Hit or Miss? Test Taking Behavior in Multiple Choice Exams

Ş. Pelin Akyol *  James Key†  Kala Krishna‡

November 17, 2013

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.

*The Pennsylvania State University, e-mail: spd161@psu.edu
†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

The multiple choice test structure is commonly used to evaluate the knowledge of candidates in a wide variety of situations. Its main advantages are to allow more broad evaluation of candidate’s knowledge in a short time, it is easy to grade, and there is no subjective effect of the grader in the evaluation. Because of these properties, it is preferred in both high and low stake exams in many countries. Examples include the University Entrance exam in India, Turkey, Japan, China, and the SAT exam in the US. A disadvantage of such exams is that candidates may attempt to guess the answer without having any knowledge on it. In other exam types, such as short answer based exams, such uneducated responses are unlikely to reap any benefit. As a response to this problem, test administers may apply negative marking for wrong answers. Grading methods in multiple choice tests may be designed in such a way that the expected score from randomly guessing a question is equal to the expected score from skipping the question. This grading method is fair only under the assumption that candidates either know the answer, or they do not. However, if they have partial knowledge about the question, the candidate’s decision to guess/attempt or skip the question will not only depend on their knowledge, but also on their degree of risk aversion. This problem may undermine the validity and the fairness of test scores, reducing the efficacy of the testing mechanism.

The multiple choice test results are used to allocate students to colleges, to measure effectiveness of schools, teachers, or to allocate an open position. Baker et. al. (2) criticize the use of test results of students to evaluate the value-added of teachers and schools, among other reasons, because of the measurement error that will be generated by random guessing. Baldiga (3) shows in an experimental setting that conditional on students’ knowledge of the test material, those who skip more questions tend to perform worse. In light of this finding, if certain groups are favored in multiple choice tests, analysis made by using these scores will be upward biased towards these groups.

In this paper we model and estimate students’ multiple choice test taking behavior in order to understand the effects which characteristics of students have on test scores by

\footnote{In Turkey, public sector jobs are allocated according to the score obtained in a multiple choice central exam, called KPSS.}
using data from Turkish University Entrance Exam (ÖSS). The ÖSS is a highly competitive, centralized examination that is held once a year. In Turkey, college admission depends only on the score obtained from the ÖSS, and the high school GPA. However, the ÖSS score has the highest weight. In the ÖSS, for each correct answer the student gains one point, and for each wrong answer 0.25 points are deducted, while no points are awarded/deducted for skipping a question. As it is a very competitive and high-stake exam, students expend significant time and effort to get prepared for this exam. Therefore, we assume that they are aware of the scoring method. These properties of the ÖSS exam provide us with a convenient environment in which to investigate the exam taking behavior of students.

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

Risk attitudes of students are an important factor in the decision to attempt whenever there is uncertainty associated with the outcome. In the literature, females are shown to be more risk averse than males (see Eckel and Grossman). To test the hypothesis that female students skip more question than males since they are more risk averse, Ben-Shakhar

---

2 Every year only one third of the students are assigned to colleges.
3 This GPA is normalized at the school year level
4 ÖSS score has at least 75% weight in scores used to allocate students to colleges.
5 Essentially testing for the Framing Effect
and Sinai (4) investigates test taking strategies of students in Hadassah and PET tests in Israel and find that females do, in fact, tend to skip more questions. Baldiga (3) explores the gender differences in skipping/guessing behavior when students are uncertain about answers. They conduct an experiment to disentangle whether a gender gap in the tendency to skip questions exists, and if so, whether this gap is driven by differential confidence in knowledge of the material, differences in risk preferences, or differential responses to high pressure testing environment. Tannenbaum (14) also investigates the effect of gender differences in risk aversion on multiple choice test results. Tannenbaum (14) shows that female students are more risk averse, so they skip more often than male students. He finds that risk aversion is able to account for the 40% of the gender differences in performance in multiple choice exams. Tannenbaum (14) is the closest paper to ours. However, our conclusions conflict somewhat: we allow risk aversion to differ between various groups (gender, predicted scores) whereas he estimates a unique risk aversion for each gender.

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

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

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

In Turkey, college admission is based on an annual, nationwide, central university entrance exam governed by the Student Selection and Placement Center (ÖSYM). High school seniors and graduated students can take the exam. There is no restriction on retaking, i.e., students are allowed to take the exam repeatedly over the years. However, they are not allowed to carry over their scores: the score obtained in a year can be used only in that year. All departments, with the exception of those that require special talent (such as art schools) accept students based on a weighted average score of university entrance exam and high school grade point average.

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

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

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

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

---

\(^6\)For more detail on track choice in high school in Turkey see Akyol and Krishna.

\(^7\)In the calculation of ÖSS scores, firstly raw scores in each field are normalized to mean 50 and standard deviation 10 by using mean and standard deviation of scores in the corresponding field. Then these normalized scores are multiplied by the weights presented in Table 1.
and exam performance as follows

\[ Y_{\text{ÖSS}}X_i = \text{ÖSS}X_i + \alpha_{\text{AOBP}}X_i \]

where \( X \in \{ \text{SAY, SÖZ, EA, DIL} \} \), and \( \alpha \) is a constant that changes according to the student’s track, preferred department and whether the student was placed (accepted) into a regular program in the previous year or not. ÖSYM publishes the lists of departments open to students according to their tracks. When students choose a program from this list, \( \alpha \) will be 0.5, while if it is outside the open list, \( \alpha \) 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, \( \alpha \) will be 0.65. If the student was placed in a regular university program in previous year, the student is punished and \( \alpha \) will be equal to either 0.25, 0.1, or 0.375. For those students, the \( \alpha \) coefficient is equal to half of the regular \( \alpha \). This punishment structure gives students an incentive to stay in their track, and to accept a position when offered.

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

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

We received a random sample of around 40,000 students from each track (Social Science, Turkish-Math, Science). In this study we focus on the social science track students (from both regular high schools and Anatolian high schools). This is due to the unique patterns seen in these tests that enable identification of key parameters, patterns which do not exist in the science and math tests. This arises due to the style of questions being asked.\(^8\) As a result we focus on the students for whom these tests matter the most: the social science

\(^8\)In particular, math and science exams have many questions where the student must solve the problem to find the answer: either the student successfully solves the question or they fail to solve the question and have no information regarding what the correct answer is likely to be.
Table 10 presents the summary statistics.

We observe substantial differences in test outcomes between first time takers and second time takers. Figure 1 shows that second time takers achieve much higher scores, for the entirety of the distribution. An aim of this paper will be to examine whether or not this is due to the second time group having a better distribution of ability compared to first time takers, or if is simply due to a change in test taking behavior, for example their willingness to guess when uncertain.9

In addition, we see the difference between male and female students’ scores in Figures 2 and 3. Males tend to have a wider distribution, with more mass on the upper and lower tails. It is notable that the difference on the upper tail is more pronounced for second time takers; first time male test takers differ from their female counterparts mainly by having more students with low scores.

