Hit or Miss? Test Taking Behavior in Multiple Choice Exams*

Pelin Akyol † James Key‡ Kala Krishna§

April 15, 2019

Abstract

This paper models and estimates the decision to answer questions in multiple choice tests with negative marking. Negative marking reduces guessing and increases accuracy while reducing the expected score of the more risk averse. Using data from the Turkish University Entrance Exam, we find that students' attitudes towards risk differ according to their gender and ability with women (and those with high ability) being significantly more risk averse. However, the impact on scores of such differences is small so that differences in risk aversion are of little consequence.

JEL Classification: I21, J24, D61, C11

Keywords: Multiple Choice Exams, Negative Marking, Risk Aversion, Bayesian Updating.

*We would like to thank Paul Grieco, Sung Jae Jun, Stephen Yeaple and Mark Roberts for their helpful comments on an earlier draft. We would also like to thank seminar and conference participants at the WEAI 11th International Conference, 11th World Congress of the Econometric Society, 30th Congress of the European Economic Association, Conference on the Economics of Health, Education, and Worker Productivity, Massey University Albany, Otago University, Victoria University Wellington, Monash University, Hacettepe University, Sabanci University and Bilkent University.

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

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

§Penn State University, CES-IFO and NBER , 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. Such exams are widely used in practice being seen as objective, fair[1] and low cost, especially when large numbers of exam takers are involved (Frederiksen [1984] and Becker and Johnston [1999]). 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 the effectiveness of schools, teachers, to enter the civil service, and to allocate open positions.[2] Furthermore, scores in such exams are likely to be important determinants of future wages and occupations (Ebenstein et al. [2016]). Its main advantages are that it allows a broader evaluation of the candidate's knowledge in a short time, it is easy to grade which matters more when large numbers of test takers are involved, 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. A disadvantage of such exams is that candidates may attempt to guess the answer without having any knowledge of the answer[3](see Budescu and Bar-Hillel [1993] and Kubinger et al. [2010]). 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 (guessing penalty). 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,

[1]A fair exam is one where the only relevant student characteristic is the students' knowledge of the material.

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

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

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.[4] This problem 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 item response theory models (IRT), Rasch model (one parameter logistic model (1PLM)), 2PLM and 3PLM. IRT models would give biased results if there is a penalty for guessing. In this case, skipping decision would depend on risk aversion and the ability of the test taker which will violate unidimensionality[5] assumption of IRT (Ahmadi and Thompson [2012]).

In this paper, we argue that the standard approaches to examining the performance of people taking multiple choice exams are inadequate and often misleading as they do not take into account skipping behavior properly. Thus, improving on existing methods is vital for understanding what lies behind performance differences of different groups and for policymaking.[6]

We specify and estimate what we believe is the first 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 focus is on the trade-off between precision and fairness. Negative marking reduces guessing, thereby increasing accuracy. However, it reduces the expected score of the more risk averse, discriminating against them. 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.

[4] A possible change in the exam grading method is removing penalties for wrong answers. This leads all students to answer all questions which would increase the noise associated with the score. (see Bereby-Meyer et al. [2002] and Kubinger et al. [2010])

[5] Unidimensionality refers to the existence of a single trait or construct underlying a set of measures (Gerbing and Anderson [1988]).

[6] For example, women tend to perform worse in multiple choice exams featuring negative marking.

Our work contributes to the literature in two ways. First, we provide an innovative way to identify differences in risk preferences (with negative marking for incorrect answers) even without having question-by-question responses for students.[7] Such penalties for guessing makes risk averse examinees less willing to guess and more likely to skip questions with consequent adverse effects on the expected score. Our approach uses the insight that skipping behavior, or a lack of it, make certain scores a lot more likely. Second, we provide a richer model than the standard Rasch model used in the literature, see for example Pekkarinen [2014].[8] 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. The individual fixed effect is thought of as ability, and the question fixed effect as the difficulty of the question. 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 student's ability. To understand why this bias occurs, consider, for example, a setting where there is negative marking and all questions extremely difficult so that students have little idea about the right answer, and all students have the same ability, though some are risk averse (and so are more likely to skip a question when they are unsure about it) while others are not. Say the risk averse group answers 20 of 80 questions getting 10 right, while the risk neutral one answers 40 of 80 questions getting 15 right. In this case, the Rasch model would estimate the probability of answering correctly for the risk averse group as 1/8 and that for the risk neutral group answering 40 questions and getting 15 right as 3/16. However, the difference in the two would be due to differences in risk aversion rather than ability.

Such differences in risk aversion are likely to exist: students close to a cutoff for a highly desirable school may well be very risk averse, and it has been argued, see for example Eckel

[7]Of course, our approach can also be used with question by question responses. The extended model is presented in the Online Appendix.

[8]While the Rasch model is often used to account for differences in difficulty across questions in a test, we ignore this aspect in this paper due to data limitations, namely, we do not observe individual item responses. On the other hand, the Rasch model does not deal with skips.

and Grossman [2008b], Charness and Gneezy [2012] and Croson and Gneezy [2009], that women are more risk averse than men. To disentangle ability and risk aversion, and obtain unbiased estimates of both risk aversion and ability, we need to specify a complete setting, one 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 in the ÖSS, and the high school GPA[9], with at least 75% of the weight being given to the ÖSS score. Each question has five possible answers; for each correct answer the student obtains one point, and for each wrong answer she is penalized 0.25 points, while no points are awarded/deducted for skipping a question.[10] 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 the large literature, see Croson and Gneezy [2009], that suggests 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 the impact of these differences by gender and ability. 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

[9]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.

[10]Thus, the expected score from a random guess is over the five possible answers is zero.

scores, than otherwise similar men. This tends to make them under-represented at both the top and the bottom end of the score distribution. The consequences of under representation at the top are particularly relevant when university entrance is very selective. This is certainly the case in many developing countries where 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 60.1% of students in the top 5% are male.[11] We find, for example, that if there was no negative marking, so that guessing when in doubt was optimal, irrespective of risk aversion, women's representation in the top 5% of placements increases by 0.3 percentage points. They go from being 39.9% of the population in these placements to being 40.2% of them.[12] Thus, though the actual mean score change is small, its impact on the gender gap at top institutions is not trivial. The differences in risk preferences account for roughly 10% of the gender gap.[13]

We also try to quantify the importance of using our approach as opposed to the standard Rasch model. We use the estimated model (which assumes all questions are equally difficult as we do not have item response data) to generate individual level data on question-by-question performance. 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 mistakenly judged by the Rasch model to be of 5% higher ability than females.[14] We present an extended

[11] We call this over-representation of male students in the top 5% of scorers (of 4.2% in this case) the gender gap.

[12] For the top 20%, the gender gap is smaller, and the reduction in over-representation of males when penalties are eliminated is also smaller, 0.1% versus 0.3%.

[13] The bottom is pushed up while the top is pulled down by greater risk aversion on the part of women but as those at the bottom do not go to college, only the top part matters for the gender gap.

[14] We set the ability of all students as the median in estimation results, and risk aversion as in the 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 featuring substantially harsh penalties (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, where the penalties are not as harsh, males of median ability correctly answer 1.83% more questions than females of equivalent ability. This is a substantial difference.

model in Online Appendix B.6.1 which can be used to provide estimates taking advantage of item-by-item responses.

