Hit or Miss? Test Taking Behavior in Multiple Choice Exams

Ş. Pelin Akyol * James Key† Kala Krishna‡

May 11, 2016

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

In this paper we argue that the standard approaches to examining the performance of people taking multiple choice exams is usually inadequate, often misleading, and sometimes just plain wrong. This is important as such exams are widely used in practice being seen as objective, fair and low cost, especially when large numbers of exam takers are involved so that improving on existing methods is vital for policy making. University entrance exams in a number of countries including Turkey, Greece, Japan and China use multiple choice exams. 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. Such exams are also widely used to measure effectiveness of schools, teachers, to enter the civil service, and to allocate open positions.[1]

We specify and estimate a structural model of students' exam taking behavior and explore how different characteristics of students like their ability and risk aversion affect their exam performance. Our objective is to understand how students behave when taking these exams, whether exam taking behavior seems to differ across groups, what seems to lie behind any such differences and to understand the consequences of these differences and their implications for public policy. Our approach also lets us shed light on how, when, and why existing approaches give misleading results and we argue that this is due to their failure to explicitly model behavior and build their estimation around the model.

Our work contributes to the literature in two ways. First, we provide an innovative way to identify differences in risk preferences if there is negative marking that relies on a mass of students answering all the questions, even without having question by question responses for students.[2] Second, we provide a richer model than the standard Rasch model used in the literature, see for example Pekkarinen [2014]. The Rasch model, using what is termed "item response theory", boils down to predicting the probability of a correct answer using a logit setup, with individual and question fixed effects. By allowing for skipping, and relating this to risk aversion, we use all the information in the data in contrast to the standard Rasch model. By ignoring information on skipping, the Rasch model gives biased estimates of a students ability. To understand why this bias occurs, consider for example, a setting where all questions are equally difficult so that there are no question specific differences. In this case, the Rasch model would estimate the "ability"

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

[2] Our approach can also be used with such question by question responses.

or probability of answering a question correctly, of someone answering 20 of 80 questions and getting half right, as 1/8 and that of someone answering 40 questions and getting 20 right as 1/4. However, the difference in the two could be largely due to differences in risk aversion rather than ability. To disentangle the two and obtain unbiased estimates of both risk aversion and ability, we need to specify a more complete setting, that includes the choice of skipping the question as done here. It is worth noting that despite the interest in such exams in the Psychology, Education, and Economics literature, there is little formal modeling and estimation based on the behavior of individual students.

We use administrative data from the Turkish University Entrance Exam (ÖSS) in our work. The ÖSS is a highly competitive, centralized examination that is held once a year. It is selective as only about a third of the exam takers are placed at all and admission to top programs is extremely competitive. College admission depends on the score obtained from the ÖSS, and the high school GPA[3], with at least 75% of the weight being given to the ÖSS score. For each correct answer the student obtains one point, and for each wrong answer he is penalized 0.25 points, while no points are awarded/deducted for skipping a question. Students expend significant time and effort to prepare for this exam and have a good understanding of how the system works.

Our results show that students' attitudes towards risk differ according to their gender and expected score. Women seem to be more risk averse at all score levels than men. This is in line with a large literature that suggest that part of why women perform worse than men in college and in the job market is due to their behavior which is less assertive and more risk averse. Students with low expected scores also tend to be less risk averse. This makes sense as there is a cutoff score to qualify for possible placement and most likely a jump up in utility upon becoming eligible for placement.

We then run counterfactual experiments to investigate how these differences impact on women and the disadvantaged. Since women tend to guess less often than they should if they were maximizing their expected score, they tend to have lower scores, and less variance in their scores, than otherwise similar men. This tend to make them under represented at both the top and the bottom end of the score distribution. This is particularly relevant, since in many developing countries only a small fraction of students are able to proceed to university. In the baseline model, males are over-represented in the top 5%: 55.9% of all test takers are male but 59.8% of students in the top

[3]The high school GPA is normalized at the school-year level using school level exam scores to make GPAs comparable across schools in each year. This also removes the incentive to inflate high school grades.

5% are male. We find, for example, that if there was no negative marking so that there was no reason not to guess when in doubt,women's representation in the top 5% of placements increases by 0.3%. They go from being 39.9% of the population in these placements to being 40.2% of them.[4]Thus, though the actual mean score change may look small, its impact on the gender gap at top institutions is not trivial. The differences in risk preferences accounts for a small part (10%) of the gender gap,[5] though looking at the overall gap is misleading as the bottom is pushed up while the top is pulled down by greater risk aversion on the part of women.

We also try to quantify the importance of using our approach as opposed to the standard Rasch model. We use the estimated model to generate individual level data on performance, question by question. As women tend to skip more often, the Rasch model tends to under estimate their ability compared to a more structural model such as ours which explicitly accounts for skipping behavior. The difference is quite substantial. Taking males and females of equivalent ability, but different risk aversion, can lead to males being judged by the Rasch model to be of 5% higher ability than females.[6]

1.1 Related Literature

The psychology and education literature has long been interested in developing test designs that generate fair results. A disadvantage of the multiple choice framework is that candidates may get the right answers just by guessing, especially if they can eliminate some options.[7] In other exam types, such as short answer based exams, pure guessing is unlikely to help. Negative marking for incorrect answers is applied to deter such guessing. If the expected payoff from randomly guessing is equal to zero, the incentive to guess can be removed, at least for risk neutral agents. However, if risk neutrality does not prevail, then such an adjustment may be problematic. If agents are risk

[4]Looking at students scoring in the top 20%, the gender gap is smaller, but eliminating skipping reduces the over-representation of males by 0.1%

[5]The gender gap measured as the over-represenation of male students in the top 5% of scorers.

[6]We set ability to be median, as found in estimation results, and risk aversion to be set similar to estimation results (equivalent to cutoffs in the baseline regime of 0.26 vs 0.25 females vs males). Under the Finnish university entrance exam scoring system (Pekkarinen [2014]), males correctly answer 5.13% more questions than females. Under the Rasch model, the number of correct answers is a sufficient statistic for ability. Even in the Turkish scoring system, with less harsh penalties, males of median ability correctly answer 1.83% more questions than females of equivalent ability. This is a substantial difference.

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

loving, they may still guess. Moreover, if agents differ in their risk preferences, negative marking of this kind may create differences in guessing behavior across groups that may undermine the validity and the fairness of test scores, reducing the efficacy of the testing mechanism as well as biasing the estimates obtained by the standard Rasch model.

Baker et al. [2010] criticize the use of test results of students to evaluate the value-added of teachers and schools partly because of the measurement error 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* suggesting that such exams will be biased against groups who skip questions rather than guess.

Burgos [2004] investigates score correction methods that reward partial knowledge by using prospect theory. They compute a fair rule which is also strategically neutral so that an agent with partial knowledge will answer, while one without any knowledge will not. Similarly, Bernardo [1998] analyzes the decision problem of students in a multiple choice exam to derive a "proper scoring rule", i.e., one that truthfully elicits the probability of each answer being correct. Espinosa and Gardeazabal [2010] models students' optimal behavior in a multiple choice exam and derives the optimal penalty that maximizes the validity of the test, i.e., maximizes 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. Using simulations, the paper argues that the optimal penalty is relatively high. Even though the penalty discriminates against risk averse students, this effect seems to be small compared with the measurement error that it prevents, especially for high ability students. None of these papers attempts to estimate ability and risk aversion of agents or to test the implications of their models empirically as we do.

On the empirical side, one line of work is experimental (see Eckel and Grossman [2008] for a survey of some of this work with a focus on the differences in risk aversion by gender). Most recently, Espinosa et al. [2013] looks at data from an experiment to show that penalizing wrong answers or rewarding skipping are not the same and that the differences between such scoring rules come from risk aversion. Their results suggest that skipping behavior depends on the scoring rule, knowledge, gender, and other covariates.

Baldiga [2013] explores the gender differences in performance in multiple choice exams. Lab experiments are designed to see whether a gender gap in performance exists, and if so, whether this gap is driven by differential confidence in knowledge of the material, differences in risk preferences (elicited in one of the treatments), or differential responses to high pressure testing environment.

She finds that confidence differences and the pressure of the environment do not seem to be important and that a greater tendency to skip questions remains important in explaining the gender gap even after controlling for differences in risk aversion. She speculates that these differences may be due to sociological (women are encouraged to be less assertive) or behavioral factor (women put a higher cost on getting a wrong answer than men). However, as with all lab experiments, because the stakes are low, it is hard to expect their behavior to perfectly reflect that in high stakes exams. In addition the data set is small with about 400 observations.

The other line of work relies on non experimental data. Risk attitudes of students are an important factor in the decision to attempt a question 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 women do, in fact, tend to skip more questions.

Tannenbaum [2012], in a closely related but flawed paper uses data on the SATs and question by question responses of students. He investigates the effect of gender differences in risk aversion on multiple choice test results. He exploits differences in the number of choices offered in a question to tease out risk aversion, since the more the options, the lower the penalty for a wrong answer. He shows that, as might be expected, a higher penalty (fewer choices) makes skipping more likely and makes the correct answer being chosen more likely, conditional on not skipping and controlling for individual fixed effects. He also shows that women are more likely to skip at a given penalty level, than men, and that their skipping behavior is more responsive to increases in the penalty. He argues that greater risk aversion on the part of women is able to account for 40% of the gender differences in performance in multiple choice exams. He backs this out from estimates that use the variation in the number of possible answers as an instrument. We question this result as his identifying assumption is that the variation in the number of possible answers has the same effect on agents of all abilities. But, Espinosa and Gardeazabal [2010], shows we that differences in risk aversion have small effects on performance, especially for higher ability students. Thus, we know that this identifying assumption is wrong. Since differences in risk aversion have larger effects on performance for lower ability agents, and as the ability distribution for women in the SATs seems to be slightly worse than that of men, the risk aversion difference is magnified in Tannenbaum's estimates. Our results show a far smaller part of the gender gap being accounted for by differences in *risk aversion*

even though risk aversion is significantly greater for women.[8] We attribute the difference in our results to the mis-specification in Tannenbaum's work: in effect his identification assumption does not hold. As a check, we use our model to simulate data on responses where questions have different numbers of answers, but where there is no difference in risk aversion between groups but there is a difference in ability. We show that the approach used by Tannenbaum mistakenly estimates a difference in risk aversion where none exists.

In some ways our data is more limited than that in Tannenbaum [2012], though in other ways, it is better. We do not have question by question responses as are available to him. As a result, we cannot directly look at the probability of skipping and getting a correct answer as done in his paper. Despite this, we are able to use information on the distribution of scores in the presence of negative marking to infer skipping tendencies and ability distributions as well as risk aversion, while allowing them to differ across groups. Thus, one of our contributions is to provide a way to estimate structural parameters of the model with limited data on question by question responses. In addition, having a structural model lets us do counterfactual exercises. Our data is much better than Tannenbaum's in that we have a considerable amount of information on past performance and socio-economic background not available to him.

In the next section, we present an overview of the data and testing environment. The particular patterns in the multiple choice tests are discussed in more detail in section three. In Section 4, the model is presented. Section 5 details the estimation strategy with the results in Section 6. Section 7 contains counterfactual experiments and a discussion of an extension, respectively. Section 8 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). Most high school seniors take the exam and there is no restriction on retaking. However, the score obtained in a year can be used only in that year. All departments, with the exception of those that require special talents (such as art schools) accept students based on a weighted average score of university entrance exam and the high school grade point average.

[8] This may be because we use data on the social studies and turkish parts of the test and allow risk aversion to differ by gender, predicted score, and exam experience while he uses data on the SAT math test and only allows for differences by gender.

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 and can choose where to spend their time. Each part of the test has 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. 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. Students' university entrance exam scores (ÖSS score by track) are calculated as a weighted average of their raw scores in each test.