3 Multiple Choice Exam Scores

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

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

First, Figures 7, 8, 9, and 10 show the distribution of scores in the social science and Turkish portions of the exam, for first time takers and second time takers in the social science track. These histograms use a bin width equal to 1.10

\footnote{We do not separate learning and selection effects in this paper.}

\footnote{While the data is not rounded to the nearest integer, we do this so that the reader may see the overall distribution of scores - as will be seen shortly, this can be difficult to see when bins are not used.}
are roughly bell shaped, as would be expected. As noted in previous work (see Frisancho et al. (11)), the distributions appear to shift to the right as we move from first time takers to second time takers.

Next, we show the histograms of scores of social science track students in the math and science portions of the exam in Figures 11, 12, 13, and 14. Two facts immediately stand out in these diagrams. First, there are a lot of scores in the zero region. Secondly, even if the large spike at zero is removed, the distribution of scores place almost all of the weight at very low scores, relative to those seen in the Turkish and social science sections of the exam.

There are two explanations for both of the two preceding facts. First, the math and science test scores have relatively little weight in the ÖSS score of social science track students. These scores are multiplied by 0.4, whereas social science and Turkish scores are multiplied by 1.8 - a substantial difference. As a result, students have little incentive to expend time and effort on these questions during the exam. Accordingly, many students opt to not even look at these sections, resulting in many observations of zero scores. Furthermore, students are explicitly told that if they are social science track students, they should spend less time in math and science. If they do attempt questions, it will be relatively few, with less than advantageous results. A second reason is that these students are not well prepared in math and science - since the 9th grade they have been engaged in the social science track curriculum, meanwhile their study efforts would optimally be directed towards the exams that matter most: Turkish and social science.

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

There are 46 particular scores that are highly relevant: those which correspond to at-

\[11\] In the exam booklet there is a note before the social science/Turkish part of the exam that says: "If you want higher score in ÖSS-SÖZ, it may be better for you to spend more than 90 minutes on verbal part of the exam."
tempting all questions. These are spaced 1.25 points apart, starting at -11.25, finishing at 45 points.

Examining the distributions of actual scores (not placed into bins of width one), we can see some striking patterns. Figures 15 through 18 show the distributions for the social science and Turkish portions of the exam, with very prominent spikes along the distribution. It is no coincidence that the spikes appear evenly placed; they correspond to scores which could be achieved while attempting all questions. This will be very important in identifying the behavior, and implies that many students are skipping very infrequently. Additionally, there is no discernable pattern as to which students obtain such scores - these spikes remain present when examining the distributions of different groups, for example high income female first time takers with low GPAs

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

4 Model

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

We model the test taking procedure as if each test takes place separately— we do not allow for outcomes in one section of the test to have any bearing on other sections. In addition, each question is approached simultaneously, so that outcomes (or beliefs regarding outcomes) of one question have no impact on other questions.

---

12It is possible that there is some unobserved heterogeneity that determines which students lie on these spikes.
13This could correspond to students only answering the easiest question in the exam.
14We are explicitly ignoring time restrictions, whereby a quick performance in one section of the exam might afford the student additional time in another section, allowing the student to more carefully examine each question.
Signals for each of the five answers depend on whether or not the answer is actually the correct answer, and are drawn as follows:

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

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

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

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

$$X = (x_1, x_2, x_3, x_4, x_5) \quad (1)$$

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

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

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

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

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:

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

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

10
This can be further simplified to:

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

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

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

Note that the sum of beliefs for each of the five answers adds up to unity. Without loss of generality, we assume that \( \beta \geq \alpha \), which leads to positive relationship between the value of the signal, and the likelihood that the answer is correct. \[^{15}\][\(^{16}\)]

Proposition 1. The outcome of the model is the same for all \( A > 0 \)

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

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

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

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

where \( B = A^\alpha \). Similarly, the incorrect answers’ signals have the following distribution:

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

\[^{15}\text{A higher shape parameter for a Pareto distribution shifts probability mass to the left.}\]
\[^{16}\text{If a student were to draw from distributions with } \beta < \alpha, \text{ smaller signals would be associated with the correct answer.}\]
So this re-scaling of the signals preserves all of information contained in the original signals, and the resulting signals have a distribution identical to one where the correct answers come from a Pareto distribution with the shape parameter equal to one, and the incorrect answers have shape parameter equal to $\beta/\alpha$. As shown in the earlier proposition, the scale parameter is irrelevant. □

Accordingly, we can, without loss of generality, take $A = 1$ for all students, and take $\alpha = 1$ for all students. As a result, the structure of a student’s signals can be represented by the shape parameter of the incorrect answer: $\beta$. A higher value of $\beta$ 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 $\beta$ students are what would be referred to as high ability students.\(^{17}\)

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

Finally, there is a subtle effect of $\beta$ on tests. Students with high ability, i.e. a high value of $\beta$, will be more confident in their choices. Even with the same signals, as we increase $\beta$, 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 $x_{\text{max}} = \max\{x_1, \ldots, x_5\}$. The student with the higher value of $\beta$ has a higher belief that $x_{\text{max}}$ is drawn from the correct answer.

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

\(^{17}\)Signal distributions for a student with ability $\beta = 3$ are shown in Figure 23.
\[
\frac{d \log(\text{Belief})}{d \gamma} = \log x_{\max} - \frac{x_1^\gamma \log x_1 + x_2^\gamma \log x_2 + x_3^\gamma \log x_3 + x_4^\gamma \log x_4 + x_5^\gamma \log x_5}{x_1^\gamma + x_2^\gamma + x_3^\gamma + x_4^\gamma + x_5^\gamma}
\] (9)

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

\[
\frac{d \text{Belief}}{d \gamma} \geq 0
\] (10)

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

Once students have observed signals for each of the five possible answers to the question, they are faced with six possible alternatives: either choosing one of the five answers, or skipping the question. Skipping the question does not affect their ÖSS 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 \( 0.2 \times 1 - 0.8 \times 0.25 = 0 \).

