a_n})$, where a_i denotes student i's aptitude and r_{a_i} denotes the rank of her aptitude among all students (with 1 being the highest rank). Ties in aptitudes are randomly broken.

6. A vector of students' exam scores $s = (s_1, \ldots, s_n)$ and a corresponding vector of exam score ranks $r_s = (r_{s_1}, \ldots, r_{s_n})$, where s_i denotes student i's exam score and r_{s_i} denotes the rank of her exam score among all students (with 1 being the highest rank). Ties in scores are randomly broken.

7. A list of strict college preferences $P_C = (P_{c_1}, \ldots, P_{c_m})$. The preference relation P_c of college c is a linear order over $I \cup \{c\}$, where iP_ci' means that college c strictly prefers student i to student i' and c denotes leaving a seat empty. Colleges prefer any student to leaving a seat empty. Colleges have identical preferences over students, which are determined by students' aptitude ranking; i.e., $iP_ci' \Leftrightarrow r_{a_i} < r_{a_{i'}}, \forall c \in C$.

8. A strict priority ordering of students at every college that is determined by students' exam score ranking: student i has a higher priority than student i' at every college if and only if $r_{s_i} < r_{s_{i'}}$. All colleges have the same priority ordering.

Student i's exam score is given by $s_i = a_i + \xi_i$, where ξ_i is the measurement error of an exam. I assume that a student's aptitude is the mean and the mode of her exam score distribution; that is, one exam score is an unbiased but noisy measure of aptitude.

Assumption 1. *For student i, an exam's measurement error ξ_i follows a distribution on the real line with mean 0 and non-zero standard deviation.*[12]

Similarly, student i's exam score rank is given by $r_{s_i} = r_{a_i} - \epsilon_i$, where ϵ_i is the measurement error of the exam in terms of rank.[13]

Below I make a simplifying assumption following the previous literature.

Assumption 2. *Students have identical preferences over colleges.*[14]

[12] The distribution of ξ_i can be continuous or discrete. Assumption 1 is essentially an assumption on s_i and a_i because ξ_i is derived from $\xi_i = s_i - a_i$. Empirically, it can be easily satisfied with normalization. See Section 4 for more details.

[13] Here I use $\epsilon_i = r_{a_i} - r_{s_i}$ instead of $\epsilon_i = r_{s_i} - r_{a_i}$ to be consistent with the definition of overplacement in Section 2.3.

[14] It reflects the reality of many college admissions markets that students' preferences over colleges are correlated. Following the previous literature, here I simplify the environment by assuming homogeneity. Although relaxing such an assumption is a well-motivated extension, it is not the focus of this paper.

Without loss of generality, assume in addition that a college with a smaller index is more desirable; i.e., $c_j P_i c_{j'} \Leftrightarrow j < j', \forall i \in I$.

The outcome of a matching market is known as a matching. Formally, a matching is a function $\mu : I \cup C \to 2^I \cup C$ such that for any $i \in I$ and any $c \in C$, (i) $\mu(i) \in C \cup i$, (ii) $\mu(c) \in 2^I$, (iii) $\mu(i) = c$ if and only if $i \in \mu(c)$, and (iv) $|\mu(c)| \leq q_c$. Thus, $\mu(i)$ denotes the assignment of student i under matching μ, and $\mu(c)$ denotes the set of students who are matched to college c under matching μ.

In the matching literature, stability is used as a central criterion to evaluate a matching outcome. Under the current structure, we can either set such a criterion according to colleges' aptitude-based preferences, or according to their exam-based priorities. They are called stability with regard to aptitude and stability with regard to exam score, respectively.[15] The definitions are given below.

Definition 1. A matching μ is *stable with regard to aptitude* ("aptitude-stable") if and only if there is no student–college pair (i, c) such that student i prefers college c to her assignment $\mu(i)$ and either (1) college c has empty seats under μ, or (2) at least one of the students in $\mu(c)$ has a lower aptitude than student i.

Definition 2. A matching μ is *stable with regard to exam score* ("score-stable") if and only if there is no student–college pair (i, c) such that student i prefers college c to her assignment $\mu(i)$ and either (1) college c has empty seats under μ, or (2) at least one of the students in $\mu(c)$ has a lower exam score than student i.

Clearly, aptitude-stability is more socially desirable than score-stability since it respects colleges' true preferences, which are assumed to be based on students' aptitudes instead of their scores in one exam. Below I use a simple example to illustrate the environment.

Example 1. *Suppose there are three students* $I = \{i_1, i_2, i_3\}$ *and three colleges* $C = \{c_1, c_2, c_3\}$ *in the market. Each college has only one slot to fill* $q = \{1, 1, 1\}$*. Students have homogeneous preferences over colleges:* $c_1 P_i c_2$ *and* $c_2 P_i c_3$*,* $i = 1, 2, 3$*. On a single exam, a student's performance is consistent with her aptitude with probability* 0.5*; she overperforms with probability* 0.25*, and underperforms with probability* 0.25*. Table 1 shows the score distributions, which are independent across students.*

[15] Similar concepts are named ex-post and ex-ante fairness in a school choice setting.

Table 1: Score Distributions and Aptitudes (Example 1)

Student	Aptitude	Score (Prob.)	Overperform (0.25)	Consistent (0.50)	Underperform (0.25)
i_1	$a_1 = 12$	$s_1 =$	16	12	8
i_2	$a_2 = 15$	$s_2 =$	19	15	11
i_3	$a_3 = 9$	$s_3 =$	13	9	5

A student's aptitude is given by the mean and mode of her score distribution. The second column of Table 1 implies students' aptitude ranks $r_a = (2, 1, 3)$*, which determine every college's aptitude-based preferences. So the unique aptitude-stable matching is*

$$\begin{matrix} i_1 & i_2 & i_3 \\ c_2 & c_1 & c_3 \end{matrix}.$$

However, students may exhibit any ranking in their exam scores.[16] *With probability* $\frac{17}{64}$*, the exam's measurement error leads to the realized score ranks* $r_s = (1, 2, 3)$*. In this case, the unique score-stable matching is*

$$\begin{matrix} i_1 & i_2 & i_3 \\ c_1 & c_2 & c_3 \end{matrix},$$

which is not aptitude-stable because both i_2 *and* c_1 *prefer each other to their current assignments.*

Example 1 shows how an exam's measurement error could distort students' realized score ranking from their aptitude ranking and thus prevent the score-stable matching from achieving aptitude-stability. Below I compare different mechanisms (combinations of a matching algorithm and a timing of preference submission) and examine which one is less likely to be effected by such noise from a single exam and is more likely to yield an aptitude-stable matching.

2.2 Two Matching Algorithms

To select a matching for the college admissions problem defined above, a systematic procedure, called a "matching algorithm," allocates students to colleges depending on students' submitted preferences and colleges' priority ordering. In terms of timing, preference submission could occur before or after the exam, but the actual matching algorithm is always conducted after preference submission and the revelation of exam results.

In the literature of college admissions and school choice problems, three matching algorithms are most widely discussed: the Boston algorithm (BOS), the Gale-Sharply Deferred

[16] Table 10 in Appendix C shows the probability of every possible score ranking.

Acceptance algorithm (DA), and the Top Trading Cycles algorithm (TTC). In the current setting where all colleges share the same priority ordering, TTC reduces to a Serial Dictatorship algorithm (SD) and is equivalent to DA (Kesten, 2006). Therefore, BOS and SD are the two competing algorithms considered in this paper.[17]

2.2.1 The Boston Algorithm (BOS)

Each student submits her preferences by ranking all colleges. Every college has the same strict priority ordering of students, which is determined by students' exam score ranking.

Step 1: Only the 1st choices of all students are considered. For each college, consider the students who have listed it as their 1st choice; assign seats of the college to these students one at a time following their priority ordering until either there are no seats left or there are no students left who have listed it as their 1st choice.

In general, Step k ($k \geq 1$) can be described as follows.

Step k: Only the kth choices of the remaining students (who have not been assigned a seat previously) are considered. For each college with still available seats, consider the students who have listed it as their kth choice; assign the remaining seats to these students one at a time following their priority ordering until either there are no seats left or there are no students left who have listed it as their kth choice.

The procedure terminates after any step k when every student is assigned a seat at some college, or if the only students who remain unassigned listed no more than k choices.

2.2.2 The Serial Dictatorship Algorithm (SD)

Each student submits her preferences by ranking all colleges. Every college has the same strict priority ordering of students, which is determined by students' exam score ranking.

Step 1: The student with the highest priority is considered. She is assigned a seat at the college of her 1st choice.

Step 2: The student with the second highest priority is considered. She is assigned a seat at her 1st choice if that college still has empty seats left; otherwise, she is assigned a seat at her 2nd choice.

In general, Step k ($k \geq 2$) can be described as follows.

Step k: The student with the kth highest priority is considered. She is assigned a seat at her most preferred college that has an empty seat.

[17] I choose SD instead of DA to be more consistent with the field environment of Chinese college admissions. See Section 4.1 for more details.

The procedure terminates when every student has been considered, or when no college seats remain.

2.3 Three Timings of Preference Submission

Apart from matching algorithms, the timing of preference submission can also largely affect the strategic behaviors of market participants, thus influencing the market outcome. Inspired by college admissions in China, I focus on three different timings, under which students are asked to submit their preferences at different stages, or different information statuses. The timings and stages are named as follows.

The "ex-ante," "interim," and "ex-post" stages refer to: before the exam, after the exam but before the revelation of exam results, and after the revelation of exam results.[18] Under the timings named "pre-exam," "halfway," and "post-score," students submit their preferences at the ex-ante, interim, and ex-post stages, respectively. The following assumption specifies the information status at the ex-post stage.

Assumption 3. *At the ex-post stage (after the revelation of exam results), it is common knowledge that every student knows her own exam score rank.*

When submitting preferences under the pre-exam and halfway timings, students do not observe the exam results, which means they do not know their priority ordering at each college. In some situations (discussed in Section 2.4), their strategies may depend on their guessed exam results. Therefore, under the pre-exam timing, a component is added to the college admissions problem defined in Section 2.1:

9. A vector of students' guessed exam scores $\hat{s}^{EA} = \left(\hat{s}_1^{EA}, \ldots, \hat{s}_n^{EA}\right)$ and a corresponding vector of students' guessed exam score ranks $\hat{r}_s^{EA} = \left(\hat{r}_{s_1}^{EA}, \ldots, \hat{r}_{s_n}^{EA}\right)$ at the ex-ante stage, where $\hat{s}_i^{EA}$ and $\hat{r}_{s_i}^{EA}$ denote student i's guessed score and guessed rank.

Under the halfway timing, the following component is added instead:

9'. A vector of students' guessed exam scores $\hat{s}^{IN} = \left(\hat{s}_1^{IN}, \ldots, \hat{s}_n^{IN}\right)$ and a corresponding vector of students' guessed exam score ranks $\hat{r}_s^{IN} = \left(\hat{r}_{s_1}^{IN}, \ldots, \hat{r}_{s_n}^{IN}\right)$ at the interim stage, where $\hat{s}_i^{IN}$ and $\hat{r}_{s_i}^{IN}$ denote student i's guessed score and guessed rank.

In the current setting, overconfidence, defined as a bias in self-evaluation, can be measured in score or rank. Following Moore and Healy (2008), I refer to score overconfidence as "overestimation," and rank overconfidence as "overplacement."

Definition 3. Student i's *overestimation* at the ex-ante stage is given by $\delta_i^{EA} \equiv \hat{s}_i^{EA} - E\left[s_i\right] = \hat{s}_i^{EA} - a_i$ and at the interim stage by $\delta_i^{IN} \equiv \hat{s}_i^{IN} - E\left[s_i\right] = \hat{s}_i^{IN} - a_i$.

[18] The interim stage is defined as a distinct information status, because a student may obtain additional information during the exam.

Definition 4. Student i's *overplacement* at the ex-ante stage is given by $\theta_i^{EA} \equiv r_{a_i} - \hat{r}_{s_i}^{EA}$ and at the interim stage by $\theta_i^{IN} \equiv r_{a_i} - \hat{r}_{s_i}^{IN}$.[19]

Naturally, a student is said to exhibit underconfidence when the above measures take negative values.

2.4 Theoretical Predictions

This section gives the theoretical predictions for strategies of market participants and the stability of matching outcomes under different combinations of matching algorithms and timings of preference submission. I refer to the SD (BOS) algorithm under pre-exam, halfway, or post-score timing as "PreExam-SD," "Halfway-SD," or "PostScore-SD" mechanism ("PreExam-BOS," "Halfway-BOS," or "PostScore-BOS" mechanism).

I first state the results for PreExam-, Halfway-, and PostScore-SD, which are largely drawn from the previous literature.

Proposition 1. *(1) PreExam-SD, Halfway-SD, and PostScore-SD are strategy-proof. (2) PostScore-SD always yields the score-stable matching. (3) PreExam-SD and Halfway-SD yield the score-stable matching in the truth-telling equilibrium.*

It is well-established in the literature that SD is strategy-proof for any realized priority ordering over students, which means truth-telling is a weakly dominant strategy for every student, regardless of her knowledge about the priority ordering at the time of preference submission. Hence, no matter which timing of preference submission is adopted, SD always implements the score-stable matching outcome in the truth-telling equilibrium (see Appendix A for a more detailed proof).

In contrast, students have strong incentives to misrepresent their true preferences under BOS. The following definition specifies a strategy in preference submission at the ex-post stage.

Definition 5. A student i is said to adopt a *score-based sorting strategy* if she lists college c_j as her first choice in preference submission such that $\sum_{k=1}^{j-1} q_k < r_{s_i} \leq \sum_{k=1}^{j} q_k$.[20]

In the current setting, score-based sorting means listing one's score-stably matched college as the first choice. Recall the environment in Example 1, where every college has only one

[19] First, note that overplacement is defined as $E[r_{s_i}] - \hat{r}_{s_i}^{EA}$ (or $E[r_{s_i}] - \hat{r}_{s_i}^{IN}$) instead of $\hat{r}_{s_i}^{EA} - E[r_{s_i}]$ (or $\hat{r}_{s_i}^{IN} - E[r_{s_i}]$), because a smaller value of rank means being better in aptitude or exam score. Second, there is a slight abuse of terminology in the definitions of δ_i^{IN} and θ_i^{IN}. At the interim stage, a student has obtained some additional information, say a signal t, about her performance on the exam. Therefore, strictly speaking, overestimation and overplacement should be measured as $\hat{s}_i^{IN} - E[s_i|t]$ and $E[r_{s_i}|t] - \hat{r}_{s_i}^{IN}$, respectively. The current definitions are adopted since it is more relevant in this environment to discuss how the *aptitude* ranking is distorted by students' guessed exam results. Third, because the measurement error of an exam has zero mean, overconfidence evaluated relative to posterior beliefs has the same average level as that evaluated relative to prior beliefs.

