Why Exclude Test Scores from Admission Decisions? *

Yucheng Liang† Wenzhuo Xu‡

[Preliminary and Incomplete]

Latest version will be updated here

February 28, 2024

Abstract

One major argument in support of test-optional and test-blind college admission policies is that standardized test scores inaccurately reflect students' abilities and are biased against those with fewer resources. This argument goes against standard economic reasoning as information, even if noisy or biased, never has negative value. In an experiment, we show that participants who are tasked with admitting students for advanced educational resources are indeed willing to exclude noisy or biased test scores from their admission criteria. This result is primarily driven by procedural fairness concerns and an underestimation of the usefulness of these scores.

*This study is approved by CMU IRB in Protocols 2023_00000333 and 2023_00000337. The RCT registry ID is AEARCTR-0012709. Kate Yixin Huang, Jack Tianrui Lin, and Allison Tribendis provided excellent research assistance. Errors are ours.

†Carnegie Mellon University. Email: ycliang@cmu.edu. Corresponding author.

‡Carnegie Mellon University. Email: wenzhuox@andrew.cmu.edu.

1 Introduction

Deciding how to use information about individuals for high-stakes allocation decisions is a critical question across various domains. From hiring, loan approval, to pre-trial release decisions, the criteria used can significantly impact individuals' lives and broader societal outcomes. These decisions often involve integrating the objective of forming accurate assessment of individuals with considerations of fairness, bias, and the potential for perpetuating existing inequalities.

College admission represents a realm where the decision on how to use available information has undergone significant change and sparked intense debate. In this domain, one of the most notable trends over the last decade is the shift from test-optional and test-blind policies being quite rare to becoming prominent forms of admission policies in the United States. Test-optional policies allow students to choose whether to submit their standardized test scores such as SAT or ACT, while test-blind institutions do not consider these scores at all. By 2024, more than 2000 US colleges have adopted a test-optional or test-blind policy (FairTest, 2024).

A central argument underpinning the shift towards test-optional and test-blind policies revolves around the adequacy and bias of standardized test scores as an input to the admission process. Critics argue that SAT and ACT do not accurately reflect a student's full range of abilities or potential for success. Moreover, because these scores may be heavily affected by factors such as access to expensive test preparation services, less privileged students, such as those from low-income households, may be disadvantaged. In a survey conducted in 2021 (Harris Poll, 2021), 42% of US participants disagree that standardized tests correctly measure a student's academic knowledge and skills, while 51% think that they are inherently biased in favor of affluent students.

While these concerns over standardized test scores are likely valid, it is still puzzling why policies that voluntarily give up information enjoy such wide adoption and public support. Despite their imperfections, SAT and ACT are unlikely to be completely uninformative. In fact, research shows that they help predict student success even after controlling for other information in college applications (UC Academic Senate, 2020). Given that colleges have the freedom to use test scores in a nuanced and context-aware manner, why would the general public prefer them to forego this

information?

In this paper, we study information choice in an experiment where participants select students for educational opportunities. We investigate whether the inadequacy and bias in test scores lead people to exclude them from the admission process and explore the rationales behind the exclusion. Our experimental paradigm has several advantages in answering the research questions. First, participants make decisions which have real impacts on students. This incentivizes truthful and thoughtful preferences by mitigating indifference and offsetting social desirability concerns. % NEEDS EDIT Second, the tightly-controlled experiment allows us to exogenously vary the adequacy and bias in test scores and other aspects of the environment. These variations help us causally identify the determinants of information preferences.

In the main experiment, each participant ("spectator") is tasked with selecting students from an introductory data science course to enroll in an advanced course. These students come from different family income backgrounds and have all completed two tests for the introductory course. No student received any additional test preparation for Test 1, but for Test 2, the experiment has 4 different test prep situations: No Prep, Inadequate Prep, Biased Prep, and Inadequate and Biased Prep. A test prep is inadequate if it boosts test scores without improving students' skills, and it is biased if only higher-income students receive the prep. Each spectator is randomly assigned to make admission decisions under one test prep situation with full knowledge of its adequacy and bias.