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

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

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

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

In order to answer a question, with answer $n$, the signal drawn for answer $n$, $x_n$, must satisfy two conditions. First, it must be the highest signal. Second, it must be high enough, given the other signals, so that the belief is greater that the cutoff required to attempt the question. We define the following function as the minimum signal $x_n$ required to attempt with the $n^{th}$ answer, given the other signals:\footnote{As there will always exist an answer with probability of being correct greater than or equal to 0.2, we do not consider cutoffs below 0.2, as they would result in the same behavior: always attempting the question, never skipping}

$$K(\{x_i\}_{i \neq n}) = \max \left( \max \{x_i\}_{i \neq n}, \left( \frac{c}{1 - c} \left( \sum_{i \neq n} x_i^2 \right) \right)^{1/\gamma} \right)$$  \hspace{1cm} (11)

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

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

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

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

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

\footnote{A diagram showing choices conditional on signal observations for a simplified two answer setup is shown in Figure 24}
These lead to three functions that describe the probabilities of each of the three possible outcomes of a question, conditional on student ability $\beta$, and cutoff $c$:

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

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

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

where $P_S(\cdot) = 1 - P_C(\cdot) - P_W(\cdot)$. Table 3 provides these, in addition to the average points earned per question, for various parameter values$^{20}$ Consistent with the literature, as the probability to skip increases, the average points per question decreases (for a fixed ability)$^{21}$

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

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

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

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

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

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

$^{20}$A $\beta$ of 3 is later found to be approximately median.
$^{21}$Of course the average points per question attempted increases.
Accordingly, we can find the probability that a student with ability \( \beta \) and cutoff \( c \) obtains a score of \( s \). Suppose that there are \( k \) possible ways of obtaining such a score: \((y_j \text{ correct}, \ (x_j - y_j) \text{ incorrect, } (45 - x_j) \text{ skipped}) \) where \( j = 1, \ldots, k \). Thus, we obtain a mapping from \((\beta, c)\) to the probability of getting score \( s \):

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

5 Estimation Strategy

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

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

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

\[\text{To adjust for school quality, we adjust the GPA of student within a school based on the performance of the school in the exam. We observe normalize GPA for each students, which is able to be converted to a ranking within the school. As we observe the mean and variance of exam scores for each school, we can easily convert the GPA to a measure that reflects the quality of the school.}\]
This estimation is conducted separately for each gender/attempt number.

We divide students into groups, according to gender, attempt number, and the range into which their predicted ÖSS-SÖZ score lies: (0, 90), [90, 100), [100, 110), [110, 120), [120, 130), [130, 140), and [140, ∞). For each group, we examine the two subjects separately. While these intervals may not contain equal numbers of students, it will allow us to make comparisons across genders and attempt numbers. For each group, we take the cutoff \( c \), and the distribution of ability \( \beta \) within the group. The ability of each student is given by \( 1 + e^{X} \), where \( X \) is distributed normally with mean \( \beta_X \) and variance \( \sigma_X^2 \). This ensures that each student has an ability greater than 1, and results in a log normal shaped distribution (shifted 1 unit to the right).

Under the assumptions we made, the probability to get each score is approximated through simulation. For student \( n \), we take a draw from \( N(\beta_X, \sigma_X^2) \) and label it \( X_r \). For this value of \( X_r \), we find \( \beta_r = 1 + e^{X_r} \), the student’s ability parameter. As we now have \((\beta_r, c)\) for student \( n \), we can find the probability to get score \( s \), which is defined as \( M_n(\beta_r, c; s) \) in the previous section. We repeat this process \( S \) times and get simulated probability:

\[
\hat{M}_n(\beta, c; s) = \frac{1}{S} \sum_{r=1}^{S} M_n(\beta_r, c; s)
\]

We can now find for each student what is the probability to obtain each of the 220 possible scores for the test in question.

We compare this to the actual distribution observed in the data, by aggregating the students. Specifically, we look at moments related to the intensity of the spikes, and the shape of the distribution. More specifically, the difference between the mass of students with scores corresponding to attempting all questions (i.e. 45, 43.75,...) and the mass of students with scores corresponding to skipping a single question (i.e. 44, 42.75,...) captures in the intensity of the spikes. For example, \( 0.4 - 0.3 = 0.1 \). If the spikes are very prominent, this difference will be large; if they are non-existent, this difference will be minimal In addition, we use the mean of scores, and the variances of scores.

\[23\]The only tests of interest are Turkish and social science
\[
\left( \frac{1}{N} \sum_{n=1}^{N} g(\beta, c) \right) \left( \frac{1}{N} \sum_{n=1}^{N} g(\beta, c) \right)'
\] (22)

Accordingly, the estimates of the cutoff \( c \) and ability distribution parameters \( \mu, \sigma^2 \) for each group are estimated by minimizing the distance between the simulated moments and the observed moments.

\[
\hat{c}, \hat{\mu}, \hat{\sigma}^2 = \arg \min \left( \frac{1}{N} \sum_{n=1}^{N} g(\beta(\mu, \sigma), c) \right) \left( \frac{1}{N} \sum_{n=1}^{N} g(\beta(\mu, \sigma), c) \right)^\prime W_N \left( \frac{1}{N} \sum_{n=1}^{N} g(\beta(\mu, \sigma), c) \right)
\] (23)

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 (13) Gourieroux and Monfort (1997) (12) Duffie and Singleton (7)) 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.

5.1 Identification

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

This is illustrated in figure [25], where the score distribution for a student (with a fixed, approximately median, ability of \( \beta = 3 \)) is shown for various cutoff levels. A cutoff of \( c = 0.225 \) gives virtually all of the mass to the attempt all scores. As the risk aversion cutoff
increases to 0.3, the spikes all but disappear.

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

The parameters of the distribution of the ability of a group of students, $\beta_X$ and $\sigma^2_X$, are identified by the distribution of scores. An increase in the mean parameter $\beta_X$ moves the score distribution to the right, increasing the mean, while an increase in the variance parameter $\sigma^2_X$ increases the variance of the score distribution. This is due to strong relationship between ability and exam scores.

6 Results

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

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

Also noteworthy is the observation that the risk aversion cutoffs tend to decrease among high scoring students between first time takers and second time takers, whereas they tend to

---

24 Estimates for the second time takers in the ÖSS score range less than 90 are not obtained due to insufficient observations.
increase among the low scoring students. This pattern may be due to selection; high scoring risk loving students may self select into taking the exam a second time, while risk averse poor performing students may have dropped out of the system. However, it goes beyond the scope of this paper to disentangle the reasons behind this pattern.\textsuperscript{25}

Figures \textsuperscript{26} through \textsuperscript{33} show the simulated distributions compared to observed distributions for the various groups. While the estimation procedure was designed only to match subgroups of the sample, the entire simulated distribution fits the data relatively well, with some exceptions: it systematically under-predicts scores which correspond to skipping multiple questions.\textsuperscript{26} In addition, the skipping behavior is overestimated among low scoring students - this is likely due to such students correctly anticipating their low expected score and acting accordingly, whereas in the estimation many of these are restricted to acting with a (high) cutoff corresponding to their fitted score.

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

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

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

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

\textsuperscript{25}Additionally, such a pattern is consistent with a model where students do not perfectly know their true ability, and update their belief after taking the exam, see Lemma 1.