[20] The concept is also named rank bias in the literature.

seat, and students' realized score ranks are $r_s = (1, 2, 3)$. Then we say all students exhibit score-based sorting if the submitted first choices of i_1, i_2, and i_3 are given by c_1, c_2, and c_3. Hence, BOS will have every student accepted in Step 1 of the procedure and achieve score-stability. Below, Proposition 2 shows that score-based sorting is an equilibrium strategy under PostScore-BOS, and the score-stable matching is implemented in equilibrium (the formal proof is given in Appendix A).

Proposition 2. *(1) Under PostScore-BOS, there is a Nash equilibrium where every student exhibits score-based sorting. (2) PostScore-BOS always implements the score-stable matching.*

Under PreExam- and Halfway-BOS, students do not observe their exam score ranks at the time of preference submission; their strategies are thus affected by their guessed exam results at the ex-ante and interim stages, respectively. As a counterpart of score-based sorting, guess-based sorting is defined below.

Definition 6. Under the pre-exam (or halfway) timing, a student i is said to adopt a *guess-based sorting strategy* if she lists college c_j as her first choice in preference submission such that $\sum_{k=1}^{j-1} q_k < \hat{r}_{s_i}^{EA} \leq \sum_{k=1}^{j} q_k$ (or $\sum_{k=1}^{j-1} q_k < \hat{r}_{s_i}^{IN} \leq \sum_{k=1}^{j} q_k$).

For PreExam- or Halfway-BOS to implement an aptitude-stable matching, it is crucial that every student's guessed rank perfectly reflects her aptitude rank, which should be commonly known to the market. Therefore, the previous literature makes the following assumption and gives the prediction stated in Proposition 3.

Assumption 4. *It is common knowledge that no student exhibits any over- or under-placement.*

Proposition 3. *If Assumption 4 (common knowledge of no over- or under-placement) holds for preference submission at the ex-ante stage (or at the interim stage) and every student exhibits guess-based sorting, PreExam-BOS (or Halfway-BOS) yields the aptitude-stable matching.*

The proof of the above proposition is straightforward. Under Assumption 4, every student who exhibits guess-based sorting lists her aptitude-stably matched college as the first choice. Under BOS, everyone is accepted in Step 1 and aptitude-stability is achieved.

However, if Assumption 4 fails, that is if students exhibit over- or under-placement, the conclusion in Proposition 3 will change significantly. Based on the fact that one is not aware of her own bias, and evidence on the false-consensus effect, I make the following assumption instead.[21]

[21] Under the false-consensus effect, people tend to believe that others are similar to them; see Ross, Greene, and House (1977) for a seminal contribution and Marks and Miller (1987) for a survey. Evidence on such an effect is also found in the lab experiment for this study (Section 3.3; Result 4).

Assumption 4′. Every student believes that she exhibits no over- or under-placement and that other students exhibit no over- or under-placement.

The beliefs specified in Assumption 4′ will be false with the presence of over- or under-placement. Since students' strategies under PreExam- and Halfway-BOS hinge on these beliefs, the matching outcome will be affected as well. This provides the intuition for Proposition 4.[22]

Proposition 4. *If Assumption 4 is replaced by 4′ for preference submission at the ex-ante stage (or at the interim stage) and every student exhibits guess-based sorting, PreExam-BOS (or Halfway-BOS) may fail to achieve aptitude-stability.*

On the other hand, truth-telling under PreExam-, Halfway-, and PostScore-SD and score-based sorting under PostScore-BOS do not depend on students' guessed exam results. Therefore, the market outcomes under these four mechanisms are less easily affected by over- or under-confidence. Below I illustrate the theoretical predictions in the setting of Example 1.

Example 1 (Cont.) **(i)** *Recall that students' aptitude ranks are* $r_a = (2, 1, 3)$ *and their exam score ranks are* $r_s = (1, 2, 3)$*. Suppose their guessed ranks at the ex-ante stage are given by* $\hat{r}_s^{EA} = (1, 1, 2)$*. Then both PreExam- and PostScore-SD yield the score-stable matching in the truth-telling equilibrium; under PostScore-BOS, if* i_1*,* i_2*, and* i_3 *all exhibit score-based sorting by submitting* c_1*,* c_2*, and* c_3 *as their first choices respectively, the score-stable matching is again implemented. However, under PreExam-BOS, if* i_1*,* i_2*, and* i_3 *all exhibit guess-based sorting by submitting* c_1*,* c_1*, and* c_2 *as their first choices, the following matching is implemented:*

$$\begin{matrix} i_1 & i_2 & i_3 \\ c_1 & c_3 & c_2 \end{matrix}.$$

(ii) *Now suppose the exam's measurement error is given by* $\epsilon = (0, 0, 0)$ *and thus* $r_s = r_a = (2, 1, 3)$*; all else stays the same. Then PreExam-BOS yields the following matching with everyone exhibiting guess-based sorting:*

$$\begin{matrix} i_1 & i_2 & i_3 \\ c_3 & c_1 & c_2 \end{matrix}.$$

Part (i) of the example indicates that under PreExam- and Halfway-BOS, overconfidence has two effects on the matching procedure.[23] First, it directly skews the sorting in preference

[22] Note that neither Propositions 3 nor 4 can give equilibrium predictions for a general environment, because students' strategic choices under PreExam- and Halfway-BOS depend on their cardinal utilities and risk attitudes. Here they only serve as a guideline for the subsequent experimental analysis, where these claims are examined using subjects' strategic behaviors and market outcomes in the lab.

[23] Given the definitions of overconfidence at the interim stage (see Definitions 3 and 4; Footnote 19), in this model the halfway timing is theoretically equivalent to the pre-exam timing. Therefore, Example 1 also has implications for Halfway-BOS and Halfway-SD.

submission: under the influence of overplacement, i_3 submits c_2 as her first choice, while i_1 submits c_1 and ends up competing with i_2 in Step 1 of BOS. Second, it brings back the noise from the exam's measurement error: due to the first effect, BOS needs to resolve the competition between i_1 and i_2 according to their exam scores and as a result of the exam's measurement error, i_1 is matched with c_1 although i_2 has a higher aptitude. Therefore, PreExam- and Halfway-BOS can be directly affected by the presence of self-evaluation biases and meanwhile, indirectly by the noise from a single-exam evaluation system.

As for individual welfare, under the setting of Part (i), i_2 is unbiased but is punished since she is allocated to c_3 instead of her aptitude-stable match c_1, while both i_1 and i_3 are rewarded for being overconfident (i_1 is matched to c_1 instead of c_2; i_3 is matched to c_2 instead of c_3). On the other hand, from Part (ii) of the example, we can see the same level of overplacement hurts i_1 but benefits i_3.

Hence, regarding the effects of overconfidence on the market outcome and on individual welfare, the prediction from the model is ambiguous since it depends on the distribution of overconfidence and the realization of the exam's measurement error. To further explore these issues, I conduct a lab experiment where subjects' preferences are induced by monetary incentives. Such a controlled setting allows me to closely observe their strategic choices, examine market stability, and analyze individual welfare.

3 A Lab Experiment

To investigate strategic behaviors and market outcomes under different mechanisms, I design an experiment with various combinations of matching algorithms and timings of preference submission. Compared to other experimental studies in the literature, the distinguishing feature of this design lies in how subjects obtain information regarding their aptitudes. In Lien et al. (2015) and Jiang (2014), the exam component is abstracted away from the experiment; instead, each student is provided with her score distribution (that is, the distribution her score will be drawn from), together with the score distributions of all the other students in the market. There is thus much less scope for over- or under-confidence since subjects are provided with perfect information of their aptitude ranking. In my design, the exam component is introduced as a real-effort task; subjects evaluate themselves at the ex-ante and interim stages using feedback from multiple practice exams, or "mock tests." Such a setting resembles the field environment and allows us to examine the existence of self-evaluation biases.

3.1 Experimental Design

Each experimental market consists of five students and five colleges. Each subject plays the role of a student; colleges are simulated in the environment since they are not strategic. Colleges

are labeled as c_1, c_2, c_3, c_4, and c_5; each has only one slot to fill. All students have the same induced preferences over colleges: when matched with c_1, c_2, c_3, c_4, or c_5, a student receives a payoff of \$20, \$15, \$10, \$5, or \$0, respectively.

Students' priority ordering at each college is determined by their score ranking in an exam. The exam consists of 20 IQ multiple choice questions, and students have 3 minutes to work. One's score equals the number of correct answers; there is no penalty for wrong answers.[24] In order to obtain a strict score ranking and thus a strict priority ordering, ties are broken randomly. When exam results are revealed, each subject can observe the scores and ranks of all five students.

At the ex-ante stage (before the exam), each student is asked to guess her exam score and the rank of her score in the market. Similarly, a guess of score and a guess of rank are again elicited at the interim stage (after the exam but before the revelation of exam results). These guesses are not observable to other students.

Table 2: Treatment Design

Timing	**The BOS Algorithm**	**The SD Algorithm**
Pre-exam	PreExam-BOS	PreExam-SD
Halfway	Halfway-BOS	Halfway-SD
Post-score	PostScore-BOS	PostScore-SD

The experiment has a three-by-two treatment design (Table 2); varying the matching algorithm (BOS or SD) and the timing of preference submission (pre-exam, halfway, or post-score). Under the pre-exam timing, preference submission follows the guess at the ex-ante stage and precedes the exam; under the halfway timing, it follows the guess at the interim stage and precedes the revelation of exam results; and under the post-score timing, it comes after the revelation of results. At the end of every treatment, an algorithm is used to match students with colleges, based on students' submitted preferences and their exam score ranking.

Treatments using the same algorithm (in the same column of Table 2) are implemented within-subject. Every subject makes three sequences of decision making. As illustrated in Figure 1, all sequences include the same six components (an exam, a guess at the ex-ante stage, a guess at the interim stage, the revelation of exam results, preference submission, and a matching procedure), but differ in the timing of preference submission. To ensure a relatively clean treatment effect, the three timings appear in a random order, and no feedback is given in between regarding other students' submitted preferences or the final matching outcomes.

Before the three treatments, subjects are given three "mock tests;" each takes the same

[24] Such a design aims to reduce the gender gap. Baldiga (2013) shows that when there is a penalty for wrong answers, women answer significantly fewer questions than men and thus do significantly worse conditional on their knowledge.

form as the exam (20 questions over 3 minutes). The results, including the scores and ranks of all five students, are revealed at the end of every mock test. This process is for subjects to learn about their aptitudes, as well as the relative standing of their aptitudes, in taking the exam. Such a design provides three mock tests and three exams (one in each treatment) for every subject. I use the average of these six performances as the measure of aptitude.[25]

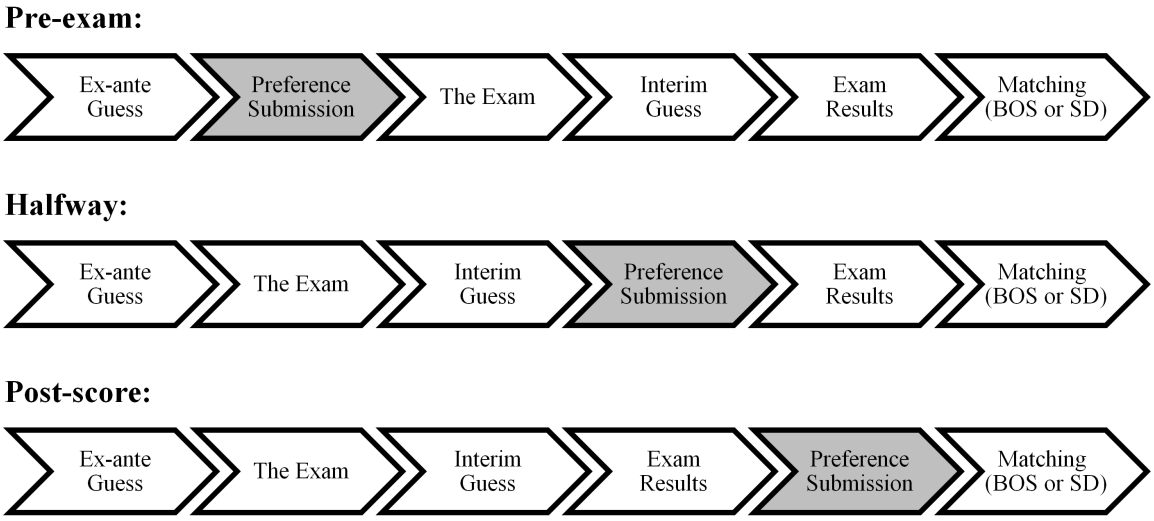


Figure 1: Timings of Preference Submission

3.2 Experimental Procedure

Each session of the experiment consists of three parts. In the first part, either BOS or SD is described and illustrated with an example, followed by five practice rounds of preference submission with randomly assigned ranks (designed to familiarize subjects with the matching algorithm). The second part is the main experiment, three mock tests followed by three treatments. At the end of the second part but before giving any feedback on matching outcomes, I elicit beliefs about other participants' overconfidence level using a question like: "The computer will now randomly choose one of the other participants in the room. During the experiment, this participant has given a total of 6 guesses about his/her rank in the exam. Please give your guesses regarding the correctness of his/her responses by guessing the value of (his/her actual rank - his/her guessed rank) for each guess." The third part elicits risk attitudes using a variation of the lottery game from Holt and Laury (2002).[26]

[25]There is some evidence of learning during the three mock tests. The main results remain unchanged if I exclude all or some of the mock tests from the measure of aptitude.

[26]Subjects are asked to make 20 choices between paired lotteries; each pair consists of a "safe" option and a "risky" option. Following Holt and Laury (2002), the total number of safe choices (ranging from 0 to 20) is used as an indicator of risk aversion. A majority of subjects chose the safe option when the probability of the higher payoff was small, and then crossed over to the risky option without ever going back to the safe option. Only 7 out of 95 subjects exhibited back-and-forth behavior.

Subjects are randomly divided into groups of five and are re-grouped for every practice round of preference submission, every mock test, and every treatment. At the end of the experiment, one mock test, one guess (either a guess of one's own score or rank or a guess of another participant's overconfidence level), and the matching outcome in one treatment are randomly chosen for payment. A subject receives \$0.25 for each correct answer in the chosen mock test, plus \$2 if the chosen guess is correct, together with a payoff of \$20, \$15, \$10, \$5, or \$0, if she is matched with c_1, c_2, c_3, c_4, or c_5 in the chosen treatment. The final payment also includes the payoff from the lottery game, a show-up fee of \$3, and a \$1-payment for completing a questionnaire.

The experiment was conducted in February 2015 at the Experimental Economics Laboratory of The Ohio State University. There were seven sessions in total. One session had 10 subjects; one had 20; and the other five sessions were conducted with 15 subjects. Out of 95 subjects (41 females and 64 males), there were 60 participants for treatments using BOS, and 45 participants for treatments using SD.[27] Each session lasted approximately 75 minutes. The average payment, including a show-up fee, was about \$18.28.

3.3 Experimental Results

Below I describe the statistics on market stability and individual welfare and then analyze patterns of strategic behavior. I also investigate how these results are influenced by the exam's measurement error and subjects' overconfidence levels.