1.1 Related Literature

The psychology and education literature has long been interested in developing test designs that generate fair results. 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.[15] 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.[16] 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

[15] Prospect theory describes the way people choose between alternatives when the probabilities associated with them are known taking a behavioral approach such as loss aversion.

[16] Proper scoring rules have been developed to reward partial knowledge where students report the subjective probability of each choice being correct rather than choose one answer so that more information is revealed. There are different types of proper scoring rules, quadratic, spherical, and logarithmic. (Bickel [2010]). The comparisons and the details of these methods are beyond the scope of this paper. In practice, the application of these methods is problematic, especially in large scale exams. Its complexity means that its rules may not be internalized by all students which could create another source of inequality.

high ability students. None of these attempt to estimate ability and risk aversion of agents or to test the implications of their models empirically as we do. We use a simple Bayesian setting where students of better ability get a more informative signal about the correct answer. This feature, together with risk aversion, allows us to use the skipping behavior of agents, as well as their accuracy, to estimate ability and risk aversion.

On the empirical side, there are two broad approaches: the first is experimental, and the second empirical. See Eckel and Grossman [2008a] for a survey of some of the work on the experimental side, especially work focusing 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 in an experimental setting. Lab experiments are cleverly 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 a 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 factors (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 a field experiment, Funk et al. [2016] find that although females are more risk averse relative to males, the differences in risk aversion does not have any effect on the differences in exam scores. Riener and Wagner [2017] investigates the gender differences in skipping tendencies among secondary students in Germany. They incentivize

a randomly selected group of students by offering non-monetary rewards which increased the stakes associated with the exam. They find that gender gap exists only for difficult questions in the non-incentivized group. This finding is consistent with the notion that the effects of risk aversion are most prominent in situations where students are not confident in their answer.

The other line of work relies on non-experimental data. However, even when an explicit model being used, the approach tends to be reduced form as the focus is on whether the patterns predicted by the model are observed in the data rather than on the estimation of (the structural parameters of) the model itself.

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 in many settings (see Eckel and Grossman [2008a], Agnew et al. [2008] and Charness and Gneezy [2012]). To test the hypothesis that female students skip more questions 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.

Espinosa and Gardeazabal [2010], show that differences in risk aversion have small effects on performance, especially for higher ability students. This is exactly what simple economics would predict: higher ability students are more able to tell the right from the wrong answers and as a result are less affected by differences in risk aversion.

We do not have question-by-question responses. As a result, we cannot directly look at the probability of skipping and getting a correct answer. 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.

In the next section, we present an overview of the data and testing environment. The particular patterns seen in the multiple choice tests are discussed in more detail in Section 3. 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 Section 8 concludes. Additional tables and figures are presented in Appendix A. Online Appendix B.2 extends the estimation to different student types.[17]

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.[18] 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 normalized high school grade point average. Students know their high school GPA by the time of the university entrance exam. However, to remove the effects of grade inflation across schools the GPA is normalized with the minimum and maximum GPA, and the mean exam performance of students in their school, which students would not know prior to the exam. Similarly, students' scores in the test are normalized according to the performance of all the students. Students submit their preference list after they observe their scores. Students are centrally allocated according to the preference list they submit and priority is given by score. A student is thus allocated to his highest ranked choice still available when his turn comes. Thus, utility depends on score.

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

[17] While the main body of the paper focuses on students in the Social Science track, this section examines those in the Language and Turkish-Math tracks.

[18] It is important to note that to retake the exam students need to wait one more year. So there is a significant opportunity cost associated with retaking.

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 time spent and expenditure on preparation. We have around 40,000 students from each track (Social Science, Turkish-Math, and Science).

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.[19] Students are given 180 minutes for 180 questions and can choose how to allocate their time between different test sections; all four sections are taken at the same 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 B.1[20] shows the test weights to obtain the track score, by track. For the Social Science track students, the placement score (Y-ÖSS) is a composite of the track score (ÖSS-SÖZ) and the standardized GPA[21] (AOBP). The track score in Social Science gives the highest weight (1.8) to the Turkish and Social Science sections of the tests, while Math and

[19]There is also a separate multiple choice language exam which is held one week after the main exam. This exam is taken by students who aim to get into college programs such as English literature.

[20]The weights presented in this table are decided by ÖSYM.

[21]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.

Science have a relatively low weight (0.4).[22]

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 score (Y-ÖSS) is calculated as follows:

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

where X$\in$\{SAY,SÖZ,EA,DIL\}, δ is a weight on the standardized 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.[23] There are penalties if a student switches tracks, or refuses a placement.

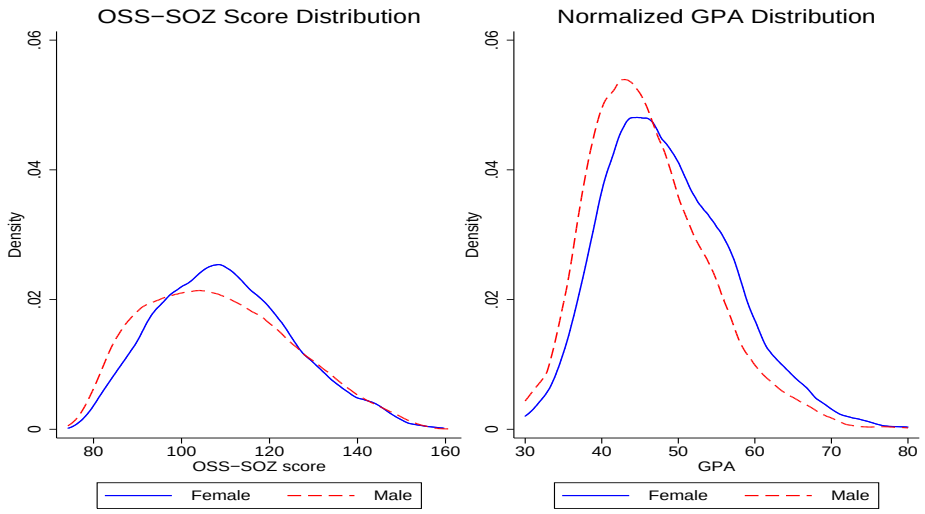
Here, we focus on first time taker Social Science track students. The retaking decision is endogeneous, therefore including repeat takers into our analysis will not be informative without controlling for the selection problem. As Turkish and social science tests have the highest weight on the social science track students' score. For questions in these subjects, students are likely to have partial knowledge, choosing the best answer. This is exactly what we model and use to estimate risk aversion. In contrast, many math and science questions involve solving the problem so that students are either quite certain of the answer or have failed to solve the question (thereby having no information regarding which answer is correct). In the former case, everyone answers, and in the latter, everyone skips when even slightly risk averse, which makes identification of risk aversion impossible. Table A.1 presents the summary statistics by gender for Social Science track students.