\textsuperscript{26}This issue is addressed in more detail in a later section.
As shown in figures 34 and 35, second time takers have higher abilities across the distribution, for both social science and Turkish. However, we can conclude definitively why this is the case.

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

In figures 36 and 37, we see how the genders differ the first time they take the exam. The lower portion of the social science ability distribution is indistinguishable, however males have a considerably better distribution for the top portion, compared to females. This is not the case with the Turkish portion—females are much better for all points in the distribution. This provides an interpretation of the observed differences in ÖSS-SÖZ scores. Males are overall worse at Turkish, but the best males make up for it in social science.

The second time takers exhibit a similar pattern in figures 38 and 39, however the advantage of males amongst the top social science students is more pronounced, and their disadvantage in Turkish decreases at the high end. As seen in the ÖSS-SÖZ score distributions, the comparison between male and female low ability students is similar to that of first time takers; whereas the high ability males gain a more favorable distribution (relative to females) in the second attempt.

27 See figures 2 and 3
7 Counterfactuals

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

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

The objects of interest in these experiments are as follows:

- The relationship between ability in social science ($\beta_{SS}$) and social science section exam score percentile
- The relationship between ability in Turkish ($\beta_T$) and social science section exam score percentile
- The relationship between gender and social science section exam score percentile
- The relationship between attempt number and social science section exam score percentile

The first two objects are clearly important. The ÖSS exam system is an allocation mechanism, presumably designed to give the students with high abilities access to the most desirable institutes of higher education. However, the exam score is an imperfect measure of student ability. The students who score in the top one percent are not necessarily those in the top one percent as 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 (8)). In addition, we have found that females tend to have higher thresholds for attempting to answer a question, i.e. they are more risk averse. Since the testing regime in question is forcing students to accept an element of risk when choosing to answer a question, the preferences regarding risk affect the distribution of final exam scores. This may tend to favor male test takers, leading in essence to a systemic bias in the testing procedure. In addition, the distributions of abilities are considerably different across gender; the regime may end up attenuating these differences.

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

The seven possible regimes used in this counterfactual experiment are:

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

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

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

4. All students attempt all questions. This is equivalent to assuming that all students are risk neutral/loving \(^{28}\)

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

\(^{28}\)This is also more or less identical to removing the penalty for answering incorrectly - both 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 \(45\).
6. The penalty for answering incorrectly is increased from 0.25 points to 0.5 points

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

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

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

As a result, the cutoffs for attempting a question must be different. To convert cutoffs from the five answer case, we first assume a CARA utility function, and solve for the risk aversion parameter that generates a given cutoff. This is repeated for each student. We then convert the risk aversion parameter to a cutoff in the four answer case.\(^{29}\)

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

For each of the seven possible regimes, we find the resulting distributions of scores for the entire sample of students, and segment scores into bins of five percent.\(^{30}\)\(^{31}\) For each of the twenty bins, ranging from the lowest scores to the highest, we find four objects of interest: share of males, share of first time takers, average social science ability and average Turkish ability.\(^{32}\)

Figures 40 through 43 show how the four objects of interest differ across the different regimes. As expected, given their greater exam score variance, the male fraction is u-shaped,\(^ {29}\) 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.\(^ {30}\) Five percent of the number of students\(^ {31}\) The rationale behind segmenting into percentiles, not scores, is to see the effects on the resulting allocation of students to university programs.\(^ {32}\) The average of log(\(\beta\)) is used for each subject.
as shown in figure 40. However, there is no discernable pattern in the differences between
the seven regimes throughout most of the range. While there are some small differences
around the median, all seven are fairly similar. We do see some small differences in the top
performing students: “Attempt All” gives a lower Male fraction (the more abundant Males
are seemingly disadvantaged by the added variance in scores) whereas the “Higher Penalty”
regime has a very slightly higher Male fraction, i.e. the small difference in risk aversion
begins to have a small effect.

This insensitivity is even more apparent when examining the other graphs. Second time
takers are dominant in the high exam scores percentiles, as would be expected, with the
seven curves lying on top of each other in figure 41.

Figures 42 and 43 have higher ability students in the higher score percentiles, with no
differences across the six cases featuring risk aversion. We do, however, see that the “Attempt
All” regime gives slightly lower average abilities amongst the highest performing students,
consistent with penalties and risk aversion allowing an improved separation of students by
abilities

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.

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 Extension

While this paper has focused on the social science track students, and in particular on the social science and Turkish tests, a variant of the model can be constructed to be applied to the science and math exams.

If one were to apply the model to the science and math exams (for the science track students) one would find a very high belief threshold for answering questions. This is due to the absence of spikes\(^{33}\) throughout most of the distribution\(^{34}\). However, this is not likely to be the case: instead of science students being extremely risk averse, it is simply much more reasonable to posit that the style of questions asked is inherently different.

For example, consider a math question involving geometry, where the student is asked to find the value of a particular angle. There are five possible answers provided, but it would not be advisable strategy to look at each answer and contemplate the degree to which it is likely to satisfy the implied system of equations. Rather, the student would simply solve for \(x\), see that her solution is found in the list of possible answers, and choose that as her answer. Or, they would fail to solve the question, and realize that each of the answers has a twenty percent probability of being correct, and so they would skip the question.

Clearly, the situation described is not well modeled with the distributional assumptions of the baseline model. Instead, it is reasonable for there to be some proportion of the questions in an exam modeled as in the standard model, while the other questions could have a bimodal distribution for the signal of the correct answer. If the student is not able to solve the question, the signal for the correct answer will be found close to the lower mode, while if they solve the question, the signal will be close to the upper mode. Higher ability students would have a greater probability mass around the higher mode, i.e. they are more

\(^{33}\) Exceptional probability mass being placed on scores corresponding to attempting all questions

\(^{34}\) With the exception of scores in the 40-45 range
likely to solve the question.

Such an extension would explain why we see so much skipping behavior, even among high scoring students. While in the standard case a high ability student is able to be confident enough most of the time, with the bimodal case they would in essence usually skip a question unless they solve the question: their belief about the likelihood of being correct is either very low or very high.

In addition to enabling analysis of the science track exams, such an extension will improve the fit of the model in the social science track exams. While the model does a reasonable job in matching the patterns seen in the score distributions, it is not capable of matching certain areas: namely the scores which could only be obtained by skipping multiple questions. The reasoning for this is that skipping multiple questions is highly unlikely in the model for all but the lowest ability students. As a result, the model under predicts the likelihood of such scores.

The simple explanation is that there are some questions in the social science and Turkish exams that fall into the category of the bimodal distribution of the correct answer. Either students know the correct answer, or they have no idea. While such questions are evidently not the majority, there are enough to impact the resulting distributions.

