Hit or Miss? Test Taking Behavior in Multiple Choice Exams

Ş. Pelin Akyol * James Key† Kala Krishna‡

February 15, 2015

Abstract

This paper models and estimates students' decision to guess/attempt or skip the question in a multiple choice test in order to understand the role that student characteristics play. We do this using data from the Turkish University Entrance Exam, a highly competitive, high stakes exam. In particular, we investigate students' behavior according to their gender, predicted score and experience in the exam. Our results show that students' attitudes towards risk differ according to their gender, predicted score and exam experience: female students behave in a more risk averse manner relative to male students, and high scoring students are more risk averse. However, our counterfactual analysis suggests that although different testing regimes can lead to different score distributions, the relationships between exam score percentiles and student characteristics are relatively invariant.

*Bilkent University, e-mail:pelina@bilkent.edu.tr

†University of Western Australia, e-mail:james.key@uwa.edu.au

‡The Pennsylvania State University, NBER, IGC and CES-IFO, e-mail:kmk4@psu.edu

1 Introduction

Multiple choice tests are commonly used to evaluate candidates in a wide variety of situations. Their main advantages are that they allow a broader evaluation of a candidate's knowledge in a short time and are objectively graded at a very low cost. Consequently they tend to be preferred in both high and low stake exams, especially when there are a large number of exam takers which allows the higher cost of developing the exam to be spread over many units. Examples include University entrance exams in a number of countries including Turkey, Greece, Japan and China. In the US, the Scholastic Aptitude Tests (SATs) and Graduate Record Exams (GREs) that are taken before applying to undergraduate and graduate schools are also mostly of this form. Multiple choice test results are widely used, not only to allocate students to colleges, but to measure effectiveness of schools, teachers, to enter the civil service, and to allocate open positions.[1]

A disadvantage of such exams is that candidates are more able to guess the correct answer, especially if they can eliminate some options.[2] In other exam types, such as short answer based exams, such uneducated responses are unlikely to reap any benefit. As a response to this problem, test administers may use negative marking for wrong answers to deter such guessing. Grading methods in multiple choice tests are often designed in such a way that the expected score from randomly guessing a question is equal to the expected score from skipping the question. However, the candidate's decision to guess/attempt or skip the question will depend not only on their knowledge, but also on their degree of risk aversion. This may undermine the validity and the fairness of test scores, reducing the efficacy of the testing mechanism.

Baker et al. [2010] criticize the use of test results of students to evaluate the value-added of teachers and schools, among other reasons, because of the measurement error that will be generated by random guessing. Baldiga [2013] shows in an experimental setting that conditional on students' knowledge of the test material, those who skip more questions tend to perform worse. In light of this finding, if certain groups are favored in multiple choice tests, using these scores will bias the results against groups who skip questions rather than guess.

In this paper we model and estimate students' multiple choice test taking behavior. We do so with a view to understanding how students seem to behave when taking these exams and whether

[1] In Turkey, public sector jobs are allocated according to the score obtained in a multiple choice central exam, called KPSS.

[2] For example, with no knowledge of the subject and four options on each question, a student would on average get 25% correct.

exam taking behavior seems to differ across groups and what the consequences of such differences seem to be. We use administrative data from the Turkish University Entrance Exam (ÖSS) to do so. The ÖSS is a highly competitive, centralized examination that is held once a year.[3] In Turkey, the college admission depends only on the score obtained from the ÖSS, and the high school GPA.[4] However, the ÖSS score has the highest weight.[5] In the ÖSS, for each correct answer the student gains one point, and for each wrong answer 0.25 points are deducted, while no points are awarded/deducted for skipping a question. As it is a very competitive and high-stake exam, students expend significant time and effort to get prepared for this exam. Therefore, we assume that they are aware of the scoring method. These properties of the ÖSS exam provide us with a convenient environment in which to investigate the exam taking behavior of students.

Psychology and education literature have long been interested in multiple choice tests to characterize optimal test designs that generate fair results, with a valid measurement method. Burgos [2004] investigates the score correction methods that will award partial knowledge by using prospect theory. They model the behavior of a representative agent, so they assume away the heterogeneity in risk aversion and ability of agents. Similarly, Bernardo [1998] analyzes the decision problem of students in a multiple choice exam to derive a fair grading rule. Espinosa and Gardeazabal [2010] model the students' optimal behavior in a multiple choice exam to find the optimal penalty that will increase the validity of the test, i.e., increase the correlation between students' knowledge and the test score by simulating their model under distributional assumptions on students' ability, difficulty of questions and risk aversion. Any of these papers attempt to test their results empirically. Espinosa and Gardeazabal [2005] tests students' rationality with an experiment by using equivalent scoring rules, but while one punishes the wrong answer, and the other one awards skipping.[6] They find evidence that students are expected utility maximizers.

Risk attitudes of students are an important factor in the decision to attempt whenever there is uncertainty associated with the outcome. In the literature, females are shown to be more risk averse than males (see Eckel and Grossman [2008]). To test the hypothesis that female students skip more question than males since they are more risk averse, Ben-Shakhar and Sinai [1991] investigates test taking strategies of students in Hadassah and PET tests in Israel and find that females do, in fact, tend to skip more questions. Baldiga [2013] explores the gender differences in skipping/guessing

[3]Every year only one third of the students are assigned to colleges.

[4]This GPA is normalized at the school year level

[5]ÖSS score has a minimum weight of 75% in the score used to allocate students to colleges.

[6]Essentially testing for the Framing Effect

behavior when students are uncertain about answers. They conduct an experiment to disentangle whether a gender gap in the tendency to skip questions exists, and if so, whether this gap is driven by differential confidence in knowledge of the material, differences in risk preferences, or differential responses to high pressure testing environment. Tannenbaum [2012] also investigates the effect of gender differences in risk aversion on multiple choice test results. Tannenbaum [2012] shows that female students are more risk averse, so they skip more often than male students. He finds that risk aversion is able to account for the 40% of the gender differences in performance in multiple choice exams. Tannenbaum [2012] is the closest paper to ours. However, our conclusions conflict somewhat: we allow risk aversion to differ between various groups (gender, predicted scores) whereas he estimates a unique risk aversion for each gender.

We contribute to the literature by using a structural model to estimate students' exam taking behavior. We investigate the effects of different characteristics of students on their exam performance. Particularly, we investigate students' behavior according to their gender, expected score and experience in the exam. Our model allows us to run counterfactual analysis to investigate the predicted change in score distributions in a variety of situations.

Our results show that students' attitudes towards risk differ according to their gender, expected score and experience in the exam. Female students behave in a more risk averse manner relative to male students, and we see a similar pattern for the first time takers relative to second time takers. While students differ in terms of how they approach the exam, our counterfactual analysis suggests that under different testing regimes there are little differences in aggregate outcomes.

In the next section, we present an overview of the data and testing environment. The particular patterns of the multiple choice tests are discussed in more detail in section three. In the fourth section, the model is presented. Section five details the estimation strategy with results in section six. Sections seven contains counterfactual experiments and a discussion of an extension, respectively. Section eight concludes.

2 Background and Data

In Turkey, college admission is based on an annual, nationwide, central university entrance exam governed by the Student Selection and Placement Center (ÖSYM). High school seniors and graduated students can take the exam. There is no restriction on retaking, i.e., students are allowed to take the exam repeatedly over the years. However, they are not allowed to carry over their scores:

the score obtained in a year can be used only in that year. All departments, with the exception of those that requires special talent (such as art schools) accept students based on a weighted average score of university entrance exam and high school grade point average.

The university entrance exam is held once a year all over the country at the same time. It is a multiple choice exam with four tests, Turkish, social science, math, and science. Students are given 180 minutes for 180 questions. Each test includes 45 questions, and each question has 5 possible answers. Students get one point for each correct answer, and they lose 0.25 points for each wrong answer. If they skip the question, they receive 0 points. Students' raw test scores are calculated by deducting $\frac{1}{4}$ of the number of incorrect answers from the number of correct answers. The university entrance exam is a paper-based exam. All students receive the same questions, and they do not receive any feedback on whether their answer is correct or not during the exam.

Students choose one of the Science, Turkish-Math, Social Science, or Language tracks at the beginning of high school.[7] Students' university entrance exam scores are calculated as a weighted average of their raw scores in each test.

Table 1 shows the test weights according to each track. For the social science track students, the Turkish and social science tests have the highest weight, while math and science have a relatively low effect on the ÖSS-SÖZ score.[8]

Students are required to pass the threshold of 105 points to submit preferences (submit an application) to 2-year college programs, while they need 120 points to apply to 4-year college programs. Students' allocation score (Y-ÖSS) is calculated based on their high school and exam performance as follows

$$\text{Y\_ÖSS\_X}_i = \text{ÖSS\_X}_i + \alpha \text{AOBP\_X}_i$$

where X$\in$\{SAY,SÖZ,EA,DIL\}, and α is a constant that changes according to the student's track, preferred department and whether the student was placed (accepted) into a regular program in the previous year or not. ÖSYM publishes the lists of departments open to students' according to their tracks. When students choose a program from this list, α will be 0.5, while if it is outside the open list, α will be 0.2. If the student has graduated from a vocational high school, and prefers

[7] For more detail on track choice in high school in Turkey see Akyol and Krishna [2014]

[8] In the calculation of ÖSS scores, firstly raw scores in each field are normalized to mean 50 and standard deviation 10 by using mean and standard deviation of scores in the corresponding field. Then these normalized scores are multiplied by the weights presented in Table 1.

a department that is compatible with his high school field, α will be 0.65. If the student was placed in a regular university program in previous year, the student is punished and α will be equal to either 0.25, 0.1, or 0.375. For those students, the α coefficient is equal to half of the regular α. This punishment structure gives students an incentive to stay in their track, and to accept a position when offered.

The data used in this study comes from multiple sources. Our main source of data is the institutional data of the 2002 university entrance exam takers. This data set includes students' raw test scores in each test, weighted test scores, high school, track, high school GPA, gender, and number of previous attempts.

The second source of data is the 2002 university entrance exam candidate survey. This survey is filled by all students while they are making their application for this exam. This data set has information on students' family income, education level, and expenditure on preparation.

We received a random sample of around 40,000 students from each track (Social Science, Turkish-Math, Science). In this study we focus on the social science track students. This is due to the unique patterns seen in these tests that enable identification of key parameters, patterns which do not exist in the science and math tests. This arises due to the style of questions being asked.[9] As a result we focus on the students for whom these tests matter the most: the social science track students.

Table 9 presents the summary statistics.

We observe substantial differences in test outcomes between first time takers and second time takers. Figure 1 shows that second time takers achieve much higher scores, for the entirety of the distribution. An aim of this paper will be to examine whether or not this is due to the second time group having a better distribution of ability compared to first time takers, or if is simply due to a change in test taking behavior, for example their willingness to guess when uncertain.[10]

In addition, we see the difference between male and female students' scores in Figure 2. Males tend to have a wider distribution, with more mass on the upper and lower tails. It is notable that the difference on the upper tail is more pronounced for second time takers; first time male test takers differ from their female counterparts mainly by having more students with low scores.

[9] In particular, math and science exams have many questions where the student must solve the problem to find the answer: either the student successfully solves the question or they fail to solve the question and have no information regarding what the correct answer is likely to be.

[10] We do not separate learning and selection effects in this paper

3 Multiple Choice Exam Scores

In this section we examine students' scores in each section of the ÖSS exam: the Turkish, social science, math and science. Recall that each section of the exam has 45 questions. For each question, there are five possible answers; answering correctly gains the student a single point, skipping the question (not giving an answer) gives zero points, but attempting the question and answering incorrectly results in a loss of a quarter point.

The scoring structure results in each multiple of 0.25 between -11.25 and 45 (with the exception of certain numbers above 42) being a possible outcome of an exam section. For example, attempting all questions and being incorrect with each question results in a score of $-\frac{45}{4} = -11.25$, while getting everything correct nets the student 45 points.

First, Figures 3 and 4 show the distribution of scores in the social science and Turkish portions of the exam, for first time takers and second time takers in the social science track. These histograms use a bin width equal to 1.[11]It can be seen that the distributions are roughly bell shaped, as would be expected. As noted in previous work (see Frisancho et al. [2012]), the distributions appear to shift to the right as we move from first time takers to second time takers.

Next, we show the histograms of scores of social science track students in the math and science portions of the exam in Figures 5 and 6 . Two facts immediately stand out in these diagrams. First, there are a lot of scores in the zero region. Secondly, even if the large spike at zero is removed, the distribution of scores place almost all of the weight at very low scores, relative to those seen in the Turkish and social science sections of the exam.

There are two explanations for both of the two preceding facts. First, math and science test scores have relatively little weight in the ÖSS score of social science track students. These scores are multiplied by 0.4, whereas social science and Turkish scores are multiplied by 1.8 - a substantial difference. As a result, students have little incentive to expend time and effort on these questions during the exam. Accordingly, many students opt to not even look at these sections, resulting in many observations of zero scores. Furthermore, students are explicitly told that if they are social science track students, they should spend less time in math and science.[12] If they do attempt questions, it will be relatively few, with less than advantageous results. A second reason is that

[11]While the data is not rounded to the nearest integer, we do this so that the reader may see the overall distribution of scores - as will be seen shortly, this can be difficult to see when bins are not used.

[12]In the exam booklet there is a note before the social science/Turkish part of the exam that says: "If you want higher score in ÖSS-SÖZ, it may be better for you to spend more than 90 minutes on verbal part of the exam."

these students are not well prepared in math and science - since the ninth grade they have been engaged in the social science track curriculum, meanwhile their study efforts would optimally be directed towards the exams that matter most: Turkish and social science.