Table 4 shows the test weights according to each track. For the social science track students where ÖSS-SÖZ is used , the Turkish and social science tests have the highest weight, while math and science have a relatively low weight.[9] Students with scores above 105 points can submit preferences (submit an application) to 2-year college programs, while 120 points are needed to apply to 4-year college programs. The placement or allocation score (Y-ÖSS) is calculated as follows

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

where X$\in$\{SAY,SÖZ,EA,DIL\}, α is a weight on the standardized[10] high school GPA. The placement score varies by track, preferred department and whether the student was placed (accepted) into a regular program in the previous year or not.[11] There are penalties if a student switches tracks, or refuses a placement. The distributions of ÖSS-SÖZ scores as well as normalized GPAs for first

[9] In the calculation of ÖSS scores, firstly raw scores in each track are normalized so the mean is 50 and the standard deviation is 10. Then these normalized scores are multiplied by the weights presented in Table 4. According to the data, the equation is ÖSS-SÖZ=0.8755*rawTurkish+0.8755*rawSS+0.187*rawMath+0.l187*rawScience+78.89

[10] The standardized GPA is the GPA normalized by the performance of the school in the university entrance exams which adjusts in effect for different grading practices across schools.

[11] The α used is chosen according to fairly complex rules. For example, 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 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.

time takers are depicted in Figures 4 and 5. Note that men seem to do a bit better than women in the exam at the higher end of the ÖSS-SÖZ score distribution, but considerably worse at the low end of the distribution. Looking at internal assessment performance, women do considerably better with higher normalized GPAs.

The data used in this study comes from multiple sources. Our main source of data is administrative data from the test administering body (ÖSYM) and the high schools on a random sample of roughly 10% of the 2002 university entrance exam takers. This data 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 have around 40,000 students from each track (Social Science, Turkish-Math, Science). Here we focus on social science track students for reasons that will become apparent shortly. Table 10 presents the summary statistics for social science students.

3 Multiple Choice Exam Scores

We begin by taking a first look at students' scores in the Turkish, social science, math and science exams. Recall that each section of the exam has 45 questions. The scoring structure results in each multiple of 0.25 between -11.25 and 45 (with the exception of certain numbers above 42) being possible outcomes in an exam.[12] For example, attempting all questions and getting all wrong results in a score of $-\frac{45}{4} = -11.25$..

Most social science students do not even attempt the math and science parts of the exam and those that do fare badly as the mean score is close to zero. This could be because math and science test scores have relatively little weight (.4 each) in the ÖSS score of social science track students. Turkish and social studies scores, in contrast, have a weight of 1.8. Students are also explicitly advised to spend less time on the math and science test.[13] In addition, these students are poorly prepared in math and science as they have not done much of it since the ninth grade and the

[12] Recall that 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.

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

questions are very challenging.

Obtaining a particular raw subject score could happen in only one way or in many ways. For example, 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 incorrect, which is not possible given there are only 45 questions.

There are 46 particular scores that are worth noting: those which correspond to attempting all questions. These are spaced 1.25 points apart, starting at -11.25, and ending at 45 points. The distributions of raw subject scores in social science and Turkish for first time takers, see Figures 6 and 7, have very prominent spikes. It is no coincidence that the spikes appear evenly placed; they correspond to the 46 scores that occur after attempting all questions and come from the fact that there is a mass point of students, of differing abilities, who answer all the questions. This is a big part of our identification strategy as we do not have question by question data for students. Math and science score distributions for social studies track do not exhibit this behavior as most students obtain a score of zero. Nor do any of the subject score distributions for the science track students exhibit this pattern of spikes across the entire support of the distribution. These spikes are only there for the top part of the distribution consistent with only the very best students attempting all the questions. For this reason, we choose to use the social studies track in our estimation. In addition, as they do not spend much time on the science and math parts of the exam, there is less of a worry about modeling time constraints that relate the different components of the exam.

4 Model

We model the test taking behavior as follows. When a student approaches a question, he gets 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 and each question is answered in isolation. We do not allow for outcomes in one section of the test to have any bearing on other sections. For example, it may be that a student feels he is doing well and this makes him more confident so that he does better on the rest of the exam. Nor do we allow for any time pressure that results in

skipping. For example, a student may know he is bad at one component of the test and so skip the entire section in order to have more time on another.

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 $\beta > 0$. Thus, the density of the signal for an incorrect answer is $\frac{\beta A_I}{x^{\beta+1}}$.The mean signal is $\frac{\beta A_I}{\beta - 1}$ decreasing in β.
- Correct answer - draw a signal from a distribution F, where F is Pareto with support $[A_C, \infty)$ and shape parameter equal to $\alpha > 0$, so that the density of the signal is $\frac{\alpha A_C}{x^{\alpha+1}}$. The mean signal is $\frac{\alpha A_I}{\alpha - 1}$ decreasing in α.

Assumption 1. $A_I = A_C = A$.

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 conditional on X, 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:

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

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

This can be further simplified to:

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

Thus, the choice of A is irrelevant. For this reason we will set it at 1 from here on. 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:

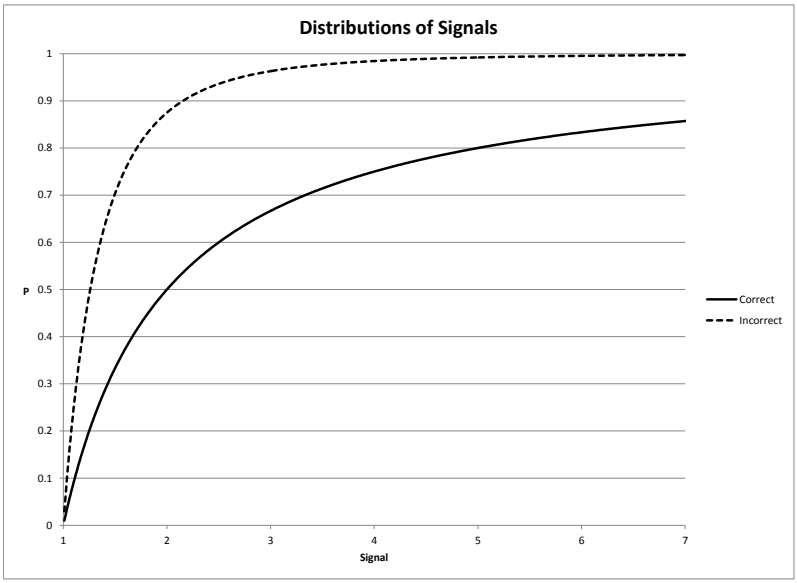
$$\text{Prob}(i \text{ is correct}|X) = \frac{x_i^{\gamma}}{\sum_{m=1}^{5} x_m^{\gamma}} \tag{6}$$

Note that the sum of beliefs for each of the five answers adds up to unity. We assume that $\beta \geq \alpha$, so that the mean signal for the incorrect answer is lower than that for the correct answer. Thus, the higher the signal, the greater the likelihood that the answer is correct.[14] A higher shape parameter for a Pareto distribution shifts probability mass to the left so that the signals would generally be smaller. Hence, if we fixed α, a higher γ(i.e., a higher β) would correspond to greater ability. In fact, it is worth emphasizing that it is the difference in the distributions of the signals of correct and the incorrect answers that captures ability. Someone who thinks all anwers are great is as bad as someone who thinks none of the answers are great: it is the ability to distinguish between the right and the wrong answers that indicates ability. This is why the mean signals mean nothing: it is only the difference in their means that matters. In addition, we assume that the lower bound for signals for both correct and incorrect distributions is the same.[15] Given these assumptions, we can rescale so that the correct answer is drawn from a distribution where $A = 1$ and the shape parameter is also 1, while the signal drawn for an incorrect answer is drawn from a distribution where $A = 1$ and the shape parameter is $\frac{\beta}{\alpha} > 1$.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. Signal distributions for a student with ability $\beta = 3$ are shown in Figure 1.

[14] If a student were to draw from distributions with $\beta < \alpha$, smaller signals would be associated with the correct answer and we would reverse our interpretation of the signal.

[15] The assumption that the lower bound for the correct one is higher, i.e., $A_C > A_I$, would mean that it is possible for student to be sure that an answer is wrong: i.e. to rule out a wrong answer. It is also possible for a student to be sure he had the right answer: this would be the case when all but one answer had a score between A_I and A_C.

Figure 1: Distributions of signals for a student with $\beta = 3$, approximately median



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 one 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:

Assuming $A_C < A_I$ would make no sense if $\beta \geq \alpha$.

$$\frac{d\log(\text{Belief})}{d\gamma} = \log x_{\max} - \frac{x_1^\gamma \log x_1 + x_2^\gamma \log x_2 + x_3^\gamma \log x_3 + x_4^\gamma \log x_4 + x_5^\gamma \log x_5}{x_1^\gamma + x_2^\gamma + x_3^\gamma + x_4^\gamma + x_5^\gamma} \tag{7}$$

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

$$\frac{d\text{Belief}}{d\gamma} \geq 0 \tag{8}$$

with the inequality strict unless $x_1 = x_2 = x_3 = x_4 = x_5$. □

Once students have observed the signals for each of the five possible answers to the question, they are faced with six possible alternatives: 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.

The relationship between ÖSS score and utility need not be linear. Utility could be convex in score: for example students must score above 120 in order to be qualified to submit preferences to a four year college program. If being able to do so makes a student much better off, students could be risk loving when their scores are near this cutoff. It could be concave in regions where schools are similar in quality so that higher scores bring little gain.

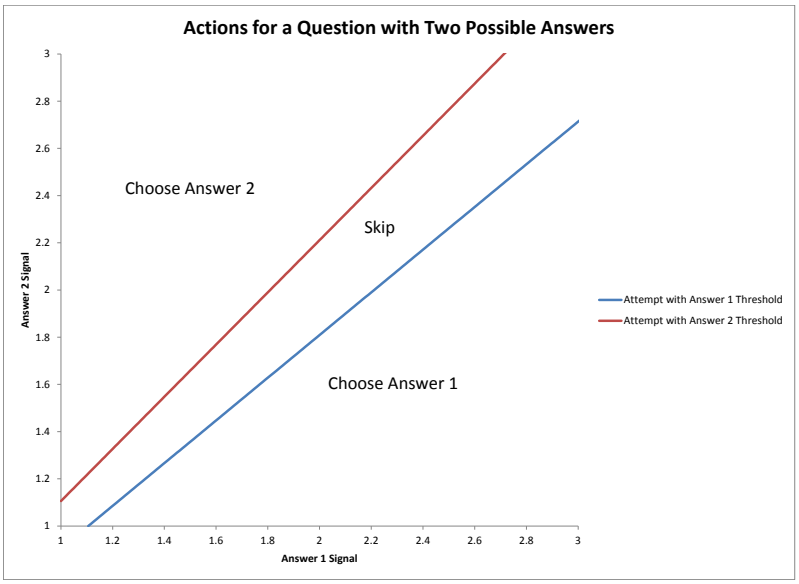
To incorporate this we allow students to have a cutoff for the belief, below which they will skip the question. If the student believes that the best answer (highest signal) has a probability of being correct greater than this cutoff, he will attempt the question, choosing the best answer. This cutoff lies in the interval [0.2 1].[16] 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 preferences.

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.

[16]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

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 that the belief that it is correct is greater that c, the cutoff required to attempt the question. We define the following function as the minimum signal x_n required to attempt the question using the n^{th} answer, given the other signals:[17]

Figure 2: Action conditional on signals for a simple two answer model (parameter values: $\beta = 3$ and cutoff $= 0.55$



$$K(\{x_i\}_{i\neq n}) = \max\left(\max\{\{x_i\}_{i\neq n}\}, \left(\frac{c}{1-c}(\sum_{i\neq n} x_i^{\gamma})\right)^{1/\gamma}\right) \tag{9}$$

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:

$$\int_{x_5=A}^{\infty}\int_{x_4=A}^{\infty}\int_{x_3=A}^{\infty}\int_{x_1=A}^{\infty}\int_{x_2=K(x_1,x_3,x_4,x_5)}^{\infty} 1 dG(x_2)dF(x_1)dG(x_3)dG(x_4)dG(x_5) \tag{10}$$

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:

[17] A diagram showing choices conditional on signal observations for a simplified two answer setup is shown in Figure 2

$$\int_{x_5=A}^{\infty}\int_{x_4=A}^{\infty}\int_{x_3=A}^{\infty}\int_{x_2=A}^{\infty}\int_{x_1=K(x_2,x_3,x_4,x_5)}^{\infty} 1dF(x_1)dG(x_2)dG(x_3)dG(x_4)dG(x_5) \tag{11}$$

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:

$$\text{Prob(Correct)} = P_C(\beta, c) \tag{12}$$

$$\text{Prob(Wrong)} = P_W(\beta, c) \tag{13}$$

$$\text{Prob(Skip)} = P_S(\beta, c) \tag{14}$$

where $P_S(\cdot) = 1 - P_C(\cdot) - P_W(\cdot)$. Table 6 provides these, in addition to the expected score from 45 questions, for various parameter values. This paper is concerned with the loss of points as a result of being risk averse. As such, the difference in expected score, between the student characterized by (β, c) and $(\beta, 0.2)$ is included.[18] Note from Table 3 that for a given β, as c rises, the probability of skipping a question increases while the average points per question decreases.[19] Note also, that this fall is increasing as the cutoff rises and less so for higher ability (β). This makes sense as it is score maximizing to set c at .2 and better agents suffer less from being more risk averse (having a higher c) as they get higher signals and so are less affected by it. In particular, students that have ability $\beta > 3$ (i.e. the best 50 percent of students) see virtually no impact on their score going from risk neutral to a cutoff of 0.25, despite the 25% increase in certainty required to attempt a question. Intuitively, this is due to such students being at least 25% sure for the vast majority of questions.