[22]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 B.1. According to the data, the equation is ÖSS-SÖZ=0.8755*rawTurkish+0.8755*rawSS+0.187*rawMath+0.l187*rawScience+78.89

[23]The δ 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, δ will be 0.5, while if it is outside the open list, δ 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, δ will be 0.65. If the student was placed in a regular university program in previous year, the student is punished and δ will be equal to either 0.25, 0.1, or 0.375. For those students, the δ coefficient is equal to half of the regular δ.

Examining Social Science track students, the distributions of track scores (ÖSS-SÖZ) as well as normalized GPAs for first time takers are depicted in Figure 1. First, note that women seem to dominate men in terms of GPAs. However, this does not carry over to exams. Men seem to do a bit better than women in the exam at the top of the track score distribution, but considerably worse at the low end. One explanation for this pattern could be greater risk aversion on the part of women which makes them skip questions with a positive expected value.

Figure 1: Score Distributions



As a robustness check, we also estimate the model with data from the Language *track* and from the Turkish-Math *track*. We only use the Language *test* in the former and the Turkish *test* in the latter. This is presented in Appendix B.2.

3 Multiple Choice Exam Scores

We begin by taking a first look at students' scores in the Turkish, Social Science, Math and Science tests. 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.[24] 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 makes sense as these students are poorly prepared in Math and Science as they have not done much of it since the ninth grade and the questions are very challenging. Also, Math and Science questions involve solving the problem and are not amenable to guessing. Finally, Math and Science test scores have relatively little weight (0.4 each) in the track score for students in the Social Science track. Turkish and Social Science scores, in contrast, have a weight of 1.8. Students are explicitly advised to spend less time on the Math and Science test.[25]

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, similarly a score of 42.5 could only have arisen through attempting all questions, getting 43 questions correct and 2 incorrect. On the other hand, 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 220 possible raw scores one can reach, however if a student attempts all the questions, not skipping any, there are only 46 raw scores that can occur. 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 the first time takers, as seen in Figures 2 and 3, have very prominent spikes.[26] It is no coincidence that the spikes appear evenly placed; they correspond to the 46 scores that occur after attempting all questions and come

[24]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.

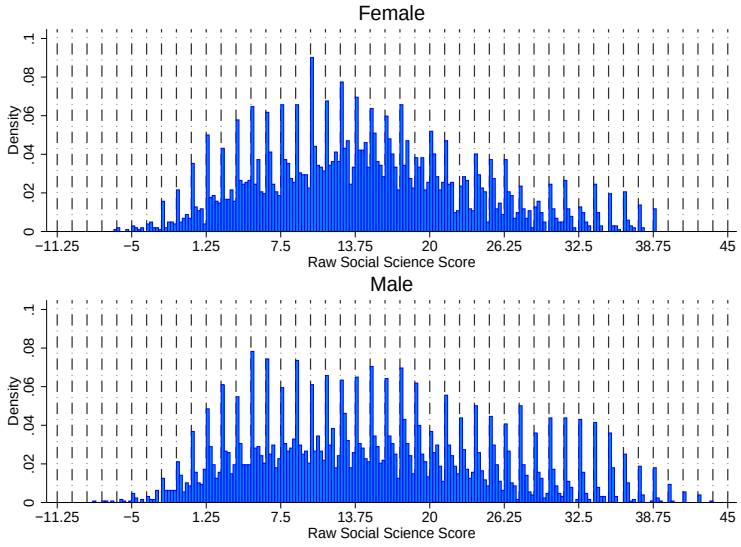
[25]In the exam booklet there is a note before the Social Science/Turkish part of the exam that says: "If you want a higher score in ÖSS-SÖZ, it may be better for you to spend more than 90 minutes on verbal part of the exam."

[26]Gridlines spaced 1.25 marks apart correspond with these spikes.

from the fact that there is a mass of students, of differing abilities, who answer all the questions. This is an important part of our identification strategy as explained below as we do not have question-by-question data for students.

Math and Science test score distributions for Social Science track students do not exhibit this pattern 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.[27] As Social Science track students do not spend much time on the Science and Math parts of the exam, we assume away the time constraint and restrict our attention to only the Social Science and Turkish sections of the exam for Social Science track students.

Figure 2: Distribution of Social Science Test Scores

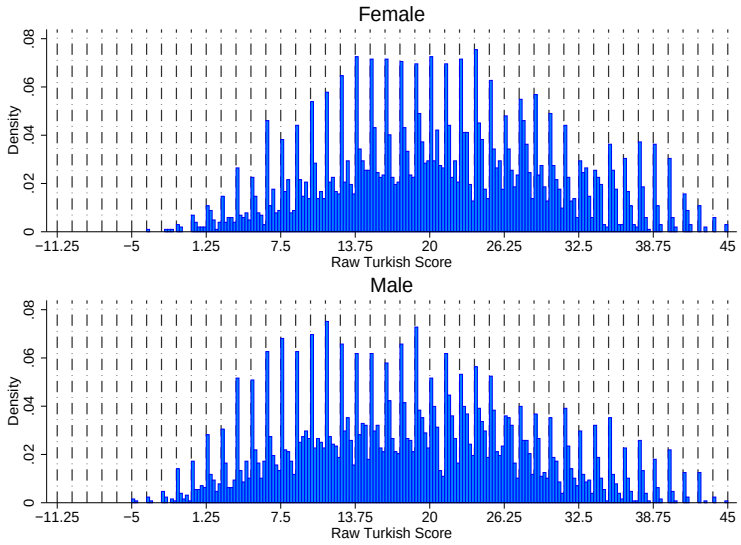


4 Model

Given the complex relationship between scores, admission outcomes and expected utilities of those outcomes, we do not seek to obtain an explicit utility function (as a function of exam

[27] These figures are available on request.

Figure 3: Distribution of Turkish Test Scores



score) in this paper. Instead, for the sake of tractability we assume that student i acts as if they have a CARA utility function over exam score y, $U_i(y) = 1 - e^{-\tau_i y}$, and they answer each test and each question in isolation.[28] Having utility increase with the score makes sense as a higher score increases the number of programs the student is eligible for, and so gives more options to a student. We do not allow for outcomes in one section of the test to have any bearing on other sections. Expressed alternatively, we do not allow a students perceived performance in previous questions to impact behavior in subsequent questions.[29] Nor do we allow for any time pressure that results in skipping: students do not skip questions in order to improve their performance in other questions.[30]

Students decide whether to answer a question or skip it, to maximize their expected utility, depending on their probability of answering the question correctly, P_C. Formally, we

[28]We assume that questions are equally difficult. However, if one has item response data, this assumption can be relaxed (see Appendix B.6).

[29]Even if students are able to perceive that they are answering better than expected ex ante, they are not able to discern if this is due to luck in being asked questions which happen to be well suited to their individual strengths, of if the exam is simply easier than average (implying that scores will be normalized downwards).

[30]In examining students in the Social Science track we believe this is appropriate, as these students overwhelmingly skip the Science and Math sections of the test, as recommended by examiners, allowing them ample time to focus on the Social Science and Turkish questions.

can write the problem of the student as follows:

$$\max \; I(answer)\left[P_C U(1) + (1 - P_C)U(-k)\right] + I(skip)U(0)$$

where k is the penalty applied for the wrong answer, and $k \geq 0$.[31]

So, the student will answer the question, if

$$P_C U(1) + (1 - P_C)U(-k) > U(0)$$

$$P_C > \frac{U(0) - U(-k)}{U(1) - U(-k)} = c$$

where c will be called attempt cutoff. If the student's probability of answering the question correctly is above this cutoff, he will answer the question, otherwise he will choose to skip it. c rises with the degree of risk aversion as shown below.