A score can correspond to a single outcome (by outcome we mean the number of correct, wrong and skipped questions), multiple outcomes, or none at all. It is clear that there is only one way that a student could obtain -11.25 or 45; a score of 42.5 could only have arisen through attempting all questions, getting 43 questions correct and 2 incorrect. A score of 40 has two possible origins: 40 correct and 5 skips, or 41 correct and 4 incorrect. It is impossible to achieve a score of 42.25: the student must have at least 43 questions correct, and at least 3 questions skipped, which is not possible given there are only 45 questions.

There are 46 particular scores that are highly relevant: those which correspond to attempting all questions. These are spaced 1.25 points apart, starting at -11.25, finishing at 45 points.

Examining the distributions of actual scores (not placed into bins of width one), we can see some striking patterns. Figures 7 and 8 show the distributions for the social science and Turkish portions of the exam, with very prominent spikes along the distribution. It is no coincidence that the spikes appear evenly placed; they correspond to scores which could be achieved while attempting all questions. This will be very important in identifying the behavior, and implies that many students are skipping very infrequently. Additionally, there is no discernible pattern as to which students obtain such scores - these spikes remain present when examining the distributions of different groups, for example high income female first time takers with low GPAs.[13]

Math and science score distributions do not exhibit this behavior, most students obtain a score of zero. There are no other apparent patterns to be gleaned from dis-aggregating the distribution, apart from a small spike at 1 in three of the four curves.[14]

4 Model

We model the test taking behavior as follows. When a student approaches a question, he observes a signal for each of the five possible answers. The vector of signals for the question is then transformed into a belief. This belief is the likelihood that each answer is in fact the correct answer. The student then decides whether or not to answer the question, and if so, which answer to choose.

We model the test taking procedure as if each test takes place separately— we do not allow

[13] It is possible that there is some unobserved heterogeneity that determines which students lie on these spikes.

[14] This could correspond to students only answering the easiest question in the exam.

for outcomes in one section of the test to have any bearing on other sections.[15] In addition, each question is approached simultaneously, so that outcomes (or beliefs regarding outcomes) of one question have no impact on other questions.

Signals for each of the five answers depend on whether or not the answer is actually the correct answer, and are drawn as follows:

- Incorrect answers - draw a signal from a distribution G, where G is Pareto with support $[A_I, \infty)$ and shape parameter $\beta > 1$.
- Correct answer - draw a signal from a distribution F, where F is Pareto with support $[A_C, \infty)$ and shape parameter equal to $\alpha > 1$.

Assumption 1. *Both F and G have support* $[A, \infty)$*, where* $A > 0$*.*

Suppose that the student observes five signals, given by the following vector:

$$X = (x_1, x_2, x_3, x_4, x_5) \tag{1}$$

where x_i is the signal that the student receives when examining answer i. What then is the student's belief regarding the likelihood that each answer is correct? Using Bayes' rule, the probability that answer i is correct can be expressed as:

$$\text{Prob}(\text{Answer } i \text{ is correct}|X) = \frac{\text{Prob}(X|\text{Answer i is correct}) \times 0.2}{\text{Prob}(X)} \tag{2}$$

Expressing the numerator in terms of the densities of the two distributions, F and G, for the case where $i = 1$:

$$\text{Prob}(X|\text{Answer 1 is correct}) = \frac{\alpha A^\alpha}{x_1^{\alpha+1}} \frac{\beta A^\beta}{x_2^{\beta+1}} \frac{\beta A^\beta}{x_3^{\beta+1}} \frac{\beta A^\beta}{x_4^{\beta+1}} \frac{\beta A^\beta}{x_5^{\beta+1}} \tag{3}$$

In essence, the density of $F(\cdot)$ at x_1 (as this is conditional on 1 being correct) multiplied by the product of the density of $G(\cdot)$ at the other signals.

It follows, by substituting equation 3 into equation 2, that the probability that answer i is correct, conditional on X, can be expressed as:

[15] We are explicitly ignoring time restrictions, whereby a quick performance in one section of the exam might afford the student additional time in another section, allowing the student to more carefully examine each question

$$\text{Prob}(i \text{ is correct}|X) = \frac{\frac{\alpha A^{\alpha}}{x_i^{\alpha+1}} \prod_{j \neq i} \frac{\beta A^{\beta}}{x_j^{\beta+1}}}{\sum_{m=1}^{5} \left(\frac{\alpha A^{\alpha}}{x_m^{\alpha+1}} \prod_{n \neq m} \frac{\beta A^{\beta}}{x_n^{\beta+1}} \right)} \tag{4}$$

where $i, j, m, n \in \{1, ..., 5\}$.

This can be further simplified to:

$$\text{Prob}(i \text{ is correct}|X) = \frac{\frac{1}{x_i^{\alpha+1}} \prod_{j \neq i} \frac{1}{x_j^{\beta+1}}}{\sum_{m=1}^{5} \left(\frac{1}{x_m^{\alpha+1}} \prod_{n \neq m} \frac{1}{x_n^{\beta+1}} \right)} \tag{5}$$

Letting $\gamma = \beta - \alpha$, so that $\frac{1}{x_i^{\alpha+1}} = \frac{1}{x_i^{\beta+1}} x_i^{\gamma}$, the expression further simplifies to:

$$\text{Prob}(i \text{ is correct}|X) = \frac{x_i^{\gamma}}{\sum_{m=1}^{5} x_m^{\gamma}} \tag{6}$$

Note that the sum of beliefs for each of the five answers adds up to unity. Without loss of generality, we assume that $\beta \geq \alpha$, which leads to positive relationship between the value of the signal, and the likelihood that the answer is correct.[16], [17]

Proposition 1. *The outcome of the model is the same for all $A > 0$*

Proof: Replace A with cA. Now, for any $P \in [0, 1)$, the signal generated by the correct answer, $F^{-1}(P)$, will be c times as large and similarly for signals generated for the incorrect answers. Compare this to a situation where the student arbitrary decides to inflate all signals by c. This will clearly have no impact on the decisions/outcome probabilities of a rational agent, but mirrors what would be seen when replacing A by cA. □

Proposition 2. *The outcome of the model is the same for all (α, β) that satisfies $\frac{\beta}{\alpha} = k \geq 1$*

Proof: Let the student transform all signal vectors X to Y, such that $y_i = (x_i)^{\alpha}$. The correct answer now has a signal distribution of:

$$F(y_i) = 1 - \frac{B}{y_i} \tag{7}$$

where $B = A^{\alpha}$. Similarly, the incorrect answers' signals have the following distribution:

$$G(y_i) = 1 - \left(\frac{B}{y_i} \right)^{\frac{\beta}{\alpha}} \tag{8}$$

[16] A higher shape parameter for a Pareto distribution shifts probability mass to the left

[17] If a student were to draw from distributions with $\beta < \alpha$, smaller signals would be associated with the correct answer.

So this re-scaling of the signals preserves all of information contained in the original signals, and the resulting signals have a distribution identical to one where the correct answers come from a Pareto distribution with the shape parameter equal to one, and the incorrect answers have shape parameter equal to β/α. As shown in the earlier proposition, the scale parameter is irrelevant. □

Accordingly, we can, without loss of generality, take $A = 1$ for all students, and take $\alpha = 1$ for all students. As a result, the structure of a student's signals can be represented by the shape parameter of the incorrect answer: β. A higher value of β draws the the mass of the distribution towards the minimum, $A = 1$, allowing the student to more clearly separate the incorrect signals from the signal given by the correct answer. In other words, higher β students are what would be referred to as high ability students.[18] However, they are high ability in terms of their ability to distinguish correct from incorrect, in this particular exam. While this should be highly correlated with most reasonable measures of ability, it is limited to the exam in question. It likely incorporates general aptitude for multiple choice test.

The effect of a higher β on test outcomes can be decomposed into three effects. First, the correct answer has a higher probability of generating the highest signal. Increasing β shifts the CDF of the incorrect answers' signals to the left, and the student's best guess (the answer with the highest signal) will be correct more often. Secondly, when the correct answer actually gives the highest signal, the probability with which the student believes that it comes from the correct answer increases as the weighted sum of the incorrect signals decreases. If the first answer is the correct answer, lowering $\sum_{i=2}^{5} x_i^{\gamma}$ increases the student's belief that answer 1 is correct.

Finally, there is a subtle effect of β on tests. Students with high ability, i.e. a high value of β, will be more confident in their choices. Even with the same signals, as we increase β, the student's belief that the highest signal comes from the correct answer increases. This is formally stated below:

Lemma 1. *Suppose there are two students: one with ability parameter $\beta = b_1$ and the other with ability parameter $\beta = b_2 > b_1$. Suppose that the two students receive identical signals X for a question. Let $x_{\max} = \max\{x_1, ..., x_5\}$. The student with the higher value of β has a higher belief that $x_{\max}$ is drawn from the correct answer.*

Proof: The belief is given by $\frac{x_{\max}^{\gamma}}{\sum_{m=1}^{5} x_m^{\gamma}}$. Taking logs, and differentiating with respect to γ, yields the following expression:

[18] Signal distributions for a student with ability $\beta = 3$ are shown in Figure 11

$$\frac{d\log(\text{Belief})}{d\gamma} = \log x_{\max} - \frac{x_1^\gamma \log x_1 + x_2^\gamma \log x_2 + x_3^\gamma \log x_3 + x_4^\gamma \log x_4 + x_5^\gamma \log x_5}{x_1^\gamma + x_2^\gamma + x_3^\gamma + x_4^\gamma + x_5^\gamma} \tag{9}$$

Since $\log x_{\max} \geq \log x_i$, and $x_i > 0$,

$$\frac{d\text{Belief}}{d\gamma} \geq 0 \tag{10}$$

With the inequality strict unless $x_1 = x_2 = x_3 = x_4 = x_5$. Since $\gamma \equiv \beta - \alpha$, the student with the highest value of β has the strongest belief ($\alpha = 1$ for both students).□

Once students have observed signals for each of the five possible answers to the question, they are faced with six possible alternatives: either choosing one of the five answers, or skipping the question. Skipping the question does not affect their test score, answering correctly increases the score by 1, while answering incorrectly decreases the score by 0.25 points. Note that the expected value of a random guess is $0.2 * 1 - 0.8 * 0.25 = 0$.

If a student were to choose an answer, they would choose the one which was most likely to be correct. A slightly higher score is clearly preferred. In this model, the answer which is most likely to be correct is the one with the highest value of x_i. Also, this answer trivially has a probability of being correct (conditional on observed signals and the student's ability) greater than or equal to twenty percent.

However, the relationship between ÖSS score and utility need not be linear. It is reasonable to suggest that there may be a degree of risk aversion present, both student's general attitudes towards risk and the structure of the relationship between ÖSS score and university admission. On the other hand, certain areas could well exhibit risk loving behavior: students must score above 120 in order to be qualified to make a preference submission to a four year college program.

As such, we stipulate that students have a cutoff for the belief. If the student believes that the best answer (highest signal) has a probability of being correct greater than the cutoff, he will attempt the question, choosing the best answer. However, if all answers have a probability lower than this cutoff, then he will skip the question. This cutoff lies in the interval [0.2 1].[19] A higher value for the cutoff implies a higher degree of risk aversion, while a cutoff of 0.2 would be supported by risk neutral/risk loving preferences.

[19]There will always exist an answer with probability of being correct greater than or equal to 0.2, therefore we do not consider cutoffs below 0.2, as they would result in the same behavior: always attempting the question, never skipping

Consider a student with ability parameter β and attempt threshold $c \in (0.2, 1)$. From these two parameters, we are able to calculate the probability that they would skip a given question, the probability of answering correctly, and the probability of answering incorrectly.

In order to answer a question, with answer n, the signal drawn for answer n, x_n, must satisfy two conditions. First, it must be the highest signal. Second, it must be high enough, given the other signals, so that the belief is greater that the cutoff required to attempt the question. We define the following function as the minimum signal x_n required to attempt with the n^{th} answer, given the other signals:[20]

$$K(\{x_i\}_{i\neq n}) = \max\Big(\max\{\{x_i\}_{i\neq n}\}, \Big(\frac{c}{1-c}(\sum_{i\neq n} x_i^{\gamma})\Big)^{1/\gamma}\Big) \tag{11}$$

Suppose that answer number 1 is the correct answer. The chance that answer number 2 is selected by the student, that is, provided as the answer, is:

$$\int_{x_5=A}^{\infty}\int_{x_4=A}^{\infty}\int_{x_3=A}^{\infty}\int_{x_1=A}^{\infty}\int_{x_2=K(x_1,x_3,x_4,x_5)}^{\infty} 1 dG(x_2)dF(x_1)dG(x_3)dG(x_4)dG(x_5) \tag{12}$$

So that the chance of the student submits an incorrect answer is the value of the above equation multiplied by the four possible incorrect answers. Similarly, the probability that the student submits a correct answer (in this case, answer number 1) is:

$$\int_{x_5=A}^{\infty}\int_{x_4=A}^{\infty}\int_{x_3=A}^{\infty}\int_{x_2=A}^{\infty}\int_{x_1=K(x_2,x_3,x_4,x_5)}^{\infty} 1 dF(x_1)dG(x_2)dG(x_3)dG(x_4)dG(x_5) \tag{13}$$

The probability that the student skips the question can be obtained similarly, by finding for each answer the probability that it gives the highest signal, yet is below the threshold to attempt.