Now we are in a position to write out the likelihood of seeing a particular score.

First we derive the probability of attempting x questions. 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:

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

[19] Of course the average points per question attempted increases. More risk averse students will be less likely to attempt a question, other things equal, but more likely to get it right, conditional on having attempted it.

$$\text{Prob(Answer } x \text{ questions)} = \binom{45}{x}(P_C + P_W)^x(P_S)^{45-x} \tag{15}$$

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

$$\text{Prob(Answer } y \text{ of } x \text{ questions correctly)} = \binom{x}{y}(\frac{P_C}{P_C + P_W})^y(\frac{P_W}{P_C + P_W})^{y-x} \tag{16}$$

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

$$\text{Score}(x, y) = y - \frac{(x - y)}{4} \tag{17}$$

Finally, 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:

$$\text{Prob}(Score = s) = M(\beta, c; s) \tag{18}$$
$$= \sum_{j=1}^{k} \binom{45}{x_j}(P_C + P_W)^{x_j} P_s^{45-x_j} \binom{x_j}{y_j}(\frac{P_C}{P_C + P_W})^{y_j}(\frac{P_W}{P_C + P_W})^{y_j - x_j}$$

For example, the probability of getting a score of 40 can be obtained as follows. It comes from 40 correct, 5 skipped and 0 incorrect, or 41 correct and 4 wrong and zero skipped so $k = 2$. The first case corresponds to $x_j = 40$ and $y_j = 40$. The second to $x_j = 45$ and $y_j = 41$.[20]

5 Estimation Strategy

In our model, students' scores depend on students' ability (β) and risk aversion cutoff, c. In our data set we observe only the student's total score. In this section we use our model to estimate the distribution of ability and risk aversion cutoffs, c.

Estimation of the parameters of interest, the distribution of student ability (β_T, β_{SS}) and risk aversion cutoff c, is conducted separately for each gender. In addition, we recognize that the relationship between ÖSS-SÖZ score and utility is not necessarily constant throughout the range

[20] $\{(x_j, y_j)\}_{j=1}^{k}$ is the set of combinations (x_j, y_j) such that $x_j, y_j \in \mathbb{N}^0$, $x_j, y_j \leq 45$ and $1.25y_j - 0.25x_j = s$

of scores: the degree of risk aversion may be different. In particular, we might 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.

For this reason we allow cutoffs to vary by gender and attempt number, and allow cutoffs to depend on the interval in which the student's predicted ÖSS-SÖZ score lies, for example 120-130. The next section explains how we predict ÖSS-SÖZ scores.

5.1 Predicted ÖSS-SÖZ Score

The risk taking behavior of students is likely to depend on their score: the utility derived from obtaining a given score is driven by the effect on the placement potential. Two different scores may give the same utility if both result in the student failing to gain admission. However, we cannot use students' actual exam scores as it is an endogenous object that is affected by students' risk taking behavior in the exam. If, for example, students who skip more questions get lower scores systematically, grouping according to actual exam score will put those students into lower groups, and then finding higher cutoffs for those students will not be informative as grouping is done partially based on risk behavior of the students. Therefore, we predict students' scores by using their observable characteristics. Specifically, GPA (adjusted for school quality)[21], education level of both parents, income levels/monthly income of parents, preparation on the four subject areas, and the school type. We run an OLS regression separately for male and female first time takers in the social science track, and use the results to predict ÖSS-SÖZ scores for each student.[22]

5.2 Estimation

We divide students into groups, according to gender, and the range into which their predicted ÖSS-SÖZ score lies: $(0, 90)$[23], $[90, 100)$, $[100, 110)$, $[110, 120)$, $[120, 130)$, $[130, 140)$, and $[140, \infty)$[24]. For each group, we examine the two subjects jointly.[25] While these intervals may not contain equal

[21] 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.

[22] The R^2 values of the predictive regression were 56% and 58% for males and females respectively

[23] In this bin, all but two students had predicted scores above 80.

[24] Most of the students in this bin has predicted scores between 140 and 150

[25] The only test sections of interest are Turkish and social science

numbers of students, it will allow us to make comparisons across genders. We assume that each group has a common attempt cutoff, c, and joint distribution of ability $(\beta_{Turkish}, \beta_{SocialScience})$. 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 covariance matrix Σ.[26] This ensures that each student has an ability in both subjects greater than 1, and results in a log normal distribution (shifted 1 unit to the right).[27] It also allows for abilities in the two subjects to be correlated, as would typically be the case.[28]

Under the assumptions made, the probability of obtaining each score is approximated through simulation. For student n, we take a draw from $N(\mu, \Sigma)$ and label the vector as X_n. From X_n, we find $(\beta_T, \beta_{SS}) = (1+e^{X_s(1)}, 1+e^{X_s(2)})$, 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 $M(\beta_k, c; x)$ in the previous section. By taking a random draw from the joint distribution of 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 correlation 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. 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 the intensity of the spikes. Denote the set of scores that are obtainable by skipping 0 questions as $ZS = \{45 - 1.25 * i\}_{i=0}^{45}$, and the set of scores that are obtainable by skipping 1 question as $OS = \{45.25 - 1.25 * i\}_{i=1}^{45}$. Suppose a student obtains a score of x. The moment capturing the intensity of the spikes in the relevant section is defined as $SI(x) = I(x \in ZS) - I(x \in OS)$.

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 means of scores, which help pin down the means of the ability distributions that students are drawing from, and the variances/covariances of scores which help pin down the variances of the ability distributions and the correlation between ability

[26] In practice, correlation coefficients ρ were obtained rather than covariance, to assist the minimization procedure and for interpretation. The covariance obtained is therefore $cov(T, SS) = \rho\sigma_T\sigma_{SS}$.

[27] Earlier analysis showed that the distribution is skewed; in addition the likelihood of answering correctly is roughly proportional to the log of ability.

[28] Within the social science track as a whole, the scores in the Turkish and social science sections are highly correlated.

draws.

More formally, moments of an outcome o[29] is given by:

$$m(o) = (o(1), o(2), o(1)^2, o(2)^2, o(1) * o(2), SI(o(1)), SI(o(2))) \tag{19}$$

where $SI(\cdot)$ 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.

$$\hat{c}, \hat{\mu}, \hat{\Sigma} = \hat{\theta} = \arg\min_{\theta} \Big[\sum_{t=1}^{T} \Big(m(o_t) - \frac{1}{S} \sum_{s=1}^{S} m(o(u_t^s, \theta)) \Big) \Big]' W_T^{-1} \Big[\sum_{t=1}^{T} \Big(m(o_t) - \frac{1}{S} \sum_{s=1}^{S} m(o(u_t^s, \theta)) \Big) \Big] \tag{20}$$

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.

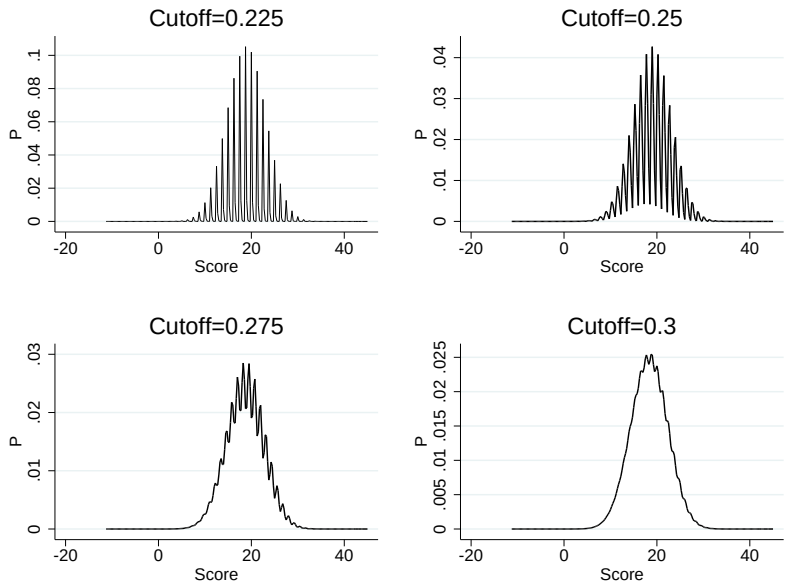
5.3 Identification

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, ceteris paribus. 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

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

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 spikes cannot be seen.

Figure 3: Distribution of scores resulting from various cutoff levels



This is illustrated in Figure 3, 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 $c = 0.225$ puts virtually all of the mass of the score distribution on values that correspond to students attempting all questions. As the risk aversion cutoff increases to 0.3, the spikes all but disappear as very few attempt all questions.

The relationship between the intensity of the spikes and the risk aversion cutoff is not constant. As we increase ability, given a cutoff c, the intensity of the spikes increases. This makes sense as high ability agents are more likely to distinguish the right answer from the wrong one and answer all questions for any given cutoff. While low ability students might have an answer with a belief above the risk aversion cutoff, this becomes increasingly rare as ability rises.

The parameters of the distribution of the ability of a group of students, (μ_T, μ_{SS}) and Σ, are identified by the distribution of scores. An increase in the mean parameter μ_T moves the Turkish score distribution to the right, increasing the mean, while an increase in the variance parameter σ_T^2 increases the variance of the Turkish score distribution. This is due to a strong relationship between ability and exam scores. Similarly with the social science section. Finally, the correlation

between Turkish and social science ability draws is obtained through the correlation of scores.

6 Results

Table 7 presents the estimates for each group of the risk aversion cutoff, the belief regarding probability of success below which a student will skip a question, in addition to the standard errors of the estimates. Figure 8 displays the cutoff estimates graphically, while Table 8 presents t-statistics for tests of the null hypothesis that the two genders have equal attempt cutoffs.

Two facts are apparent. 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 in line with the literature - males are acting in a less risk averse manner. Secondly, the cutoff is systematically lower in the predicted score ranges 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[30] and seems to rise with the predicted score bin consistent with increasing risk aversion.

Figures 9 and 10 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 the likelihood of scores which correspond to skipping multiple questions.[31] In addition, the skipping behavior is underestimated 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. Conversely, students who are actually better than the econometrician would predict (yet know their ability) are assigned a risk aversion of a student of lesser ability, which is typically low, and so will skip very infrequently. Overall, the error in assigning students to the correct risk aversion range leads to lesser spikes at the low end of the score distribution, and greater spikes at the high end of the distribution. Estimation methods which grouped students based on

[30] Cutoffs for the top students are approximately 0.26, which has meaning that these students will only answer a question if they are at least 26% sure of their answer. Significantly more than the 20% likelihood of a random guess.

[31] This can be explained by the presence of some questions that do not feature partial knowledge, i.e. students will either "know" the answer or will have no information.

actual ÖSS-SÖZ score did not feature this issue.

Estimates of the parameters governing the distribution of ability for each group are presented in Table 9. Recall that ability is parameterized as $(1+e^X, 1+e^X)$, where $X \sim N(\mu, \Sigma)$. The means and variances of the components of X in each group are presented.

As expected, groups that are predicted to have high predicted exam scores have much better distributions 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 implied by the distribution of exam scores.[32] There is another possibility, that the fitted ÖSS score is not as accurate for males.