A rudimentary method for allowing for such behavior is to allow a fraction of students to have a much higher propensity to skip, while the remainder of the students act as before, with a reasonable, low cutoff. While this clearly isn’t completely in line with the extended model, it will allow more skipping whilst retaining the spikes, allowing a better fit, and enable interpretation of the results.

As before, for each group (for example, female first time students in the ÖSS-SÖZ score range 120-130, Turkish exam) we find the cutoff (for the low type cutoff) below which they skip the question. In addition, we find the proportion of students that act as if they are very risk averse (taken as a cutoff of 0.50) for each group. As before, the first term is captured by the prominence of the scores corresponding to attempting all questions. Meanwhile, the second term is identified by how sharply the probability decreases as we move from scores

---

35 For example they all but rule out one or more incorrect answers
36 For example, scores which are 0.25 points below the spikes; for example 38.75 could be obtained by getting 40 correct and 5 wrong, 38.5 could only be obtained by getting 39 right, 2 wrong and 4 skipped.
corresponding to skipping a single question, to those corresponding to skipping two questions, and so on.

Estimates of the parameters are shown in tables 6, 7, 8, and 9. The estimates of the cutoffs (for the low type students) follow similar patterns to the original model, and are relatively similar, with the exception of the average students.

The proportion of high type students versus expected total score grouping follows a u-shaped pattern. This explains why the estimates for the average students have changed more than other groups: there was significant skipping behavior which the basic model attempted to account for. It is also what would be expected by the model. High ability students should be able to "solve" the bimodal type questions much more frequently than the average ability students, resulting in a low amount of skipping. This would result in the basic model (where the proportion of low type students is equal to one) matching relatively well. Average ability students on the other hand would be unable to solve the questions considerably more frequently, which results in this exercise finding a small probability of the low cutoff type students. Students with low ability are perhaps likely to attempt questions even if they have no information regarding the answer - unless their score gets above 105 points they cannot get into any university/college.

The resulting distribution also provides a much better fit, compared to the basic model, as seen in figures 44 through 51. In the lens of the extended model this would be due to the presence of a few questions that would only be answered (with any degree of certainty) when solved.

An additional dimension through which the model can be extended is the simultaneous nature of the exam. Time is scarce for many students, it follows that spending additional time on one section of the exam reduces the time available for other sections of the exam. Conversely, students with a high ability in one section of the exam will be able to finish that section relatively briskly, and will have more time available to ponder the other section.

This constraint could be addressed by a model in which a student possesses two ability terms - one for social science and the other for Turkish - and the ability to distinguish correct from incorrect depends on both inherent ability, and the amount of time allocated to the question. The ability to distinguish is increasing in both, however the two could be said to
be substitutes. The student then optimally decides how much time to assign each section of the exam. Preliminary evidence tends to support this: the correlation between estimated Turkish and social science ability parameters tends to be higher for low ability students - consistent with the notion that high ability students are not affected by time constraint as much as low ability students.

Estimation would proceed similar to the simple model, except instead of solving for abilities in each section separately, they would be estimated jointly, with a much more complicated objective function. In addition, the relationship between ability, time per question, and likelihood of choosing the right answer would need to be estimated.

---

37 For feasibility, it would need to be assumed that the student looks at all questions in the social science and Turkish sections, and none of the science/math questions, which is not too far from reality for the majority of students.

38 If a student is time constrained, a higher ability in Turkish will afford them more time in social science, allowing them to perform better in the social science.

39 A candidate would be a CES function, where the effective ability of a student, $\beta_e$, takes the form $\beta_e = (\beta^{t-1} + t)^{\frac{1}{t-1}}$. Here, $\beta$ would be the true ability of the student, while $t$ is the time allocated to the question.
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 (resulting in zero expected return from a random guess), we find that differences have very little bearing on aggregate outcomes.

References


10 Appendix

Table 1: Test Weights

<table>
<thead>
<tr>
<th>Track</th>
<th>Math</th>
<th>Science</th>
<th>Turkish</th>
<th>Social Science</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science Track (ÖSS-SAY)</td>
<td>1.8</td>
<td>1.8</td>
<td>0.4</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>Social Science Track (ÖSS-SÖZ)</td>
<td>0.4</td>
<td>0.4</td>
<td>1.8</td>
<td>1.8</td>
<td>0</td>
</tr>
<tr>
<td>Turkish-Math Track (ÖSS-EA)</td>
<td>0.8</td>
<td>0.4</td>
<td>0.8</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>Language Track (ÖSS-DIL)</td>
<td>0</td>
<td>0</td>
<td>0.4</td>
<td>0.4</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Table 2: Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std.Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (Male=1)</td>
<td>37650</td>
<td>0.567</td>
<td>0.496</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ÖSS-SÖZ</td>
<td>37650</td>
<td>116.737</td>
<td>18.912</td>
<td>0</td>
<td>161.166</td>
</tr>
<tr>
<td>Normalized High School GPA</td>
<td>37650</td>
<td>47.550</td>
<td>7.825</td>
<td>30</td>
<td>80</td>
</tr>
<tr>
<td>Raw Turkish Score</td>
<td>37650</td>
<td>23.983</td>
<td>9.965</td>
<td>-10</td>
<td>45</td>
</tr>
<tr>
<td>Raw Social Science Score</td>
<td>37650</td>
<td>19.229</td>
<td>10.397</td>
<td>-8.75</td>
<td>45</td>
</tr>
<tr>
<td>Raw Math Score</td>
<td>37650</td>
<td>1.853</td>
<td>4.061</td>
<td>-9</td>
<td>42.5</td>
</tr>
<tr>
<td>Raw Science Score</td>
<td>37650</td>
<td>0.197</td>
<td>1.296</td>
<td>-8.75</td>
<td>25.75</td>
</tr>
</tbody>
</table>

**Education level of Dad**

<table>
<thead>
<tr>
<th>Level</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std.Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary or less</td>
<td>37650</td>
<td>0.557</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Middle/High school</td>
<td>37650</td>
<td>0.281</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2-year higher education</td>
<td>37650</td>
<td>0.028</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>College/Master/Phd</td>
<td>37650</td>
<td>0.050</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Missing</td>
<td>37650</td>
<td>0.085</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Income Level**

<table>
<thead>
<tr>
<th>Level</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std.Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;250 TL</td>
<td>37015</td>
<td>0.459</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>250-500 TL</td>
<td>37015</td>
<td>0.393</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>500-750 TL</td>
<td>37015</td>
<td>0.095</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>750-1000 TL</td>
<td>37015</td>
<td>0.030</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

(continued on next page)
<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std.Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000-1500 TL</td>
<td>37015</td>
<td>0.012</td>
<td>0.012</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1500-2000 TL</td>
<td>37015</td>
<td>0.006</td>
<td>0.006</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>&gt;2000 TL</td>
<td>37015</td>
<td>0.005</td>
<td>0.005</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Number of Attempts**