These lead to three functions that describe the probabilities of each of the three possible outcomes of a question, conditional on student ability β, and cutoff c:

$$\text{Prob(Correct)} = P_C(\beta, c) \tag{14}$$

$$\text{Prob(Wrong)} = P_W(\beta, c) \tag{15}$$

$$\text{Prob(Skip)} = P_S(\beta, c) \tag{16}$$

[20] A diagram showing choices conditional on signal observations for a simplified two answer setup is shown in Figure 12

where $P_S(\cdot) = 1 - P_C(\cdot) - P_W(\cdot)$. Table 3 provides these, in addition to the average points earned per question, for various parameter values.[21] Consistent with the literature, as the probability to skip increases, the average points per question decreases (for a fixed ability).[22]

In each exam, the student faces 45 questions, with signals and outcomes independent across all questions in the exam. From this, we can find the probability that the student attempts $x \in \{0, ..., 45\}$ questions, skipping $45 - x$ questions:

$$\text{Prob(Answer } x \text{ questions)} = \binom{45}{x} (P_C + P_W)^x (P_S)^{45-x} \tag{17}$$

Conditional on answering x questions, the probability that $y \in \{0, ..., x\}$ questions are answered correctly is:

$$\text{Prob(Answer } y \text{ of } x \text{ questions correctly)} = \binom{x}{y} \left(\frac{P_C}{P_C + P_W}\right)^y \left(\frac{P_W}{P_C + P_W}\right)^{y-x} \tag{18}$$

A student that attempts x questions, correctly answering y questions, achieves a score in that exam of:

$$\text{Score}(x, y) = y - \frac{(x - y)}{4} \tag{19}$$

Accordingly, we can find the probability that a student with ability β and cutoff c obtains a score of s. Suppose that there are k possible ways of obtaining such a score: (y_j correct, $(x_j - y_j)$ incorrect, $(45 - x_j)$ skipped) where $j = 1, ..., k$. Thus, we obtain a mapping from (β, c) to the probability of getting score s:

$$\text{Prob}(Score = s) = M(\beta, c; s) \tag{20}$$
$$= \sum_{j=1}^{k} \binom{45}{x_j} (P_C + P_W)^{x_j} P_s^{45-x_j} \binom{x_j}{y_j} \left(\frac{P_C}{P_C + P_W}\right)^{y_j} \left(\frac{P_W}{P_C + P_W}\right)^{y_j - x_j}$$

5 Estimation Strategy

In our model, students' scores depend on students' ability (β) and risk aversion cutoff, c. In our data set we observe student's scores, but not the decomposition. In this section we use our model to estimate the distribution of ability and risk aversion cutoff.

[21] A β of 3 is later found to be approximately median.

[22] Of course the average points per question *attempted* increases.

Estimation of the parameters of interest, distribution of student ability (β_T, β_{SS}) and risk aversion cutoff c, is conducted jointly for each gender, and attempt number. In addition, we recognize that the relationship between ÖSS-SÖZ score and utility is not necessarily constant throughout the range of scores: the degree of risk aversion may be different. In particular, we could expect that students anticipating low scores would be considerably less risk averse, since scores below a cutoff result in the same outcome: an inability to submit preferences/apply to universities. This would result in a jump in the payoff function as students cross the cutoff score.

As a result, while we allow cutoffs to vary by gender and attempt number, we also allow cutoffs to depend on the interval in which the student's predicted ÖSS-SÖZ score lies, for example 120-130. To accomplish this, we first regress ÖSS-SÖZ on GPA (adjusted for school quality)[23], education level of both parents, monthly income of parents, and preparation on the four subject areas. We use the results of the regression to derive fitted values of ÖSS-SÖZ, predicted exam scores given observable characteristics, for each student. This estimation is conducted separately for each gender/attempt number.

We divide students into groups, according to gender, attempt number, and the range into which their predicted ÖSS-SÖZ score lies: $(0, 90)$, $[90, 100)$, $[100, 110)$, $[110, 120)$, $[120, 130)$, $[130, 140)$, and $[140, \infty)$. For each group, we examine the two subjects jointly.[24] While these intervals may not contain equal numbers of students, it will allow us to make comparisons across genders and attempt numbers. For each group, we take the cutoff c, and the distribution of ability β within the group. The ability of each student in subject k is given by $1 + e^{X_k}$, where (X_T, X_{SS}) is distributed normally with mean $\mu = (\mu_T, \mu_{SS})$ and variance matrix Σ. This ensures that each student has an ability in both subjects greater than 1, and results in a log normal shaped distribution (shifted 1 unit to the right).

Under the assumptions we made, the probability to get each score is approximated through simulation. For student s, we take a draw from $N(\mu, \Sigma)$ and label the vector as X_s. From X_s, we find $(\beta_{T,s}, \beta_{SS,s}) = (1 + e^{X_s(1)}, 1 + e^{X_s(2)})$, the student's ability vector. As we now have (β_T, β_{SS}, c) for student n, we can find the probability the student obtains score x in subject k, which is defined as $M_n(\beta_k, c; x)$ in the previous section. By taking a random draw from the joint distribution of

[23]To adjust for school quality, we adjust the GPA of student within a school based on the performance of the school in the exam. We observe normalize GPA for each students, which is able to be converted to a ranking within the school. As we observe the mean and variance of exam scores for each school, we can easily convert the GPA to a measure that reflects the quality of the school.

[24]The only tests of interest are Turkish and social science

scores, we can generate the simulated student's test outcome o: the Turkish score and social science score.

In order to find the relevant parameters for the group (cutoff, means of X_T, X_{SS}, variances of X_T, X_{SS} and covariance between X_T and X_{SS}), we use simulated method of moments. We compare simulated test scores to those observed in the data. Specifically, we look at moments related to the intensity of the spikes, and the shape of the distribution. More specifically, the difference between the mass of students with scores corresponding to attempting all questions (i.e. 45, 43.75,...) and the mass of students with scores corresponding to skipping a single question (i.e. 44, 42.75,...) captures in the intensity of the spikes. For example, $0.4 - 0.3 = 0.1$. If the spikes are very prominent, this difference will be large; if they are non-existent, this difference will be minimal. In addition, we use the mean of scores, and the variances/covariances of scores.

More formally, moments of an outcome o [25] is given by:

$$m(o) = (o(1), o(2), I(o), o(1)^2, o(2)^2, o(1) * o(2)) \tag{21}$$

where $I(o)$ is equal to 1 if the score corresponds to attempting every question, -1 if skipping a single question, and zero otherwise.

Accordingly, the estimates of the cutoff c and ability distribution parameters μ, Σ for each group are estimated by minimizing the distance between the simulated moments and the observed moments.

$$\hat{c}, \hat{\mu}, \hat{\Sigma} = \hat{\theta} = \arg\min_{\theta} \Big[\sum_{t=1}^{T} \Big(m(o_t) - \frac{1}{S} \sum_{s=1}^{S} m(o(u_t^s, \theta)) \Big) \Big]' W_T^{-1} \Big[\sum_{t=1}^{T} \Big(m(o_t) - \frac{1}{S} \sum_{s=1}^{S} m(o(u_t^s, \theta)) \Big) \Big] \tag{22}$$

where T is the number of observations in the data, TS is the number of simulated draws, and W_T is the weighting matrix.

With the identity matrix used as the weighting matrix, we obtain an estimate of the parameters of each group that is consistent and asymptotically normal. Applying the two step procedure,(Hansen [1982], Gourieroux and Monfort [1997], Duffie and Singleton [1993]) this estimate is used to generate a weighting matrix. Using the new weighting matrix, the procedure is repeated, and a consistent and asymptotically normal estimate is obtained.

[25] o is (Turkish score, social science score)

5.1 Identification

Identification of the risk aversion cutoff, c, is achieved through matching the intensity of the spikes. For example, if students are risk averse then they will tend to skip. Thus, at low values of c, students will have a very low probability of skipping a question: it is unlikely that the answer with the highest signal has a low enough probability of being correct to be below the risk aversion cutoff. As a result, almost all of the probability mass of a given student's distribution will be located on scores corresponding to attempting all questions. As the risk aversion cutoff increases, students become more and more likely to skip *some* questions, resulting in more mass lying on scores unreachable by attempting all questions (i.e. some questions must be skipped), while the spikes still remain prominent. Increasing the risk aversion cutoff further results in enough skipping activity so that spike cannot be seen.

This is illustrated in figure 13, where the score distribution for a student (with a fixed, approximately median, ability of $\beta = 3$) is shown for various cutoff levels. A cutoff of $c = 0.225$ gives virtually all of the mass to the attempt all scores. As the risk aversion cutoff increases to 0.3, the spikes all but disappear.

The relationship between the intensity of the spikes and the risk aversion cutoff is not constant. For a fixed cutoff c, as we increase ability, the intensity of the spikes increases. While low ability students might have a high chance of having a highest belief below the risk aversion cutoff, it becomes increasingly rare as we move to the high ability students.

The parameters of the distribution of the ability of a group of students, β_X and σ_X^2, are identified by the distribution of scores. An increase in the mean parameter β_X moves the score distribution to the right, increasing the mean, while an increase in the variance parameter σ_X^2 increases the variance of the score distribution. This is due to a strong relationship between ability and exam scores. Similarly, the covariance between ability in Turkish and social science is obtained through the correlation of scores.

6 Results

Table 4 contains the estimates of the risk aversion cutoff, the belief regarding probability of success below which a student will skip a question, for the various groups,[26] in addition to the standard

[26]Estimates for the second time takers in the ÖSS score range less than 90 are not obtained due to insufficient observations.

errors of the estimates.

Two facts are apparent. While males and females have roughly similar cutoffs, males tend to have lower risk aversion cutoffs, especially for students whose predicted score is above the threshold that allows them to submit a preference list. This is even more so among second time takers. This is in line with the literature - males are acting in a less risk averse manner. Secondly, the cutoff decreases systematically in the score range below 120. This matches what we know about the payoff structure. For low scores, students should be much less risk averse since any score below 105 will not allow the student to submit preferences for any school, and any score below 120 will not permit the student to submit preferences for four year college programs. Above 120, the cutoff remains relatively high: a lower score could see the student forced to attend a much less desirable institution.

Also noteworthy is the observation that the risk aversion cutoffs tend to decrease among high scoring students between first time takers and second time takers, whereas they tend to increase among the low scoring students. This pattern may be due to selection; high scoring risk loving students may self select into taking the exam a second time, while risk averse poor performing students may have dropped out of the system. However, it goes beyond the scope of this paper to disentangle the reasons behind this pattern.[27]

Figures 14 through 17 show the simulated distributions compared to observed distributions for the various groups. While the estimation procedure was designed only to match subgroups of the sample, the entire simulated distribution fits the data relatively well, with some exceptions: it systematically under-predicts scores which correspond to skipping multiple questions.[28] In addition, the skipping behavior is overestimated among low scoring students - this is likely due to such students correctly anticipating their low expected score and acting accordingly, whereas in the estimation many of these are restricted to acting with a (high) cutoff corresponding to their fitted score.

Estimates of the parameters governing the distribution of ability for each group are presented in Table 5. Recall that ability is parametrized as $1 + e^X$, where $X \sim N(\mu, \sigma^2)$. The mean and standard deviation of X in each group are presented.

As predicted, groups that are predicted to have high exam scores have much better distributions

[27] Additionally, such a pattern is consistent with a model where students do not perfectly know their true ability, and update their belief after taking the exam, see Lemma 1.

[28] This issue is addressed in more detail in a later section.

of ability for both Turkish and social science. However, there is significant variance in the distributions, reflective of the fact that the fitted score is an imperfect measure of overall student ability. We see that females tend to have higher ability in Turkish, but lower ability in social science, when compared to males in the corresponding group. This implies that males tend to have a comparative advantage in social science.

In addition, we observe that males tend to have higher variance in their distribution of ability. In fact, the variance is greater for all groups. This has two interpretations. First, the distribution of abilities is more dispersed among males, which is also implies by the distribution of exam scores. There is another possibility, that the fitted ÖSS score is not as accurate for males. There could be a number of causes for this, such as GPA being a poor predictor of exam scores, etc.

Looking at the distributions of ability across the various groups, we see similar patterns. As shown in figure 18, second time takers have higher abilities across the distribution, for both social science and Turkish. However, we can conclude definitively why this is the case.

There are two possibilities for this difference in ability: selection and learning. It is possible that students tend to learn between their first and second attempts, so that they increase their ability in the social science and Turkish sections of the exam. However, it is also possible that the perceived change in the distribution is simply due to a selection effect. Consider the students who choose to continue. They could very well be different from students who choose not to retake the exam. It could be that the best students do not retake the exam: they are admitted into a university program and so have no reason to take the exam. On the other side of the distribution, the worst students may have very little incentive to return, as they have almost no chance of meeting the threshold required to apply to programs. If the second effect is important, selection could very well result in a better distribution of student abilities among second time takers.

In figure 19, we see how the genders differ the first time they take the exam. The lower portion of the social science ability distribution is indistinguishable, however males have a considerably better distribution for the top portion, compared to females. This is not the case with the Turkish portion - females are much better for all points in the distribution. This provides an interpretation of the observed differences in ÖSS-SÖZ scores. Males are overall worse at Turkish, but the best males make up for it in social science.

The second time takers exhibit a similar pattern in figure 20; however the advantage of males amongst the top social science students is more pronounced, and their disadvantage in Turkish

decreases at the high end. As seen in the ÖSS-SÖZ score distributions[29], the comparison between male and female low ability students is similar to that of first time takers; whereas the high ability males gain a more favorable distribution (relative to females) in the second attempt.