3.3.1 Market Outcomes

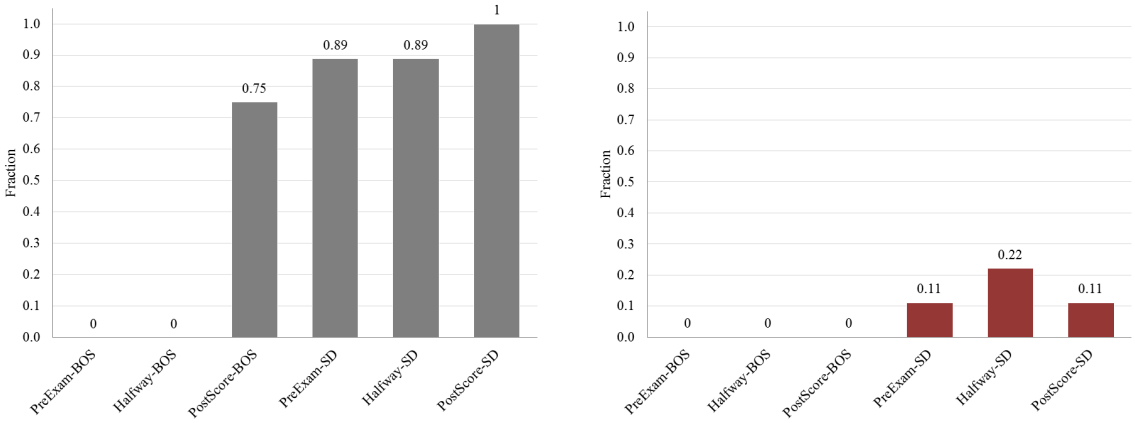
When evaluating a matching mechanism, we are interested in how frequently an aptitude-stable outcome is produced on a market level (Result 1). For more detail, I also examine the proportion of aptitude-stably matched student-college pairs (Result 2).

Result 1. *(i) Score-stability is achieved in all markets under PostScore-SD, most markets under PreExam-SD, Halfway-SD, and PostScore-BOS, but no markets under PreExam-BOS or Halfway-BOS. (ii) Aptitude-stability is rarely achieved under any mechanism.*

Figure 2 summarizes, for each treatment, the fraction of markets that yield the score-stable or aptitude-stable matching.[28] As shown in Figure 2b, aptitude-stability is only observed in 2 out of 9 markets under Halfway-SD, and 1 out of 9 markets under PreExam- and PostScore-SD, which could be considered as coincidences because the aptitude-stable matching also happens to be score-stable in these four markets.

[27] For treatments using SD, the pilot data exhibit less variation since there exists a dominant strategy (truth-telling). Therefore, the power calculation prior to the experiment requires fewer data points.

[28] There are a total of 12 markets under PreExam-, Halfway-, and PostScore-BOS, and a total of 9 markets under PreExam-, Halfway-, and PostScore-SD.



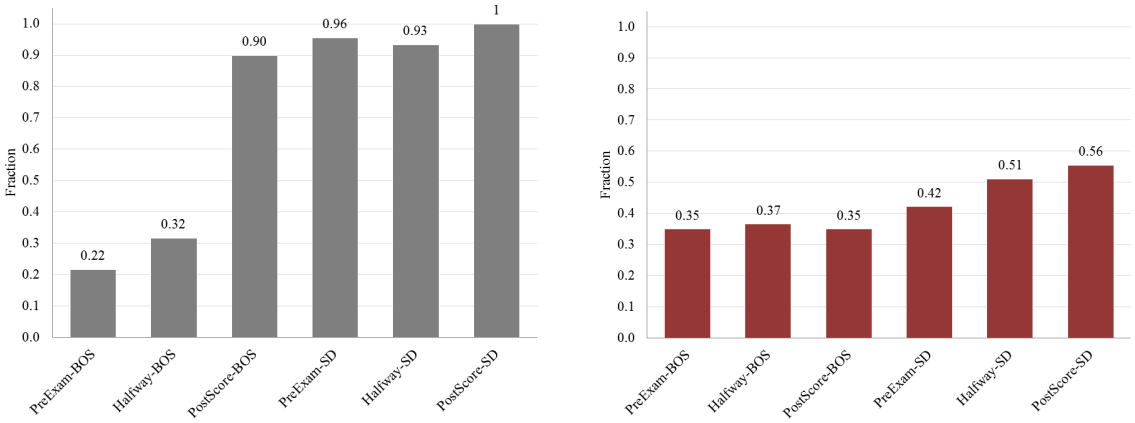
(a) Score-Stable Matchings

(b) Aptitude-Stable Matchings

Figure 2: Score-Stability and Aptitude-Stability

To measure how severely the market outcome is distorted from score-stability (aptitude-stability) under each mechanism, I calculate the proportion of students who are allocated to their score-stably (aptitude-stably) matched colleges, that is, the proportion of score-stably (aptitude-stably) matched pairs.

Result 2. *(i) PreExam-BOS and Halfway-BOS yield a smaller proportion of score-stably matched pairs than the other four mechanisms. (ii) PreExam-BOS and Halfway-BOS yield a smaller proportion of aptitude-stably matched pairs than Halfway-SD and PostScore-SD.*



(a) Score-Stably Matched Pairs

(b) Aptitude-Stably Matched Pairs

Figure 3: Score-Stably and Aptitude-Stably Matched Pairs

According to Figure 3a, there is a smaller proportion of score-stably matched pairs under

PreExam-BOS or Halfway-BOS compared to the other four mechanisms ($p < 0.001$).[29] As for the aptitude-stably matched pairs presented in Figure 3b, PreExam-BOS yields a smaller proportion of such pairs than Halfway-SD ($p = 0.05$) and PostScore-SD ($p = 0.016$). Similarly, there is also a smaller proportion under Halfway-BOS compared to Halfway-SD ($p = 0.076$) and PostScore-SD ($p = 0.026$).[30]

To sum up, the results regarding market outcomes confirm the theoretical predictions on score-stability but largely contradict the predictions on aptitude-stability under Assumption 4 (common knowledge of no over- or under-placement). In particular, not only do PreExam-BOS and Halfway-BOS fail to achieve aptitude-stability but they also create more severe distortions from aptitude-stability compared to Halfway-SD and PostScore-SD. To trace the reason behind such a result, I first examine the existence and heterogeneity of subjects' self-evaluation biases, and then analyze how such biases influence preference submission and market outcomes under different mechanisms.

3.3.2 Overconfidence

Since rank is a much more relevant notion than score in the current setting, below I use overplacement as the primary measure of overconfidence. Recall r_{a_i} ("*AptitudeRank*") refers to a student's rank of aptitude; ϵ_i ("*ExamError*") refers to an exam's measurement error in terms of rank; θ_i^{EA} ("*OverconfidenceEA*") and θ_i^{IN} ("*OverconfidenceIN*") are defined as a subject's level of overplacement at the ex-ante stage and the interim stage.[31]

Result 3. *At both ex-ante and interim stages, (i) students exhibit overconfidence on average; (ii) men exhibit more overconfidence than women.*

The average level of overplacement is 0.50 rankings at the ex-ante stage and is 0.26 at the interim stage. Since both values are significantly greater than zero ($p < 0.001$, t tests), students exhibit overconfidence at both stages. Moreover, θ_i^{EA} is significantly larger than θ_i^{IN} on average ($p < 0.001$, paired t and sign test). Figure 4 compares the distributions of θ_i^{EA} and θ_i^{IN} to the distribution of the exam's measurement error ϵ_i. While all three variables exhibit similar variances, ϵ_i has a significantly larger mass on zero compared to θ_i^{EA} ($p < 0.001$, McNemar's test) or θ_i^{IN} ($p = 0.042$, McNemar's test). This provides us with some intuition behind Result 2: compared to the exam's measurement error, the behavioral error due to

[29] All the proportion tests comparing PreExam-BOS (or Halfway-BOS) to PreExam-SD, to Halfway-SD, and to PostScore-SD yield a p-value smaller than 0.001. The McNemar's test comparing PreExam-BOS (or Halfway-BOS) to PostScore-BOS yields a p-value smaller than 0.001.

[30] The p-values are from proportion tests.

[31] In Section 3.3.2, data from all six treatments are pooled together because in each treatment, subjects' guessed exam results are elicited at both ex-ante and interim stages. Moreover, as shown in Table 3, there is generally no significant treatment effect on overconfidence.

self-evaluation biases could lead to more mismatched pairs, that is, more severe distortions from aptitude-stability.[32]

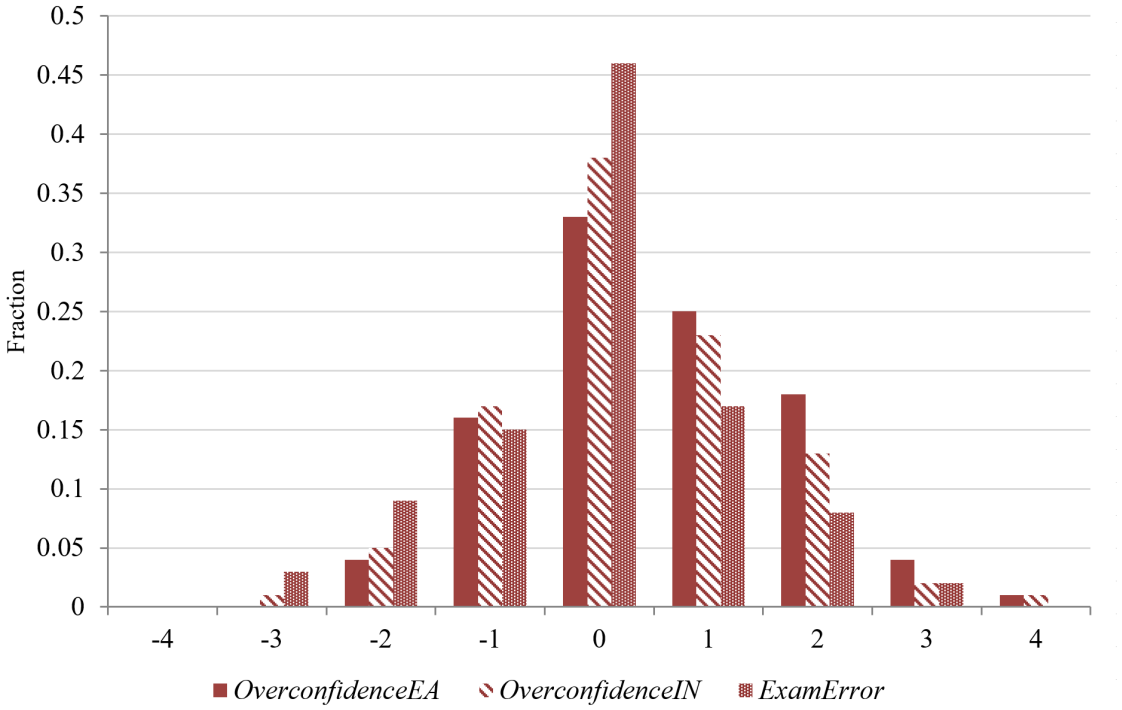


Figure 4: Distribution of *OverconfidenceEA*, *OverconfidenceIN*, and *ExamError*

To understand which factors can influence and thus predict a student's overconfidence level, I run an OLS regression of *OverconfidenceEA* and *OverconfidenceIN*, with the data clustered by subject. The results are displayed in Table 3 and briefly summarized as follows. First, men are more overconfident than women ((the marginal effect of *Female* at the mean of *RiskAverse* is -0.184 at the ex-ante stage and is -0.113 at the interim stage; such a negative effect is more significant for more risk-averse subjects). Second, at the ex-ante stage, those who are less risk averse tend to be more overconfident, and the coefficient of the interaction term $Female \times RiskAverse$ indicates this effect is mainly driven by men. Third, students with lower aptitudes (larger values of *AptitudeRank*) exhibit more overconfidence.[33] There are no consistent treatment effects (except the 10%-level significance of *PreExam-BOS* and the 5%-level significance of *PostScore-BOS* on *OverconfidenceIN*).

Recall at the end of the experiment, each subject is asked to estimate the levels of over-

[32]Since overconfidence directly skews the sorting in preference submission, not only the heterogeneity but also an overall tendency in self-evaluation biases will lead to distortions from aptitude-stability. On the other hand, the exam's measurement error in terms of rank has zero mean by construction.

[33]The correlation between aptitude and overconfidence may be partially driven by a ceiling effect: the student with the highest aptitude rank cannot have a positive level of overconfidence. See a similar remark on the field result in Footnote 45.

placement θ_j^{EA} and θ_j^{IN} for a randomly drawn other subject j; I refer to the estimates for θ_j^{EA} and θ_j^{IN} as "*GuessedOtherEA*" and "*GuessedOtherIN*". Result 4 suggests that subjects are not aware of the general tendency of overplacement.

Result 4. *At both ex-ante and interim stages, subjects underestimate other students' average level of overconfidence.*

Without any significant treatment effect, the average level of *GuessedOtherEA* is 0.03 and that of *GuessedOtherIN* is -0.21. Comparing to the mean of θ_i^{EA} (0.50) and that of θ_i^{IN} (0.26), we conclude that on average, subjects underestimate others' overconfidence level at both ex-ante and interim stages ($p < 0.001$, t tests). Such a result provides evidence for Assumption 4′, which could be explained by the unawareness of one's own bias, together with the false-consensus effect.

Table 3: Predicting Factors of Overconfidence (OLS)

Dep. Var.	*OverconfidenceEA*		*OverconfidenceIN*	
Female	-1.008**	(0.430)	-0.974*	(0.500)
RiskAverse	-0.074***	(0.020)	-0.063***	(0.021)
Female×*RiskAverse*	0.067**	(0.033)	0.070*	(0.037)
AptitudeRank	0.603***	(0.044)	0.525***	(0.046)
PreExam-BOS	0.053	(0.167)	0.321*	(0.183)
Halfway-BOS	-0.114	(0.167)	0.154	(0.188)
PostScore-BOS	0.053	(0.175)	0.388**	(0.185)
Halfway-SD	0.022	(0.182)	0.067	(0.183)
PostScore-SD	-0.133	(0.136)	0.133	(0.140)
Constant	-0.330	(0.316)	-0.714**	(0.326)
Observations	315		315	

Notes: Robust standard errors are shown in parentheses, allowing for clustering by subject. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. In the regression, *Female* is a dummy variable that equals 1 for a female subject and 0 otherwise; *RiskAverse* is the total number of safe choices made by a subject during risk attitude elicitation; *Female*×*RiskAverse* is the interaction term between *Female* and *RiskAverse*; *PreExam-BOS*, *Halfway-BOS*, *PostScore-BOS*, *Halfway-SD*, and *PostScore-SD* are dummy variables for the corresponding treatments. The descriptive statistics of key variables are summarized in Table 11 of Appendix C.

3.3.3 Overconfidence and Preference Submission

Recall in Section 2.4, I discussed three strategies under different treatments: truth-telling under PreExam-, Halfway-, and PostScore-SD; score-based sorting under PostScore-BOS; and

guess-based sorting under PreExam- and Halfway-BOS. From the results below, we can see all three strategies are common in the experimental data.