In this section, we will construct a model that allows us to structurally estimate these attempt cutoffs as well as the ability distributions of the students (since ability affects the probability of answering the question correctly). We model 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 an 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.

Signals for each of the five answers depend on whether or not the answer is actually correct. Signals for incorrect answers are drawn from a distribution G, where G is Pareto with support $[A_I, \infty)$ and shape parameter $\beta > 0$. Thus, the density of the signal x for an incorrect answer is $\frac{\beta A_I}{x^{\beta+1}}$. The mean signal is $\frac{\beta A_I}{\beta - 1}$ which is decreasing in β. Signals for correct answers are drawn 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 $\frac{\alpha A_I}{\alpha - 1}$

[31] Recall that test takers possess CARA utility functions.

which is decreasing in α.

Assumption 1 $A_I = A_C = A$.

This assumption rules out complete certainty that an answer is correct or incorrect.[32]

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:

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

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, this is 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:

[32]Assuming 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. Assuming $A_C < A_I$ would make no sense if $\beta \geq \alpha$ (implying that some answers are so bad they must be true!).

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

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:

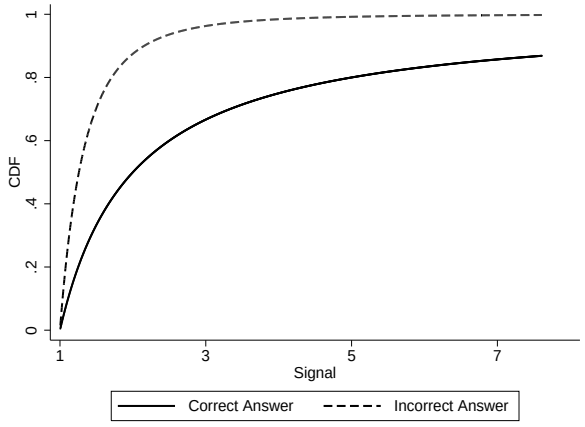
$$\text{Prob}(i \text{ is correct}|X) = \frac{x_i^{\gamma}}{\sum_{m=1}^{5} x_m^{\gamma}} \tag{6}$$

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, x_i, the greater the likelihood that answer i is correct.[33] A higher shape parameter for a Pareto distribution shifts probability mass to the left so that the signals from incorrect answers 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 incorrect answers that captures ability. Someone who thinks all answers are great is as bad as someone who thinks none of the answers are great: it is the extent to which one can 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. Given these assumptions, we can rescale so that the *correct answer* is drawn from a distribution

[33] 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.

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 $\beta = 3$ (approximately median) are shown in Figure 4.

Figure 4: Distributions of signals for a student with $\beta = 3$



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. Second, 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.[34] This is formally stated below:

Proposition 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:

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

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

[34] By signals we refer to the observed vector of x values. To see why ability matters, consider a vector of signals (3,1.2,1.1,1.3,1.2). A high ability student would interpret this as being favourable towards the first answer. A student with no ability, i.e. $\beta = 1$, obtains no information from the signals and can only conclude that all answers have an equal likelihood of being correct.

has a probability of being correct (conditional on observed signals and the student's ability) greater than or equal to twenty percent.

As explained students 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].[35] As shown below, a higher value for this cutoff implies a higher degree of risk aversion, while a cutoff of 0.2 would be supported by risk neutral preferences.

Proposition 2 *There is a monotonically increasing relationship between the risk aversion parameter, τ, and the attempt cutoff, c.*

Proof. Proof is presented in Appendix B. ■

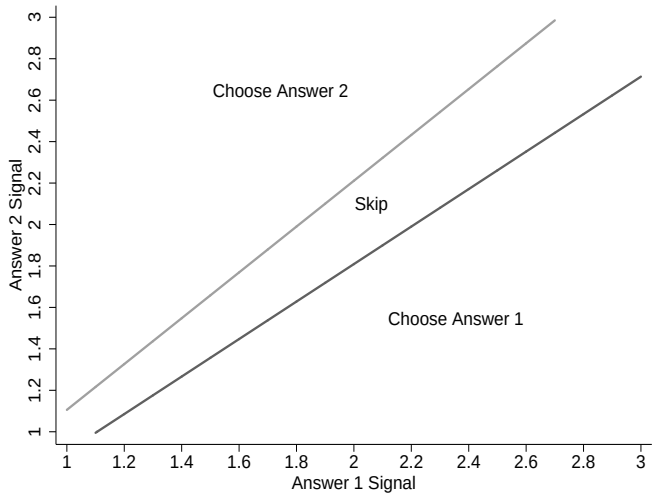
Consider a student with ability parameter β (recall that $\alpha = 1$) and attempt cutoff $c \in (0.2, 1)$. In order to answer a question, with answer i, the signal drawn for answer i, x_i, must satisfy two conditions. First, it must be the highest signal. Second, it must be high enough (given the other signals, and ability β) that the belief that it is correct is greater than c, the cutoff required to attempt the question. A diagram showing choices conditional on signal observations for a simplified two answer setup (with $\beta = 3$ and $c = .55$) is shown in Figure 5. If the signal for j is sufficiently high, then j is selected. In between the two lines, where signals are very close to each other, the best option is to skip the question. This skip region is larger the greater the risk aversion of the agent (the greater the value of c).

5 Estimation Strategy

In our model, students' scores depend on students' ability (β) and attempt cutoff, c, which captures attitudes towards risk. In our data set we observe only the student's score in each part of the exam, and not the outcome question by question. In this section we explain

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

Figure 5: Actions for a Question with Two Possible Answers



how we can use our model to estimate the distribution of ability and attempt cutoffs, c, which captures the extent of risk aversion.

Estimation of the parameters of interest, the distribution of student ability $\beta = (\beta_T, \beta_{SS})$ and attempt 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 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 attempt cutoffs to vary by gender, and allow them to depend on the interval in which the student's predicted Social Science track score (ÖSS-SÖZ) lies, for example 120-130. This predicted score in effect proxies for ability. We explain how we predict the Social Science track score (ÖSS-SÖZ) in Appendix B.

5.1 Estimation

We divide students into groups, according to gender, and the range into which their predicted track score (ÖSS-SÖZ) lies: $(0, 90)$, $[90, 100)$, $[100, 110)$, $[110, 120)$, $[120, 130)$, $[130, 140)$, and $[140, \infty)$.[36] These groups do not contain equal numbers of students, but do contain at least 100 students.[37] For each group, we examine the two subjects (Social Science and Turkish) jointly as we allow correlation in the ability of a student in the two. We assume that students in each score group have a common attempt cutoff, c, and draw from the joint distribution of ability $(\beta_{Turkish}, \beta_{SocialScience})$. The ability of each student in subject $k \in (T, SS)$ is given by $1 + e^{\psi_k}$, where (ψ_T, ψ_{SS}) are distributed normally with mean $\mu = (\mu_T, \mu_{SS})$ and covariance matrix Σ.[38] 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).[39] It also allows for abilities in the two subjects to be correlated, as would typically be the case.[40]

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 ψ_n. From ψ_n, we find $(\beta_T, \beta_{SS}) = (1 + e^{\psi_s(1)}, 1 + e^{\psi_s(2)})$, the student's ability vector. As we now have (β_T, β_{SS}, c) for student n, we can generate the simulated student's test outcome, namely the Turkish score and Social Science score.