7 Counterfactuals

Having recovered the parameters regarding the risk aversion exhibited by students in the multiple choice tests, in addition to estimates regarding the distribution of ability (as measured by β for each subject, the parameter in the Pareto distribution that governs dispersion of signals), we are now able to perform counterfactual experiments.

In these experiments, we will compare outcomes of a number of testing regimes, and student behaviors. For example, how would exam outcomes differ if all students attempted (answered) every question, as would happen if the penalty for answering incorrectly were removed? This is relevant because it is fully feasible to change the testing regime, and there is the possibility that the regime has an effect on the outcomes: males and females, first and second time takers, act differently. In addition, the rationale behind penalties is to reduce the amount of random guessing, therefore reducing score variance and improving the effectiveness of tests.

The objects of interest in these experiments are as follows:

- The relationship between ability in social science (β_{SS}) and social science section exam score percentile
- The relationship between ability in Turkish (β_T) and social science section exam score percentile
- The relationship between gender and social science section exam score percentile
- The relationship between attempt number and social science section exam score percentile

The first two objects are clearly important. The ÖSS exam system is an allocation mechanism, presumably designed to give the students with high abilities access to the most desirable institutes of higher education. However, the exam score is an imperfect measure of student ability. The students who score in the top one percent are not necessarily those in the top one percent as

[29] See figure 2

measured by ability. The testing regime, and restrictions on behavior, may affect the dispersion of possible scores for students, affecting how precisely the system identifies the best students.

The third relation of interest, gender vs. exam scores, is also important. It is recognized in the literature that males are less risk averse than females in test situations (see Eckel and Grossman [2008]). In support of this, we have found that females tend to have higher thresholds for attempting to answer a question, i.e. they are more risk averse. Since the testing regime in question is forcing students to accept an element of risk when choosing to answer a question, the preferences regarding risk affect the distribution of final exam scores. This may tend to favor male test takers, leading in essence to a systemic bias in the testing procedure. In addition, the distributions of abilities are considerably different across gender; the regime may end up attenuating these differences.

Finally, we also see different observed attitudes to risk across attempt numbers, as well as vastly different distributions of abilities. As a result, the regime may be able to influence the proportion of students that are first time takers in a given exam score percentile. While this paper does not go into much detail on the topic, the costs/benefits of delaying entry into university are likely to be important.

The seven possible regimes used in this counterfactual experiment are:

1. The baseline model, as estimated in the previous section

2. Preferences of males and females are switched, so that the cutoff used by a male i^{th} time taker in exam subject j, with ÖSS-SÖZ score interval k is switched with that used by a female i^{th} time taker in exam subject j, with ÖSS-SÖZ score interval k, and vice versa

3. Preferences of first and second time takers are switched, so that the cutoff used by a male 1^{st} time taker in exam subject j, with ÖSS-SÖZ score interval k is switched with that used by a male 2^{nd} time taker in exam subject j, with ÖSS-SÖZ score interval k, and vice versa

4. All students attempt all questions. This is equivalent to assuming that all students are risk neutral/loving, and identical to removing the penalty for answering incorrectly. Both would cause rational students to answer every question. Scores would, however, need to be rescaled to reflect the absence of such a penalty: instead of ranging from -11.25 to 45, they would range from 0 to 45.

5. Each question has only 4 answers to choose from, with the penalty for an incorrect answer adjusted accordingly

6. The penalty for answering incorrectly is increased from 0.25 points to 0.5 points

7. The penalty for answering incorrectly is increased from 0.25 points to 1 point

While the second is clearly eliciting the gender effect on outcomes, and the third the effect of experience (through test behavior), the fourth counterfactual seeks to examine the effect of having penalties for incorrect answers, as opposed to the simple, standard approach of a single point for each question answered correctly.

The fifth requires more explanation. In the default regime, there are five answers, with a single point for correct answers and a quarter point lost for incorrect answers. This results in an expected gain of zero from a random guess; accordingly, we set the penalty equal to one third of a point in the four answer scenario, resulting in a random guess having an expected gain of zero.

As a result, the cutoffs for attempting a question must be different. To convert cutoffs from the five answer case, we first assume a CARA utility function, and solve for the risk aversion parameter that generates a given cutoff. This is repeated for each student. We then convert the risk aversion parameter to a cutoff in the four answer case.[30]

The sixth counterfactual is designed to elicit more skipping from students, in order to increase the impact that differences regarding risk preference have on exam outcomes. The seventh continues, increasing the penalty even further. Similar to the four-answer counterfactual, new cutoffs are obtained for both counterfactuals.

For each of the seven possible regimes, we find the resulting distributions of scores for the entire sample of students, and segment scores into bins of five percent.[31][32] For each of the twenty bins, ranging from the lowest scores to the highest, we find four objects of interest: share of males, share of first time takers, average social science ability and average Turkish ability.[33]

Figures 21 through 24 show how the four objects of interest differ across the different regimes. As expected, given their greater exam score variance, the male fraction is u-shaped, as shown in figure 21. However, there is no discernible pattern in the differences between the seven regimes throughout most of the range. While there are some small differences around the median, all seven are fairly similar. We do see some small differences in the top performing students: "No Penalty"

[30]For example, a cutoff of 0.240 in the five answer case implies risk aversion coefficient of 0.383 (CARA utility), which results in a cutoff of 0.300 in the four answer case.

[31]Five percent of the number of students

[32]The rationale behind segmenting into percentiles, not scores, is to see the effects on the resulting allocation of students to university programs.

[33]The average of $\log(\beta)$ is used for each subject

gives a lower Male fraction (the more abundant Males are seemingly disadvantaged by the added variance in scores) whereas the "Penalty: 1" regime has a very slightly higher Male fraction, i.e. the small difference in risk aversion begins to have a small effect. Of course an alternative explanation is that the increased penalty increases the accuracy of the test, reducing the dulling effect that random guessing has on the prevalence of males in the top part of the distribution. The increased prevalence of males at the lowest percentiles when the penalty is equal to 1 supports this notion.

This insensitivity is even more apparent when examining the other graphs. Second time takers are dominant in the high exam scores percentiles, as would be expected, with the seven curves lying on top of each other in figure 22.

Figures **??** and 24 have higher ability students in the higher score percentiles, with no differences across the six cases featuring risk aversion. We do, however, see that the "Attempt All" regime gives slightly lower average abilities amongst the highest performing students, consistent with penalties and risk aversion allowing an improved separation of students by abilities. Similarly, the increased penalty regimes appear to be more effective at separating students by ability.

The impact of the increased penalties on average abilities of combined score quantiles is most evident for the top quantiles. Although the difference is small, it is not negligible. We can in fact quantify the increased separation of students by ability. For the top $x\%$ of students (by combined score), we can find the average log ability in the two subjects, both for the baseline model and the counterfactual with quadrupled penalty. For small x, these numbers will be lower in the baseline model. We can then extend the baseline model, increasing the amount of questions, and then see how the quantiles compare. Specifically, we examine the case where $x = 13.5\%$, as 13.5% of first and second time takers in the social science track are admitted to university. We find that an additional 25 questions (70 in total) must be asked in each section in order for the baseline model to have a comparable admitted class, compared to the 45 question, quadrupled penalty version.[34] So it seems that the increased validity resulting from the increased penalty is significant from a practical point of view.

Although the seven regimes can lead to considerable different score distributions, the relationship between gender, attempt number and ability, and exam score percentiles is relatively invariant. The reasoning for this is relatively straightforward. While there may be differences in attitudes to risk, and resulting test taking behavior, the implications of these differences happen to be rather

[34] Alternatively, if the penalty were quadrupled, the number of questions in each section could be reduced to only 27 yet would retain equivalent validity

small due to the characteristics of situations when these differences are relevant. While a difference in the cutoff of 0.23 versus 0.25 may be considerable given that the risk neutral cutoff is 0.2, and implies considerably different attitudes to risk, the effect on scores is small for two reasons. Firstly, there is a relatively low chance that a student has a belief lying between 0.23 and 0.25 for a given question. Secondly, if the belief does lay in that region, the expected gain from answering (and hence that from having a low cutoff) is at most 0.0625 points. Even when the penalty is raised, leading to more skipping behavior, the total effect on allocations is minor. Essentially, differences in skipping behavior are not common, and are restricted to certain situations; in these situations there is little difference between skipping and attempting.

However, a degree of caution should be applied when applying this result to other tests with different students. Here, the lack of an effect is the result of a relatively low degree of risk aversion overall, in addition to an exam where students are able to be confident enough to answer a vast majority of the time. While there is no obvious reason why the first might be particular to this group of students, it is very reasonable to suggest that the second depends very much on the style of the exam, questions asked and so on.

8 Conclusions

This paper investigates the factors that affect students' exam taking behavior in multiple choice tests. By constructing a structural model of a student's decision to attempt/skip a question in a multiple-choice exam, we estimate the risk aversion cutoff and ability distributions of students. We do so by dividing students into different groups according to their gender, experience in the exam, and the predicted ÖSS score, which depends on their background characteristics, high school performance, and the quality of the high school which they attended. Crucially, we allow different groups of students to have a different risk aversion and ability distribution in each part of the test.

Our results suggest that there are significant differences in different groups in the way they approach the exam. Female students act in a more risk averse manner in all groups relative to males. We also find that students with low expected scores have a lower risk aversion cutoff, which is consistent with the pay-off structure.

While our findings suggest that females behave in a more risk averse manner, which theoretically leads to a disadvantage in tests which impose a penalty, we find that differences have very little bearing on aggregate outcomes. In fact, imposing penalties primarily improves the effectiveness of

tests: separating the low ability students from the high ability students.

References

Akyol, S. P. and Krishna, K. (2014). Preferences, selection, and value added: A structural approach applied to turkish exam high schools.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., and Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*, volume 278. Economic Policy Institute Washington, DC.

Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science*.

Ben-Shakhar, G. and Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *The Journal of Educational Measurement*, 28(1):23–35.

Bernardo, J. (1998). A decision analysis approach to multiple-choice examinations. *Applied Decision Analysis*, IV:195–207.

Burgos, A. (2004). Guessing and gambling. *Economics Bulletin*, 4(4):1–10.

Duffie, D. and Singleton, K. J. (1993). Simulated moments estimation of markov models of asset prices. *Econometrica*, 61(4):pp. 929–952.

Eckel, C. C. and Grossman, P. J. (2008). Men, women, and risk aversion: Experimental evidence. *Handbook of Experimental Economics*, 1(113):1061–1073.

Espinosa, M. P. and Gardeazabal, J. (2005). Do students behave rationally in multiple-choice tests? evidence from a field experiment. *mimeo*.

Espinosa, M. P. and Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, 54(5):415–425.

Frisancho, V., Krishna, K., and Yavas, C. (2012). Learning gains among repeat takers of the turkish college entrance exam. *mimeo*.

Gourieroux, C. and Monfort, A. (1997). *Simulation-based econometric methods*. Oxford University Press.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Tannenbaum, D. I. (2012). Do gender differences in risk aversion explain the gender gap in sat scores? uncovering risk attitudes and the test score gap. *mimeo*.

9 Appendix

Table 1: Test Weights

	Math	Science	Turkish	Social Science	Language
Science Track (ÖSS-SAY)	1.8	1.8	0.4	0.4	0
Social Science Track (ÖSS-SÖZ)	0.4	0.4	1.8	1.8	0
Turkish-Math Track (ÖSS-EA)	0.8	0.4	0.8	0.3	0
Language Track (ÖSS-DIL)	0	0	0.4	0.4	1.8

Table 2: Summary Statistics

Variable	Obs.	Mean	Std.Dev.	Min	Max
Gender (Male=1)	37650	0.567	0.496	0	1
ÖSS-SÖZ	37650	116.737	18.912	0	161.166
Normalized High School GPA	37650	47.550	7.825	30	80
Raw Turkish Score	37650	23.983	9.965	-10	45
Raw Social Science Score	37650	19.229	10.397	-8.75	45
Raw Math Score	37650	1.853	4.061	-9	42.5
Raw Science Score	37650	0.197	1.296	-8.75	25.75
Education level of Dad					
Primary or less	37650	0.557		0	1
Middle/High school	37650	0.281		0	1
2-year higher education	37650	0.028		0	1
College/Master/Phd	37650	0.050		0	1
Missing	37650	0.085		0	1
Income Level					
<250 TL	37015	0.459		0	1
250-500 TL	37015	0.393		0	1
500-750 TL	37015	0.095		0	1
750-1000 TL	37015	0.030		0	1
1000-1500 TL	37015	0.012		0	1
1500-2000 TL	37015	0.006		0	1

(continued on next page)

Variable	Obs.	Mean	Std.Dev.	Min	Max
>2000 TL	37015	0.005		0	1
Number of Attempts					
1st attempt	35470	0.261		0	1
2nd attempt	35470	0.250		0	1
3rd attempt	35470	0.242		0	1
4th attempt	35470	0.154		0	1
5th attempt	35470	0.092		0	1
Prep School Expenditure					
No prep school	37577	0.262		0	1
Scholarship	37577	0.008		0	1
<1000 TL	37577	0.223		0	1
1000-2000 TL	37577	0.095		0	1
>2000 TL	37577	0.034		0	1
Missing	37577	0.378		0	1
Time Spend in Preparation in 11th Grade					
Turkish Preparation	37650	0.439	0.874	0	3
Social Science Preparation	37650	0.465	0.905	0	3
Math Preparation	37650	0.301	0.653	0	3
Science Preparation	37650	0.125	0.393	0	3