Result 5. *Regardless of the timing of preference submission, more than 80% of the students report their preferences truthfully under SD, while more than 80% of the students misrepresent their preferences under BOS.*

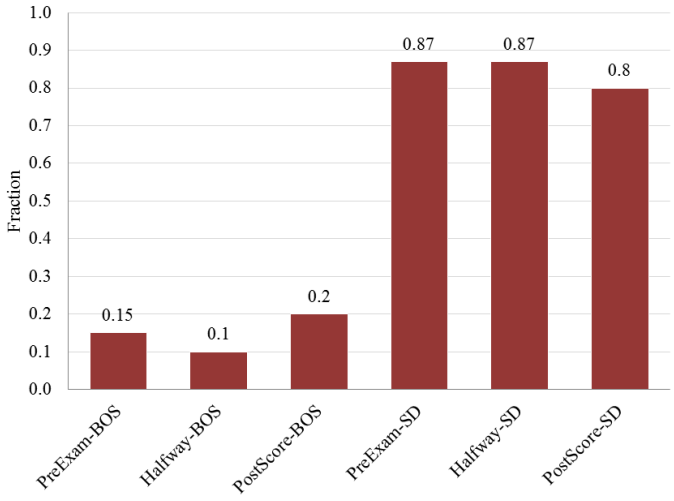


Figure 5: Truth-telling in Preference Submission

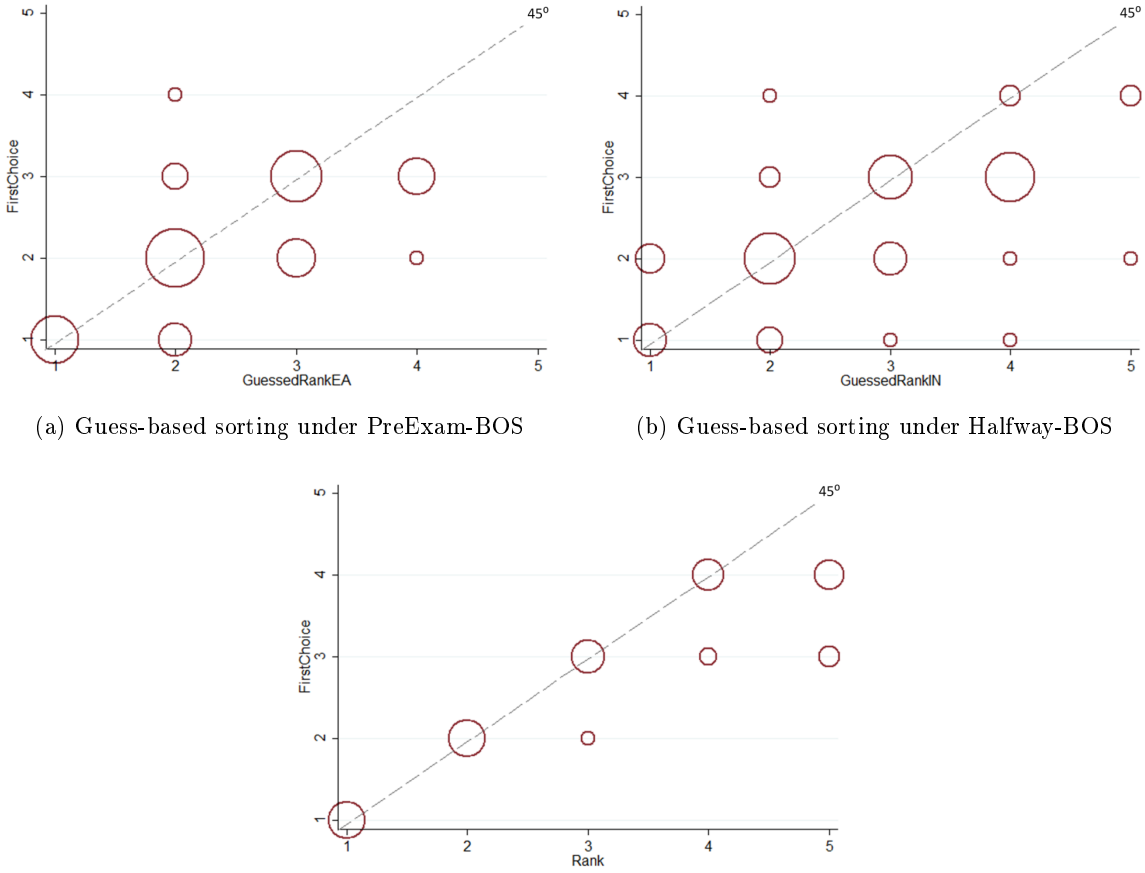
Figure 5 summarizes the proportions of truth-telling subjects under different mechanisms. As predicted by the model, truthful revelation dominates under SD, while preference misrepresentation dominates under BOS. The following result describes the general patterns of misrepresentation under BOS.

Result 6. *(i) Under PostScore-BOS, about 70% of the students exhibit score-based sorting. (ii) Under PreExam-BOS and Halfway-BOS, students tend to exhibit guess-based sorting or adopt slightly more aggressive strategies than guess-based sorting.*

To identify score-based or guess-based sorting under BOS, I focus on the variable *FirstChoice*, which is given by the index of the college listed on top of one's submitted preferences. For example, $FirstChoice = 3$ if a subject chooses college c_3 as her first choice. Moreover, recall the variables r_{s_i} ("*Rank*"), $\hat{r}_{s_i}^{EA}$ ("*GuessedRankEA*"), and $\hat{r}_{s_i}^{IN}$ ("*GuessedRankIN*") are defined as one's realized rank, guessed rank at the ex-ante stage, and guessed rank at the interim stage, respectively.

Under PostScore-BOS, a subject is said to exhibit score-based sorting in the experiment if $FirstChoice = r_{s_i}$, because the index of her score-stably matched college equals r_{s_i}. Figure 6c is a bubble chart that shows the relationship between *FirstChoice* and r_{s_i} under PostScore-BOS; the size of each bubble is determined by frequency. We can see that a majority of the

data is on the 45-degree line, meaning most subjects (71.67%) exhibit score-based sorting. The bubbles under the 45-degree line represents those who adopt a more aggressive strategy since $FirstChoice < r_{s_i}$, that is, the college of one's first choice is more desirable than her score-stable match. Most students with such a strategy are ranked 5th in the exam.[34]



(a) Guess-based sorting under PreExam-BOS

(b) Guess-based sorting under Halfway-BOS

(c) Score-based sorting under PostScore-BOS

Figure 6: Preference Submission under BOS

Under PreExam-BOS, a student is said to exhibit guess-based sorting if $FirstChoice = \hat{r}_{s_i}^{EA}$. 61.67% of the subjects use this strategy (see bubbles on the 45-degree line of Figure 6a). We also observe a considerable mass (30%) on $FirstChoice = \hat{r}_{s_i}^{EA} - 1$, indicating a slightly more aggressive strategy than guess-based sorting. Similar patterns are observed under Halfway-BOS (Figure 6b). While 46.67% of the students adopt guess-based sorting with $FirstChoice = \hat{r}_{s_i}^{IN}$, 35% of them exhibit $FirstChoice = \hat{r}_{s_i}^{IN} - 1$.

The aforementioned aggressive strategic choices suggest that subjects tend to be overop-

[34]This confirms the theoretical prediction that under the current setting, students ranked 5th in the exam are indifferent among all strategies in equilibrium (see the proof of Proposition 2 in Appendix A).

timistic about the extent of competition in the market. In an environment with more uncertainty, like PreExam- or Halfway-BOS, they appear to be gambling on the chance that no others will choose more desirable colleges as their first choice, leaving them the opportunity to get in. From the subsequent regression analysis (Table 4), we will be able to examine whether such behaviors are related to one's beliefs about other students' over- or under-confidence levels.

Result 7. *(i) Under PreExam-BOS and Halfway-BOS, a subject's first choice in preference submission is predicted by her aptitude rank and overconfidence level. (ii) Under PostScore-BOS, a subject's first choice is predicted by her aptitude rank and the exam's measurement error.*

Table 4 displays the results for ordered logit regressions of *FirstChoice* under (1) PreExam-BOS, (2) Halfway-BOS, and (3) PostScore-BOS.[35] Regression (3) shows significant effects of both *AptitudeRank* and *ExamError* on *FirstChoice* under PostScore-BOS. Since by definition, an exam outcome is composed of one's aptitude together with a measurement error, such a result echoes the fact that a majority of subjects exhibit score-based sorting (Result 6).

Now I focus on treatments PreExam-BOS and Halfway-BOS. First, by regressions (1) and (2), a subject with a higher level of overconfidence (an increase in *OverconfidenceEA* or *OverconfidenceIN*) or a better rank of aptitude (a decrease in *AptitudeRank*) tends to choose a more desirable college as the first choice (a decrease in *FirstChoice*). Significant marginal effects for each outcome are displayed in Table 12 of Appendix C. For example, under PreExam-BOS, a subject is 30.1% more likely to choose the best college c_1 as her first choice if her overplacement is increased by one rank; she is 26.9% less likely to choose c_1 if her aptitude is placed one rank worse in the market. Hence, not only aptitudes, but also overconfidence levels enter students' strategic choices, thus influencing the performance of these two mechanisms.

Second, by regressions (1) and (2), *GuessedOtherEA*, *GuessedOtherIN*, and *RiskAverse* do not have any significant influence on *FirstChoice*. Recall that under PreExam- and Halfway-BOS, a considerable number of subjects adopt more aggressive strategies than guess-based sorting. Apparently, such behaviors are not correlated with risk attitudes and cannot be rationalized by beliefs on others' overconfidence levels. Therefore, optimism is displayed on two levels: not only are subjects overconfident in guessing the exam outcomes, but also they tend to "shoot for the stars" in preference submission.

[35] In the regressions, I exclude all variables that a subject does not observe at the time of preference submission. For example, $OverconfidenceIN$ is excluded from regression (1), because preferences are submitted before the guess at the interim stage. The conclusions remain unchanged if these variables are included.

Table 4: First Choice in Preference Submission (Ordered Logit)

Dep. Var.	*FirstChoice*		
	(1)	(2)	(3)
	PreExam-BOS	Halfway-BOS	PostScore-BOS
OverconfidenceEA	-2.687***	-1.320***	-0.507
	(0.523)	(0.448)	(0.516)
OverconfidenceIN		-0.984**	-0.704
		(0.379)	(0.494)
AptitudeRank	2.391***	2.214***	4.691***
	(0.482)	(0.429)	(0.958)
ExamError			-4.307***
			(0.848)
GuessedOtherEA	0.246	0.120	0.030
	(0.215)	(0.207)	(0.277)
GuessedOtherIN		-0.131	0.595*
		(0.240)	(0.360)
RiskAverse	0.054	-0.049	0.043
	(0.070)	(0.065)	(0.081)
Observations	60	60	60

Notes: Standard errors are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

3.3.4 Overconfidence and Individual Welfare

To measure how much a student's welfare is distorted from the aptitude-stable matching, I define the variable *WelfareDistortion* as *AptitudeRank*—which equals the index of one's aptitude-stably matched college—minus the index of one's currently matched college. A positive value of *WelfareDistortion* indicates that a student is allocated to a college with a smaller index than her aptitude-stable match, which means the current mechanism is giving her an "unfair" advantage that cannot be justified by her aptitude.

The distributions of *WelfareDistortion* under different treatments are illustrated in Figure 7. We can clearly see a smaller mass at 0 under PreExam- and Halfway-BOS compared to Halfway- and PostScore-SD. Such a conclusion has already been drawn in Result 2, stating that PreExam- and Halfway-BOS tend to yield a smaller proportion of aptitude-stably matched pairs. In addition, the distribution under PostScore-SD exhibits a smaller variance compared to those under PreExam-BOS ($p = 0.006$, variance ratio test) and Halfway-BOS ($p = 0.040$, variance ratio test).

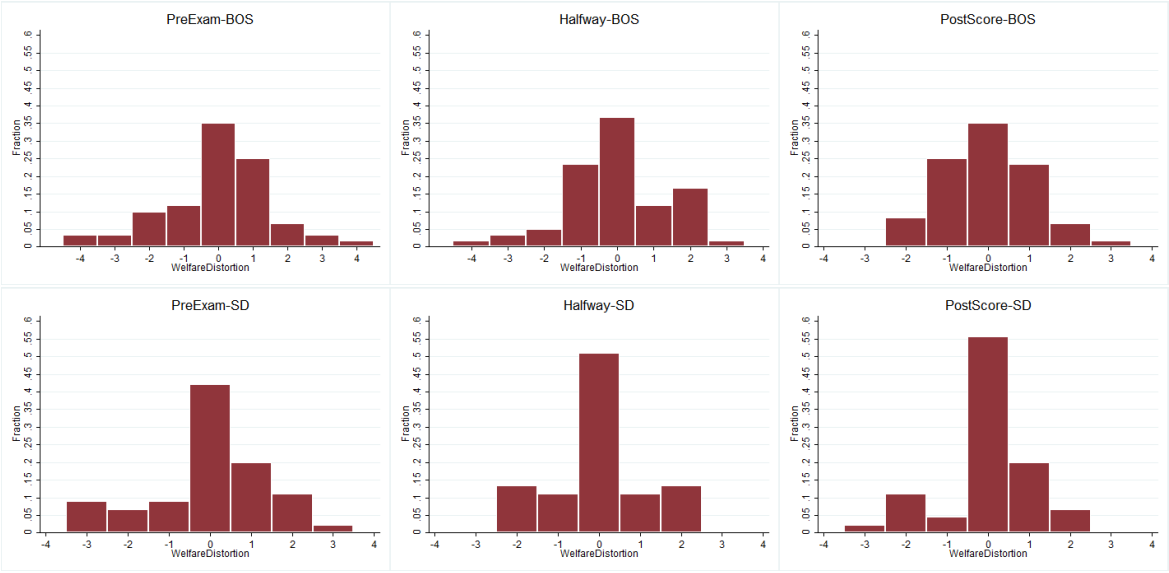


Figure 7: Distributions of Individual Welfare Distortion from Aptitude-Stability

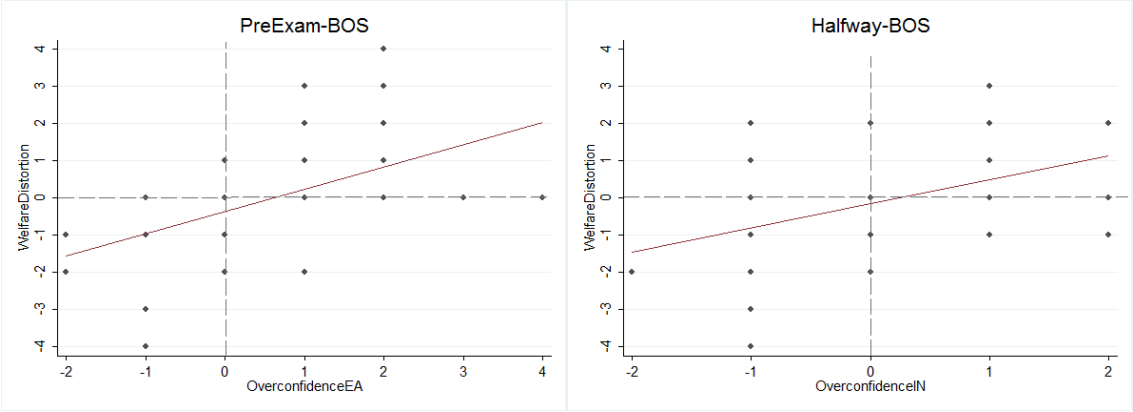
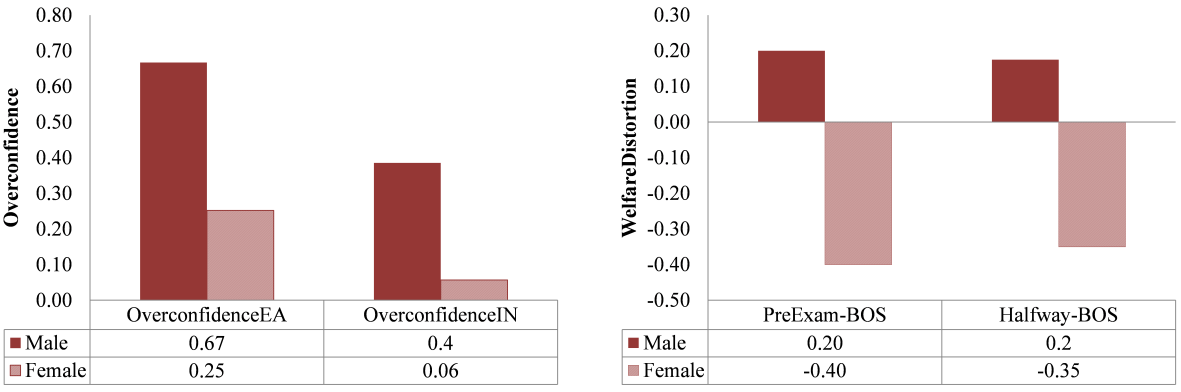