Spectators always have access to students' Test 1 scores and family income status when they make admission decisions. In addition, they may have access to students' Test 2 scores. We elicit spectators' preferences over whether to have this access under a test prep situation. We conduct this elicitation three times, once before any admission decision is made, once after six rounds of admissions, and another in the form of advice provided to other spectators right after the second elicitation. These elicitations are incentivized: the reported preferences in the first two elicitations could determine whether Test 2 scores are revealed before an admission decision, and the advice could be sent to other spectators. We also ask for justifications of the reported preferences in

open-ended questions.

The results show that the preference to have access to Test 2 scores is strongly influenced by the adequacy and bias of the test prep situation. Before making any admission decisions, 82% of spectators prefer to have access to Test 2 scores if no student received any additional test prep. In contrast, when the test prep is inadequate, biased, or both, 35%, 46%, and 56% of spectators prefer to exclude Test 2 scores from their own information sets. The preference to exclude Test 2 scores is mostly strict, and the result is robust to several treatment variations: the prevalence of information exclusion is virtually unchanged when students' family income status is not observed or when Test 1 scores are also subject to noise and bias.

Why do spectators prefer not to have access to test scores? Even if the test prep is inadequate or biased, standard consequentialism dictates that more information never does harm. We consider three deviations from standard consequentialism that could lead to the kind of information exclusion observed in our experiment. First, spectators may care about the fairness of the test as an admission procedure for non-instrumental reasons. As a result, they may feel that including inadequate or biased test scores could taint the fairness of the admission process. Second, spectators may worry that admission decisions could become more difficult when inadequate or biased test scores are present. Third, spectators may be concerned about over-relying on these test scores when they are observed, leading to mistakes in their admission decisions.

We organize these potential explanations in a unified theoretical framework and test their empirical implications in our experiment. The results favor procedural fairness concerns as the driving force behind the test scores exclusion and are inconsistent with the other two explanations. First, spectators who prefer to exclude test scores use them less when the scores are observed for admission decisions. This result is inconsistent with the hypothesis that concerns about overusing these scores are driving their exclusion. Second, preference for excluding inadequate or biased test scores are not significantly associated with indecision and response time in admission decisions where these scores are present. Thus, the results do not support the decision difficulty hypothesis. Moreover, spectators' information preferences when the admission decisions are made by them-

selves are almost the same as their advice to others. This is consistent with a non-instrumental notion of procedural fairness but again inconsistent with the other two hypotheses. For the costs of making admission decisions should be eliminated when the decisions are made by others while the potential for misusing test scores exacerbated. The procedural fairness hypothesis is further supported by results of an additional treatment. In the Performance Prediction treatment, instead of making admission decisions, spectators predict the performance of students in the advanced data science course and are paid for accuracy. For this prediction task which does not affect the students in any way, the fraction of spectators willing to exclude inadequate and biased test scores reduces by half. This result indicates that most of the test scores exclusion behavior is driven by fairness concerns when the welfare of others is at stake.

For people with a non-instrumental concern for procedural fairness, their demand for inadequate or biased test scores depends on the tradeoff between the moral cost of including these scores and their perceived usefulness. Indeed, many spectators who prefer to exclude Test 2 scores justify it by saying the scores are not useful. However, this perception is inaccurate. Many of these spectators end up using the scores in their admission decisions, and when we elicit their demand for test scores for a second time after that experience, the demand increases across the board. These results suggest that the exclusion of inadequate or biased test scores is partly driven by an initial underestimation of their usefulness, and this misperception can be mitigated with experience of making admission decisions.

Although our controlled experimental environment allows us to credibly identify the effect of interest, the design rules out potentially important motivations for test-optional and test-blind policies in the real world such as strategic and general equilibrium considerations.[1] To investigate whether the adequacy and bias are the primary considerations influencing attitudes toward test-optional and test-blind policies, we ask the spectators in our experiment whether they support these policies and why. Regardless of their support for these policies, the adequacy and bias of standardized tests are indeed the main considerations. Moreover, the support for these policies

[1]For example, these policies may help colleges alleviate social pressure and fend off legal challenges on their admission criteria. They may also reduce the financial and mental costs that students incur on test preparation.

is strongly correlated with the decision to exclude biased Test 2 scores in the experiment. These results indicate that the preferences identified in our controlled experiment are relevant to real world policy attitudes.