Figure 1: Distribution of OSS-SOZ Scores

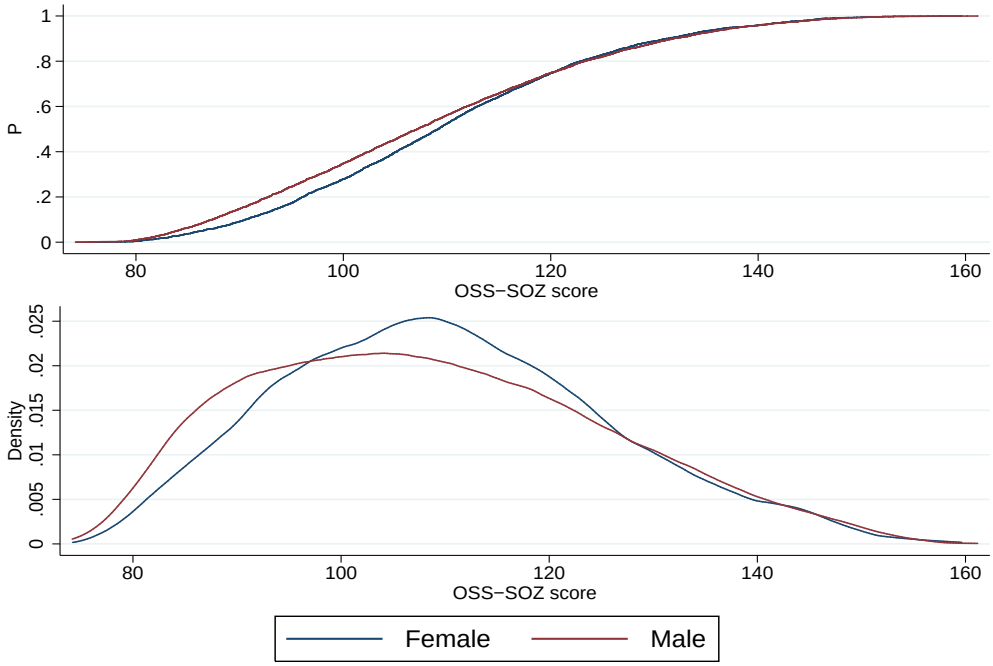

Figure 2: Distribution of OSS-SOZ Scores

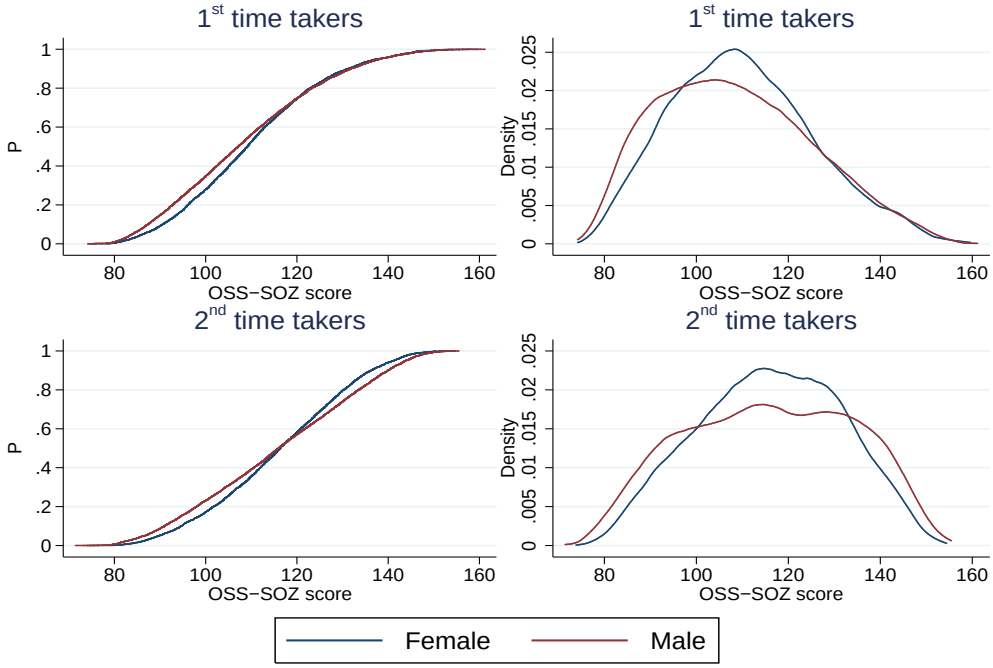

Figure 3: Social Science Test Scores (bins of width one)

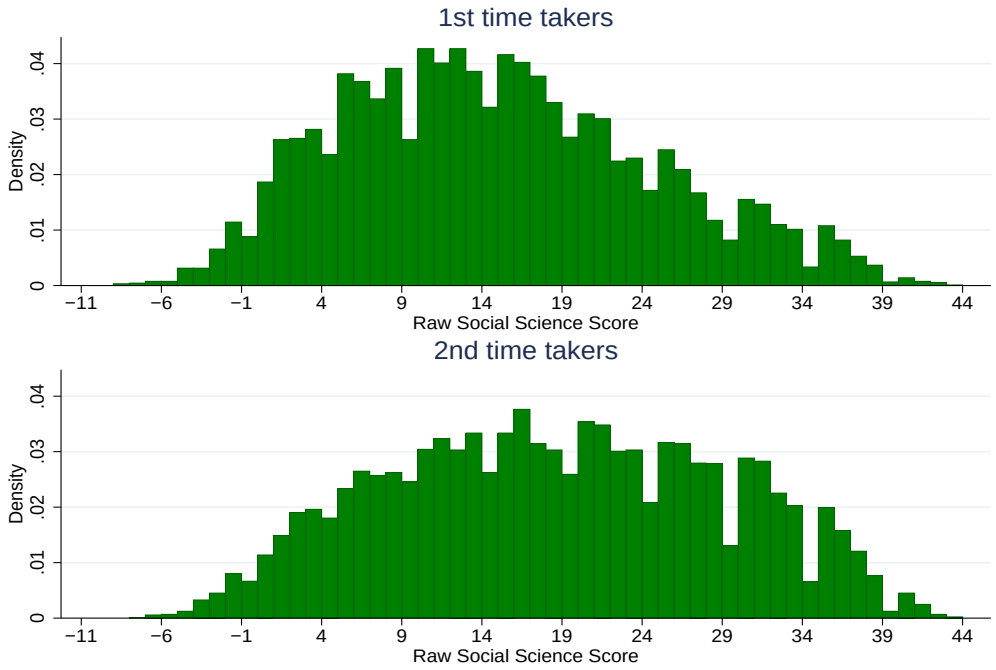

Figure 4: Turkish Test Scores (bins of width one)

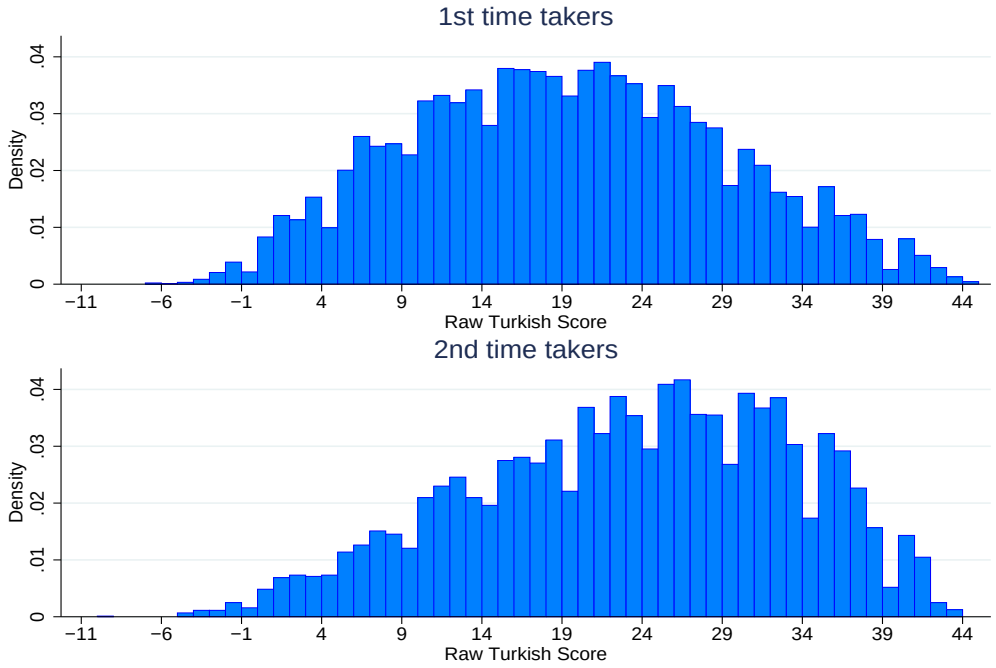

Figure 5: Math Test Scores (bins of width one)

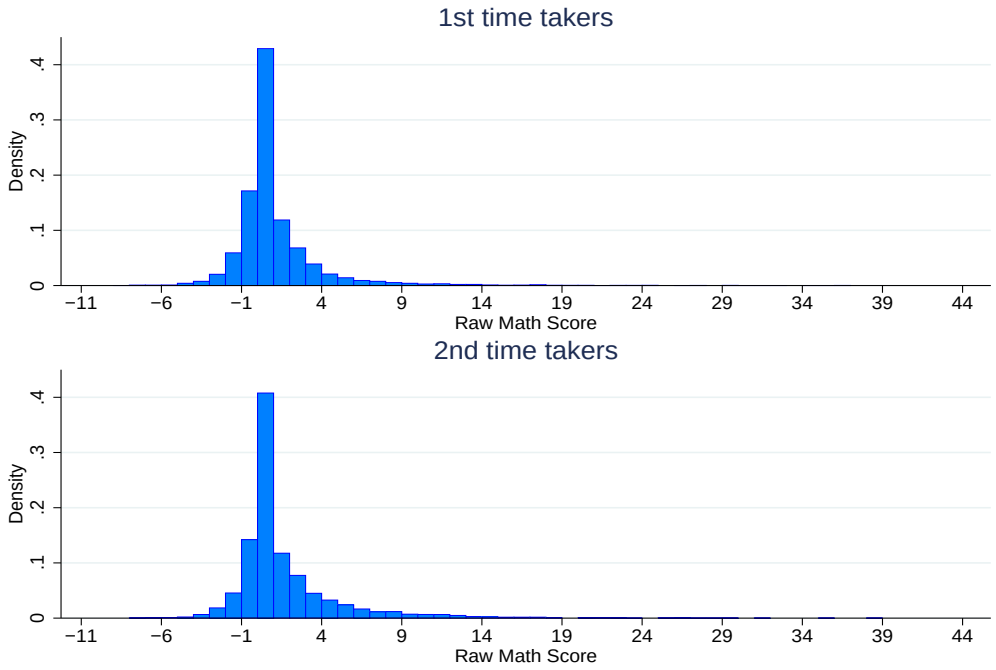

Figure 6: Science Test Scores (bins of width one)