In order to estimate the relevant parameters for the group (cutoff, means of ψ_T, ψ_{SS}, variances of ψ_T, ψ_{SS} and correlation between ψ_T and ψ_{SS}), we use simulated method of moments. For every group we postulate a cutoff, the mean of ψ_T, ψ_{SS}, the variance of ψ_T, ψ_{SS} and correlation between ψ_T and ψ_{SS}. We make 100 draws for each student in the

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

[37] With the exception of females in the lowest expected score range.

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

[39] It can be shown that the likelihood of answering correctly increases approximately linearly with respect to the log of ability, so that a log-normally distributed ability would generate the roughly normal score distribution observed.

[40] Within the Social Science track as a whole, the scores in the Turkish and Social Science sections are highly correlated, the correlation coefficient is 0.78 with the p-value 0.

group and construct the relevant moments for the group. These moments are the mean scores in the two subjects, the variance of these scores, the correlation between the scores in the two subjects, and the intensity of the spikes in the two subjects. The difference between the mass of students with scores corresponding to attempting all questions (i.e. $45, 43.75,..,-11.25$) and the mass of students with scores corresponding to skipping a single question (i.e. $44, 42.75,...,-11$) is what we mean by the intensity of the spikes.[41] If the spikes are very prominent, this difference will be large; if they are non-existent, this difference will be minimal. In a given exam, for each such pair, we find this difference and take its sum to get the overall measure of the spike intensity. This gives us two more moments to match.[42]

We compare simulated test score moments to those observed in the data and choose the parameters that minimize the objective function. Accordingly, the estimates of the vector $\theta^g = (c^g, \mu^g, \Sigma^g)$, cutoff c and ability distribution parameters for each group g, denoted by the vector $\hat{\theta}^g = \left(\hat{c}^g, \hat{\mu}^g, \hat{\Sigma}^g\right)$, are estimated by minimizing the distance between the simulated moments, $\hat{m}(g)$, and the observed moments, $m(g)$.

$$\hat{\theta}^g = \arg\min_{\theta}(\hat{m}(g) - m(g))' W_T^{-1} (\hat{m}(g) - m(g)) \tag{9}$$

where W_T is the weighting matrix. As usual, we begin by using the identity matrix as the weighting matrix thereby obtaining 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 which improves the efficiency of the estimate.

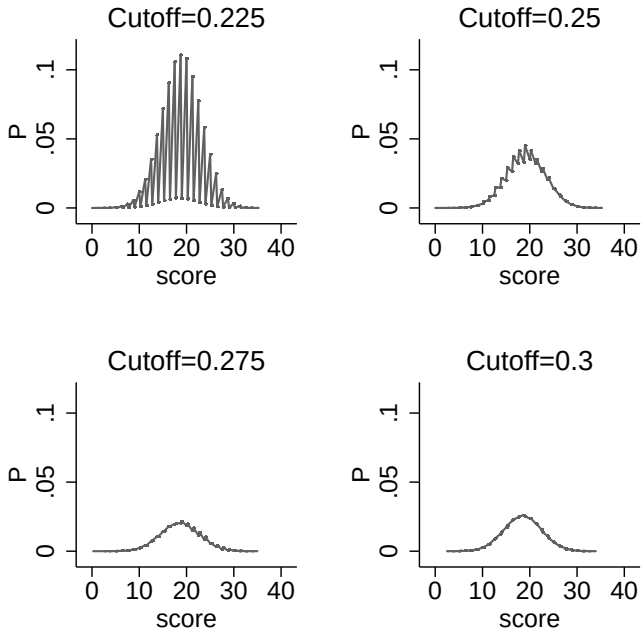
For a given c, the means of scores help pin down the means of the ability distributions

[41] There are many ways to get a particular score. For example, 35 can be reached by correctly answering 35 and skipping 10, correctly answering 36 and skipping 9 (4 incorrect) or correctly answering 37 with 8 incorrect. This multiplicity is not generating the spikes. As seen in Figure 7 below, the prevalence of spikes is clearly driven by risk aversion.

[42] There are alternative ways to measure the intensity of the spikes. It is also possible to define the spike intensity for each section - Turkish and Social Science- separately. We do not need to do so as we have a single cutoff that defines risk aversion so that a single measure suffices for identification.

that students are drawing from, and the variances/covariances of scores help pin down the variances of the ability distributions and the correlation between ability draws. The choice of c is pinned down by the extent of spikes in the score distributions for Turkish and Social Science as explained below. Identification of the attempt cutoff, c, is achieved through matching the intensity of the spikes. If students are less risk averse then they will tend to not skip, ceteris paribus. Thus, at low values of c, almost all of the probability mass of a given student's distribution will be located on scores corresponding to attempting all questions and resulting in spikes. As c increases, students become more and more likely to skip *some* questions, resulting in more mass lying on scores unreachable by attempting all questions so that spikes can no longer be seen.

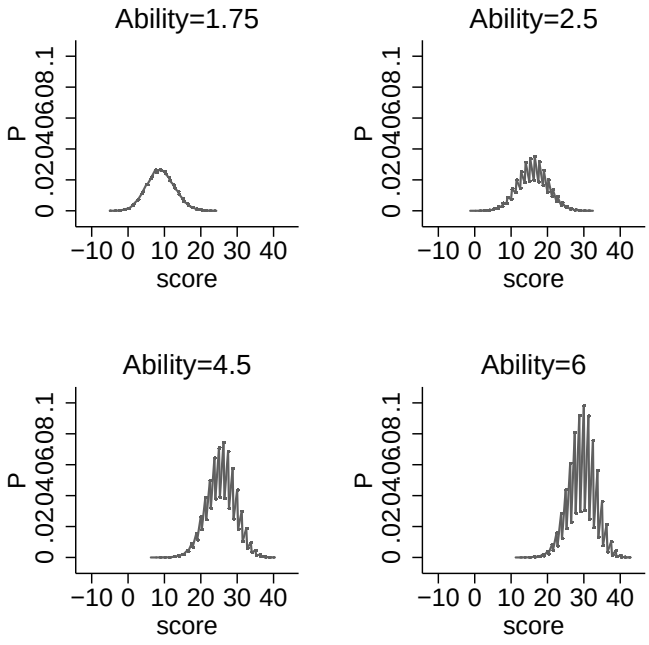
Figure 6: Distribution of scores resulting from various cutoff levels



This is illustrated in Figure 6[43], 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

[43]Only the set of scores corresponding to attempting all questions and the set of scores corresponding to skipping a single question are shown as identification is based on the relative probabilities of the two sets of scores.