Figure 8: Overconfidence and Individual Welfare Distortion

As shown in Section 2.4, being overconfident under PreExam- or Halfway-BOS could hurt or benefit a student's welfare, depending on the distribution of overconfidence and the realization of the exam's measurement error in the market. Figure 8 helps us to take a first look at the relationship between overconfidence and individual welfare in the experimental data. Both graphs exhibit a generally positive correlation, which means PreExam- and Halfway-BOS tend to reward those who are overconfident and punish those who are underconfident. Combined with the the fact that men are more overconfident than women (Result 3), we can conclude males tend to receive an unfair advantage under these two mechanisms.[36] Figure

[36]This conclusion also uses the fact that gender does not impose a direct effect on *WelfareDistortion* (see regressions (1) and (2) in Table 5).

9 clearly shows that the gender difference in overconfidence (Figure 9a) is translated into a gender penalty for women in terms of individual welfare (Figure 9b) under PreExam- and Halfway-BOS.



(a) Average Overconfidence by Gender

(b) Average *WelfareDistortion* by Gender

Figure 9: Gender Penalty under PreExam-BOS and Halfway-BOS

To obtain specific marginal effects, I run an OLS regression of *WelfareDistortion* under (1) PreExam-BOS, (2) Halfway-BOS, and (3) PostScore-BOS.[37] The main results are summarized as follows.

Result 8. *On an individual level, PreExam-BOS and Halfway-BOS create more severe and more varied distortions from aptitude-stability than PostScore-SD. Such distortions are affected by one's overconfidence level and strategic choice, as well as the exam's measurement error. Specifically, a student tends to be matched to a better college if*
(i) she exhibits a higher level of overconfidence,
(ii) she performs better in the exam, or
(iii) she adopts a more aggressive strategy in preference submission.

Regression (3) in Table 13 shows that *ExamError* significantly and positively affects *WelfareDistortion*. This means under PostScore-BOS, a student's performance on one exam has a direct influence on her welfare, which is not surprising given the strong evidence for score-based sorting (Result 6).

Recall in Section 2.4, I use Example 1 to illustrate that under PreExam- and Halfway-BOS, the presence of overconfidence can cause welfare distortions both directly (it skews the sorting in preference submission) and indirectly by bringing back the noise of the exam's measurement error (it creates conflicts in submitted preferences, thus forcing BOS to resolve

[37] Tables 13 and 14 in Appendix C present the ordered logit regressions of *WelfareDistortion* and the significant marginal effects.

them using exam scores). Such an intuition is well supported by regressions (1) and (2), where both *OverconfidenceEA* (or *OverconfidenceIN*) and *ExamError* impose a significant influence on *WelfareDistortion* under PreExam-BOS (or Halfway-BOS). On average, if a student's level of overplacement is increased by one, her *WelfareDistortion* rises by 0.493 under PreExam-BOS and by 0.789 under Halfway-BOS; if a student's score rank in the exam is increased by one, her *WelfareDistortion* increases by 0.379 under PreExam-BOS and by 0.288 under Halfway-BOS.

Table 5: Individual Welfare Distortion (OLS)

Dep. Var.	*WelfareDistortion*		
	(1) PreExam-BOS	(2) Halfway-BOS	(3) PostScore-BOS
OverconfidenceEA	0.493***	-0.217	-0.024
	(0.143)	(0.252)	(0.076)
OverconfidenceIN		0.789***	0.086
		(0.250)	(0.084)
ExamError	0.379**	0.288**	0.943***
	(0.156)	(0.123)	(0.074)
AggressiveStrategy	0.689***	0.423*	-0.025
	(0.259)	(0.224)	(0.131)
GuessedOtherEA	0.172	0.166	-0.002
	(0.105)	(0.120)	(0.046)
GuessedOtherIN		-0.089	0.043
		(0.142)	(0.059)
RiskAverse	0.020	0.057	0.024
	(0.040)	(0.039)	(0.015)
Female	-0.009	-0.304	0.025
	(0.371)	(0.330)	(0.144)
Constant	-0.772	-0.834*	-0.288
	(0.523)	(0.493)	(0.197)
Observations	60	60	60

Notes: Standard errors are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Besides biases in beliefs, we also observe a considerable proportion of students adopt strategies other than guess-based sorting under PreExam- and Halfway-BOS.[38] To measure the extent of such a deviation, I define the variable *AggressiveStrategy* as *GuessedRankEA* (or *GuessedRankIN*) minus *FirstChoice* under PreExam-BOS (or Halfway-BOS). A positive value

[38]Figure 18 in Appendix C shows the proportions of students with (i) unbiased beliefs and guess-based sorting strategies; (ii) biased beliefs and guess-based sorting strategies; (iii) unbiased beliefs and non-guess-based sorting strategies; and (iv) biased beliefs and non-guess-based sorting strategies.

of *AggressiveStrategy* indicates a strategy more aggressive than guess-based sorting, because the college of one's first choice is more desirable than her score-stable match. According to regressions (1) and (2), a more aggressive strategy tends to have a significant and positive effect on individual welfare. On average, a unit increase in *AggressiveStrategy* raises *WelfareDistortion* by 0.689 under PreExam-BOS and by 0.423 under Halfway-BOS. Under PostScore-BOS, *AggressiveStrategy* is defined as *Rank* minus *FirstChoice* and measures the extent of deviation from score-based sorting. Since a majority of subjects adopt the equilibrium strategy under PostScore-BOS, *AggressiveStrategy* does not exhibit a significant effect in regression (3).

3.4 No Effects of Additional Information

The above results clearly suggest that under PreExam- and Halfway-BOS, overconfidence serves as a major obstacle to the implementation of an aptitude-stable matching. In this section, I explore whether an improved information condition could help to reduce overconfidence and thus enhance the performance of these two mechanisms. While a detailed description is given in Appendix B, below I briefly introduce the design and summarize the main results and implications.

In the new environment, after being re-grouped at the beginning of each treatment, every subject is provided with all the past performances of her new group members, including their scores in the three mock tests and in all the exams they have taken in the previous treatments. The average score of each member is calculated and displayed as well.[39]

The data show that the provision of such additional information has very little influence over subjects' levels of overconfidence, their strategic behaviors, and the market outcomes. This is a rather negative result since it indicates that almost all of the biases observed before stem from one's belief about herself and thus cannot be reduced even with very detailed information regarding the rest of the market. Therefore, it posts a even bigger challenge to PreExam- and Halfway-BOS in some field environments. For example, in Chinese college admissions, since mock tests are mostly created and organized by different high schools, a student cannot obtain direct information on the percentile of her score in the entire market (see Section 4.2). In this case, the conclusion of this section raises the possibility that even with perfect information revelation, the performance of PreExam- and Halfway-BOS still cannot be improved because self-evaluation biases stay on the same level.

[39] Recall in each session of the original design, about 15 participants are randomly divided into groups of five and are re-grouped for every mock test and every treatment. From the results of the three mock tests, subjects should be able to obtain a relatively clear picture of their aptitude ranking. However, for each treatment they do not receive any specific information on the other four group members. Under such a setting, there exist two major sources of overconfidence: overconfidence about the group composition ("I might be grouped with less smart people in this exam") and overconfidence about one's own performance ("I can score higher in this exam"). The new information condition is essentially muting the former while keeping the latter.

4 Field Evidence from Chinese College Admissions

One of the most important features of Chinese college admissions lies in the strikingly high stakes involved in students' self-evaluation and preference submission process. Hence, the first question we should ask is whether self-evaluation biases like overconfidence continue to exist in such a setting. Secondly, I use the field data to make welfare comparisons across different mechanisms, which depend on the distribution of overconfidence and the realization of the exam's measurement error in the specific market.

In two provinces of China, where the ex-ante and the halfway timings of preference submission are adopted, I collected data regarding students' academic aptitudes, guessed and realized exam results, and demographic information. In both samples, students exhibit overall biases in self-evaluation, as well as significant heterogeneity in the magnitudes of their biases. Because students' true preferences are difficult to elicit in the field setting, in order to answer the question on welfare comparison I run a simulation using the field data and the strategic patterns observed in the lab. The results show that PreExam- or Halfway-BOS tend to create more severe and more varied distortions from aptitude-stability compared to PostScore-SD. Although the assumptions made by the simulation method limit its ability to give general predictions for the complex field environment, the results suggest a potential explanation for the recent reforms in China's college admissions policy: most districts are currently in transition from a mechanism that resembles PreExam-BOS into a mechanism more similar to PostScore-SD.

4.1 Chinese College Admissions

According to statistics released by the Ministry of Education of the People's Republic of China, in the year 2014, about 9,390,000 applicants competed for seats at 2,246 higher education institutions. While the admission rate for these institutions was around 74%, it fell to 39% for universities that offer a bachelor's degree, and to about 2% for the top 39 universities.

In this centralized matching procedure, every college has an identical priority ordering over students, which is fully determined by their score ranking on a single standardized exam: the college entrance exam, also known as gaokao. Each year, high school graduates take the exam held by their residential districts and submit a preference list over colleges. Each district makes its own admissions policy. Students can choose colleges outside their own districts. But because for each district, the capacity (or "quota") of each college is predetermined and announced in advance, the college admissions market in every district is an independent market.

Since its introduction in 1952, the centralized procedure has undergone frequent reforms.[40] In addition, the specific matching mechanism varies across the country, mainly in two dimensions: the matching algorithm and the timing of preference submission. There are two primary

[40] See Chen and Kesten (2013) for a more detailed description.

the matching algorithms – a sequential and a parallel algorithm – and three primary timings of preference submission – pre-exam, halfway, and post-score.

The sequential algorithm is very similar to BOS; it tries to accommodate as many students as possible into their reported first choices. A parallel algorithm, on the other hand, is a combination of BOS and SD. Under such an algorithm, students' preference lists are composed of three or four tiers. While SD is applied within each tier, BOS is applied between tiers. Chen and Kesten (2013) show that although a parallel algorithm is still not fully strategy-proof, it is more strategy-proof than a sequential algorithm.[41]

In recent years, Chinese college admissions are in the transition from a sequential mechanism with pre-exam or halfway timing to a parallel mechanism with post-score timing. By 2014, the parallel algorithm had been introduced to almost all districts in China. As for the timing of preference submission, in 2014, pre-exam was only used in Shanghai and Beijing; halfway was only used in Xinjiang; all the remaining districts used post-score.[42] Therefore, I collected data from Shanghai and Xinjiang to investigate the pre-exam and halfway timings, respectively.

4.2 Data

Shanghai was under the pre-exam timing in 2014. The data collection mainly took place in May at Shanghai Pengpu High School. Students submitted their preferences about three weeks before the college entrance exam. In the meantime, their guessed exam results were elicited for the purpose of this study.[43] To evaluate students' academic aptitudes, scores from seven mock tests were also collected. Follow-up data on exam results were obtained after the revelation in June. The sample size is 95, including 40 male and 55 female students.

In Xinjiang, the data were collected in June at No. 6 High School of Kuerle City. The procedure resembles that in Shanghai: students' guessed exam results were elicited at the time of their preference submission. However, under the halfway timing, it took place after the college entrance exam but before the revelation of exam results. Three mock test scores and the exam results were also obtained.[44] The sample size is 119, including 54 male and 65 female students.

In the field environment, mock tests are mostly created and organized by different high

[41] Chen and Kesten (2013) characterize a parallel mechanism as a combination between BOS and Deferred Acceptance (DA). In the context of Chinese college admissions with a unique priority ordering of students, SD is equivalent to DA. Here I choose to use the SD specification since it is more similar to the official description of a parallel mechanism.

[42] In 2015, Xinjiang and Beijing also changed to the post-score timing.

[43] The official preference submission procedure was conducted online. However, students were also asked to submit written copies to the school in order to get feedback and advice from their teachers. The forms used to elicit their guessed exam results were distributed together with the empty preference lists for them to fill out.

[44] Since mock tests are organized by different high schools, the number of mock tests varies across schools and districts.

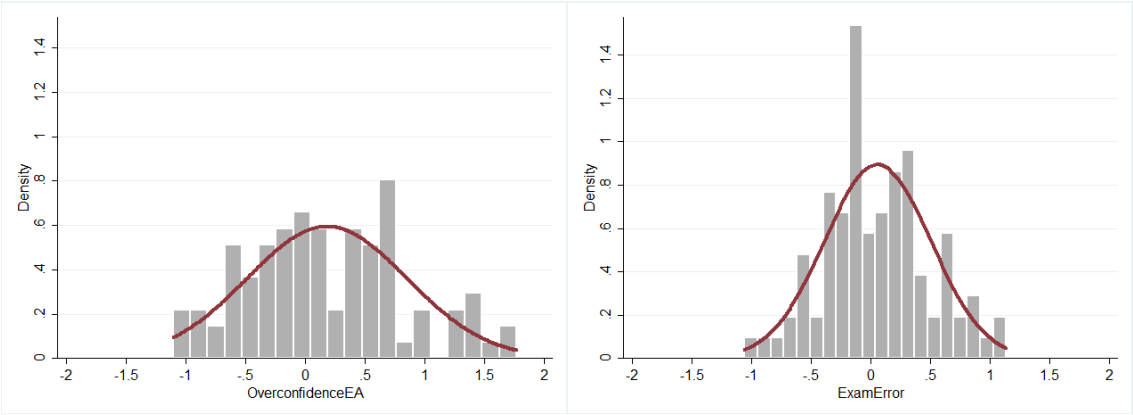
schools. Thus, students cannot obtain direct information on the relative standing of their aptitudes in the entire market. Since the exam score distribution stays relatively stable from year to year, most students infer such information by fitting their guessed scores into the distribution from the previous year. Therefore, below I mainly discuss the results in terms of scores rather than ranks and use overestimation as the primary measure of overconfidence.