<table>
<thead>
<tr>
<th>Attempt</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std.Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st attempt</td>
<td>35470</td>
<td>0.261</td>
<td>0.261</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2nd attempt</td>
<td>35470</td>
<td>0.250</td>
<td>0.250</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3rd attempt</td>
<td>35470</td>
<td>0.242</td>
<td>0.242</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4th attempt</td>
<td>35470</td>
<td>0.154</td>
<td>0.154</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5th attempt</td>
<td>35470</td>
<td>0.092</td>
<td>0.092</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Prep School Expenditure**

<table>
<thead>
<tr>
<th>Expenditure</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std.Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>No prep school</td>
<td>37577</td>
<td>0.262</td>
<td>0.262</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Scholarship</td>
<td>37577</td>
<td>0.008</td>
<td>0.008</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>&lt;1000 TL</td>
<td>37577</td>
<td>0.223</td>
<td>0.223</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1000-2000 TL</td>
<td>37577</td>
<td>0.095</td>
<td>0.095</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>&gt;2000 TL</td>
<td>37577</td>
<td>0.034</td>
<td>0.034</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Missing</td>
<td>37577</td>
<td>0.378</td>
<td>0.378</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Time Spend in Preparation in 11th Grade**

<table>
<thead>
<tr>
<th>Preparation</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std.Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turkish Preparation</td>
<td>37650</td>
<td>0.439</td>
<td>0.874</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Social Science Preparation</td>
<td>37650</td>
<td>0.465</td>
<td>0.905</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Math Preparation</td>
<td>37650</td>
<td>0.301</td>
<td>0.653</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Science Preparation</td>
<td>37650</td>
<td>0.125</td>
<td>0.393</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
Figure 1: Distribution of OSS-SOZ scores, first vs second time takers

Figure 2: Distribution of OSS-SOZ scores, first time takers, male vs female
Figure 3: Distribution of OSS-SOZ scores, second time takers, male vs female

Figure 4: Density of OSS-SOZ scores, first vs second time takers
Figure 5: Density of OSS-SOZ scores, first time takers, male vs female

Figure 6: Density of OSS-SOZ scores, second time takers, male vs female
Figure 7: Histogram of scores from social science test, first time takers, bins of width one

Figure 8: Histogram of scores from social science test, second time takers, bins of width one
Figure 9: Histogram of scores from Turkish test, first time takers, bins of width one

Figure 10: Histogram of scores from Turkish test, second time takers, bins of width one
Figure 11: Histogram of scores from math test, first time takers, bins of width one

Figure 12: Histogram of scores from math test, second time takers, bins of width one
Figure 13: Histogram of scores from science test, first time takers, bins of width one

Figure 14: Histogram of scores from science test, second time takers, bins of width one
Figure 15: Distribution of scores from social science test, first time takers
Figure 16: Distribution of scores from social science test, second time takers

Figure 17: Distribution of scores from Turkish test, first time takers
Figure 18: Distribution of scores from Turkish test, second time takers

Figure 19: Distribution of scores from math test, first time takers
Figure 20: Distribution of scores from math test, second time takers

Figure 21: Distribution of scores from science test, first time takers
Figure 22: Distribution of scores from science test, second time takers
Figure 23: Distributions of signals for a student with $\beta = 3$, approximately median
Figure 24: Action conditional on signals for a simple two answer model (parameter values: \( \beta = 3 \) and cutoff = 0.55)
Table 3: Question outcomes for various parameter values: probabilities of skipping, being correct, being incorrect, and the average points per question

<table>
<thead>
<tr>
<th>β</th>
<th>Cutoff</th>
<th>Prob(S)</th>
<th>Prob(C)</th>
<th>Prob(I)</th>
<th>PPQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.2</td>
<td>0</td>
<td>0.405</td>
<td>0.595</td>
<td>0.257</td>
</tr>
<tr>
<td>2</td>
<td>0.225</td>
<td>0.012</td>
<td>0.403</td>
<td>0.585</td>
<td>0.257</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>0.085</td>
<td>0.386</td>
<td>0.529</td>
<td>0.254</td>
</tr>
<tr>
<td>2</td>
<td>0.275</td>
<td>0.192</td>
<td>0.359</td>
<td>0.449</td>
<td>0.247</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>0.303</td>
<td>0.328</td>
<td>0.370</td>
<td>0.235</td>
</tr>
<tr>
<td>2</td>
<td>0.325</td>
<td>0.403</td>
<td>0.297</td>
<td>0.300</td>
<td>0.222</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0</td>
<td>0.535</td>
<td>0.465</td>
<td>0.419</td>
</tr>
<tr>
<td>3</td>
<td>0.225</td>
<td>0.003</td>
<td>0.534</td>
<td>0.463</td>
<td>0.419</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>0.030</td>
<td>0.528</td>
<td>0.442</td>
<td>0.418</td>
</tr>
<tr>
<td>3</td>
<td>0.275</td>
<td>0.081</td>
<td>0.515</td>
<td>0.404</td>
<td>0.414</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
<td>0.143</td>
<td>0.498</td>
<td>0.360</td>
<td>0.408</td>
</tr>
<tr>
<td>3</td>
<td>0.325</td>
<td>0.208</td>
<td>0.478</td>
<td>0.315</td>
<td>0.399</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0</td>
<td>0.619</td>
<td>0.381</td>
<td>0.524</td>
</tr>
<tr>
<td>4</td>
<td>0.225</td>
<td>0.001</td>
<td>0.619</td>
<td>0.380</td>
<td>0.524</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>0.017</td>
<td>0.616</td>
<td>0.368</td>
<td>0.524</td>
</tr>
<tr>
<td>4</td>
<td>0.275</td>
<td>0.049</td>
<td>0.608</td>
<td>0.344</td>
<td>0.522</td>
</tr>
<tr>
<td>4</td>
<td>0.3</td>
<td>0.091</td>
<td>0.596</td>
<td>0.314</td>
<td>0.517</td>
</tr>
<tr>
<td>4</td>
<td>0.325</td>
<td>0.137</td>
<td>0.581</td>
<td>0.281</td>
<td>0.511</td>
</tr>
</tbody>
</table>
Figure 25: Distribution of scores resulting from various cutoff levels
Table 4: Estimates of Risk Aversion Cutoff