Figure 7: Distribution of scores resulting from different ability levels



$c = 0.225$ puts virtually all of the mass of the score distribution on values that correspond to students attempting all questions. As the attempt cutoff increases to 0.3, the spikes all but disappear as very few attempt all questions. While there are multiple ways through which a particular score, say 35, may be obtained, ultimately this does not impact our identification strategy as long as these spikes are evident in the data.

The relationship between the intensity of the spikes and the attempt 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 are not likely to have an answer with a belief above the attempt cutoff, this becomes increasingly common as ability rises. This is shown in Figure 7[44], where the attempt cutoff is set to 0.25 [45].

The parameters of the distribution of the ability of a group of students, (μ_T, μ_{SS}) and

[44]Only the set of scores corresponding to attempting all questions and the set of scores corresponding to skipping a single question are shown as identification is based on the relative probabilities of the two sets of scores.

[45]Ability ranges from approximately the 20^{th} to the 80^{th} percentiles, as estimated

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 reflected in the correlation of scores.

6 Results

Figure 8 displays estimates of the cutoff (the belief below which a student skips) graphically.[46] Male and female attempt cutoffs are significantly different in all bins except the lowest and highest.[47] Two facts are apparent. Males tend to have lower attempt 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 as discussed previously. 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[48] and seems to rise with the predicted score bin consistent with increasing risk aversion.

Figures 9 and 10 show the simulated score distributions compared to observed distributions for the various groups. The figure clearly consists of spikes for attempting all questions. To help with visualization, these are presented in different grid lines corresponding to scores which could be obtained by skipping no questions. While the estimation procedure was designed only to match subgroups of the sample, the entire simulated distribution fits the data

[46] Full results, including standard errors, are shown in Table A.2.

[47] Test statistics are shown in Table A.3.

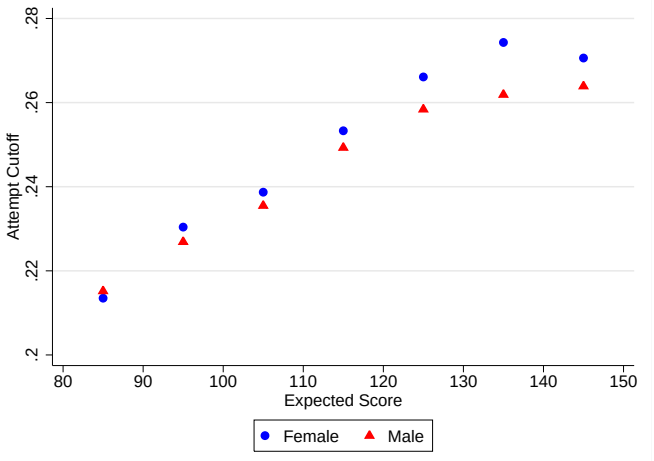
[48] Cutoffs for the top students are approximately 0.26, which has meaning that these students will only answer a question if they are at least 26% sure of their answer. Significantly more than the 20% likelihood of a random guess.

relatively well overall. It is worth noting that estimation methods which grouped students based on *actual* track (ÖSS-SÖZ) score did better here.

Estimates of the parameters governing the distribution of ability for each group are presented in Table A.4. Recall that ability is parametrized as $(1 + e^{\psi}, 1 + e^{\psi})$, where $\psi \sim N(\mu, \Sigma)$. The means and variances of the components of ψ in each group are presented.

As we estimate the distributions for students in the Social Science track, differences in ability distributions could come from selection into this track as well as differences given selection**.** For example, if the track was seen as female friendly in some way, it might attract more women, especially weak ones, with better women going into less female friendly tracks. With this qualification, we see that females tend to have higher ability in Turkish, but slightly lower ability in Social Science, when compared to males in the corresponding group.[49] This is consistent with males having a comparative and absolute advantage in Social Science and is consistent with findings in the literature that women have the advantage in Language skills (See Lynn [1992]).

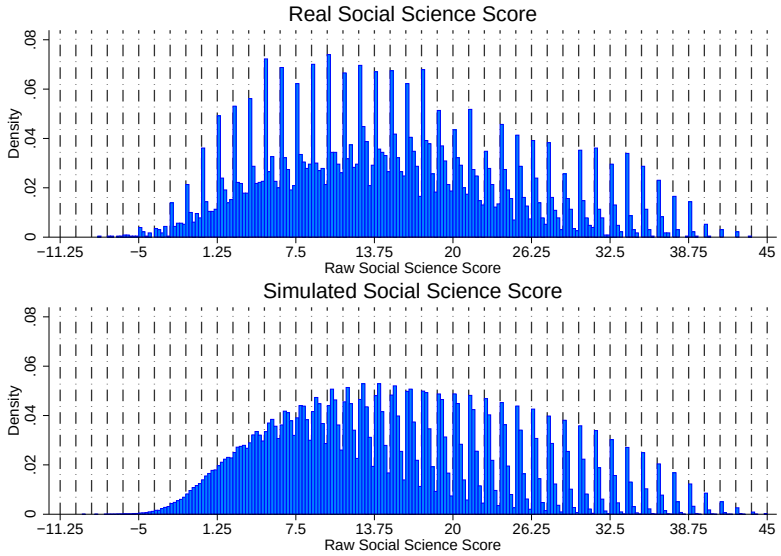
Figure 8: Estimates of Attempt Cutoffs: Social Science Track



In addition, we observe that males tend to have higher variance in their distribution of

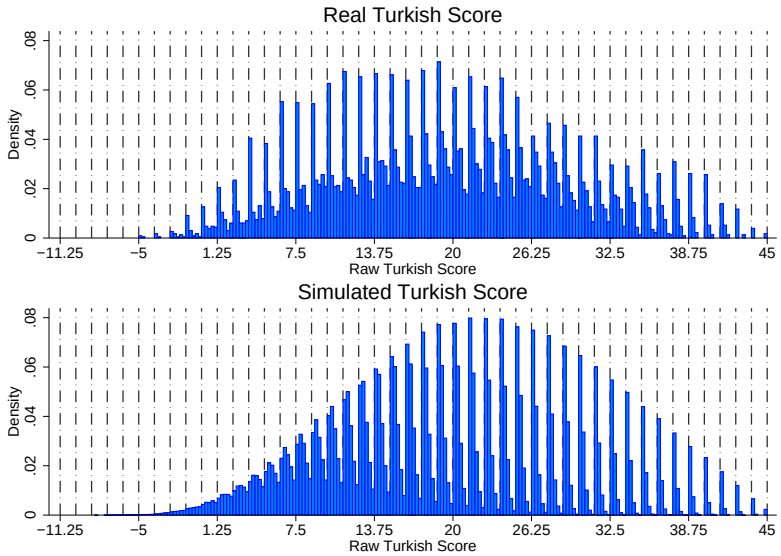
[49]The estimated ability distributions for the two sections of the test are depicted in Figure A.1-A.2b in Appendix A.

Figure 9: Data vs Simulated Score Distribution: Social Science



ability. In fact, the variance is greater for all groups.[50] The correlation between ability in Turkish and Social Science seems to be higher for each decile for females, as seen in Table A.5. 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

Figure 10: Data vs Simulated Score Distribution: Turkish



[50]The estimated ability distributions for Turkish and Social Sciences by gender reflect this higher variance as in Figures A.1 and A.2a.

also explain the higher variance for males.