4.3 Results

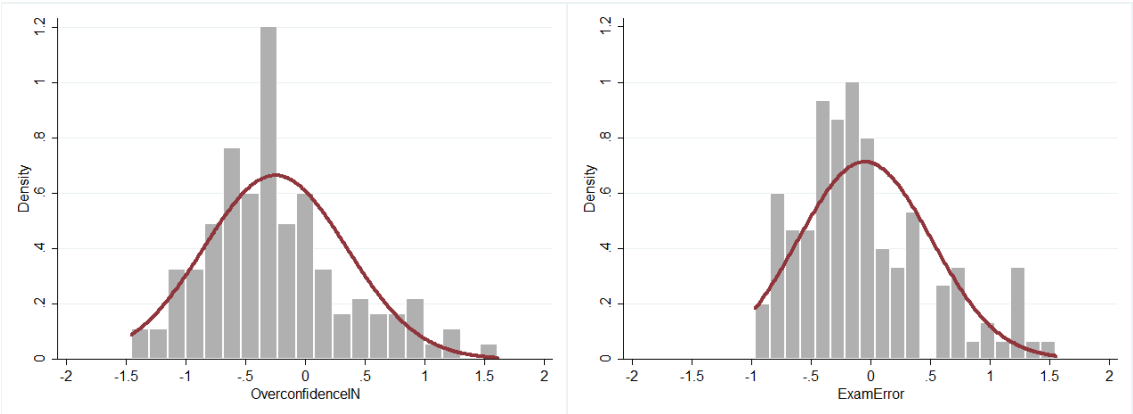
Since the difficulty and thus the score distribution can vary across mock tests and the college entrance exam, for each sample all mock test scores are normalized so that the distributions have the same mean and variance as the exam score distribution. A student's *Aptitude* is then evaluated as the average of all her mock test scores and her exam score. Moreover, the variable *ExamError* refers to the exam's measurement error in terms of score; *OverconfidenceEA* and *OverconfidenceIN* are given by a student's level of overestimation at the ex-ante stage and the interim stage. Since the total score differs in Shanghai and Xinjiang, for the convenience of comparison, I report the relative values of *Aptitude*, *ExamError*, *OverconfidenceIN*, and *OverconfidenceEA* to the standard deviation of *Aptitude*.

Result 9. *(i) Under the pre-exam timing in Shanghai, students exhibit overconfidence at the ex-ante stage; the variance of such biases is larger than the variance of the exam's measurement error. (ii) Under the halfway timing in Xinjiang, students exhibit underconfidence at the interim stage; the variance of such biases is not significantly different from the variance of the exam's measurement error.*

Under the pre-exam timing in Shanghai, students' average level of overestimation at the ex-ante stage is given by 0.185 times the standard deviation of *Aptitude*, which is significantly greater than 0 ($p = 0.004$, t test). Figure 10a compares the distribution of *OverconfidenceEA* with the distribution of *ExamError*. We can see that *OverconfidenceEA* exhibits a larger variance than *ExamError* ($p < 0.001$, variance ratio test). On the other hand, under the halfway timing in Xinjiang, the mean of *OverconfidenceIN* is -0.249, significantly lower than 0 ($p < 0.001$, t test). Moreover, *OverconfidenceIN* and *ExamError* do not exhibit significant different variances (Figure 10b).



(a) Shanghai (Pre-exam Timing)



(b) Xinjiang (Halfway Timing)

Figure 10: Levels of Overconfidence and the Exam's Measurement Error

The following two results summarize the predicting factors of a student's overconfidence level in Shanghai and Xinjiang.

Result 10. *Under the pre-exam timing in Shanghai, students are heterogeneous in overconfidence at the ex-ante stage, which can be partially predicted by:*
(i) gender: female students are more overconfident than male students;
(ii) aptitude level: those who have lower aptitudes exhibit more overconfidence.

Result 11. *Under the halfway timing in Xinjiang, students are heterogeneous in overconfidence at the interim stage, which can be partially predicted by:*
(i) gender: male students are more overconfident than female students;
(ii) aptitude: those who have lower aptitudes exhibit more overconfidence;
(iii) ethnic group: students of the Uyghur ethnic group exhibit more overconfidence.

I run an OLS regression of *OverconfidenceEA* using the Shanghai sample and a regression of *OverconfidenceIN* using the Xinjiang sample. The results are displayed in Table 6. In

Shanghai, the average overestimation for female students is higher than that for male students by 0.289 at the mean of *Aptitude*; the interaction term *Female*×*Aptitude* indicates such a difference is smaller for those who have higher aptitudes. Moreover, students with lower aptitudes exhibit more overconfidence: when one's aptitude is increased by 1, a male student's overestimation decreases by 0.241 on average, while a female student's overestimation decreases by 0.514 on average.

Table 6: Overconfidence in Shanghai and Xinjiang (OLS)

Dep. Var.	*OverconfidenceEA* δ_i^{EA} (Shanghai)		*OverconfidenceIN* δ_i^{IN} (Xinjiang)	
Female	3.521***	(1.272)	-1.550***	(0.518)
Uyghur			2.238***	(0.519)
Aptitude	-0.241***	(0.077)	-0.249***	(0.063)
Female×*Aptitude*	-0.273**	(0.107)	0.235***	(0.085)
Uyghur×*Aptitude*			-0.294***	(0.086)
Constant	2.872***	(0.927)	1.034***	(0.374)
Observations	95		119	

Notes: Standard errors are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

In Xinjiang, the marginal effect of *Female* at the mean of *Aptitude* is -0.151; such a negative effect is less significant as *Aptitude* increases. Second, when one's aptitude is improved by 1, a male student's overestimation decreases by 0.249, while a female student's overestimation decreases by 0.014 on average.[45] Third, since Xinjiang Region has a significant population of the Uyghur minority ethnic group, about 50% of students in the sample took the exam in Uyghur language. The result shows these student exhibit a higher average overestimation by 0.483 at the mean of *Aptitude,* compared to the other students, mostly of the Han majority.

From Results 9 to 11, we can clearly see that distributions and predicting factors of students' self-evaluation biases could vary dramatically across different markets. We even observe overall underconfidence in Xinjiang and women being more overconfident than men in Shanghai, which contradicts most findings in the literature of overconfidence. Therefore, in studying a specific market, we may need to tailor our choice of matching mechanism to the behavioral attributes of its participating agents. Based on the data collected in Shanghai and Xinjiang,

[45]The correlation between aptitude and overconfidence in both Shanghai and Xinjiang can be partially driven by a ceiling effect: since every exam has a perfect score, a student with the highest possible aptitude cannot have a positive level of overconfidence.

a simulation is conducted to help us compare market outcomes under different mechanisms.

According to the lab experiment results, a student's strategy in preference submission under PreExam- and Halfway-BOS is mainly affected by her aptitude and overconfidence level (Result 7). Therefore, I first run a multinomial logit regression of several strategic patterns observed in the lab such as guess-based sorting, aggressive guess-based sorting etc. Then, using the field data on aptitude and overconfidence, I predict every student's probability of playing each strategy. Finally, a simulation gives us the matching outcome of an artificially built market, with a structure similar to the ones in the lab environment.[46]

Although such a simulation method has its limitations in predicting for the complex field environment, it still provides us with important intuitions regarding whether the observed biases could generate similar welfare distortions as in the lab. The results are summarized below. Here I mainly compare PreExam- or Halfway-BOS to PostScore-SD, which as suggested by the lab data, ensures score-stability.

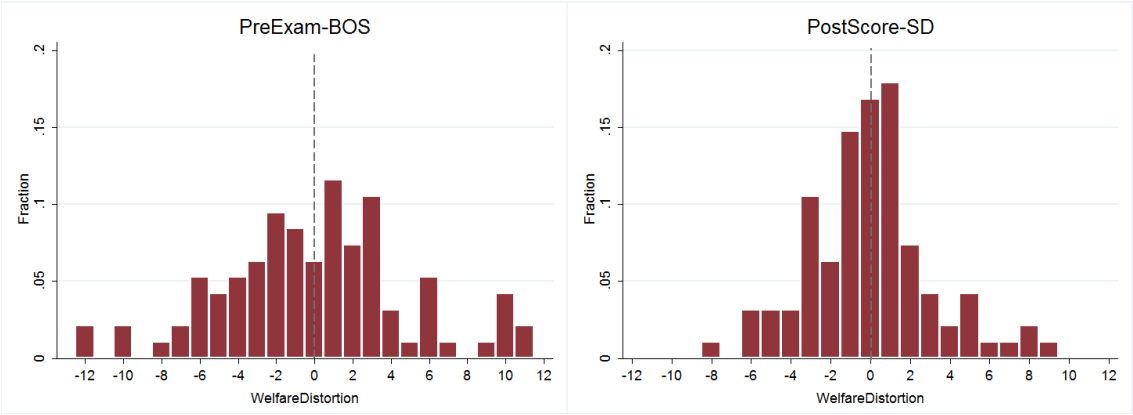
Result 12. *(1) On an individual level, PreExam-BOS in the simulated Shanghai market (or Halfway-BOS in the simulated Xinjiang market) creates more severe and more varied distortions from aptitude-stability than PostScore-SD. (2) A student tends to be matched to a better college if she exhibits a higher level of overconfidence.*

The distributions of *WelfareDistortion* are illustrated in Figure 11. Similar to the patterns observed in the lab, in both simulated markets, PostScore-SD yields a larger mass at 0 than PreExam-BOS ($p = 0.012$, McNemar's test) or Halfway-BOS ($p = 0.003$, McNemar's test). Moreover, the distribution under PostScore-SD also has a smaller variance compared to PreExam- and Halfway-BOS ($p < 0.001$, variance ratio tests).

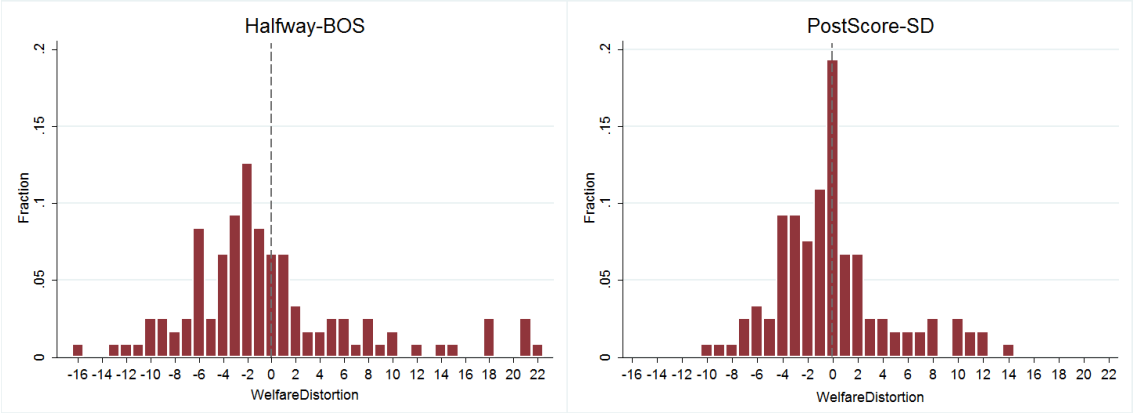
Like the lab experiment result, each simulated market exhibits a positive relationship between overconfidence and individual welfare; that is, PreExam-BOS and Halfway-BOS tend to reward those who are overconfident and punish those who are underconfident (Figure 12). Combined with the predicting factors of overconfidence in Results 10 and 11, we can conclude that PreExam-BOS tends to give an unfair advantage to females or students with lower aptitudes in Shanghai, and that Halfway-BOS gives an unfair advantage to males, students with lower aptitudes, or students of the Uyghur ethnic group in Xinjiang.

The same conclusion can be drawn from Table 7, which shows the OLS regression results of *WelfareDistortion* for (1) the Shanghai market under PreExam-BOS and (2) the Xinjiang market under Halfway-BOS. In particular, if a student's level of overestimation is increased by one standard deviation of *Aptitude*, her value of *WelfareDistortion* increases by 4.186 in Shanghai and by 9.425 in Xinjiang on average.

[46] Since each sample includes about 100 students, in each simulated market, I build about 20 colleges; every college has a capacity of 5.

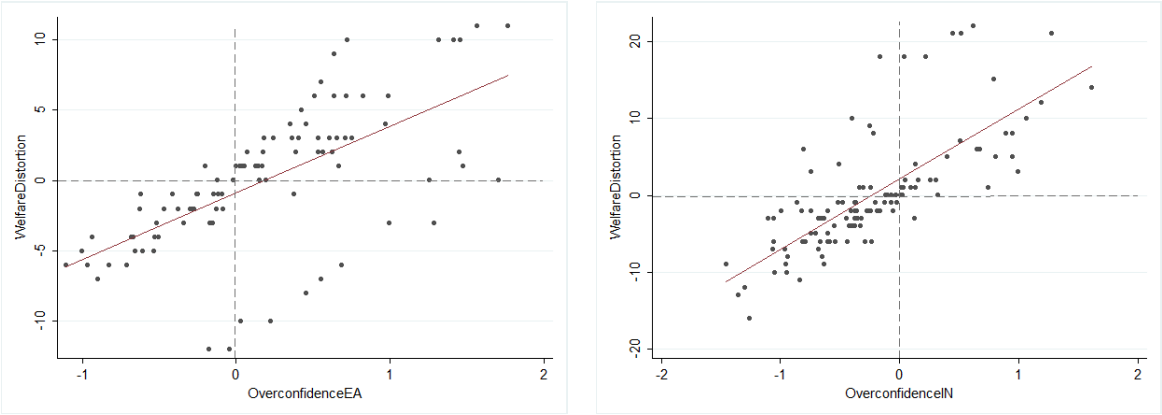


(a) Simulated Shanghai Market



(b) Simulated Xinjiang Market

Figure 11: Individual Welfare Distortion from Aptitude-Stability in Simulated Markets



(a) Shanghai (PreExam-BOS)

(b) Xinjiang (Halfway-BOS)

Figure 12: Overconfidence and Individual Welfare Distortion in Simulated Markets

Table 7: Individual Welfare Distortion in Simulated Markets (OLS)

Dep. Var.	*WelfareDistortion*			
	(1)		(2)	
	PreExam-BOS (Shanghai)		Halfway-BOS (Xinjiang)	
OverconfidenceEA	4.186***	(0.570)		
OverconfidenceIN			9.425***	(0.926)
ExamError	3.554***	(0.833)	-1.791*	(1.072)
Female	-0.822	(0.738)	-0.093	(1.270)
Uyghur			1.475	(1.590)
Female×*Uyghur*			0.859	(1.901)
Constant	-0.510	(0.545)	1.124	(0.998)
Observations	95		119	

Notes: Standard errors are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

To sum up, in both samples collected from Chinese college admissions, students exhibit overall biases in self-evaluation, as well as significant heterogeneity in the magnitudes of their biases. A simulation using the field data and the strategic patterns observed in the lab shows that PreExam- or Halfway-BOS tend to create more severe and more varied distortions from aptitude-stability compared to PostScore-SD. This suggests a potential explanation for the recent reforms in China's college admissions policy: most districts are currently in transition from a mechanism that resembles PreExam-BOS into a mechanism more similar to PostScore-SD.