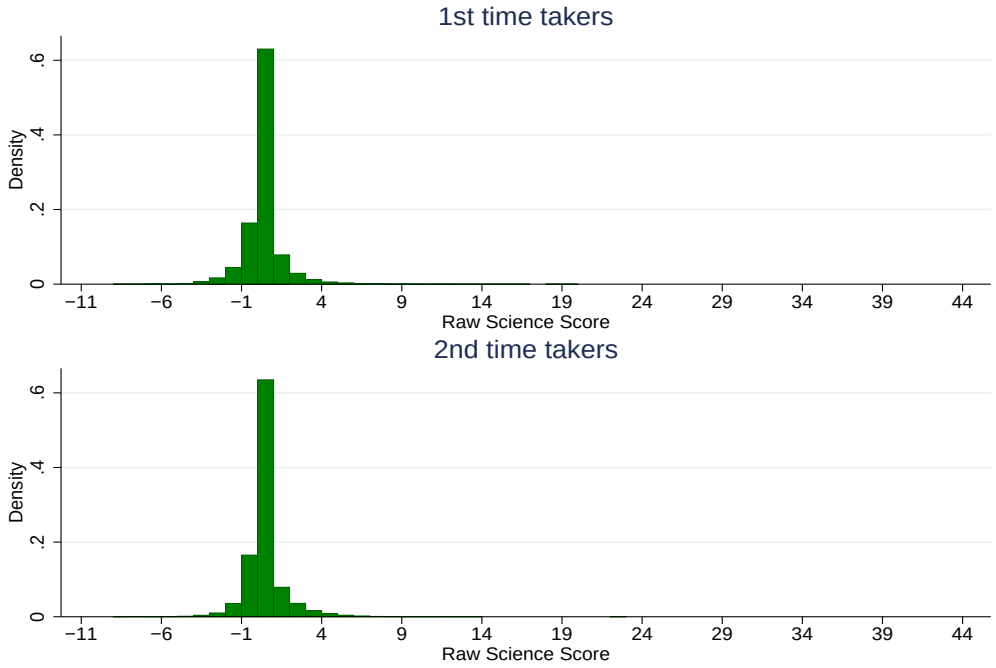

Figure 7: Distribution of Social Science Test Scores

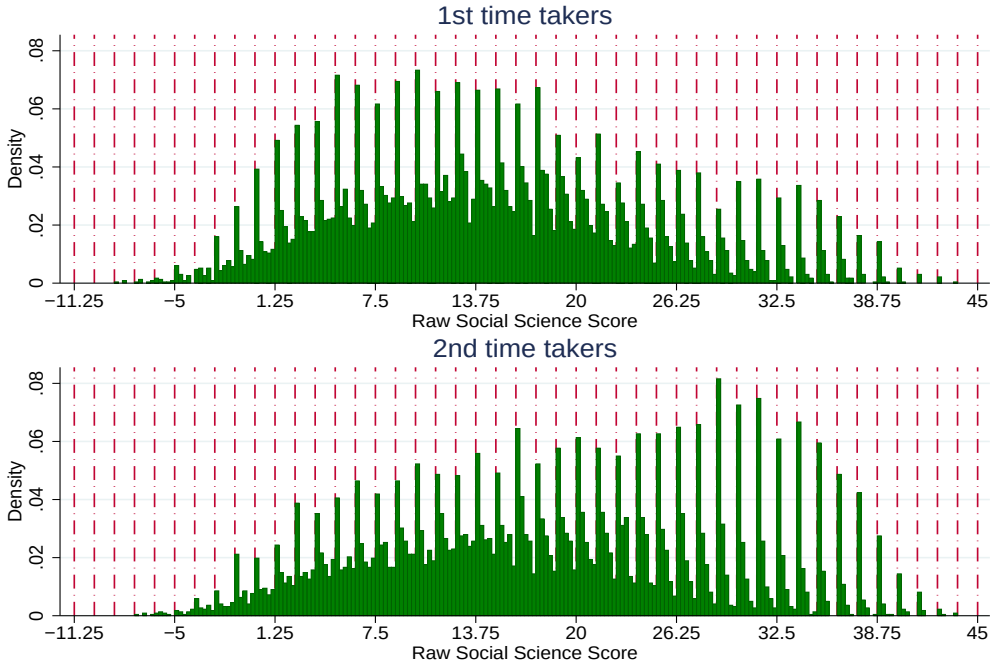

Figure 8: Distribution of Turkish Test Scores

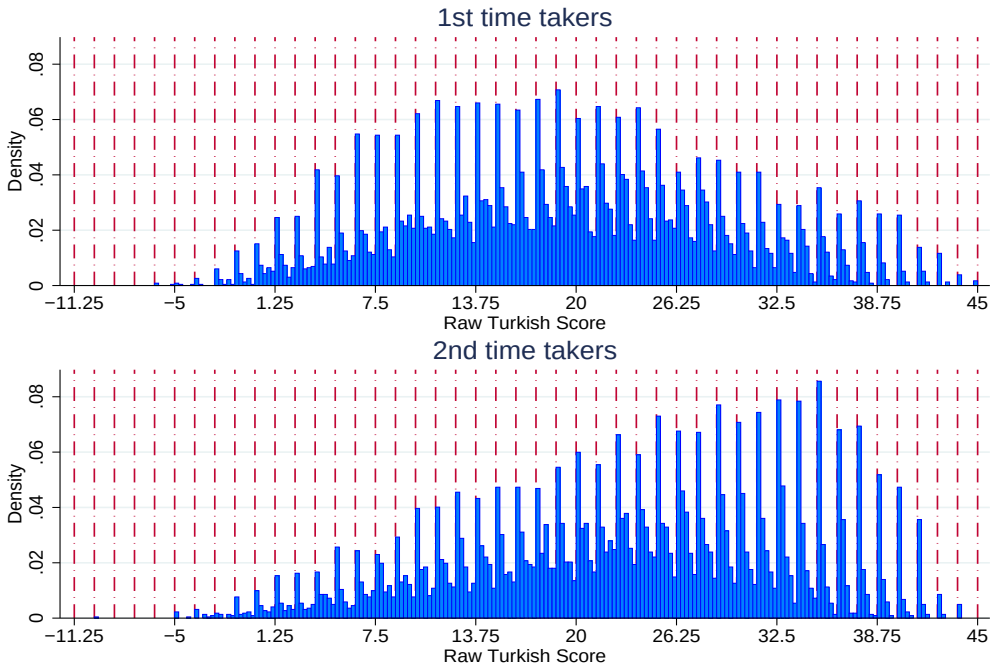

Figure 9: Distribution of Math Test Scores

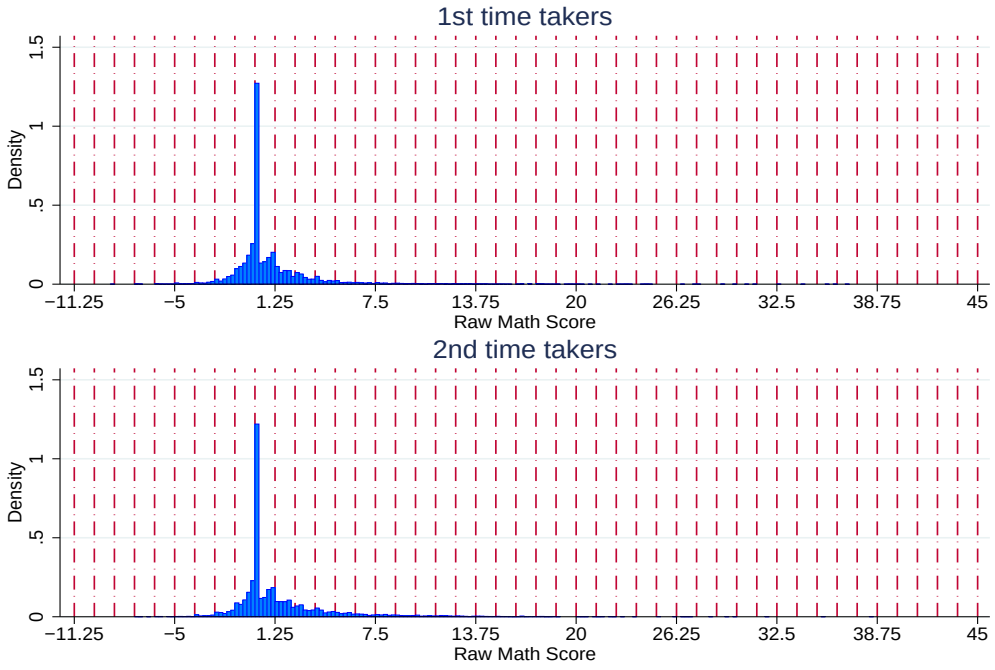

Figure 10: Distribution of Science Test Scores

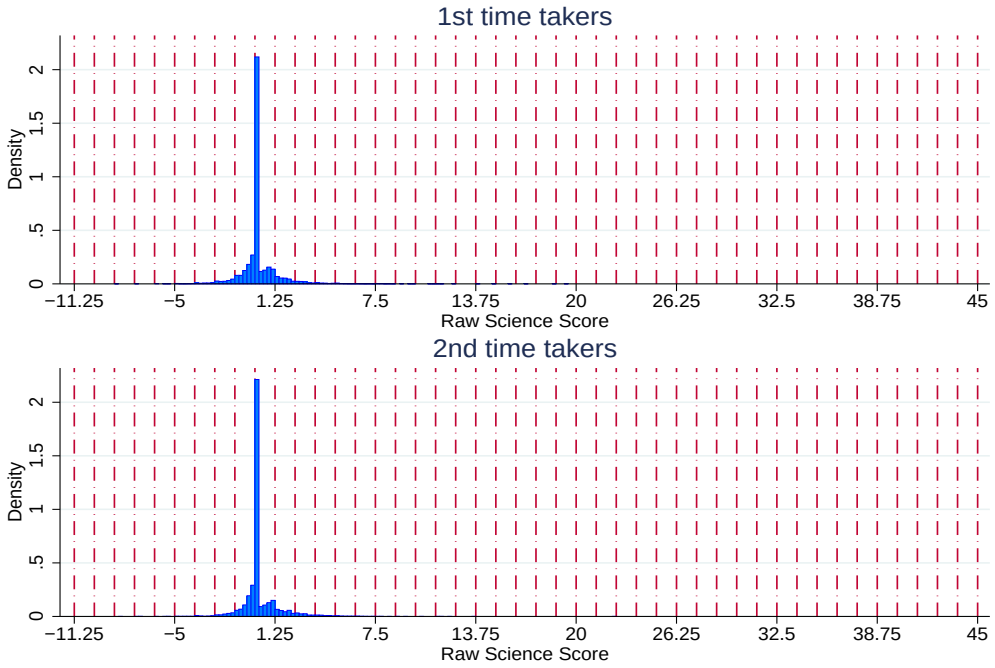

Figure 11: Distributions of signals for a student with $\beta = 3$, approximately median

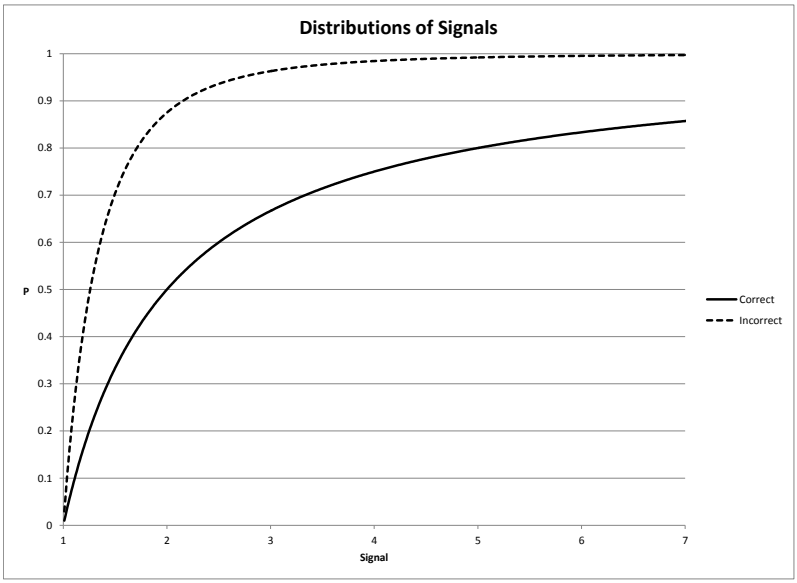

Figure 12: Action conditional on signals for a simple two answer model (parameter values: $\beta = 3$ and cutoff $= 0.55$

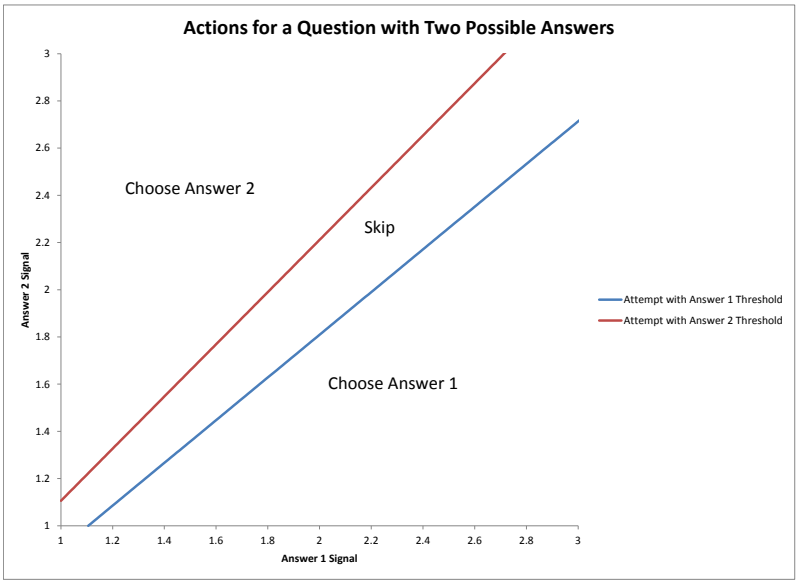

Table 3: Question outcomes for various parameter values: probabilities of skipping, being correct, being incorrect, and the average points per question

β	Cutoff	Prob(S)	Prob(C)	Prob(I)	PPQ
2	0.2	0	0.405	0.595	0.257
2	0.225	0.012	0.403	0.585	0.257
2	0.25	0.085	0.386	0.529	0.254
2	0.275	0.192	0.359	0.449	0.247
2	0.3	0.303	0.328	0.370	0.235
2	0.325	0.403	0.297	0.300	0.222
3	0.2	0	0.535	0.465	0.419
3	0.225	0.003	0.534	0.463	0.419
3	0.25	0.030	0.528	0.442	0.418
3	0.275	0.081	0.515	0.404	0.414
3	0.3	0.143	0.498	0.360	0.408
3	0.325	0.208	0.478	0.315	0.399
4	0.2	0	0.619	0.381	0.524
4	0.225	0.001	0.619	0.380	0.524
4	0.25	0.017	0.616	0.368	0.524
4	0.275	0.049	0.608	0.344	0.522
4	0.3	0.091	0.596	0.314	0.517
4	0.325	0.137	0.581	0.281	0.511

Figure 13: Distribution of scores resulting from various cutoff levels

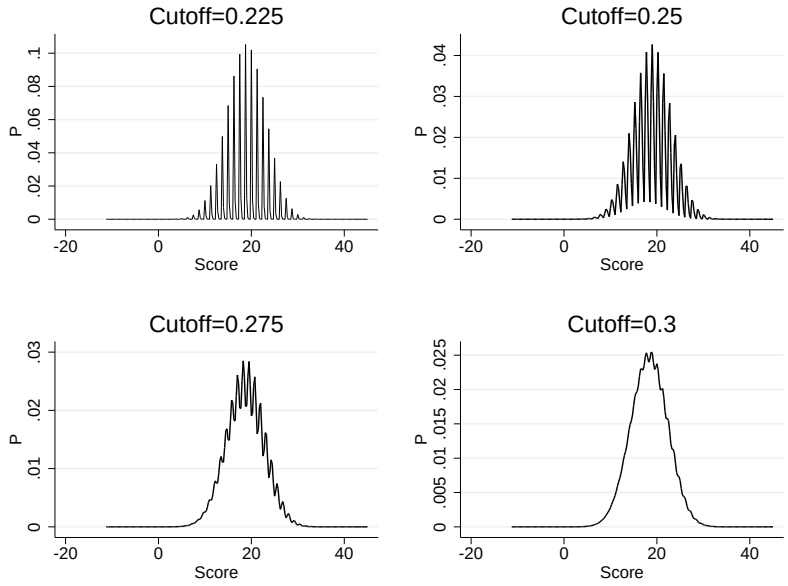

Table 4: Estimates of Risk Aversion Cutoff

	Female		**Male**	
	1st time takers	**2nd time takers**	**1st time takers**	**2nd time takers**
(0,90)	0.2429		0.2100	
	(0.0269)		(0.0026)	
[90,100)	0.2322	0.2330	0.2272	0.2388
	(0.0023)	(0.0373)	(0.0019)	(0.0658)
[100,110)	0.2396	0.2463	0.2364	0.2363
	(0.0009)	(0.0027)	(0.0010)	(0.0016)
[110,120)	0.2546	0.2529	0.2480	0.2456
	(0.0017)	(0.0013)	(0.0016)	(0.0012)
[120,130)	0.2612	0.2620	0.2594	0.2529
	(0.0037)	(0.0022)	(0.0043)	(0.0016)
[130,140)	0.2763	0.2677	0.2633	0.2562
	(0.0062)	(0.0043)	(0.0036)	(0.0020)
[140,∞)	0.2796	0.2773	0.2697	0.2596
	(0.0175)	(0.0146)	(0.0076)	(0.0096)

Standard errors are reported in parentheses.