Literature review. [TBA]

2 Experimental Design

To set up the context where spectators make admission decisions, we recruited students from a US university to take an introductory data science course. Upon finishing the course, these students took two tests, referred to as Test 1 and Test 2, covering different course content areas. Each exam consisted of five questions with a total possible score of 10. For Test 1, students did not receive any additional test preparation. However, for Test 2, we provided some students with one of two types of test prep. The first, "skill-enhancing test prep," offered insights to improve their data analysis skills, and the insights were relevant to one of the test questions. The second type, "non-skill-enhancing test prep," simply provided the answer to one test question without enhancing analytical skills. Both types of test prep could potentially increase a student's score on Test 2 by up to two points.

We recruit participants ("spectators") from Prolific to select students from the introductory course to enroll in an advanced data science course. Each spectator makes 7 rounds of admission decisions, each time selecting three out of eight students. They could also choose an option that says "I cannot decide which 3 students to admit." The spectators have access to each student's score from Test 1 and know whether the student came from a higher-income (self-reported family income $\geq$ \$100,000) or lower-income background (< \$100,000). They may, in addition, know the students' Test 2 scores.

Before the admission rounds begin, we ask spectators if they prefer to have access to Test 2 scores for their decision-making. Their responses could be yes, no, or indifferent, along with a justification for their choice. Spectators answer this question twice, each time assuming a differ-

ent scenario regarding Test 2 preparation. The first scenario is a "No Prep" condition where no student received any additional preparation for Test 2. The second scenario is randomized among spectators and includes one of the following conditions:

- **Inadequate Prep**: All students received non-skill-enhancing test prep.
- **Biased Prep**: Only students from higher-income backgrounds received skill-enhancing test prep.
- **Inadequate and Biased Prep**: Only students from higher-income backgrounds received non-skill-enhancing test prep.

We randomize the order in which spectators encounter these two scenarios.

After reporting their information preferences for the two scenarios, each spectator gets to know which one is the actual test prep situation for the students considered for admission. For this scenario, we ask spectators to confirm their previously stated information preference by completing a small real effort task, which entails typing in a sentence. A confirmation would ensure that the observability of Test 2 scores adheres to their stated preference in the majority of admission rounds.

Spectators know that one of the seven rounds consists of real students who took our introductory course while the other six rounds are fictitious. Unbeknownst to them, the real round is the last one, which is also the only one where the observability of Test 2 scores is affected by their reported information preference. Whether Test 2 scores are revealed is fixed for the fictitious rounds: they are in Rounds 4 to 6 but not in Rounds 1 to 3.

Table 1 lists the students' test scores and family income status for the six fictitious rounds under the No Prep scenario.[2] These numbers are specifically designed. The first three rounds allow us to identify spectators' social preferences such as preference for meritocracy and lower-income students. Rounds 4 and 5 allow us to observe spectators' tradeoffs between the two test scores, while Round 6 identifies the tradeoff between high Test 2 scores and lower-income status. The orders within the first and second three rounds are randomized.

[2]For the other three test prep scenarios, the only difference is that the Test 2 scores for students who received test prep are set to be one point higher.

Table 1: Student information in the six fictitious admission rounds

Round	Higher-income students	Lower-income students
1	(9,-), (8,-), (5,-), (4,-)	(9,-), (7,-), (6,-), (4,-)
2	(8,-), (8,-), (7,-), (0,-)	(6,-), (5,-), (5,-), (5,-)
3	(7,-), (7,-), (7,-), (6,-)	(7,-), (7,-), (7,-), (7,-)
4	(9,9), (6,5), (5,5), (5,5)	(8,8), (7,6), (6,8), (5,5)
5	(9,8), (8,6), (7,8), (3,4)	(9,9), (6,5), (6,5), (6,5)
6	(7,9), (7,8), (7,5), (6,5)	(7,9), (7,7), (7,5), (7,4)

Notes: The numbers in each parenthesis represent a student's scores in Test 1 and Test 2. Test 2 scores are revealed in Rounds 4 to 6 but not in Rounds 1 to 3.

After the sixth round of admission decision, we elicit spectators' information preference for a second time. This elicitation is intended to measure if experience with admission decisions affects information preferences. Right after this elicitation, we also ask spectators to advise other participants who are in the same test prep scenario on whether to request access of Test 2 scores. The advice reflects spectators' information preferences when others decide whom to admit.

The last round of admission decision is made after the advice elicitation, where the disclosure of Test 2 scores is based on a random selection of the spectators' expressed information preferences at one of the two junctures. Finally, we survey participants on their attitudes towards test-blind and test-optional admissions policies and collect demographic information.