Social Science Test

<table>
<thead>
<tr>
<th></th>
<th>Female 1st time takers</th>
<th>Female 2nd time takers</th>
<th>Male 1st time takers</th>
<th>Male 2nd time takers</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 90)</td>
<td>0.2078 (0.0004)</td>
<td>0.2128 (0.0010)</td>
<td>0.2394 (0.0042)</td>
<td></td>
</tr>
<tr>
<td>[90, 100)</td>
<td>0.2186 (0.0008)</td>
<td>0.2212 (0.0008)</td>
<td>0.2305 (0.0008)</td>
<td>0.2328 (0.0011)</td>
</tr>
<tr>
<td>[100, 110)</td>
<td>0.2280 (0.0008)</td>
<td>0.2319 (0.0009)</td>
<td>0.2416 (0.0008)</td>
<td>0.2380 (0.0009)</td>
</tr>
<tr>
<td>[110, 120)</td>
<td>0.2413 (0.0008)</td>
<td>0.2427 (0.0008)</td>
<td>0.2416 (0.0008)</td>
<td>0.2380 (0.0009)</td>
</tr>
<tr>
<td>[120, 130)</td>
<td>0.2501 (0.0005)</td>
<td>0.2483 (0.0007)</td>
<td>0.2448 (0.0007)</td>
<td>0.2421 (0.0010)</td>
</tr>
<tr>
<td>[130, 140)</td>
<td>0.2585 (0.0005)</td>
<td>0.2550 (0.0006)</td>
<td>0.2520 (0.0008)</td>
<td>0.2472 (0.0010)</td>
</tr>
<tr>
<td>[140, ∞)</td>
<td>0.2621 (0.0005)</td>
<td>0.2572 (0.0005)</td>
<td>0.2599 (0.0007)</td>
<td>0.2408 (0.0005)</td>
</tr>
</tbody>
</table>

Turkish Test

<table>
<thead>
<tr>
<th></th>
<th>Female 1st time takers</th>
<th>Female 2nd time takers</th>
<th>Male 1st time takers</th>
<th>Male 2nd time takers</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 90)</td>
<td>0.2186 (0.0006)</td>
<td>0.2143 (0.0006)</td>
<td>0.2394 (0.0042)</td>
<td></td>
</tr>
<tr>
<td>[90, 100)</td>
<td>0.2300 (0.0008)</td>
<td>0.2321 (0.0007)</td>
<td>0.2259 (0.0008)</td>
<td>0.2339 (0.0009)</td>
</tr>
<tr>
<td>[100, 110)</td>
<td>0.2374 (0.0006)</td>
<td>0.2441 (0.0009)</td>
<td>0.2251 (0.0008)</td>
<td>0.2377 (0.0008)</td>
</tr>
<tr>
<td>[110, 120)</td>
<td>0.2509 (0.0006)</td>
<td>0.2525 (0.0008)</td>
<td>0.2494 (0.0007)</td>
<td>0.2464 (0.0009)</td>
</tr>
<tr>
<td>[120, 130)</td>
<td>0.2669 (0.0005)</td>
<td>0.2610 (0.0007)</td>
<td>0.2621 (0.0006)</td>
<td>0.2529 (0.0009)</td>
</tr>
<tr>
<td>[130, 140)</td>
<td>0.2728 (0.0006)</td>
<td>0.2678 (0.0007)</td>
<td>0.2680 (0.0007)</td>
<td>0.2554 (0.0012)</td>
</tr>
<tr>
<td>[140, ∞)</td>
<td>0.2842 (0.0007)</td>
<td>0.2695 (0.0007)</td>
<td>0.2624 (0.0013)</td>
<td>0.2494 (0.0007)</td>
</tr>
</tbody>
</table>

Standard errors are reported in parentheses.
<table>
<thead>
<tr>
<th>Table 5: Estimates of Ability Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social Science Test</strong></td>
</tr>
<tr>
<td><strong>Female</strong></td>
</tr>
<tr>
<td>1st time takers</td>
</tr>
<tr>
<td>$\mu$</td>
</tr>
<tr>
<td>( (0,90) )</td>
</tr>
<tr>
<td>( [90,100) )</td>
</tr>
<tr>
<td>( [100,110) )</td>
</tr>
<tr>
<td>( [110,120) )</td>
</tr>
<tr>
<td>( [120,130) )</td>
</tr>
<tr>
<td>( [130,140) )</td>
</tr>
<tr>
<td>( [140,\infty) )</td>
</tr>
<tr>
<td>( (0.0588)</td>
</tr>
<tr>
<td>( (0.0606)</td>
</tr>
</tbody>
</table>

| **Female**                               |
| 1st time takers | 2nd time takers | 1st time takers | 2nd time takers |
| $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
|\( (0,90) \) | -0.4183 | 0.7809 | -0.8048 | 0.9512 | (0.0766) | (0.1442) | (0.0953) | (0.1858) |
|\( [90,100) \) | 0.0586 | 0.7704 | 0.3617 | 0.7782 | -0.1709 | 0.8463 | -0.0545 | 0.9881 |
|\( [100,110) \) | 0.5666 | 0.6843 | 0.6261 | 0.8481 | 0.3783 | 0.7862 | 0.4239 | 1.0484 |
|\( [110,120) \) | 0.8939 | 0.7017 | 1.1424 | 0.8771 | 0.9043 | 0.7773 | 0.9204 | 1.0383 |
|\( [120,130) \) | 1.6618 | 0.6739 | 1.6625 | 0.7852 | 1.4202 | 0.7051 | 1.4947 | 1.0073 |
|\( [130,140) \) | 2.2422 | 0.6683 | 2.1831 | 0.7169 | 2.0508 | 0.8194 | 1.9670 | 0.9673 |
|\( [140,\infty) \) | 2.9406 | 0.8726 | 2.3210 | 0.6901 | 2.5363 | 0.9870 | 2.2627 | 0.7013 |
|\( (0.0900) | (0.4054) | (0.0705) | (0.2146) | (0.0981) | (0.3566) | (0.0717) | (0.2127) |

| **Turkish Test**                          |
| **Female**                               |
| 1st time takers | 2nd time takers | 1st time takers | 2nd time takers |
| $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
|\( (0,90) \) | -0.0512 | 0.6910 | -0.0795 | 0.7809 | -0.0795 | 0.7809 | (0.0588) | (0.2146) |
|\( [90,100) \) | 0.0057 | 0.7150 | 0.0057 | 0.7150 | 0.0057 | 0.7150 | (0.0588) | (0.2146) |
|\( [100,110) \) | 0.0057 | 0.7150 | 0.0057 | 0.7150 | 0.0057 | 0.7150 | (0.0588) | (0.2146) |
|\( [110,120) \) | 0.0057 | 0.7150 | 0.0057 | 0.7150 | 0.0057 | 0.7150 | (0.0588) | (0.2146) |
|\( [120,130) \) | 0.0057 | 0.7150 | 0.0057 | 0.7150 | 0.0057 | 0.7150 | (0.0588) | (0.2146) |
|\( [130,140) \) | 0.0057 | 0.7150 | 0.0057 | 0.7150 | 0.0057 | 0.7150 | (0.0588) | (0.2146) |
|\( [140,\infty) \) | 0.0057 | 0.7150 | 0.0057 | 0.7150 | 0.0057 | 0.7150 | (0.0588) | (0.2146) |