The correlation between ability in Turkish and Social Science seems to be higher for each decile for females, as seen in Table 10. This would tend to give females an advantage in terms of being at the top: in order to gain admission students must perform well in both Turkish and social science. It would also explain the higher variance for males.

In both the model and estimation procedure we have assumed that each question is of the same level of difficulty. This assumption was necessary as we did not observe item responses. However in another paper (Akyol et al. [2016]) we construct a model and estimation procedure to examine data sets with item responses. In addition, we examine data from a mock exam, aimed to prepare students for the social science track. Having recovered question difficulty parameters (in addition to individual student ability and risk aversion), we are able to see the effect on score distributions of having identical difficulty. We first find the expected score of the median student with the original questions. We then find the level of difficulty which, if all questions were to have identical difficulty, would give the same expected score to the median student. Comparing score distributions, we find that the variance of the median student's score would be 9% higher with questions of constant difficulty, as opposed to the variance obtained with the original questions.[33]

[32] And the dispersion of predicted scores

[33] Similar results when looking at the 25^{th} and 75^{th} percentile students: 10% and 7% (respectively) higher variance with constant difficulty.

The assumption therefore leads to our correlation estimates being biased upward. With constant difficulty the simulated section scores of a student would feature more variance. So to match the high degree of correlation between scores in the data, the estimated correlation between ability must be higher than it would be otherwise. This explains why some of the correlation estimates are located at the boundary.

Looking at the distributions of ability across the various groups, we see similar patterns.

In Figure 12 and 13, 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.

7 Counterfactuals

Having recovered the parameters regarding 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 are different and 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 effectiveness of a test in this context is measured by which students are able to gain admission to university programs. Accordingly we look at the the composition of various score percentiles, for example what are the characteristics of the students who score in the top 5% of the cohort?

The objects of interest in these experiments are as follows:

- The relationship between ability in social science (β_{SS}) and exam score percentile

- The relationship between ability in Turkish (β_T) and exam score percentile
- The relationship between gender and 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 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 systematic bias in the testing procedure. In addition, the distributions of abilities are considerably different across gender; the regime may end up attenuating these differences.

The seven possible regimes used in this counterfactual experiment are:

1. The baseline model, as estimated in the previous section.
2. 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.
3. Risk preferences of females are altered, so that the cutoff used by a female student in predicted ÖSS-SÖZ score interval k is is changed to that used by a male student in predicted ÖSS-SÖZ score interval k and not vice versa (labeled as "Equal Cutoffs" in figures).
4. Each question has only 4 answers to choose from, with the penalty for an incorrect answer adjusted accordingly.
5. The penalty for answering incorrectly is increased from 0.25 points to 0.5 points.

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

7. The number of questions in each section is doubled, from 45 to 90.

The second 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. As it eliminates the impact of risk aversion, it eliminates the channel through which gender differences in risk aversion affect gender differences in outcomes.[34] The third however keeps the impact of risk aversion, but eliminates the gender differences in risk aversion. While this is not entirely feasible to perform in practice, we can still use the counterfactual exercise to quantify the effect of gender differences in risk aversion in the current system.

The fourth requires more explanation. Reducing the number of choices makes the gamble involved in answering have higher stakes. This should exacerbate the effect of different risk preferences across the genders. 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.[35] In addition, having four answers instead of five, and increasing the penalty accordingly, can have a detrimental effect, even in the absence of risk aversion. The increased penalty can lead to increased variances of scores for a given student[36]

The fifth and sixth counterfactuals are designed to elicit more skipping from students and to amplify the impact of differences in risk preference across the genders. As in the four-answer counterfactual, cutoffs are translated into implied CARA parameters and new cutoffs are obtained for both counterfactuals.

[34]It also eliminates the channel whereby random guessing is reduced, which is often the rationale for introducing penalties for incorrect answers.

[35]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.

[36]For a student of no ability, the standard deviation of the points earned for a single question is, for a student of (approximately median) ability $\beta = 3$ is 0.66 (four answers) vs 0.62 (five answers) i.e. scores are more precise when there are five answers than when there are four answers. For a student of ability $\beta = 6$ (approximately the top 10%) the standard deviation is 0.58 vs 0.56.

Finally, we examine a hypothetical exam with twice the number of questions per section, 90 compared to 45 in the actual exam. This allows us to place results in the context of a more precise exam: increasing the number of questions increases the ability of the exam to distinguish students based ability.[37] For each of the possible regimes, we find the resulting distributions of scores for the entire sample of first time students in the social science track.

We simulate students using the parameters estimated,[38] generating scores in the Turkish and social science section, adding the two to generate an exam score for each student.[39] We then segment students into bins by using the total score. The bins are constructed such that five percent of students are in each bin, so that we observe the 5% of students who perform the worst, the 5% who perform the best etc.[40][41] For each of the twenty bins, ranging from the lowest scoring students to the highest, we find three objects of interest: share of males, average social science ability and average Turkish ability.[42]

We first examine the effect of the different regimes on the ability of the exam to select the most capable students, the relationship between exam score percentile and (average log) ability in the two subjects, Turkish and social science. Figures 14 and 16 have higher ability students in the higher score percentiles, as expected, but with relatively small differences across the seven regimes.[43] In order to see more clearly what the differences are, Figures 15 and 17 show the difference between the counterfactual of interest and the baseline model. A positive value for a score percentile means that the ability of students in that score percentile is higher in the counterfactual than in the baseline model. A regime that delivers a positive value on the right (high scoring students) and a negative value on the left (low scoring students) would be preferred to the baseline model, as it more correctly identifies the strong and weak students.

As the Turkish and social science abilities show very similar patterns, they will be discussed jointly.

We see that the "Attempt All" regime is very similar to the baseline throughout the distribution,

[37] In practice it may not be feasible to have such an exam due to costs.

[38] 1000 students were simulated for each student observed in the data.

[39] We did not simulate scores from math and science as the majority of students skipped theses sections, and scores of those who attempted were close to zero.

[40] Five percent of the number of students.

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

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

[43] In Figures 14 through to 20 the top scoring groups of students are on the right side, the worst performing on the left.

perhaps slightly lower average abilities amongst the highest performing students, consistent with penalties and risk aversion allowing an improved separation of students by abilities. The "rev drag" counterfactual also shows very little difference to the quality of admitted students, consistent with the small changes in risk aversion that occurred.

The "four answers" regime is typically below the baseline on the right hand side and above it on the left hand side, indicating that having five answers delivers a better admitted class than questions with four answers. This is due to a student of fixed ability having a lower score variance with five answers than four: more answers leads to a more accurate signal.

Next we see the opposite pattern for the regimes with high penalties. Figures 15 and 17 clearly show that, although there is not considerable differences between them and the baseline (as shown in Figures 14 and 16), they dominate the baseline model. Average abilities under these regimes are lower than the baseline on the left (more accurately identifying weak students)[44] and higher than the baseline on the right (more accurately identifying strong students). The differences dwarf those between the baseline and the "four answers" regime. Furthermore, we can note that having a penalty equal to 1 performs even better than having a penalty equal to 0.5, especially when considering the average ability of the top 15% of students.

The reason for the effectiveness of these high penalty regimes is simple. They strongly discourage attempting when the student is relatively uncertain of their answer. This results in much less low-confidence attempting (i.e. fewer random guesses) and therefore score distributions of an individual student will exhibit less (relative) variance, resulting in a cleaner signal. The downside is that there will be students who are perhaps 51% sure of their answer, but will choose to skip rather than facing a 51% chance of losing a point, due to risk aversion. Even risk neutral (or risk loving) students will choose to skip a question unless they are 50% sure of their answer when the penalty is quadrupled to 1. So there will be a considerable amount of partial knowledge which students will be unable to convey. Regardless, the former effect overwhelms the latter in this setting.

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

[44] With the exception of the double penalty regime for the lowest ventile. Examining more carefully, the lowest 5% actually has a higher average ability than the second lowest 5%. This is not due to any risk aversion differences (the pattern remains even if all students are given the same cutoff of 0.25). The explanation is simple: The bottom 5% tends to consist of students who attempted questions and had bad luck. Since attempting is related to a higher ability we observe this interesting pattern.

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 these students 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.[45] So it seems that the increased validity resulting from the increased penalty is significant from a practical point of view; it is possible to increase validity while reducing the amount of time which students must be able to concentrate for, making stamina less of a factor.

Finally, we examine the impact of the various regimes on the male fractions of the different score percentiles. As expected, given their greater exam score variance, the male fraction is u-shaped, as shown in Figure 18. The lowest scoring students are predominately male, as are the highest scoring students. However, there are minor differences between the seven regimes throughout most of the range.

In the baseline model, the top 10% of students (by score) are 58% male. But males are 56% of the social science track first time takers: a 2% gender gap in favor of males. Examining Figure 19, we see that risk aversion does not appear to be a major component of the gender gap. In fact, it appears that there would be no reduction in the gap if we were to eliminate skipping, or eliminate risk aversion differences: both "attempt all" and "rev. drag" give a slightly higher male fraction when looking at the top 15% of students.

While there are some small differences around the upper area, all seven are fairly similar. We do however see that the high penalty regimes have a slightly higher male fraction, i.e. the small difference in risk aversion begins to have a small effect. This is consistent with a bias in favor of male due to risk aversion. 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 high supports this notion, as does the similar pattern seen when the number of questions is doubled.

To distinguish the amplification of the gender bias caused by the quadrupled penalty from

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

the more refined ability identification, we introduce a new counterfactual: quadrupled penalty combined with changing female students' risk aversion to be that of males. If there is a gender bias present in the quadrupled penalty counterfactual, caused by different risk aversion, then the male fraction in the top percentiles should decrease, going from quadrupled penalty to 'quadrupled penalty combined with equal cutoffs'. This is what we observe in Figure 20. We see that the male fraction is lower in the top percentiles when we equalize cutoffs, but the magnitude is perhaps on the low side. There is a 0.6% gap caused by gender differences when looking at the top 5%, and an impact of 0.4% when looking at the the top 10%. Considering the rather extreme penalty that is being applied (1 point for correct, -1 for incorrect), the effects smaller than would perhaps be expected, yet are still practically significant.

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 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 Other Tracks

While the main paper has focused on the social science track (ÖSS-SÖZ) due to the importance of both the Turkish and social science sections (simultaneously placing explicitly negligible weight on the science and math sections), other tracks may be considered. As seen in Table 4, the Turkish Math track (ÖSS-EA) also places high emphasis on the Turkish section of the exam, a section which has been shown to be well described by the model.[46]

The language track is also noteworthy, as this track features a language section, which as would be expected accounts for a large part of the student's admission score. Moreover, the model likely fits the style of question which one would imagine would be asked in such an exam.

There are however some important differences between the language exam and the regular exam. First, the language exam is held separately, some time after the regular (science, math, Turkish and social science sections) exam. In addition to this, students are able to view the correct answers following the regular exam. This would give the language track students information regarding their score in the regular exam. As the social science and Turkish sections contribute to the language track score (albeit a small contribution) this information is relevant. Secondly, although the scoring system is the same for each question (1 point for correct, -0.25 for incorrect, 5 possibilities and the option to skip), there are in total 100 questions in the language exam. As previously, we only observe the total section score.

Accordingly, we estimate the model for the Turkish-Math track students, examining the Turkish section only. We then estimate the model for the language track students, examining the language exam only.

8.1 Estimation

Estimation occurs in a very similar manner to the main body of the paper. After separating first attempt students by gender and into predicted ÖSS score bins, we use simulated method of moments to obtain the distribution of ability, and the attempt cutoff for the group. As we are only examining one subject, the ability distribution is univariate. Moments to be matched are similar to before, but only examining one section.

To obtain the predicted score bins, we run a regression between score and observable characteristics, and use predicted values. While the Turkish Math section binning process is the same as

[46] A result of a heavy emphasis on partial knowledge type questions.

before, the language track is slightly different. As students are able to see the general test questions and correct answers after the exam, it is reasonable to expect students to accurately determine their score from the Turkish and social science sections of the exam, at least to a reasonable degree. We therefore use the students' actual performance in the general test when predicting their score in the language exam.

8.2 Data

Focusing on students making their first attempt, we obtain for the Turkish Math track 7972 female and 7919 male students. For the language track[47] we obtain 9280 female and 3681 male students.