Incentives. The elicitations of information preferences and admission decisions are all incentivized. Spectators are told that their admission decisions for the round that consists of real students have a chance of being implemented. For the first elicitation of information preferences, after the true test prep scenario is revealed, we ask spectators to confirm their preferences for this scenario by typing in a sentence unless they reported indifference. They are told that the confirmed preference will affect whether the majority of the admission rounds have Test 2 scores available or not. If they do not type in the sentence, they are told that there may be more rounds where the observability of Test 2 scores does not adhere to their reported preferences. For the second

elicitation of information preferences after the sixth round of admission decisions, spectators are told that one of their two reports of information preferences will determine whether Test 2 scores are revealed for the last admission decision. The advice to other spectators has a chance of being sent out for real.

Logistics. We recruit 900 spectators from Prolific on December 18, 2023, and 596 of whom participate in our main treatment. Each participant receives a fixed payment of $5 and the median time spent is 17.5 minutes. We conduct several additional robustness and mechanism treatments, one of which (Performance Prediction) includes an additional $3 incentive bonus. These additional treatments will be introduced later in the paper.

3 Conceptual Framework

Why would spectators choose to exclude freely available test score information from admission decisions? In this section, we start from a standard consequentialist model and derive a negative result: even if she cares about fair outcomes, a standard consequentialist spectator would never exclude any information. Then, we consider several deviations from this model which could lead to exclusion of test scores and derive their implications.

3.1 Set-up

Let θ be the state of the world relevant to admission, such as students' family income status, ability, and potential. The admission decision is denoted by a which is a vector of ones and zeros representing whether each student is admitted or rejected. Prior to decision, the spectator decides whether to have access to students' test scores T with the probability of its realizations denoted by $p(t)$. All the other information she observes about the students is summarized by S, which will be suppressed below for ease of notation.

The spectator's preference for admission outcomes is described by the value function $v(a, \theta)$. Note that this function is general enough to accommodate social preferences such as meritocracy,

preference for low-income students, preference for diversity, etc.

3.2 Standard consequentialism

A standard consequentialist spectator maximizes the expected value of admission outcomes given available information. Therefore, the value of test scores is

$$V(T) = \sum_t p(t) \cdot E[v(a(t), \theta) - v(a(\emptyset), \theta)|t], \tag{1}$$

where $a(t)$ is the optimal admission decision given test scores t, and $a(\emptyset)$ is the optimal decision when t is not observed. Because for each t, $a(t)$ yields a weakly higher value than $a(\emptyset)$, the value of test scores is always nonnegative. Hence, the spectator never strictly prefers to exclude T.

3.3 Procedural fairness

For allocation decisions, people may care about the fairness of the decision procedure independent of the resulting outcome. For spectators with such concerns, including inadequate or biased test scores as part of the admission criteria may taint the fairness of the whole procedure. We model procedural fairness concerns as a reduced-form moral cost $C(T)$ of including the test scores T. The value of T then becomes

$$V(T) = \sum_t p(t) \cdot E[v(a(t), \theta) - v(a(\emptyset), \theta)|t] - C(T). \tag{2}$$

With procedural fairness concerns, whether to include or exclude the test scores depends on the tradeoff between the expected usefulness of the scores and the moral cost of including them. This leads to an important comparative statics: test score exclusion increases as the expected usefulness of the scores goes down.

3.4 Decision costs

Another potential reason for test score exclusion is that admission decisions with the test scores present may be more costly than those without. It could be the mental cost of sifting through and digesting additional information. It could also be the increased social pressure if the test scores are also observed by third parties. Specifically, with the test scores present, the spectator chooses $\tilde{a}(t)$ to maximize $E[v(a,\theta)|t] - c(a,t)$. The value of test scores is then

$$V(T) = \sum_t p(t) \cdot E[v(\tilde{a}(t),\theta) - c(\tilde{a}(t),t) - v(a(\emptyset),\theta)|t]. \tag{3}$$

One implication of decision costs as a potential driver of test score exclusion is that spectators would be less willing to exclude the test scores if other people make the admission decisions and, hence, bear the decision costs.