Standard errors are reported in parentheses.
Figure 26: Score distribution from data vs simulated distribution resulting from estimates
Figure 27: Score distribution from data vs simulated distribution resulting from estimates

Figure 28: Score distribution from data vs simulated distribution resulting from estimates
Figure 29: Score distribution from data vs simulated distribution resulting from estimates

Figure 30: Score distribution from data vs simulated distribution resulting from estimates
Figure 31: Score distribution from data vs simulated distribution resulting from estimates

Figure 32: Score distribution from data vs simulated distribution resulting from estimates
Figure 33: Score distribution from data vs simulated distribution resulting from estimates
Figure 34: Distributions of social science ability, as estimated by the model
Figure 35: Distributions of Turkish ability, as estimated by the model

Figure 36: Distributions of social science ability, as estimated by the model
Figure 37: Distributions of Turkish ability, as estimated by the model

Figure 38: Distributions of social science ability, as estimated by the model
Figure 39: Distributions of Turkish ability, as estimated by the model
Figure 40: The relationship between the proportion of males and ÖSS-SOZ score quantiles from the counterfactual exercises

Figure 41: The relationship between the proportion of first time takers and ÖSS-SOZ score quantiles from the counterfactual exercises
Figure 42: The relationship between social science ability and ÖSS-SOZ score quantiles from the counterfactual exercises

Figure 43: The relationship between Turkish ability and ÖSS-SOZ score quantiles from the counterfactual exercises
Table 6: First Time - Low Type Cutoffs

<table>
<thead>
<tr>
<th></th>
<th>&lt; 90</th>
<th>90-100</th>
<th>100-110</th>
<th>110-120</th>
<th>120-130</th>
<th>130-140</th>
<th>&gt; 140</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1SS</td>
<td>0.204</td>
<td>0.217</td>
<td>0.224</td>
<td>0.235</td>
<td>0.250</td>
<td>0.251</td>
<td>0.259</td>
</tr>
<tr>
<td>M1SS</td>
<td>0.207</td>
<td>0.216</td>
<td>0.228</td>
<td>0.238</td>
<td>0.247</td>
<td>0.253</td>
<td>0.256</td>
</tr>
<tr>
<td>F1T</td>
<td>0.213</td>
<td>0.223</td>
<td>0.232</td>
<td>0.244</td>
<td>0.255</td>
<td>0.278</td>
<td>0.268</td>
</tr>
<tr>
<td>M1T</td>
<td>0.211</td>
<td>0.221</td>
<td>0.228</td>
<td>0.248</td>
<td>0.256</td>
<td>0.258</td>
<td>0.259</td>
</tr>
</tbody>
</table>

Table 7: Second Time - Low Type Cutoffs

<table>
<thead>
<tr>
<th></th>
<th>&lt; 90</th>
<th>90-100</th>
<th>100-110</th>
<th>110-120</th>
<th>120-130</th>
<th>130-140</th>
<th>&gt; 140</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2SS</td>
<td>0.209</td>
<td>0.216</td>
<td>0.224</td>
<td>0.237</td>
<td>0.251</td>
<td>0.255</td>
<td>0.258</td>
</tr>
<tr>
<td>M2SS</td>
<td>0.208</td>
<td>0.216</td>
<td>0.224</td>
<td>0.236</td>
<td>0.242</td>
<td>0.244</td>
<td>0.245</td>
</tr>
<tr>
<td>F2T</td>
<td>0.218</td>
<td>0.226</td>
<td>0.235</td>
<td>0.244</td>
<td>0.262</td>
<td>0.265</td>
<td>0.264</td>
</tr>
<tr>
<td>M2T</td>
<td>0.213</td>
<td>0.224</td>
<td>0.231</td>
<td>0.238</td>
<td>0.248</td>
<td>0.254</td>
<td>0.258</td>
</tr>
</tbody>
</table>

Table 8: First Time - Share of Low Type Students

<table>
<thead>
<tr>
<th></th>
<th>&lt; 90</th>
<th>90-100</th>
<th>100-110</th>
<th>110-120</th>
<th>120-130</th>
<th>130-140</th>
<th>&gt; 140</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1SS</td>
<td>0.835</td>
<td>0.825</td>
<td>0.420</td>
<td>0.280</td>
<td>0.596</td>
<td>0.778</td>
<td>1.000</td>
</tr>
<tr>
<td>M1SS</td>
<td>0.854</td>
<td>0.710</td>
<td>0.579</td>
<td>0.667</td>
<td>0.596</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>F1T</td>
<td>0.844</td>
<td>0.687</td>
<td>0.500</td>
<td>0.553</td>
<td>0.473</td>
<td>0.918</td>
<td>1.000</td>
</tr>
<tr>
<td>M1T</td>
<td>0.905</td>
<td>0.581</td>
<td>0.432</td>
<td>0.523</td>
<td>0.566</td>
<td>0.737</td>
<td>0.996</td>
</tr>
</tbody>
</table>
Table 9: Second Time - Share of Low Type Students

<table>
<thead>
<tr>
<th></th>
<th>&lt; 90</th>
<th>90-100</th>
<th>100-110</th>
<th>110-120</th>
<th>120-130</th>
<th>130-140</th>
<th>&gt; 140</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2SS</td>
<td>0.852</td>
<td>0.625</td>
<td>0.319</td>
<td>0.367</td>
<td>0.553</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>M2SS</td>
<td>0.816</td>
<td>0.731</td>
<td>0.610</td>
<td>0.732</td>
<td>0.810</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>F2T</td>
<td>0.828</td>
<td>0.706</td>
<td>0.580</td>
<td>0.474</td>
<td>0.748</td>
<td>0.936</td>
<td>1.000</td>
</tr>
<tr>
<td>M2T</td>
<td>0.877</td>
<td>0.713</td>
<td>0.520</td>
<td>0.372</td>
<td>0.653</td>
<td>0.892</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Figure 44: Score distribution from data vs simulated distribution from two type extended model
Figure 45: Score distribution from data vs simulated distribution from two type extended model

Figure 46: Score distribution from data vs simulated distribution from two type extended model
Figure 47: Score distribution from data vs simulated distribution from two type extended model

Figure 48: Score distribution from data vs simulated distribution from two type extended model
Figure 49: Score distribution from data vs simulated distribution from two type extended model

Figure 50: Score distribution from data vs simulated distribution from two type extended model
Figure 51: Score distribution from data vs simulated distribution from two type extended model