5 Conclusion

Many centralized college admissions markets adopt a standardized exam to evaluate students' aptitudes and determine their priorities in the matching procedure. Since every exam entails a measurement error, the exam-based priorities can only serve as a noisy proxy for colleges' aptitude-based preferences. The previous literature suggests the effect of this noise can be diminished with a "PreExam-BOS" mechanism (a Boston algorithm combined with preference submission before the exam). Using a lab experiment, I conclude otherwise: (i) since pre-exam preference submission is skewed by overconfidence, PreExam-BOS creates more severe and more varied welfare distortions than the PostScore-SD mechanism (a Serial Dictator-

ship algorithm combined with preference submission after the revelation of exam results); (ii) PreExam-BOS introduces unfairness by rewarding overconfidence and punishing underconfidence, thus serving as a gender penalty for women.

In a field investigation on actual Chinese students, I find similar behavioral biases such as overconfidence. A simulation based on the field data and the strategic patterns observed in the lab also shows that PreExam-BOS is inferior to PostScore-SD in terms of welfare distortions. This suggests a potential explanation for China's recent policy reform from a mechanism that resembles PreExam-BOS to a mechanism more similar to PostScore-SD. Admittedly, the assumptions made by the simulation method limit its ability to give general predictions for the field environment. But it is not the main purpose of this paper. Instead, this study intends to introduce a behavioral perspective and to present a tradeoff that can emerge in real markets. Further efforts should be made for a more thorough field investigation.

Although preferable to PreExam-BOS, PostScore-SD is still largely affected by the exam's measurement error. This implies the challenge of obtaining a fair market outcome when students are evaluated with a single standardized exam. Thus, it raises the need for policymakers to weigh the benefits and costs when adopting such a noisy evaluation system. In practice, similar systems exist in various environments including public school choice, college admissions, as well as labor market clearinghouses. Such prevalence calls for more research in the future. In addition to the behavioral aspects considered in this paper, more issues like heterogeneous preferences, asymmetric information, and constrained choices in preference submission should be added into the discussion.

References

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

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

Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science 60*(2), 434–448.

Barber, B. M. and T. Odean (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly journal of Economics*, 261–292.

Camerer, C. and D. Lovallo (1999). Overconfidence and excess entry: An experimental approach. *The American Economic Review*, 306–318.

Chen, Y. and O. Kesten (2013). From boston to chinese parallel to deferred acceptance: theory and experiments on a family of school choice mechanisms. *Working Paper*.

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

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 261–292.

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

Featherstone, C. and M. Niederle (2008). Ex ante efficiency in school choice mechanisms: an experimental investigation. *Working Paper*.

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

Glaser, M. and M. Weber (2007). Overconfidence and trading volume. *The Geneva Risk and Insurance Review 32*(1), 1–36.

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

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

Jiang, M. (2014). When do stable matching mechanisms fail? the role of standardized tests in college admissions. *Working Paper*.

Kesten, O. (2006). On two competing mechanisms for priority-based allocation problems. *Journal of Economic Theory 127*(1), 155–171.

Kleitman, S. and L. Stankov (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences 17*(2), 161–173.

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

Lien, J. W., J. Zheng, and X. Zhong (2015). Preference submission timing in school choice matching: testing fairness and efficiency in the laboratory. *Experimental Economics*, 1–35.

Lien, J. W., J. Zheng, and X. Zhong (2017). Ex-ante fairness in the boston and serial dictatorship mechanisms under pre-exam and post-exam preference submission. *Games and Economic Behavior 101*, 98–120.

Malmendier, U. and G. Tate (2005). Ceo overconfidence and corporate investment. *The Journal of Finance 60*(6), 2661–2700.

Marks, G. and N. Miller (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin 102*(1), 72.

Moore, D. A. and P. J. Healy (2008). The trouble with overconfidence. *Psychological Review 115*(2), 502.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 1067–1101.

Odean, T. (1999). Do investors trade too much? *American Economic Review*, 1279–1298.

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

Ross, L., D. Greene, and P. House (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology 13*(3), 279–301.

Schaefer, P. S., C. C. Williams, A. S. Goodie, and W. K. Campbell (2004). Overconfidence and the big five. *Journal of Research in Personality 38*(5), 473–480.

Stankov, L. and J. D. Crawford (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences 21*(6), 971–986.

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

Appendix A: Proofs

Proof of Proposition 1

Proof. (1) In the current setting, where all colleges have the same strict priority ordering over students, the SD algorithm is a special case of the TTC algorithm. By Abdulkadiroğlu and Sönmez (2003), TTC is strategy-proof for any realized priority ordering over students. Therefore, truth-telling is a weakly dominant strategy for every student, regardless of her knowledge about the priority ordering at the time of preference submission. This proves the strategy-proofness of the PreExam-SD, Halfway-SD, and PostScore-SD mechanisms.

(2) By Kesten (2006), since the priority structure here satisfies the acyclic condition, the matching outcome of TTC is stable according to priorities. In addition, the uniqueness of such an outcome is proved by Haeringer and Klijn (2009). Translating into terms under the current setting, PostScore-SD always yield the score-stable matching outcome.

(3) From (1), we know in the truth-telling equilibrium, students' submitted preferences stay the same under PreExam-SD and Halfway-SD as those under PostScore-SD. And colleges' priority ordering depends only on students' exam score ranking. Therefore, given the fact that a matching algorithm only considers students' submitted preferences and colleges' priority ordering, the PreExam-SD and Halfway-SD also implement the score-stable matching outcome in the truth-telling equilibrium. □

Proof of Proposition 2

Proof. Define the total number of seats at colleges c_1, c_2, ..., and c_k as $Q_k = \sum_{j=1}^{k} q_j$.

First, for a student with score rank $r_{s_i} = 1, 2, ...,$ or Q_1, it is a dominant strategy to list c_1 as her first choice. This is because she will be accepted by the best college c_1 regardless of other students' submitted preferences. Any other first choice will make her strictly worse off, because she will always be accepted by her first choice.

Given this, a student with score rank $r_{s_i} = Q_1 + 1, ...,$ or Q_2 best responds by listing c_2 as the first choice. Deviating to c_1 will get her rejected in the first step and thus cannot make her better off. Any other first choice will make her strictly worse off.

Similarly, it follows that a student with score rank $r_{s_i} = Q_2 + 1, ...,$ or Q_3 best responds by listing c_3 as the first choice, a student with score rank $r_{s_i} = Q_3 + 1, ...,$ or Q_4 best responds by listing c_4 as the first choice, and so on.

Finally, consider students with lowest score ranks $r_{s_i} = Q_M + 1, ..., n$, where $Q_M < n$ and $Q_{M+1} \geq n$. Given other students' equilibrium strategies, these students are indifferent among all strategies that list c_{M+1} above c_{M+j} $(j \geq 2)$.

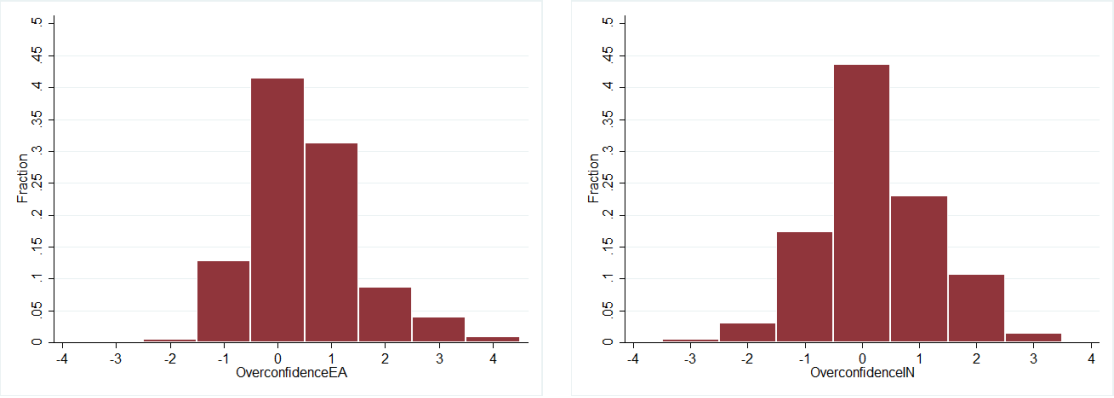
(1) From the reasoning above, in any Nash equilibrium, students with $r_{s_i} = 1, ..., Q_1$ list c_1 as the first choice; students with $r_{s_i} = Q_1 + 1, ..., Q_2$ list c_2 as the first choice; ...and

students with $r_{s_i} = Q_{M-1} + 1, ..., Q_M$ list c_M as the first choice. Any remaining choices of these students and for students with $r_{s_i} = Q_M + 1, ..., n$, any strategies that list c_{M+1} above c_{M+j} $(j \geq 2)$ can exist in a Nash equilibrium. Therefore, those Nash equilibria where students with $r_{s_i} = Q_M + 1, ..., n$ list c_{M+1} as the first choice are the ones where every student exhibit score-based sorting.

(2) Given the characterization of students' equilibrium strategies, it is easy to see that all equilibria yield the same matching, where a student with $r_{s_i} = 1, ...,$ or Q_M is assigned a seat at her first choice, and a student with $r_{s_i} = Q_M + 1, ...,$ or n is assigned a seat at c_{M+1}. Such a matching is stable according to exam-based priorities or is score-stable. □

Appendix B: Treatments with Additional Information

The provision of additional information adds a third dimension to the original treatment design. From the overwhelming truth-telling behaviors under SD regardless the timing of preference submission, we can conclude that subjects' decision-making and the market outcomes are not affected by different information conditions. Therefore, I focus on the three mechanisms using BOS; the three new treatments with additional information are named "PreExam-BOS-INFO," "Halfway-BOS-INFO," and "PostScore-BOS-INFO," respectively. All the other details of the experimental design and procedure are similar to those described in Sections 3.1 and 3.2.[47]



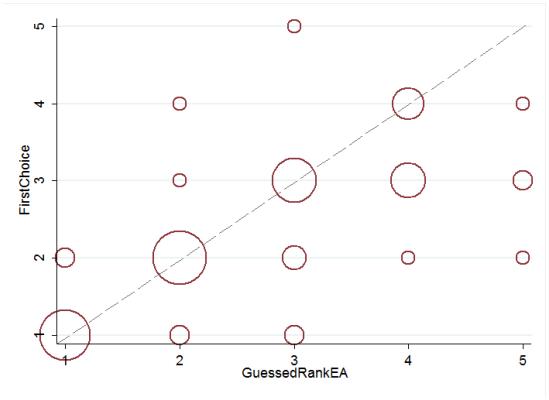
(a) Distribution of *OverconfidenceEA* θ_i^{EA} (b) Distribution of *OverconfidenceIN* θ_i^{IN}

Figure 13: Levels of Overconfidence with Additional Information

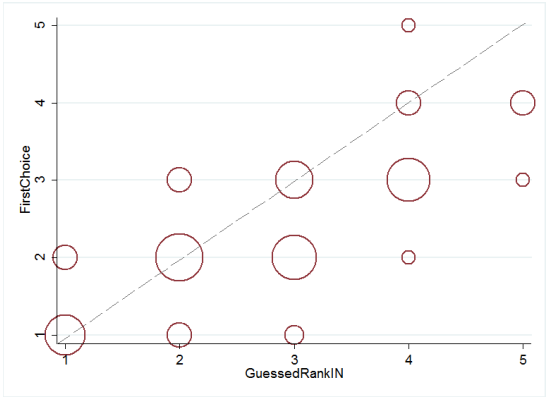
Result 13. *The provision of additional information fails to reduce the magnitude of overconfidence at either the ex-ante or interim stage.*

[47] For the new treatments, four additional sessions (one with 20 subjects, three with 15 subjects) were conducted in September 2015 at the Experimental Economics Laboratory of The Ohio State University. There were a total of 65 participants (22 females and 43 males).

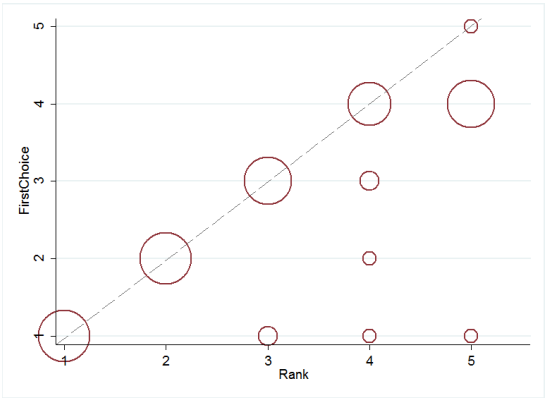
Under the new treatments, subjects' average level of overplacement is 0.51 at the ex-ante stage and is 0.24 at the interim stage.[48] Comparing to the average without additional information (0.50 at the ex-ante stage and 0.26 at the interim stage), the magnitudes of biases are clearly not reduced ($p > 0.44$).



(a) Guess-based sorting under PreExam-BOS-INFO



(b) Guess-based sorting under Halfway-BOS-INFO



(c) Score-based sorting under PostScore-BOS-INFO

Figure 14: Preference Submission under BOS

Given the above result, together with the fact that subjects' strategic choices are not significantly affected by the additional information (Figure 14), we can expect that our conclusion regarding aptitude-stability remain unchanged.

Result 14. *The provision of additional information fails to improve the performance of PreExam-BOS and Halfway-BOS: compared to PostScore-SD, they still create more severe and more varied distortions from aptitude-stability .*

On a market level, aptitude-stability is rarely achieved under any of the three new treat-

[48] Again, data from all three treatments are pooled together since there is no significant treatment effect on overconfidence. Figure 13 shows the distributions of θ_i^{EA} and θ_i^{IN}.

ments (except one market under PreExam-BOS-INFO). From Figure 15, we can see in terms of the proportion of aptitude-stably matched pairs, PreExam-BOS-INFO does not perform significantly better than PreExam-BOS ($p = 0.36$), while Halfway-BOS-INFO performs slightly worse than Halfway-BOS ($p = 0.07$).