3.5 Anticipated decision mistakes

One last explanation for test score exclusion we consider is related to anticipated decision mistakes. If a spectator is concerned that her admission decision may be biased when she observes the test scores, she may be willing to blind herself from them. Formally, the spectator anticipates that when she observes the test scores t, she will choose $\hat{a}$ to maximize $E[\hat{v}(a,\theta)|t]$ where $\hat{v}$ is a biased value function different from v. Therefore, the value of test scores,

$$V(T) = \sum_t p(t) \cdot E[v(\hat{a}(t),\theta) - v(\hat{a}(\emptyset),\theta)|t], \tag{4}$$

could be negative.

This explanation of test score exclusion also has implications for when admission decisions are made by other people. Assuming that other people's decisions are more likely to be misaligned than one's own, people should be more willing to exclude test scores for other people's admission decisions.

4 Main Results

4.1 Information preferences under different test prep scenarios

Figure 1 shows the distributions of information preferences under different test prep scenarios. Before making any admission decisions, 11% of participants prefer to exclude Test 2 scores from the admission process in the No Prep scenario, whereas the number increases to 35%, 46%, and 56% when the scores are inadequate, biased, and both, respectively. The preference to exclude Test 2 scores is mostly strict: 92% of these participants, when prompted, complete the real effort task to confirm their exclusion preference. This result indicates that the inadequacy and bias of a test make people more willing to exclude it from the admission process.

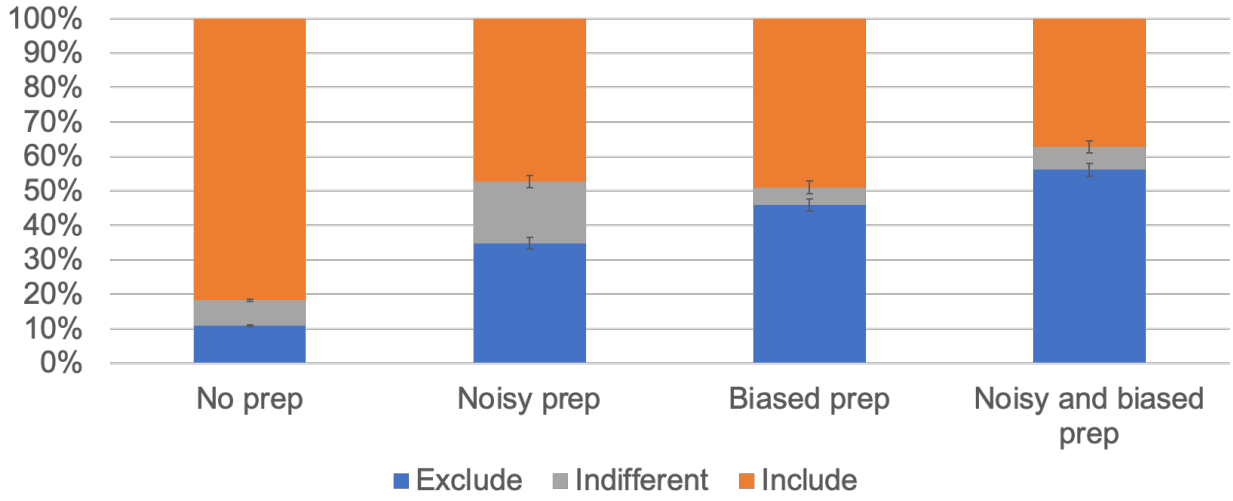

Figure 1: Information preferences across test prep scenarios (1st elicitation)

4.2 Admission decisions without Test 2 scores

What are spectators' objectives when they make admission decisions? Results from the first three rounds of decisions reveal that most spectators are meritocratic, with some showing a preference for lower-income students. In Round 1 where 2 of the 3 top Test 1 scores belong to higher-income students, 76% of spectators select all 3 top performers. In Round 2 where all top performers are from higher-income background, the percentage of purely meritocratic decisions drops to

62%. If, in addition to purely meritocratic choices, we also consider income-conditional meritocratic decisions that only admit top performers within each income group (but not necessarily across groups), the percentage increases to 89% for Round 1 and 86% for Round 2. Only 2% and 1% of spectators in these rounds exclusively admit lower-income students.

In Round 3 where 3 higher-income and all 4 lower-income students are tied for the top Test 1 score, 22% admit lower-income students exclusively whereas 53% admit from both income groups. Very few spectators (4%) only admit higher-income students. A nontrivial 24% state that they could not make a decision.