We assume that all questions are at the same level of difficulty in both the model and estimation procedure. This assumption is necessary as we do not observe item responses. Would this create biases in our estimation? We check for this using simulations.

In Appendix B.6, 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 Social Science track students for the university entrance exam. 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 questions of a uniform level of 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.[51] On the one hand, having more homogeneous questions removes one source of variability. On the other, the level of difficulty matters: if the exam is difficult, then the median student facing the average difficulty question will guess more often than when the questions are of variable difficulty. For this reason, we think of our estimates on variance as being an upper bound and this explains why some of them are located at the boundary.

We also run the model for the Language test in the Language track and the Turkish exam in the Turkish-Math track as a robustness check. These are presented in Appendix B.2. It is reassuring to note that the estimates look a lot like those above, despite the raw data looking quite different.

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

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 the regime may well affect outcomes. Our focus is on two points. First we look at the gender gap, defined as the over-representation of males at the top of the score distribution. This comes both from ability differences and from differences in behavior. In particular we will quantify the impact of risk aversion differences on test outcomes as the test format changes. Second, we look at the effectiveness of a test as the format varies. Effectiveness is defined as the extent to which performance matches abilities. The rationale behind penalties is to reduce the amount of random guessing, therefore reducing score variance and improving effectiveness. The downside is that as women seem to be significantly more risk averse than men, this accentuates the gender gap. Our focus is to understand the extent of this trade-off.

For this reason we consider seven possible regimes in our counterfactual experiments. These 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.[52]

[52] In this case, scores would 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 changed to that used by a male student in predicted ÖSS-SÖZ score interval k (labeled as "Equal Cutoffs" in the figures). Note that the second regime eliminates risk aversion differences across gender, and makes all agents risk neutral. The third regime keeps risk aversion, but eliminates the gender differences in risk aversion. While this is not feasible to perform in practice, we can use the counterfactual exercise to quantify the effect of gender differences in risk aversion in the current system.

4. Each question has only four answers to choose from, with the penalty for an incorrect answer adjusted accordingly. This will increase the impact of risk aversion and accentuate the gender gap and hinder the effectiveness of the exam. 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 group. We then convert the risk aversion parameter to a cutoff in the four answer case.[53] Note that having four answers instead of five, and increasing the penalty accordingly, can increase variances of scores for a given student even in the absence of risk aversion. [54]

[53]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.

[54]The standard deviation of the points earned for a single question is, for a student of (approximately median) ability $\beta = 3$, 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.

5. The penalty for answering incorrectly is increased from 0.25 points to 1 point. This will accentuate the gender gap but increase effectiveness as guessing is reduced. This counterfactual is 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.

6. The number of questions in each section is doubled, from 45 to 90. This will improve the effectiveness of the exam and can increase the gender gap if males are more prevalent at higher abilities. 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 on ability.

For each of the six possible regimes, we find the resulting distributions of scores for the entire sample of (first time exam takers) students in the Social Science track.

We simulate the model using the parameters estimated,[55] generating scores in the Turkish and Social Science section, adding the two to generate an exam score for each student.[56] 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.[57]

We first examine the effect of the different regimes on the ability of the exam to select the most capable students. To do so we look the relationship between exam score percentile and (average log) ability in the two subjects, Turkish and Social Science. Figures 11 and 12 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

[55] 1000 students were simulated for each student observed in the data.

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

[57] As seen in Figures A.3a and A.3b in the Appendix, the exams do sort by ability as higher score percentiles are associated with higher average ability in all the regimes studied.

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.

Figure 11: Turkish Ability (Δ vs baseline)

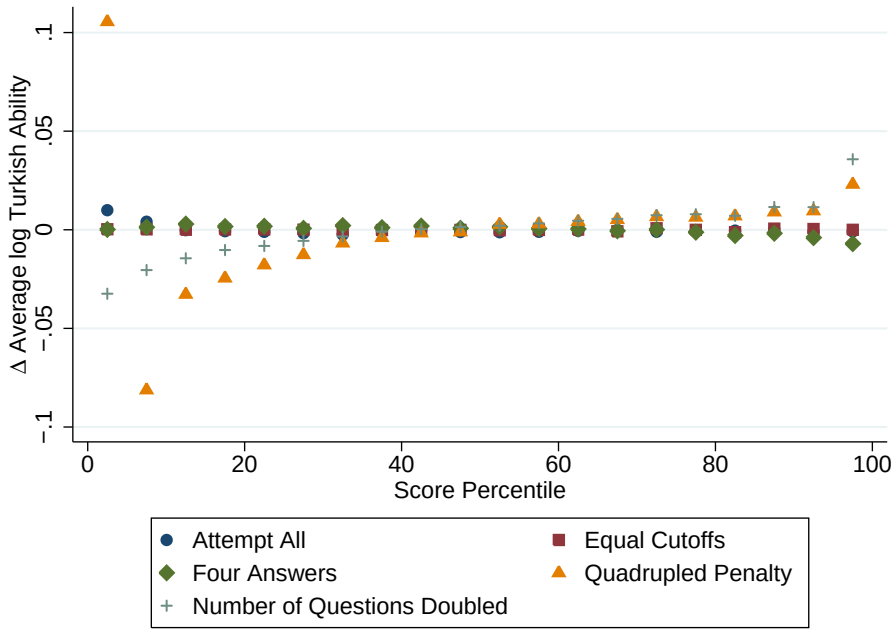
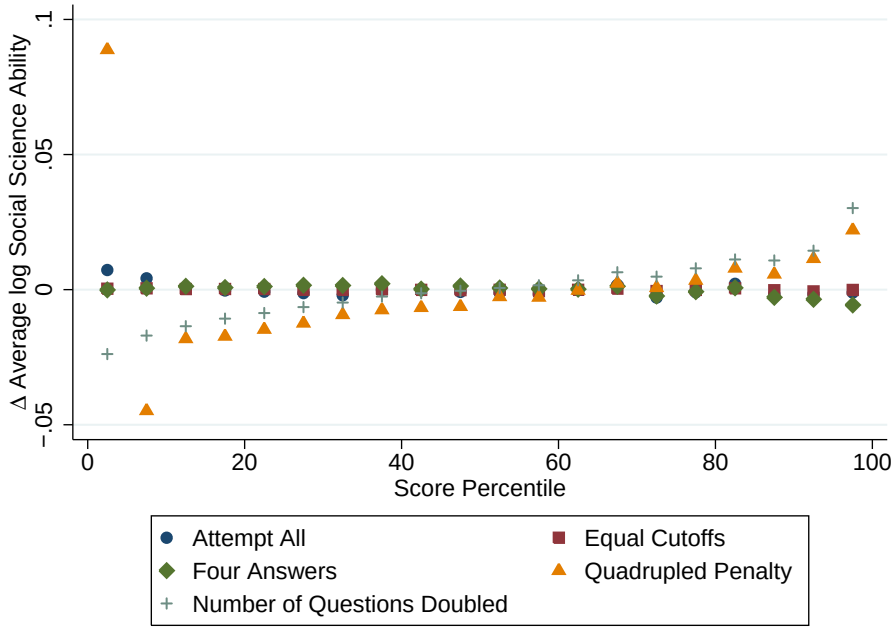


Figure 12: Social Science Ability (Δ vs baseline)



As the Turkish and Social Science abilities show very similar patterns, they will be discussed jointly. We see that the "Attempt All", "equal cutoffs", and "four answers" regimes show very little difference in terms of the quality of admitted students in Figures 11 and 12 consistent with the small differences in risk aversion estimated. Although the differences between the higher penalty regime and the baseline are small, higher penalties clearly do a better job at sorting. Average abilities under these regimes are lower than the baseline on the left (more accurately identifying weak students)[58] and higher than the baseline on the right (more accurately identifying strong students). The pattern of differences between the baseline and the "number of questions doubled" regime is similar, suggesting this also improves sorting.