Separating into score bins, we sample sizes as shown in Tables 11 and 12.

The aggregate score patterns can be seen in Figures 21 and 22 in the appendix. While the Turkish exam section of the Turkish Math track students looks relatively similar to previous histograms, the language track students illustrate a much different pattern. This is due to the large number of questions of the language section: 100 compared to 45 in other sections. As a result, the medium ability students will tend to skip enough questions for the spikes to diminish greatly. While in the other exam sections there were 45 opportunities for a student to skip a question, in the language section there are more than double the amount of chances to skip. It follows that there will be more skipped questions, which have the effect of reducing the intensity of the spikes, especially in the middle of the distribution.[48]. Estimating the model for data showing a much different aggregate pattern also serves as a robustness check.

[47] There were three different foreign language options, each with their own exam. We chose to focus on the English language students as they were the vast majority. Sample sizes are those for the English language track.

[48] This was also the location where the spikes were least intense in the social science and Turkish sections

8.3 Results

Estimates of ability distributions for the different groups. For the sake of brevity, only the attempt cutoffs are presented.

Table 1: Estimates of Attempt Cutoffs for Turkish-Math Track students - Turkish Section

	< 90	90-100	100-110	110-120	120-130	130-140	> 140
Male	0.2267	0.2338	0.2497	0.2606	0.2669	0.2711	0.2750
Female	0.2240	0.2358	0.2534	0.2660	0.2783	0.2823	0.2881

Table 2: Estimates of Attempt Cutoffs for Language Track students - Language Section

	<90	90-100	100-110	110-120	120-130	130-140	> 140
Male	0.2252	0.2297	0.2395	0.2485	0.2554	0.2596	0.2622
Female	0.2233	0.2329	0.2400	0.2525	0.2596	0.2672	0.2618

Table 3: t-stat for test: (female cutoff-male cutoff)=0

	<90	90-100	100-110	110-120	120-130	130-140	> 140
Turkish Math Track	-0.518	1.199	1.984	2.771	4.764	3.660	2.198
Language Track	-0.382	0.905	0.237	1.830	2.457	3.836	-0.089

As before, there are two important patterns. First, as seen in Figures 23 and 24, the cutoff increases as we move from students who expect to perform poorly to students who expect to perform well. This is in line with expectations, given the payoff structure. Secondly, males tend to have lower cutoffs than females: they are less risk averse.

Another important observation is that these cutoff patterns are very similar to those observed in the social science track. Even the language track, where the data exhibited very different patterns, has a similar pattern of risk aversion, providing further support for the model and empirical approach.

Finally, the magnitude of the cutoffs, and the degree of differences between male and female students, was relatively similar across tracks. In all three tracks males were significantly less risk averse in some of the expected score ranges, and never significantly more risk averse.

9 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., Key, J., and Krishna, K. (2016). Estimation of multiple choice exams.

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

Espinosa, M. P., Gardeazabal, J., et al. (2013). Do students behave rationally in multiple choice tests? evidence from a field experiment. *Journal of Economics and Management*, 9(2):107–135.

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.

Pekkarinen, T. (2014). Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization*.

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*.

10 Appendix

Table 4: 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 5: Summary Statistics

Variable	Obs.	Mean	Std.Dev.	Min	Max
Male	9273	0.558		0	1
ÖSS-SÖZ score	9273	108.219	19.030	0	161.166
Normalized High School GPA	9273	47.144	8.199	30	80
Raw Turkish Score	9273	19.268	9.778	-6.25	45
Raw Social Science Score	9273	14.805	9.640	-8.75	43.75
Raw Math Score	9273	0.851	2.884	-9	37
Raw Science Score	9273	0.115	1.260	-8.75	19.5
Education level of Dad					
Primary or less	9273	0.557			
Middle/High School	9273	0.285		0	1
2-year higher education	9273	0.023		0	1
College/Master/PhD	9273	0.046		0	1
Missing	9273	0.089		0	1
Education level of Mom					
Primary or less	9273	0.795			
Middle/High School	9273	0.128		0	1
2-year higher education	9273	0.007		0	1
College/Master/PhD	9273	0.011		0	1
Missing	9273	0.058		0	1
Prep School Expenditure					

(continued on next page)

Variable	Obs.	Mean	Std.Dev.	Min	Max
No prep school	9246	0.010		0	1
Scholarship	9246	0.208		0	1
<1000 TL	9246	0.065		0	1
1000-2000 TL	9246	0.015		0	1
>2000 TL	9246	0.364		0	1
Income Level					
<250 TL	9122	0.459		0	1
250-500 TL	9122	0.386		0	1
>500 TL	9122	0.155		0	1
Time Spent in Turkish Preparation					
<100 hours	9273	0.078		0	1
100-200 hours	9273	0.102		0	1
>200 hours	9273	0.042		0	1
Time Spent in Social Science Preparation					
<100 hours	9273	0.081		0	1
100-200 hours	9273	0.095		0	1
>200 hours	9273	0.064		0	1
Time Spent in Math Preparation					
<100 hours	9273	0.116		0	1
100-200 hours	9273	0.072		0	1
>200 hours	9273	0.017		0	1
Time Spent in Science Preparation					
<100 hours	9273	0.074		0	1
100-200 hours	9273	0.014		0	1
>200 hours	9273	0.004		0	1

Table 6: Question outcomes for various parameter values: probabilities of skipping (S), being correct (C), being incorrect (I), expected score out of 45, and the reduction in expected score as compared to a risk neutral student of the same ability

β	Cutoff	Prob(S)	Prob(C)	Prob(I)	Expected Score	Loss vs Risk Neutral
2	0.2	0	0.405	0.595	11.57	-
2	0.225	0.012	0.403	0.585	11.57	0.00
2	0.25	0.085	0.386	0.529	11.43	0.14
2	0.275	0.192	0.359	0.449	11.12	0.45
2	0.3	0.303	0.328	0.370	10.58	0.99
2	0.325	0.403	0.297	0.300	9.99	1.58
3	0.2	0	0.535	0.465	18.86	-
3	0.225	0.003	0.534	0.463	18.86	0.00
3	0.25	0.030	0.528	0.442	18.81	0.05
3	0.275	0.081	0.515	0.404	18.63	0.23
3	0.3	0.143	0.498	0.360	18.36	0.50
3	0.325	0.208	0.478	0.315	17.96	0.90
4	0.2	0	0.619	0.381	23.58	-
4	0.225	0.001	0.619	0.380	23.58	0.00
4	0.25	0.017	0.616	0.368	23.58	0.00
4	0.275	0.049	0.608	0.344	23.49	0.09
4	0.3	0.091	0.596	0.314	23.27	0.31
4	0.325	0.137	0.581	0.281	23.00	0.58