4.3 Admission decisions with Test 2 scores

How do spectators make admission decisions in Rounds 4 to 6 where Test 2 scores are revealed? We construct student information in these rounds so that for most spectators, the only nontrivial decision is to choose one of two "focal" students, one with a higher Test 2 score and the other with either a higher Test 1 score or a lower income status. This decision then allows us to identify spectators' tradeoff between the two tests and between Test 2 scores and low-income preference.

Specifically, in each round, we include two students, one higher-income and one lower-income, whose scores dominate all others and four students whose scores are lower than the rest. These dominance relationships hold even if we undo the potential impact of test preparation. As a result of this design, the vast majority of spectators (94% in Round 4, 91% in Round 5, and 87% in Round 6) admit the two dominant students and reject the dominated, and the third admitted student must come from the remaining two.

In Round 4, both focal students come from a lower-income background. One student scores 2 points higher in Test 2 but 1 point lower in Test 1. In the No Prep scenario, 39% favor the student with the higher Test 1 score, while 54% favor the student with the higher Test 2 score. The distribution is similar in the Biased Prep scenario (39% vs. 52%) and the Inadequate and Biased Prep scenario (43% vs. 51%) but tilts toward Test 1 when lower-income students receive non-skill-enhancing test prep in the Inadequate Prep scenario (63% vs. 35%). This pattern is consistent

with Bayesian updating: as lower-income students' Test 2 scores become less informative of their ability thanks to the non-skill-enhancing test prep they receive, spectators should optimally lower the decision weight placed on Test 2 scores.

Round 5 differs from Round 4 in that the two focal students come from a higher-income background. In the No Prep scenario, 36% of spectators favor the student with the higher Test 1 score, while 55% favor the student with the higher Test 2 score. The distribution shifts toward Test 1 in the Biased Prep scenario (48% vs. 44%), and even more so in the two scenarios where higher-income students receive non-skill-enhancing test prep (53% vs. 43% in the Inadequate Prep scenario, 57% vs. 33% in the Inadequate and Biased Prep scenario). Similar to the Round 4 results, it is consistent with Bayes' rule to lower the decision weight on Test 2 scores as these students receive non-skill-enhancing test prep.

5 Rationales

To understand the rationales behind the information preferences, we ask spectators to provide justifications in an open-ended question and summarize the answers using GPT 4, a large language model. The exclusion of Test 2 scores predominantly revolves around two considerations – the fairness and usefulness of the test scores. In the two scenarios where only higher-income students receive test prep, 86% (Biased) and 74% (Inadequate and Biased) of spectators who prefer to exclude Test 2 scores mention that including the scores would be unfair. This number drops to 37% in the Inadequate Prep scenario and further to 11% in the No Prep scenario. In the three scenarios with test prep, 69% (Inadequate), 73% (Biased) and 66% (Inadequate and Biased) of spectators who prefer to exclude Test 2 scores mention that the scores would not be useful for their admission decisions. For the No Prep scenario, the number drops to 52%.

To provide behavioral evidence on fairness concerns as a driving force of the exclusion of Test 2 scores, we conduct a diagnostic treatment where spectators predict the performance of students who already completed the advanced data science course. Spectators' predictions won't affect the

students in any way but are incentivized for accuracy. In this treatment where other-regarding preferences should have no bite, only 29% of participants in the prediction treatment choose to exclude Test 2 scores in the Inadequate and Biased Prep scenario. This number is significantly lower than the 56% of spectators who prefer exclusion in the main treatment under the same scenario ($p < 0.001$).

To provide behavioral evidence for the usefulness concern as a driving force of test score exclusion, we compare the use of Test 2 scores between spectators who prefer to exclude them with those who don't. Across Rounds 4 to 6 and across the four test prep scenarios, spectators who choose to exclude Test 2 scores use them less when they are available. This result is consistent with procedural fairness concerns where test score exclusion is associated with a low expectation of the usefulness of the scores. On the contrary, it is not consistent with anticipated decision mistakes as a explanation for the exclusion.

One remaining question is whether spectators correctly anticipate the usefulness of Test 2 scores when they report their information preferences. This question can be answered by comparing information preferences before and after 6 rounds of admission decisions, as spectators might learn about the usefulness of test scores while making decisions. After the admission rounds, demand for Test 2 scores increase across all scenarios. This indicates that spectators might have underestimated the usefulness of Test 2 scores when they initially decide to exclude them.

6 Discussions

[TBA]

References

[TBA]