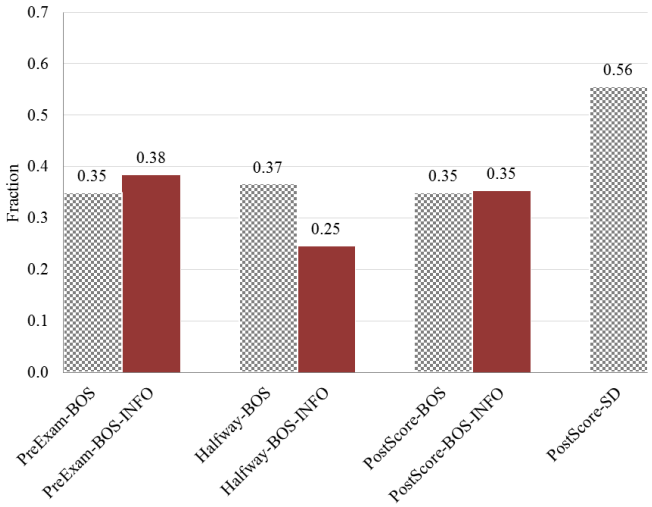


Figure 15: Aptitude-Stably Matched Pairs

Next I compare the three new mechanisms with PostScore-SD. PostScore-SD is the best-performing mechanism in Section 3.3, and as mentioned above, it is natural to assume that its performance is not affected by the change of information environment. Figure 15 shows that PostScore-SD yields a larger proportion of aptitude-stably matched pairs than PreExam-BOS-INFO ($p = 0.031$), Halfway-BOS-INFO ($p < 0.001$), and PostScore-BOS-INFO ($p = 0.014$). As for the variable *WelfareDistortion* (Figure 16), the distribution under PostScore-SD still exhibits a smaller variance compared to PreExam-BOS-INFO ($p = 0.008$) and Halfway-BOS-INFO ($p = 0.016$). Moreover, the positive relationship between overconfidence and individual welfare remain unchanged under PreExam-BOS-INFO and Halfway-BOS-INFO (see Figure 17 and Table 8).

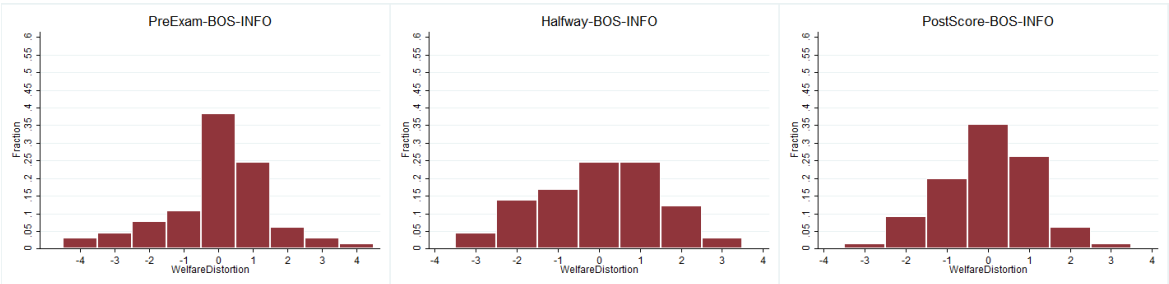


Figure 16: Distributions of Individual Welfare Distortion from Aptitude-Stability

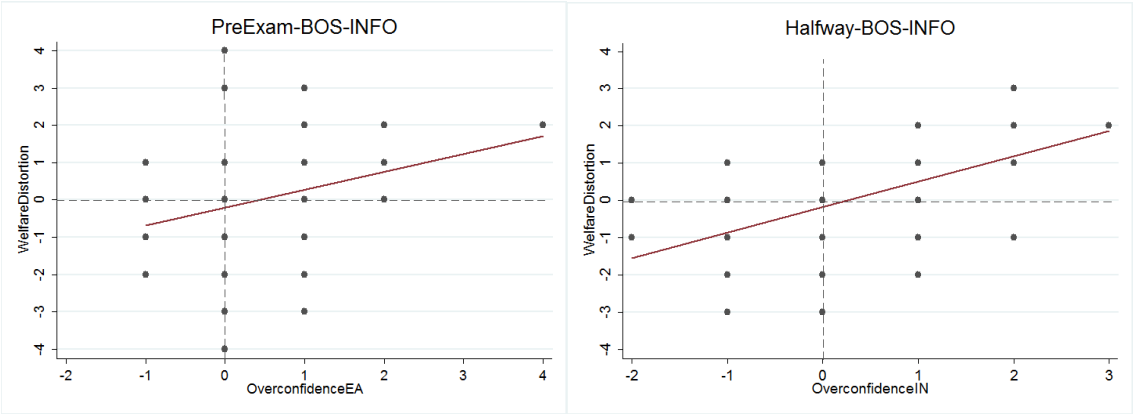


Figure 17: Overconfidence and Individual Welfare Distortion

Table 8: Individual Welfare Distortion (OLS)

Dep. Var.	*WelfareDistortion*		
	(1)	(2)	(3)
	PreExam-BOS-INFO	Halfway-BOS-INFO	PostScore-BOS-INFO
OverconfidenceEA	0.495***	0.161	-0.044
	(0.177)	(0.193)	(0.060)
OverconfidenceIN		0.553**	0.010
		(0.210)	(0.059)
ExamError	0.565***	0.631***	0.989***
	(0.150)	(0.150)	(0.038)
AggressiveStrategy	0.726***	0.255	-0.018
	(0.189)	(0.215)	(0.054)
GuessedOtherEA	0.230**	-0.119	-0.004
	(0.102)	(0.110)	(0.033)
GuessedOtherIN		0.037	-0.009
		(0.121)	(0.040)
RiskAverse	0.037	0.017	-0.002
	(0.037)	(0.035)	(0.010)
Female	0.091	0.514*	0.090
	(0.319)	(0.297)	(0.090)
Constant	-0.882*	-0.711	0.024
	(0.484)	(0.436)	(0.138)
Observations	65	65	65

Notes: Standard errors are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Appendix C: Supplementary Tables and Figures

Table 9: Standardized Exams in Representative Countries (Year: 2014)

Country	Standardized Exam	Number of Participants or Applicants	Data Source
China	National College Entrance Exam (Gaokao)	9,390,000	http://gaokao.eol.cn/
Greece	Panhellenic Exams	104,616	http://edu.klimaka.gr/
Russia	Unified State Exam	757,303	http://vestnikkavkaza.net/articles/society/57810.html
South Korea	College Scholastic Ability Test	640,619	http://www.kice.re.kr/main.do?s=suneung
Turkey	Higher Education Exam-Undergraduate Placement Exam	2,086,115	http://www.osym.gov.tr/

Notes: The statistic for Greece is from the year 2015.

Table 10: Distribution of Score Rankings (Example 1)

Score Ranking r_s	Probability
(2,1,3)	26/64
(2,3,1)	17/64
(1,2,3)	17/64
(1,3,2)	1/64
(3,2,1)	1/64
(3,1,2)	2/64

Table 11: Descriptive Statistics of Key Variables (Lab Experiment)

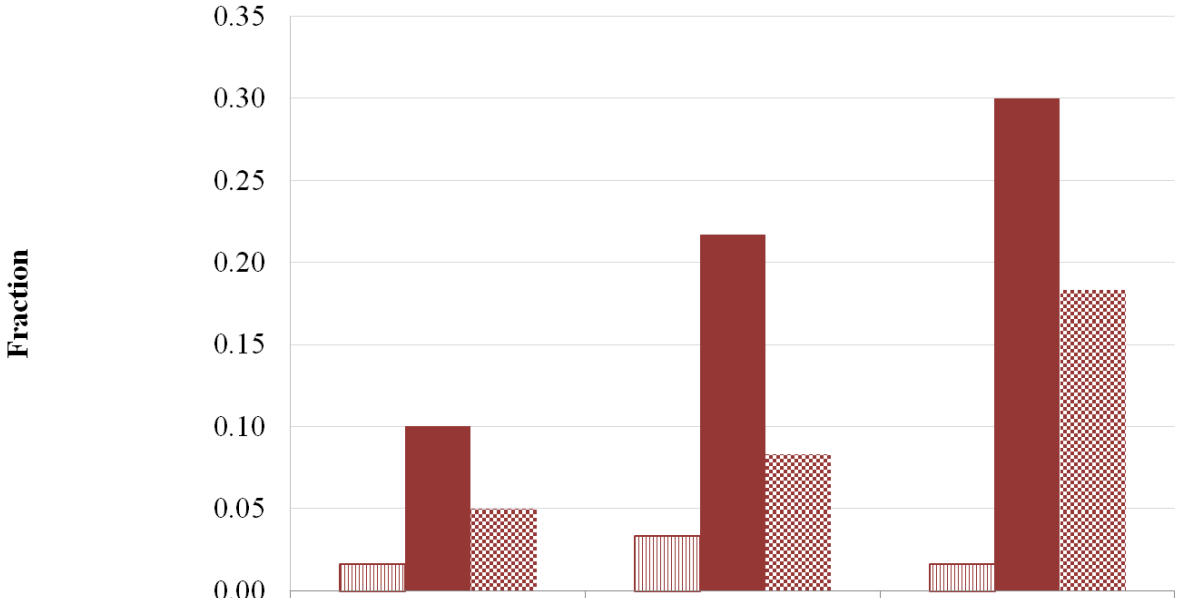
Variable	Mean	Std. Dev.	Min	Max
AptitudeRank	3	1.416	1	5
ExamError	0	1.205	-3	3
OverconfidenceEA	0.505	1.211	-3	4
OverconfidenceIN	0.257	1.192	-4	4
GuessedOtherEA	0.029	1.515	-4	3
GuessedOtherIN	-0.213	1.401	-4	3
RiskAverse	12.305	4.232	0	20
Female	0.390	0.489	0	1
Observations	315			

Table 12: Ordered Logit Marginal Effects for First Choice

PreExam-BOS	Marginal Effects for			
FirstChoice=	1	2	3	4
OverconfidenceEA	0.301***	0.190	-0.485***	-0.005
AptitudeRank	-0.269***	-0.170	0.433***	0.005

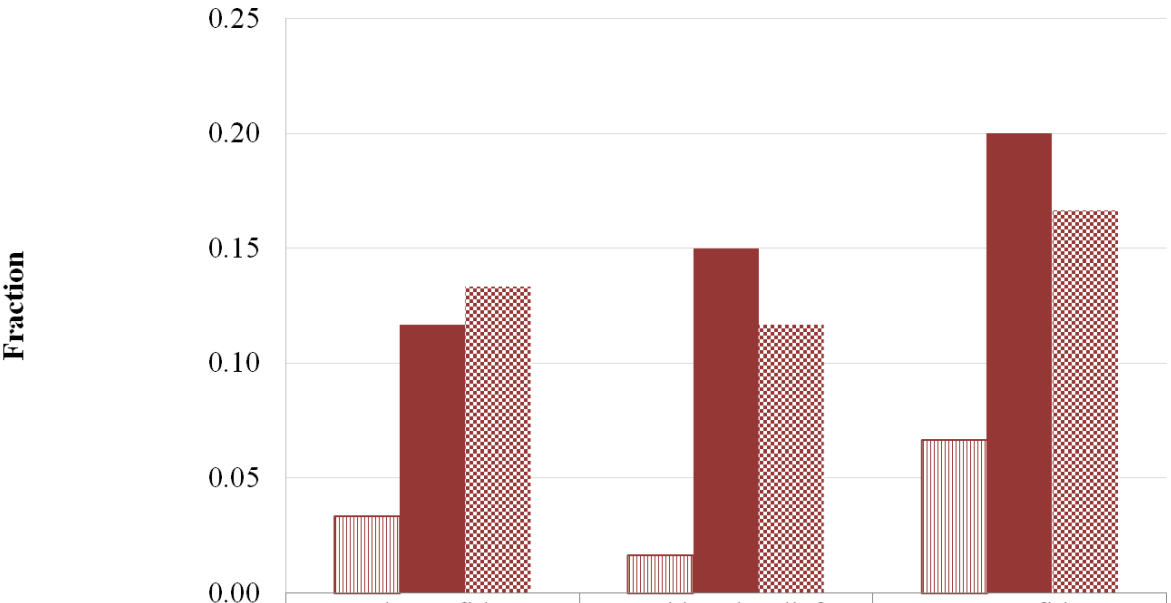
Halfway-BOS	Marginal Effects for			
FirstChoice=	1	2	3	4
OverconfidenceEA	0.079**	0.269**	-0.316***	-0.033
OverconfidenceIN	0.063**	0.215**	-0.253**	-0..025
AptitudeRank	-0.127***	-0.434***	0.510***	0.051*

PostScore-BOS	Marginal Effects for			
FirstChoice=	1	2	3	4
AptitudeRank	-0.012	-0.710***	0.656**	0.066
ExamError	0.010	0.604***	-0.558**	-0.056



	Underconfident	Unbiased Belief	Overconfident
Conservative	0.02	0.03	0.02
Guess-based Sorting	0.10	0.22	0.30
Aggressive	0.05	0.08	0.18

(a) PreExam-BOS



	Underconfident	Unbiased Belief	Overconfident
Conservative	0.03	0.02	0.07
Guess-based Sorting	0.12	0.15	0.20
Aggressive	0.13	0.12	0.17

(b) Halfway-BOS

Figure 18: Biased Beliefs and Non-guess-based Sorting Strategies

Table 13: Individual Welfare Distortion (Ordered Logit)

Dep. Var.	*WelfareDistortion*		
	(1)	(2)	(3)
	PreExam-BOS	Halfway-BOS	PostScore-BOS
OverconfidenceEA	0.733***	-0.628	-0.150
	(0.232)	(0.434)	(0.433)
OverconfidenceIN		1.632***	0.512
		(0.485)	(0.532)
ExamError	0.571**	0.547***	5.369***
	(0.238)	(0.209)	(0.925)
AggressiveStrategy	1.010***	0.851**	-0.280
	(0.389)	(0.388)	(0.776)
GuessedOtherEA	0.251	0.261	-0.079
	(0.160)	(0.205)	(0.283)
GuessedOtherIN		-0.135	0.072
		(0.249)	(0.352)
RiskAverse	0.039	0.118**	0.142**
	(0.057)	(0.067)	(0.092)
Female	0.053	-0.623	-0.045
	(0.560)	(0.562)	(0.851)
Observations	60	60	60

Notes: Standard errors are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Table 14: Ordered Logit Marginal Effects for Individual Welfare Distortion

(1) PreExam-BOS	Marginal Effects for				
WelfareDistortion=	-2	-1	0	1	2
OverconfidenceEA	-0.046**	-0.062**	-0.030	0.109***	0.032*
ExamError	-0.036*	-0.048*	-0.023	0.085**	0.025*
AggressiveStrategy	-0.064*	-0.086*	-0.041	0.150**	0.044

(2) Halfway-BOS	Marginal Effects for				
WelfareDistortion=	-2	-1	0	1	2
OverconfidenceIN	-0.043	-0.260***	0.062	0.125**	0.138**
ExamError	-0.014	-0.087**	0.021	0.042**	0.046**
AggressiveStrategy	-0.022	-0.136**	0.032	0.065*	0.072*
RiskAverse	-0.003	-0.019*	0.004	0.009	0.010*

(3) PostScore-BOS	Marginal Effects for				
WelfareDistortion=		-1	0	1	
ExamError		-0.435**	-0.115	0.551**	