Figure 4: Distribution of ÖSS-SÖZ Scores

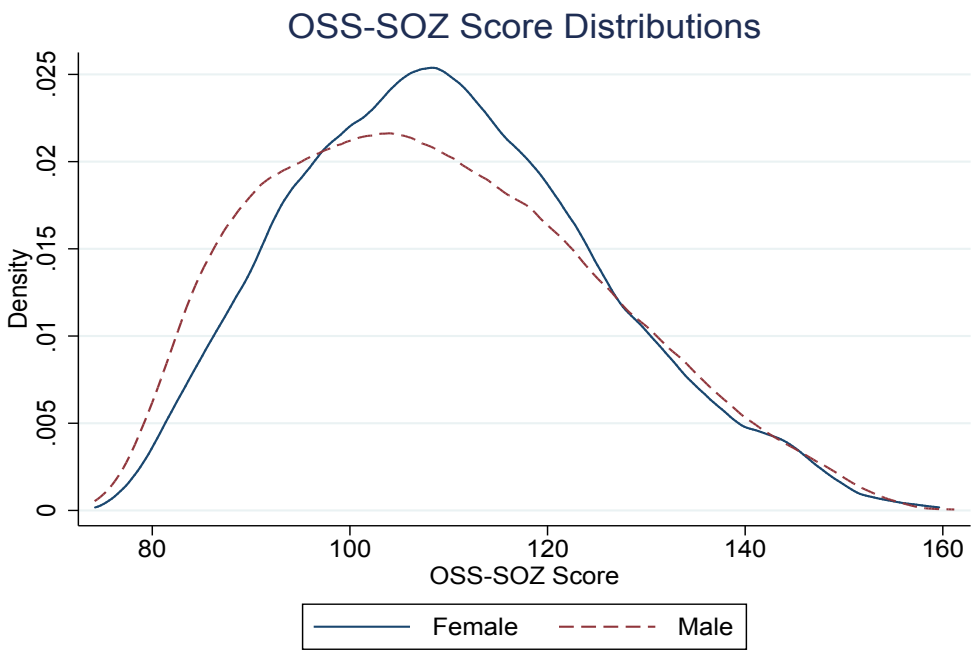


Figure 5: Distribution of Normalized GPA

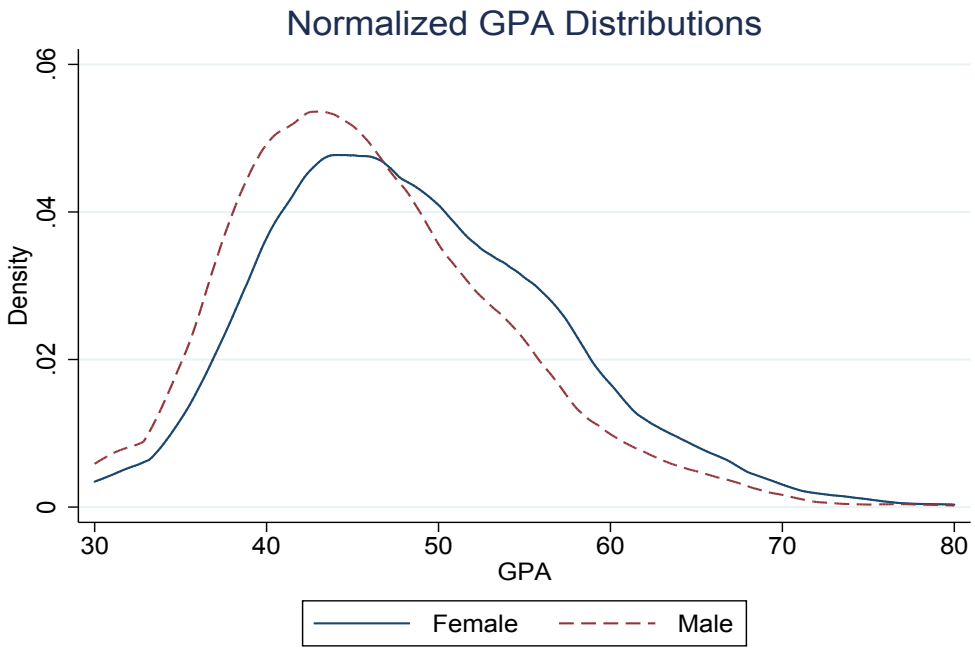


Figure 6: Distribution of Social Science Test Scores

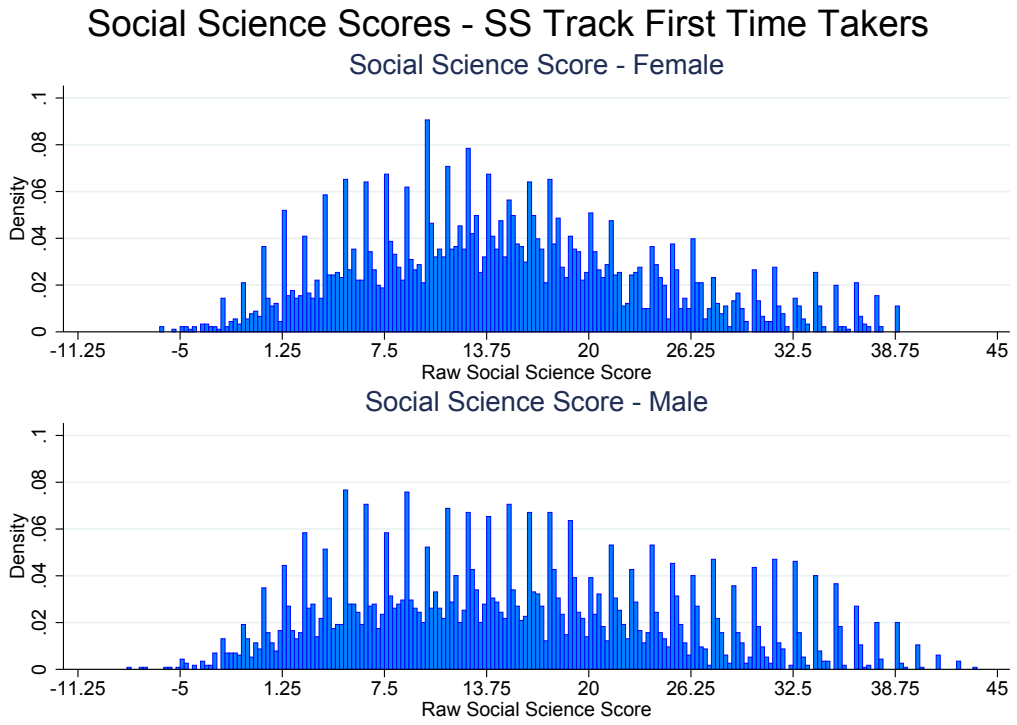


Figure 7: Distribution of Turkish Test Scores

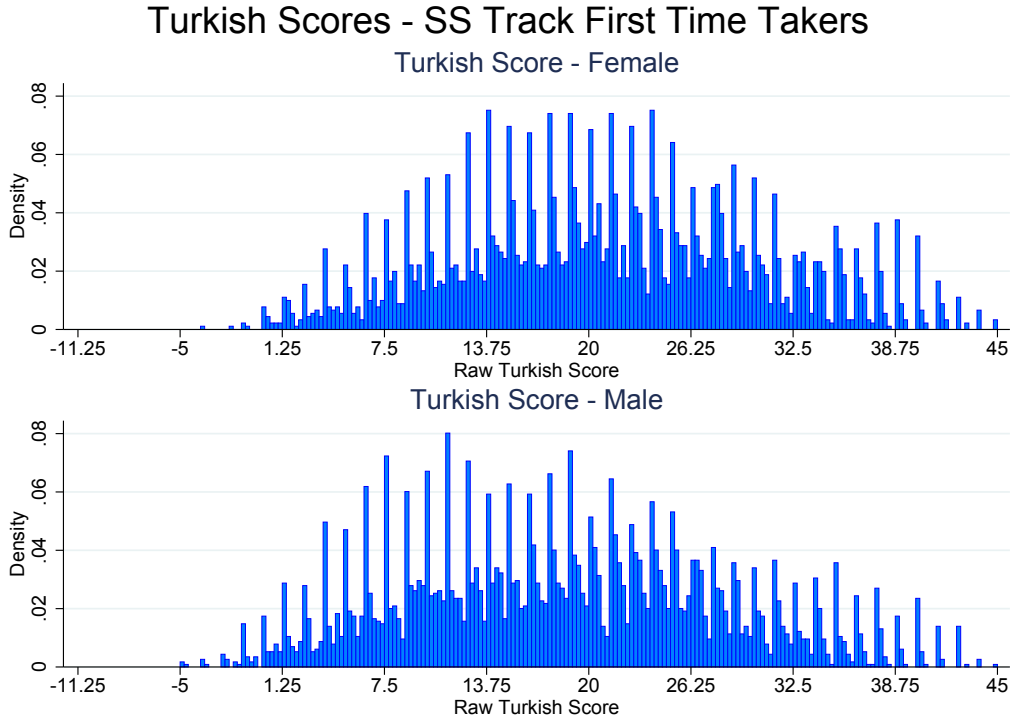


Table 7: Estimates of Risk Aversion Cutoff

	Female 1^{st} time takers	**Male 1^{st} time takers**
(0,90)	0.2135	0.2152
	(0.0028)	(0.0018)
[90,100)	0.2304	0.2269
	(0.0012)	(0.0009)
[100,110)	0.2387	0.2355
	(0.0008)	(0.0008)
[110,120)	0.2533	0.2493
	(0.0015)	(0.0012)
[120,130)	0.2661	0.2584
	(0.0028)	(0.0017)
[130,140)	0.2743	0.2619
	(0.0042)	(0.0026)
[140,∞)	0.2706	0.2639
	(0.0050)	(0.0032)

Standard errors are reported in parentheses.

Table 8: t-stat for test: (female cutoff-male cutoff)=0

	<90	90-100	100-110	110-120	120-130	130-140	> 140
Social Science Track	-0.51	2.43	2.86	2.06	2.38	2.50	1.14

Figure 8: Estimates of Attempt Cutoffs: Social Science Track

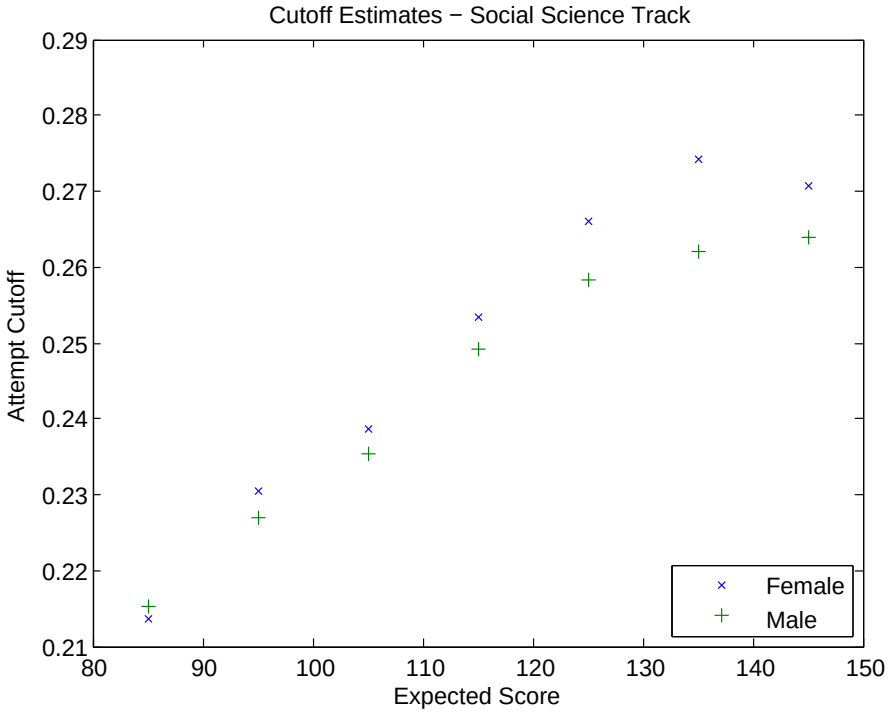


Table 9: Estimates of Ability Distribution Parameters

Social Science Test				
	Female		**Male**	
	μ	σ	μ	σ
(0,90)	-1.10	0.78	-1.38	0.97
	(0.21)	(0.18)	(0.18)	(0.12)
[90,100)	-0.65	0.71	-0.62	0.80
	(0.04)	(0.04)	(0.04)	(0.03)
[100,110)	-0.08	0.59	0.07	0.70
	(0.02)	(0.02)	(0.02)	(0.02)
[110,120)	0.49	0.56	0.69	0.65
	(0.02)	(0.02)	(0.02)	(0.02)
[120,130)	1.02	0.49	1.28	0.57
	(0.03)	(0.03)	(0.03)	(0.03)
[130,140)	1.49	0.43	1.72	0.57
	(0.05)	(0.05)	(0.05)	(0.04)
[140,∞)	1.91	0.43	2.22	0.34
	(0.06)	(0.08)	(0.05)	(0.08)
Turkish Test				
	Female		**Male**	
	μ	σ	μ	σ
(0,90)	-0.38	0.72	-0.85	0.77
	(0.16)	(0.13)	(0.11)	(0.09)
[90,100)	0.08	0.58	-0.15	0.67
	(0.03)	(0.03)	(0.03)	(0.02)
[100,110)	0.58	0.53	0.40	0.61
	(0.02)	(0.02)	(0.02)	(0.02)
[110,120)	1.10	0.52	0.93	0.57
	(0.02)	(0.02)	(0.02)	(0.02)
[120,130)	1.67	0.47	1.43	0.53
	(0.03)	(0.03)	(0.03)	(0.03)
[130,140)	2.23	0.42	1.98	0.56
	(0.05)	(0.05)	(0.05)	(0.05)
[140,∞)	2.91	0.61	2.62	0.56
	(0.09)	(0.10)	(0.08)	(0.09)

Standard errors are reported in parentheses.

Table 10: Estimates of Correlation between logs of Turkish and Social Science Ability

	Female 1^{st} time takers	**Male 1^{st} time takers**
(0,90)	1.000	0.783
	n/a	(0.108)
[90,100)	0.920	0.848
	(0.035)	(0.029)
[100,110)	0.948	0.888
	(0.024)	(0.021)
[110,120)	0.921	0.928
	(0.027)	(0.020)
[120,130)	0.990	0.900
	(0.042)	(0.036)
[130,140)	1.000	0.948
	n/a	(0.054)
[140,∞)	0.862	0.841
	(0.144)	(0.211)

Standard errors are reported in parentheses.