Table 5: Estimates of Ability Parameters

	Social Science Test							
	Female				**Male**			
	1st time takers		**2nd time takers**		**1st time takers**		**2nd time takers**	
	μ	σ	μ	σ	μ	σ	μ	σ
(0,90)	-2.432	1.660			-3.031	1.775		
	(0.399)	(0.084)			(0.211)	(0.171)		
[90,100)	-0.702	0.807	-0.573	0.814	-0.620	0.866	-0.473	1.030
	(0.037)	(0.031)	(0.111)	(0.080)	(0.039)	(0.037)	(0.180)	(0.111)
[100,110)	-0.102	0.619	-0.035	0.777	0.068	0.751	-0.036	1.010
	(0.023)	(0.018)	(0.020)	(0.027)	(0.011)	(0.019)	(0.030)	(0.032)
[110,120)	0.509	0.549	0.469	0.690	0.694	0.618	0.564	0.944
	(0.023)	(0.016)	(0.023)	(0.023)	(0.025)	(0.017)	(0.027)	(0.015)
[120,130)	1.035	0.567	1.027	0.632	1.232	0.656	1.265	0.791
	(0.032)	(0.018)	(0.029)	(0.015)	(0.034)	(0.015)	(0.031)	(0.021)
[130,140)	1.495	0.516	1.409	0.582	1.742	0.644	1.774	0.838
	(0.055)	(0.034)	(0.028)	(0.035)	(0.053)	(0.033)	(0.046)	(0.040)
[140,∞)	2.020	0.389	1.764	0.424)	2.181	0.449	1.997	0.436
	(0.069)	(0.073)	(0.084)	(0.094)	(0.055)	(0.167)	(0.069)	(0.025)
	Turkish Test							
	Female				**Male**			
	1st time takers		**2nd time takers**		**1st time takers**		**2nd time takers**	
	μ	σ	μ	σ	μ	σ	μ	σ
(0,90)	-1.263	1.427			-1.726	1.417		
	(0.245)	(0.152)			(0.199)	(0.105)		
[90,100)	0.106	0.618	0.136	0.913	-0.141	0.730	0.128	0.796
	(0.027)	(0.022)	(0.108)	(0.099)	(0.024)	(0.024)	(0.091)	(0.102)
[100,110)	0.571	0.579	0.684	0.658	0.411	0.640	0.398	0.824
	(0.020)	(0.014)	(0.026)	(0.025)	(0.016)	(0.018)	(0.026)	(0.024)
[110,120)	1.118	0.514	1.175	0.650	0.909	0.516	0.890	0.817
	(0.022)	(0.011)	(0.022)	(0.020)	(0.022)	(0.012)	(0.023)	(0.013)
[120,130)	1.681	0.552	1.703	0.583	1.453	0.609	1.511	0.725
	(0.033)	(0.023)	(0.028)	(0.016)	(0.033)	(0.019)	(0.029)	(0.013)
[130,140)	2.256	0.538	2.104	0.640	2.062	0.666	1.972	0.779
	(0.058)	(0.043)	(0.035)	(0.051)	(0.055)	(0.035)	(0.045)	(0.048)
[140,∞)	2.978	0.553	2.409	0.618	2.541	0.571	2.368	0.536
	(0.135)	(0.148)	(0.146)	(0.091)	(0.082)	(0.086)	(0.077)	(0.042)

Standard errors are reported in parentheses.

Table 6: Estimates of Covariance between Turkish and Social Science Ability

	Female		**Male**	
	1st time takers	**2nd time takers**	**1st time takers**	**2nd time takers**
(0,90)	2.255		2.259	
	(0.189)		(0.324)	
[90,100)	0.472	0.636	0.555	0.688
	(0.027)	(0.149)	(0.038)	(0.170)
[100,110)	0.330	0.469	0.425	0.782
	(0.015)	(0.031)	(0.021)	(0.041)
[110,120)	0.268	0.438	0.275	0.722
	(0.006)	(0.025)	(0.009)	(0.017)
[120,130)	0.301	0.362	0.376	0.548
	(0.008)	(0.005)	(0.010)	(0.015)
[130,140)	0.246	0.353	0.393	0.616
	(0.015)	(0.044)	(0.042)	(0.063)
[140,∞)	0.186	0.229	0.200	0.225
	(0.018)	(0.037)	(0.116)	(0.031)

Standard errors are reported in parentheses.