Figure 9: Data vs simulated distribution: social science, first time takers

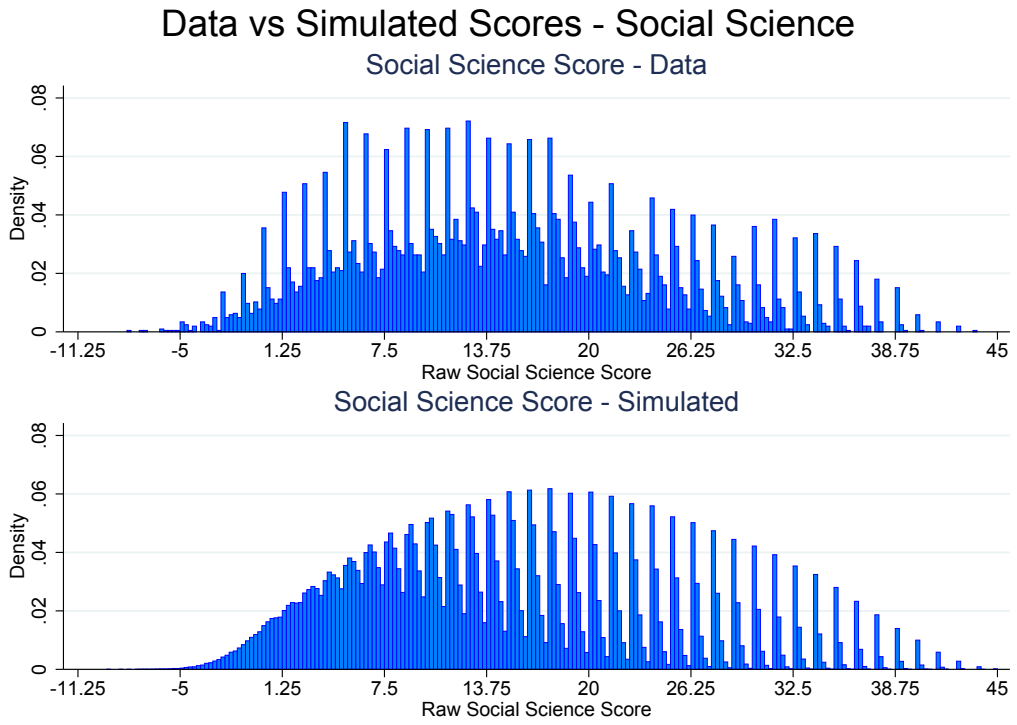


Figure 10: Data vs simulated distribution: Turkish, first time takers

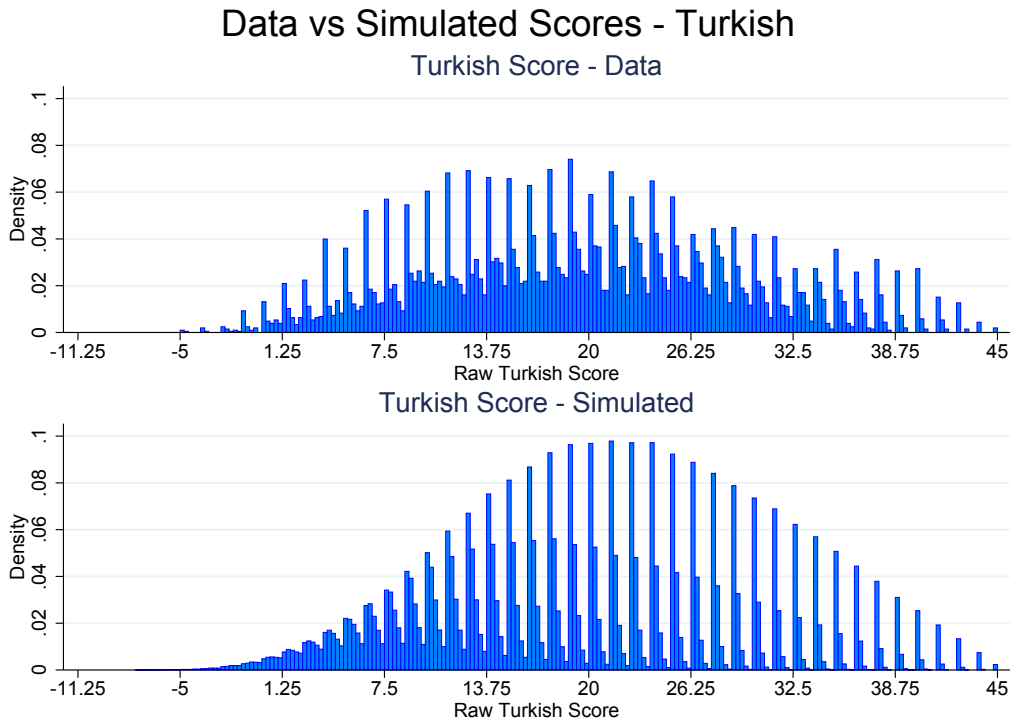


Figure 11: Distributions of Social Science and Turkish ability

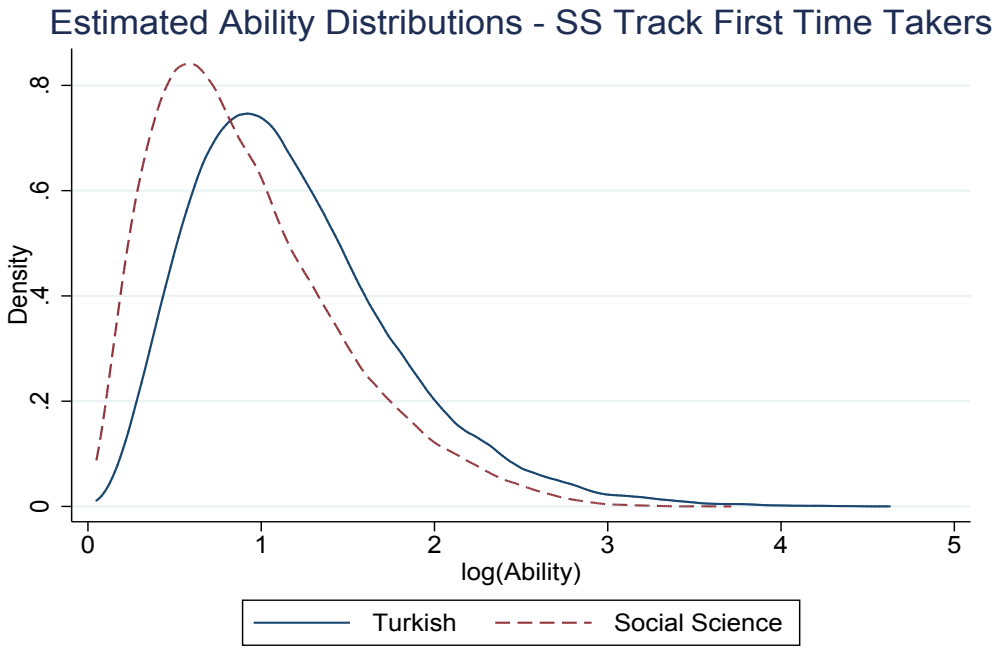


Figure 12: Distributions of Social Science ability - Female vs Male

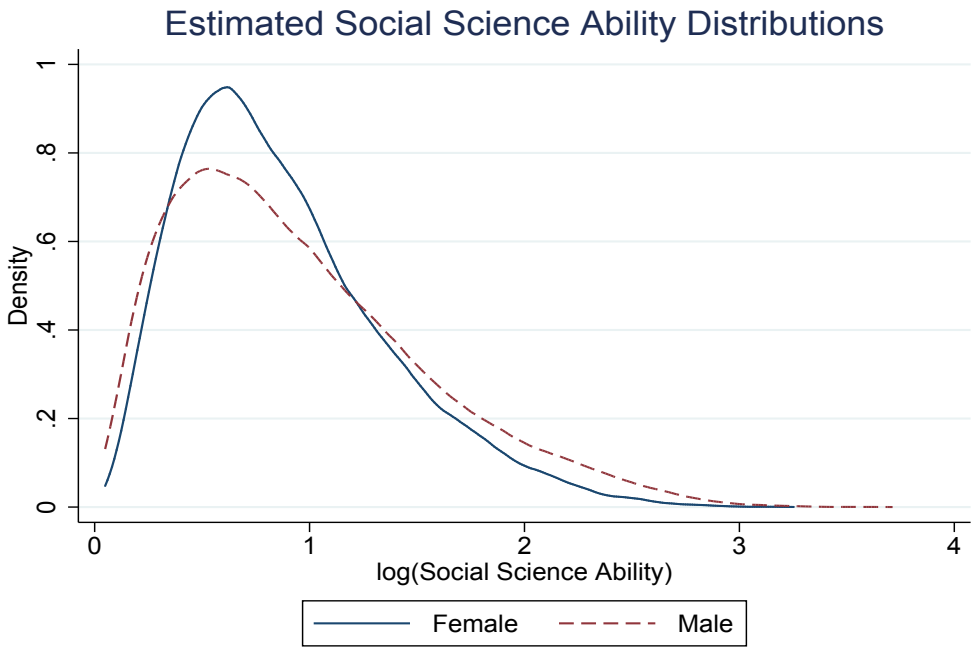


Figure 13: Distributions of Turkish ability - Female vs Male

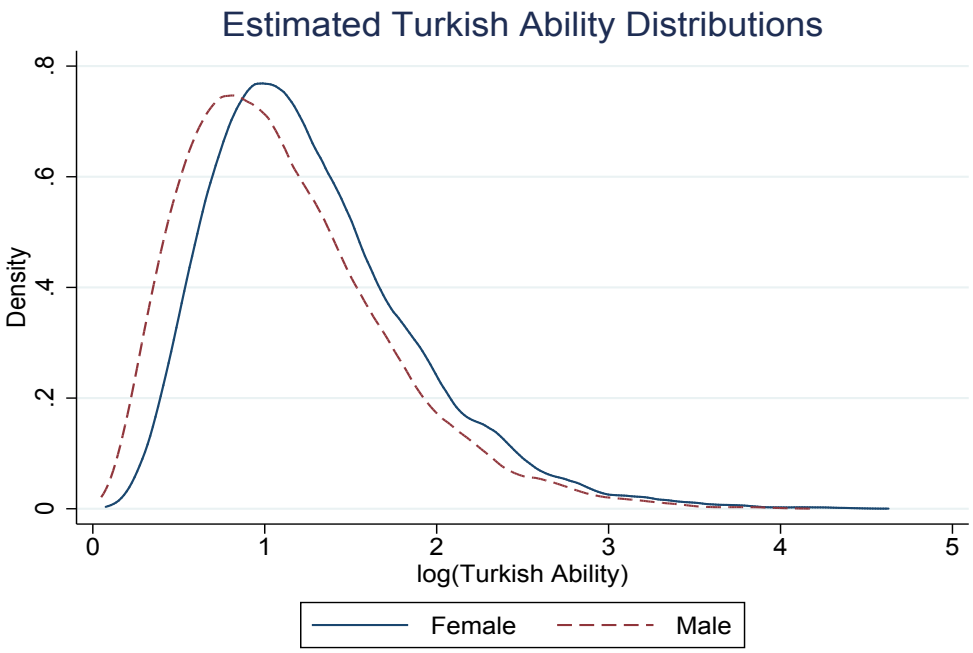


Figure 14: Turkish Ability Counterfactuals

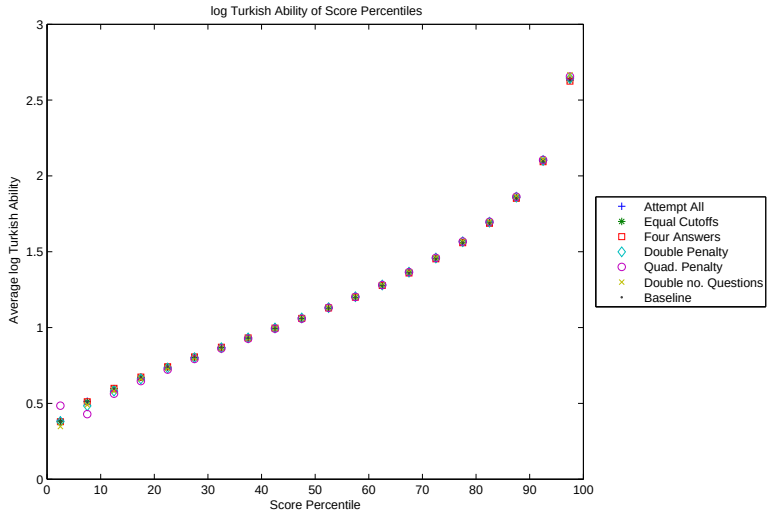


Figure 15: Turkish Ability (Δ vs baseline)

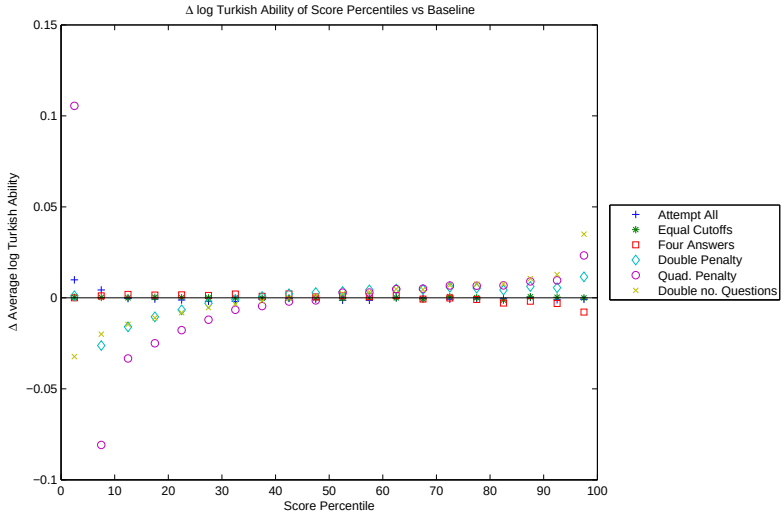


Figure 16: Social Science Ability Counterfactuals

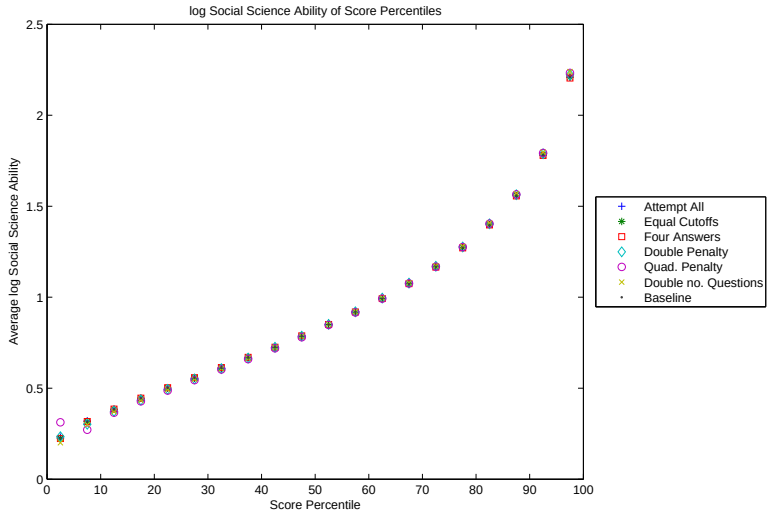


Figure 17: Social Science Ability (Δ vs baseline)

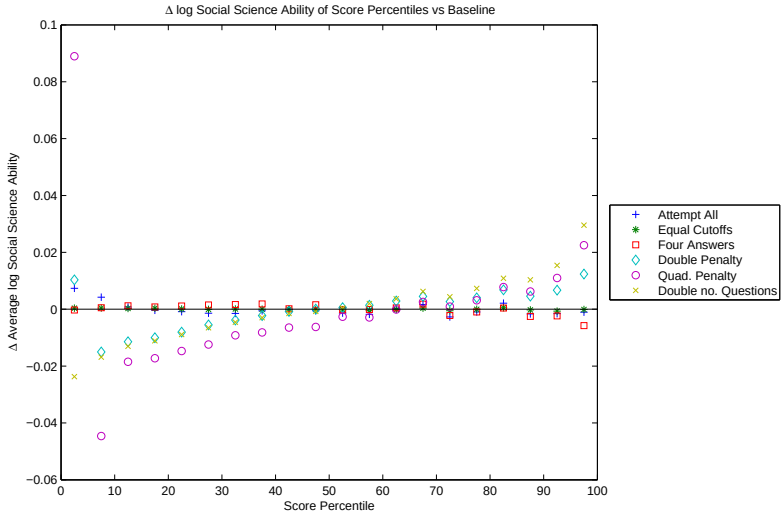


Figure 18: Male Fraction Counterfactuals

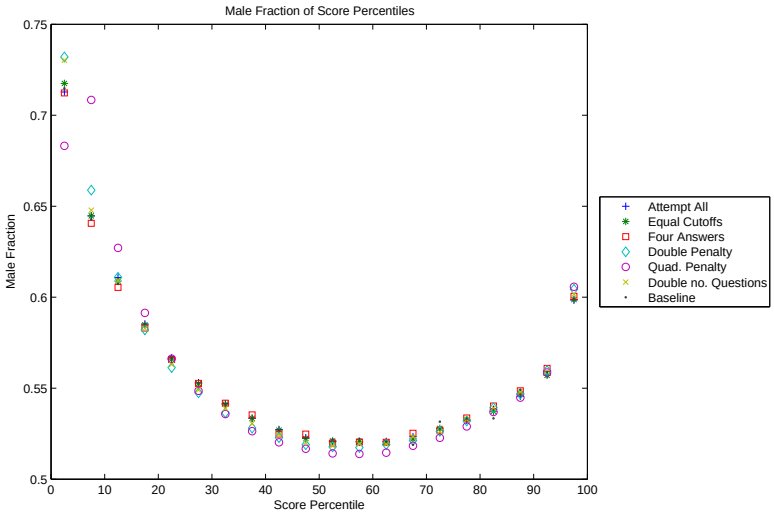


Figure 19: Male Fraction (Δ vs baseline)

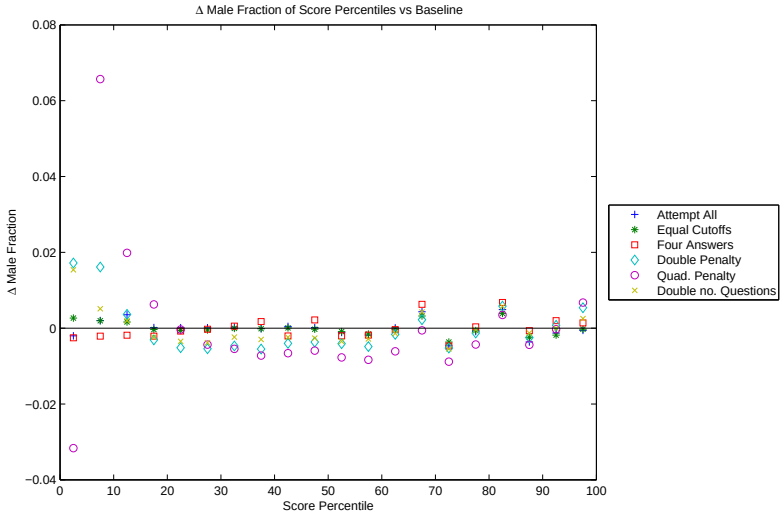


Figure 20: Male Fraction: Equal Cutoffs combined with Quadrupled Penalties

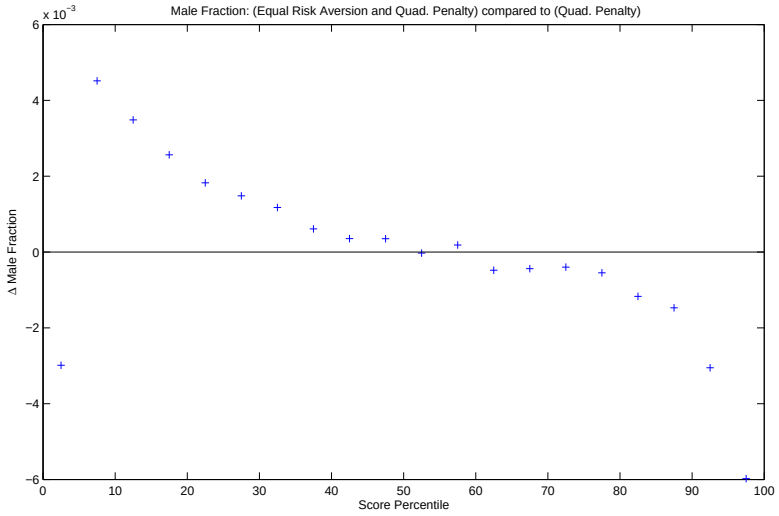


Figure 21: Turkish scores of Turkish Math track students (first attempt, male and female combined)

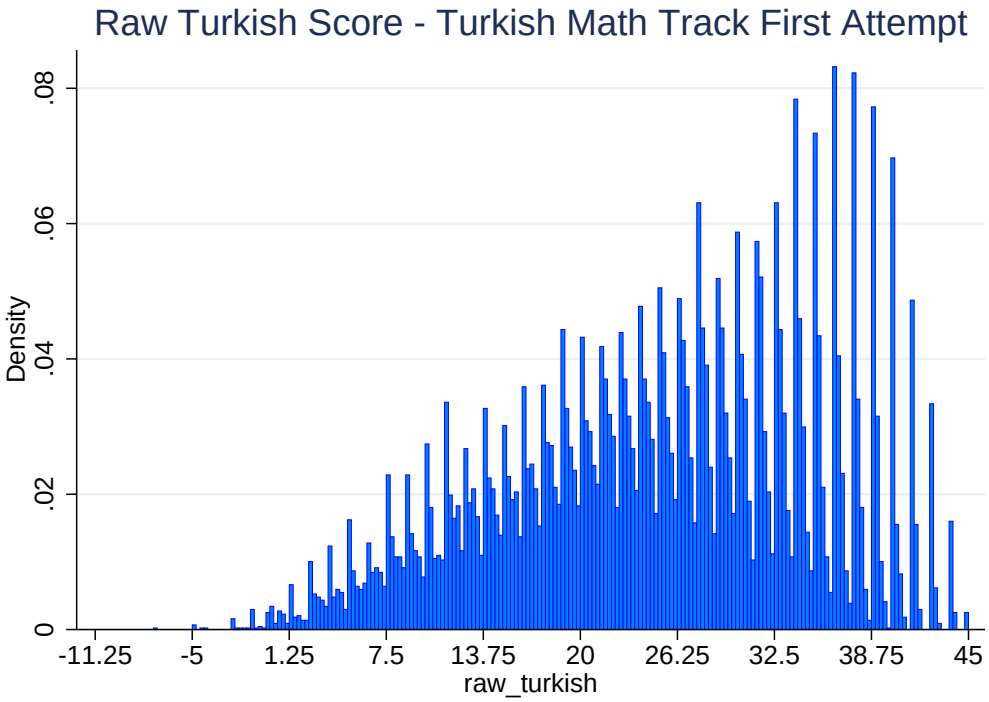


Figure 22: Language scores of Language track students (first attempt, male and female combined)

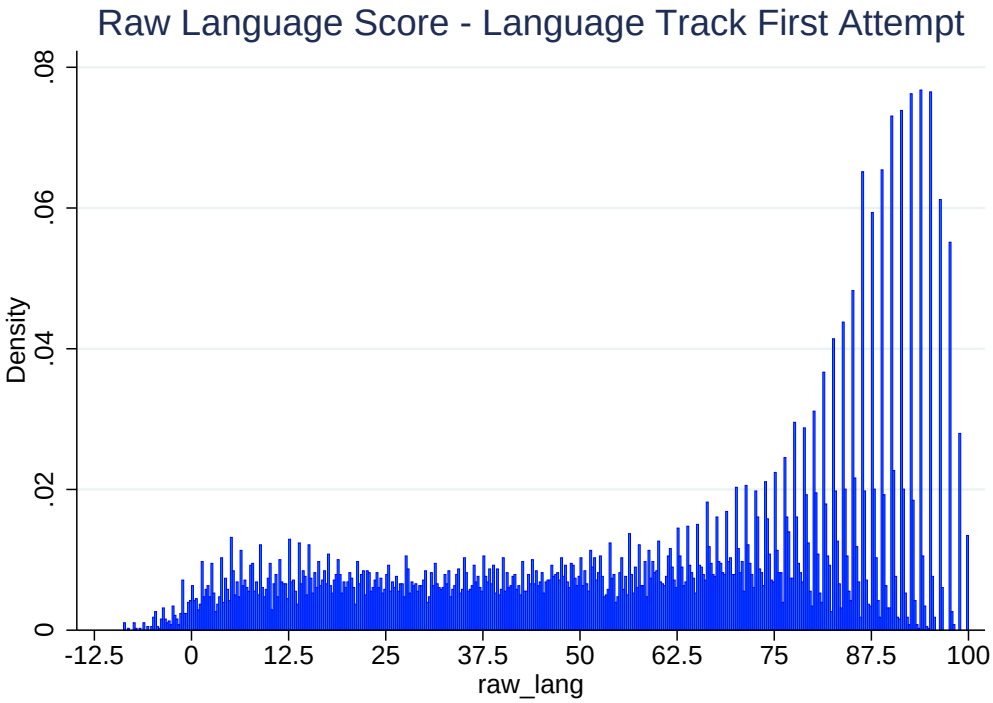


Table 11: Number of observations vs predicted score bin, Turkish-Math track

	80-90	90-100	100-110	110-120	120-130	130-140	> 140
Male	69	1224	2407	2011	1302	711	195
Female	21	825	2406	2137	1408	853	322

Table 12: Number of observations vs predicted score bin, Language track

	80-90	90-100	100-110	110-120	120-130	130-140	> 140
Male	259	417	618	735	823	642	187
Female	437	798	1299	1828	2181	2062	675

Figure 23: Estimates of Attempt Cutoffs: Turkish Math Track

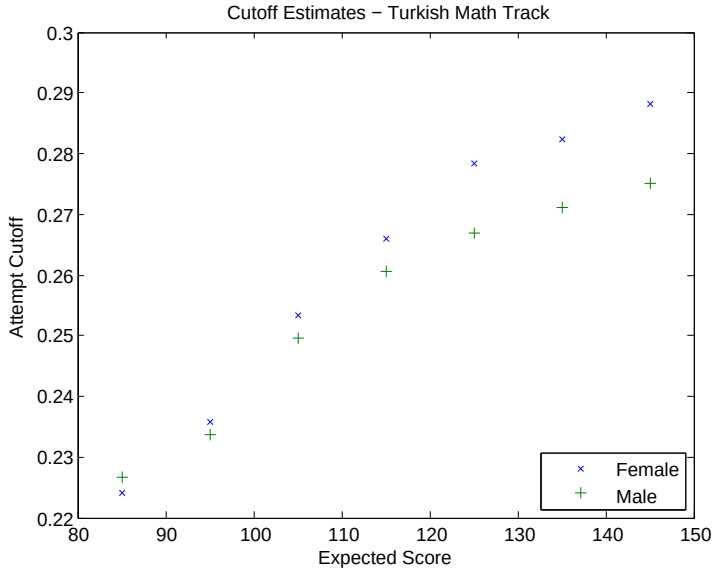


Figure 24: Estimates of Attempt Cutoffs: Language Track (English)