Figure 14: Data vs simulated distribution: social science, first time takers

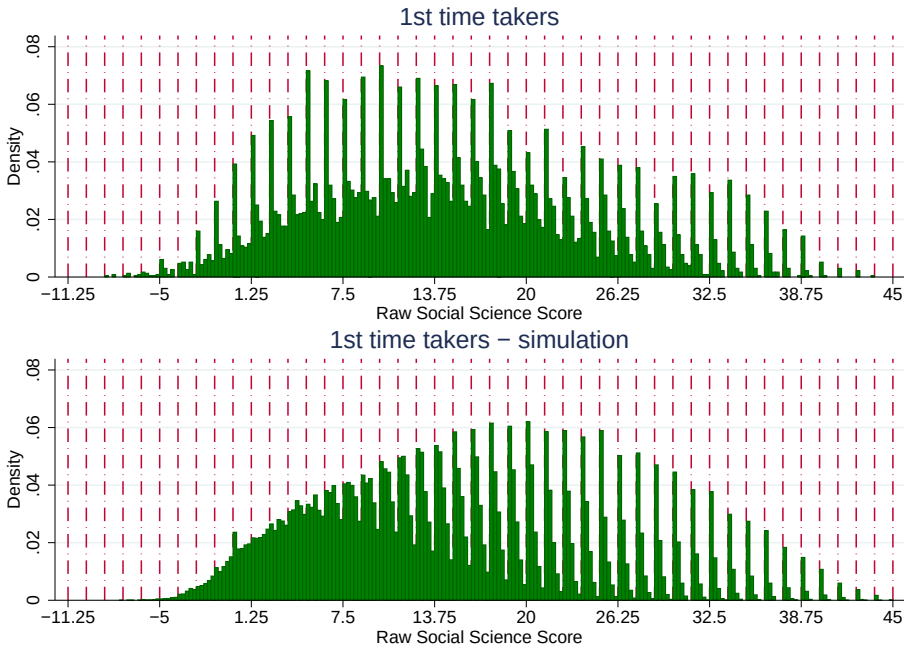

Figure 15: Data vs simulated distribution: Turkish, first time takers

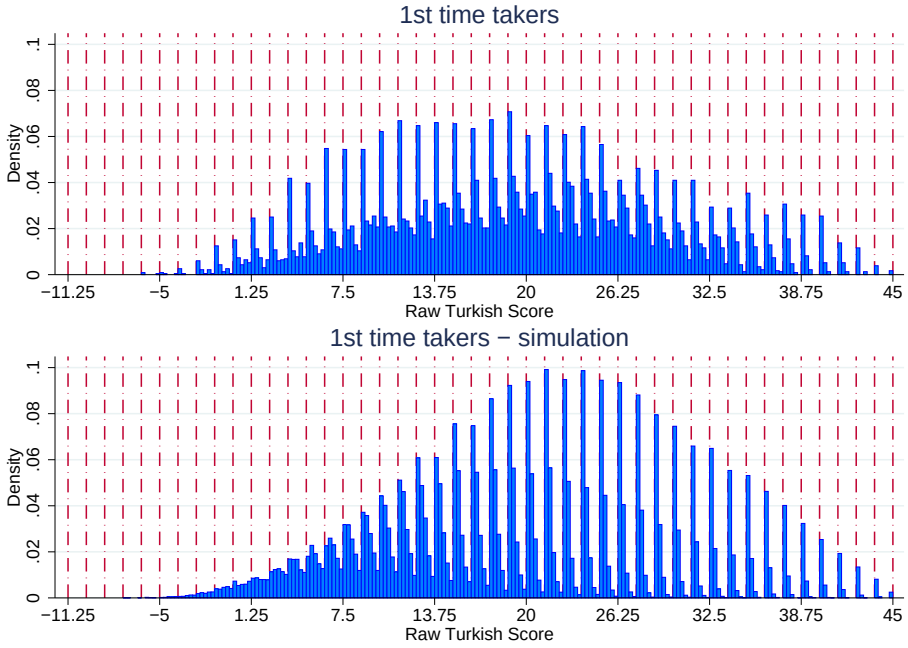

Figure 16: Data vs simulated distribution: social science, second time takers

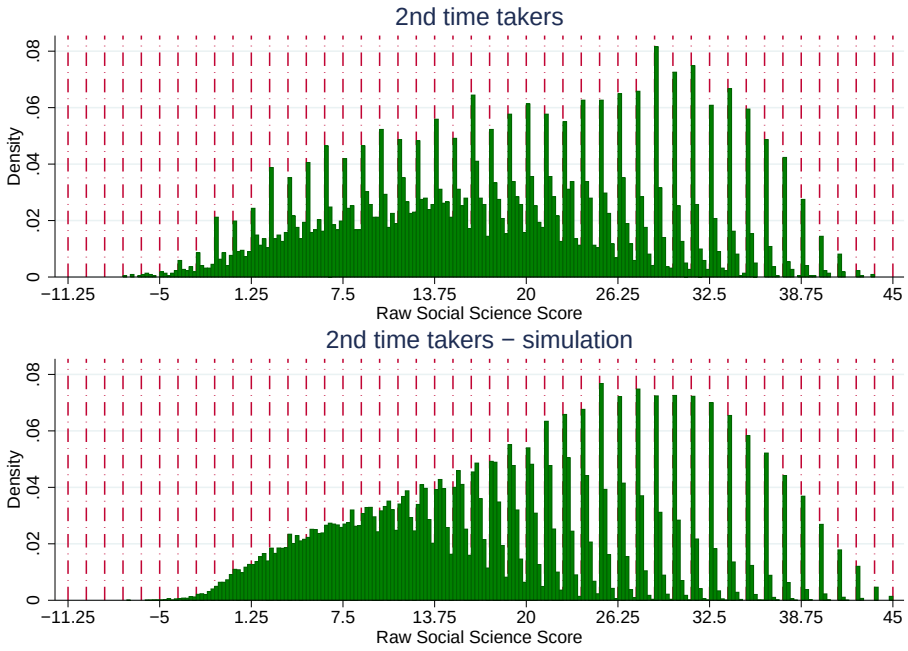

Figure 17: Data vs simulated distribution: Turkish, second time takers

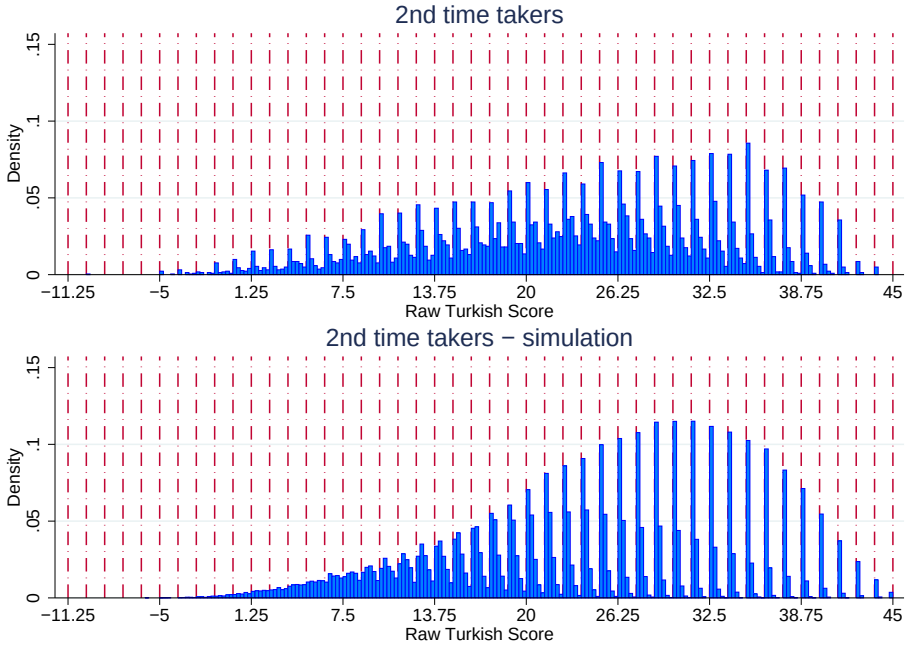

Figure 18: Distributions of Social Science and Turkish ability

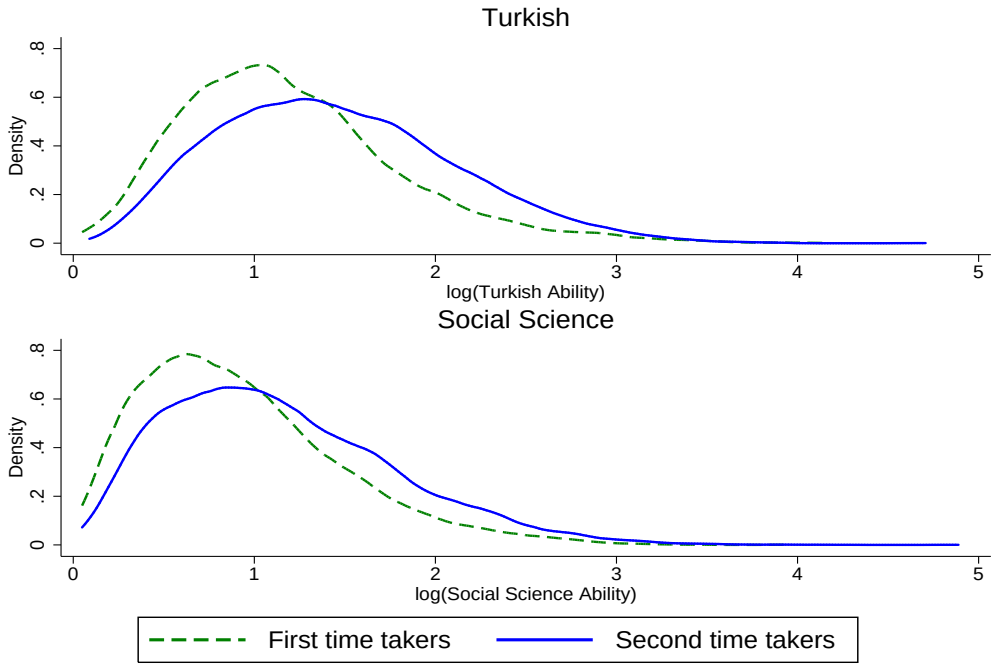

Figure 19: Distributions of Social Science and Turkish ability for First time takers

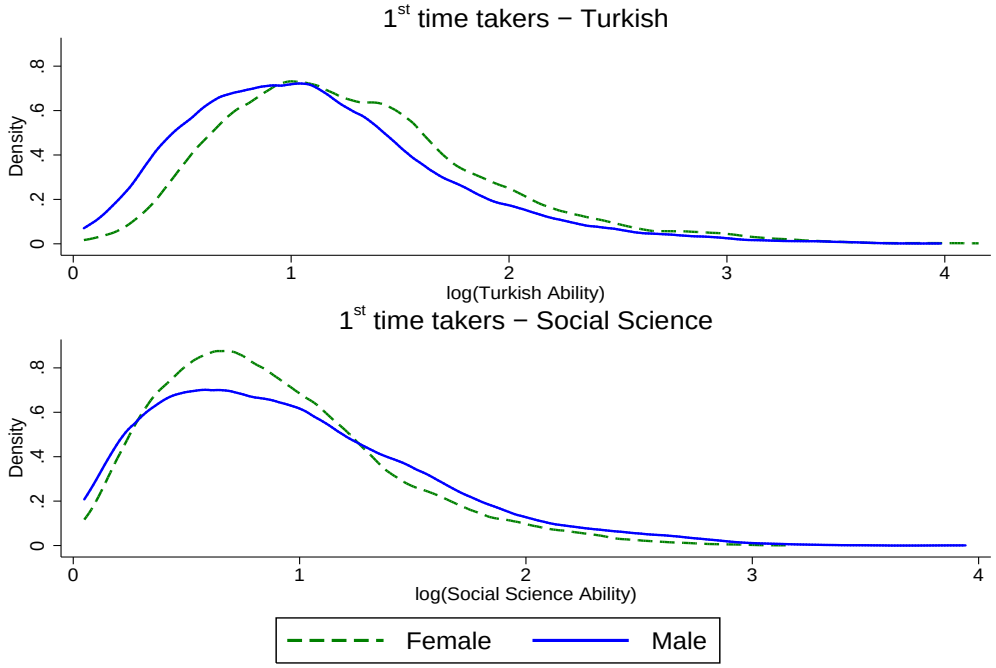

Figure 20: Distributions of Social Science and Turkish ability for Second time takers

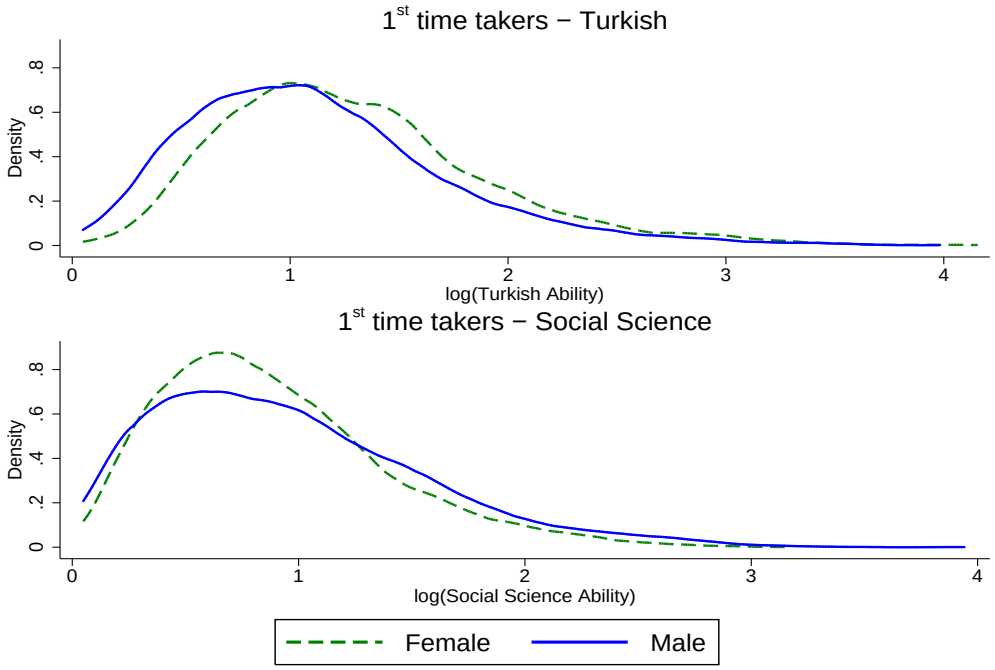

Figure 21: Counterfactual: Fraction of Males vs ÖSS-SOZ Score Quantiles

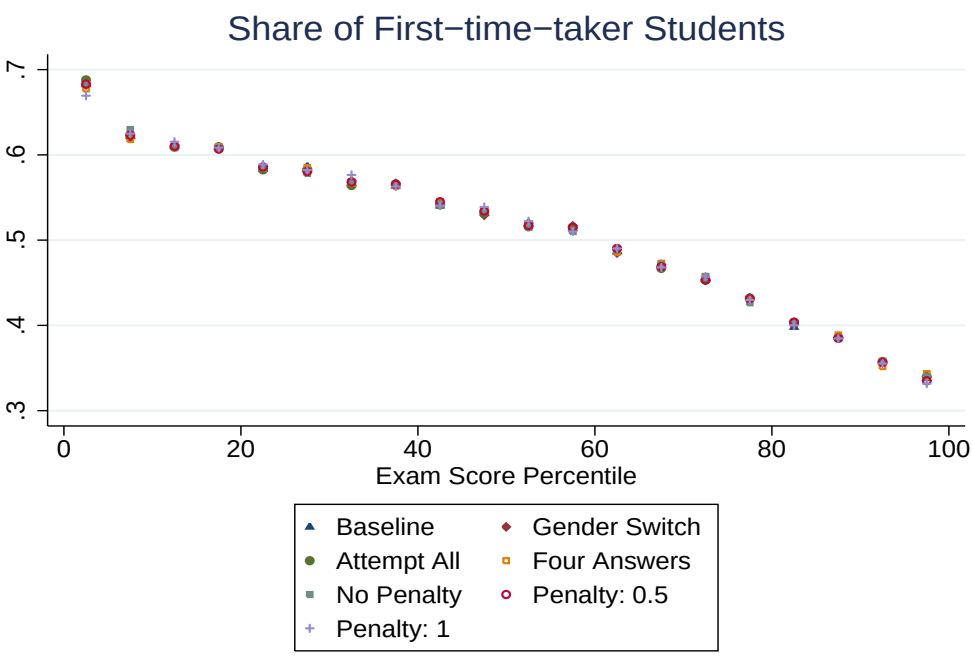

Figure 22: Counterfactual:Fraction of First-time-takers vs ÖSS-SOZ Score Quantiles

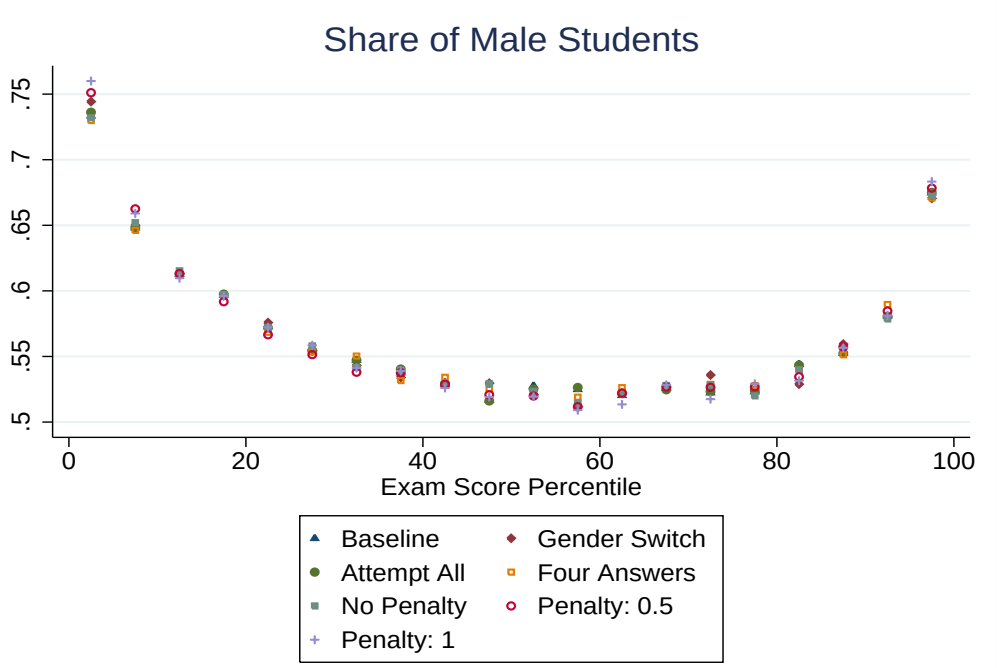

Figure 23: Counterfactual: Social Science Ability vs ÖSS-SOZ Score Quantiles

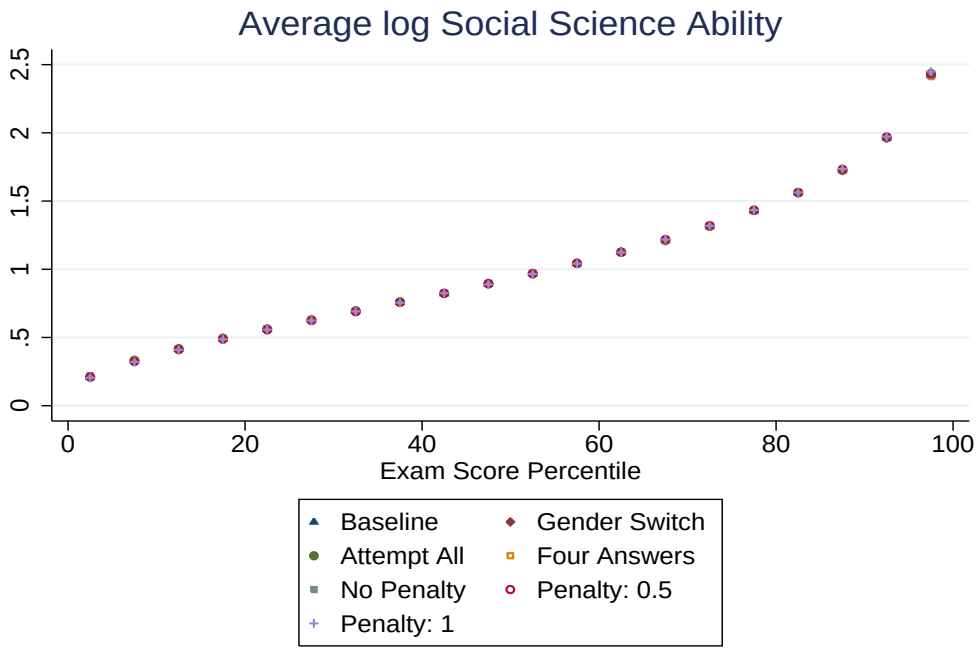

Figure 24: Counterfactual: Turkish Ability vs ÖSS-SOZ Score Quantiles