The reason for the higher effectiveness of the high penalty regime is simple. It strongly discourages attempting when the student is relatively uncertain of their answer. This results in much less guessing which reduces the variance in the score distributions of an individual student, resulting in a cleaner signal. The downside is that their partial knowledge will not be conveyed as students will skip even when they have some knowledge of the right answer and differences in risk aversion by gender will have more of an impact.

Both greater penalties for wrong answers and more questions improve the ability of the exam to sort students. How do they compare to each other? The impact of the increased penalties on average abilities of combined score quantiles is most evident for the top quantiles. Note that the top 13.5% roughly get admission in the Social Science track. 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, to the 45 question, quadrupled penalty version.[59]

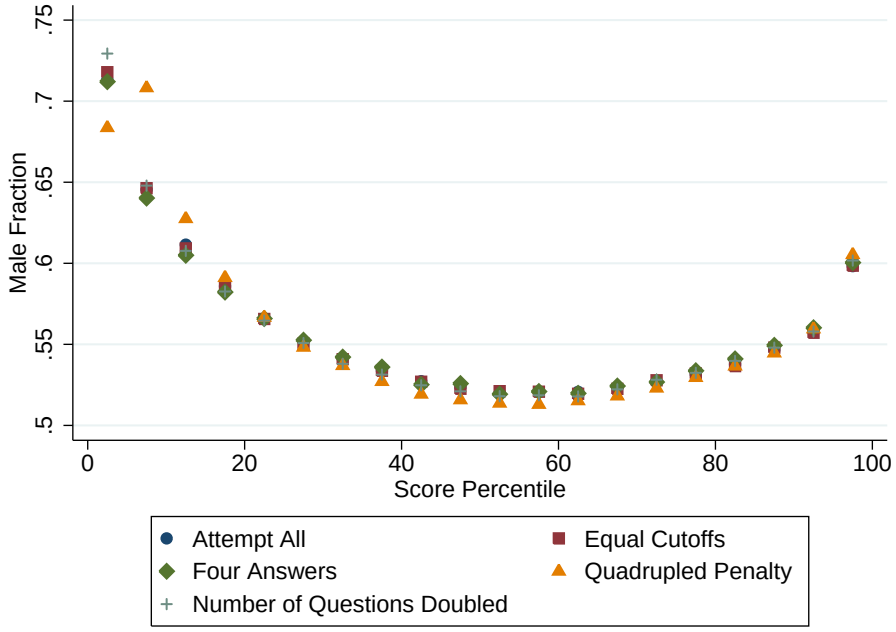
Finally, we examine the impact of the various regimes on the fraction of males in the

[58] With the exception of the quadruple penalty regime for the lowest ventile. Examining more carefully, the lowest 5% actually has a higher average ability than the second lowest 5% (see Figures A.3a and A.3b). 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.

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

different score percentiles. In particular, we want to see if there is any effect on the (over)-representation of males in the top scoring percentiles: the gender gap. Lower risk aversion raises the presence of men at the top and the bottom. Thus the male fraction is U shaped as shown in Figure 13.

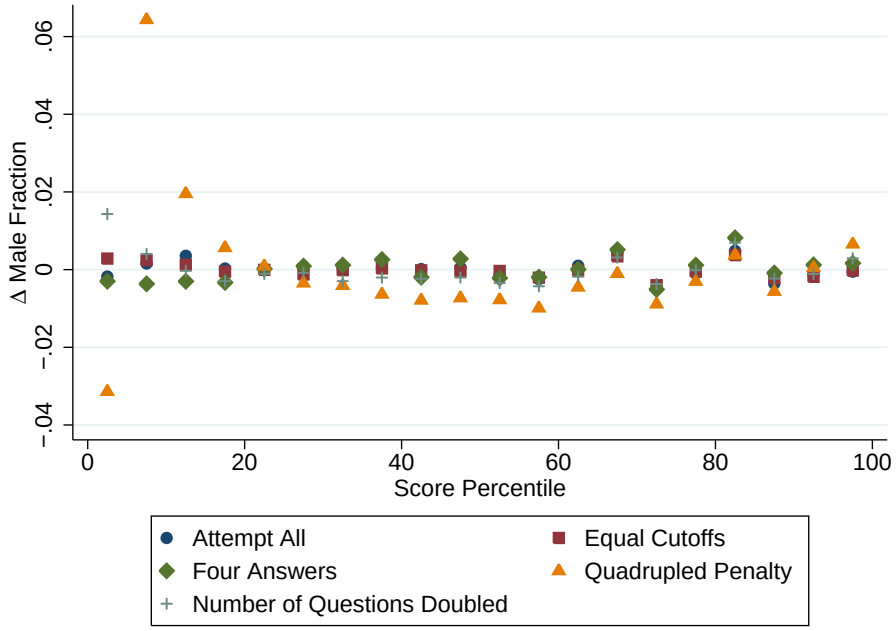
Figure 13: Male Fraction Counterfactuals



In the baseline model, the top 10% of student (by score) are 42% female. But females are 44% of the Social Science track first time takers: a 2% gender gap against females. Examining Figure 14, we see that risk aversion does not appear to be a major component of the gender gap. There is a minimal reduction in the gender gap if we were to eliminate skipping, or eliminate risk aversion differences. The impact on the gender gap of the remaining counterfactuals, while slightly higher, remains small.

Why do risk aversion differences seem to matter so little? There are 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 lie in this 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

Figure 14: Male Fraction (Δ vs baseline)



minor. Essentially, differences in choices made due to skipping behavior are not common, and when they do arise have small consequences. Intuitively, this is like saying that while ordering at a restaurant, the best option is usually clear, and when it is not, the choice made has little consequence.

The small impact of risk aversion, especially for students with high ability, is clear when examining Table 1. A student with a cutoff of 0.275 with ability $\beta = 4$ (approximately top 25%) has an expected score 0.09 lower than a risk neutral student of the same ability. The impact on students in the top 10% is even smaller.

Table 1: 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

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

8 Conclusions

This paper investigates the factors that affect students' exam taking behavior in multiple choice tests. By constructing a structural model of a student's decision to attempt/skip a question in a multiple-choice exam, we estimate structural parameters of the model. Our work has focused on two questions: the extent to which multiple choice exams with negative marking are biased against women versus their better performance in terms of their effectiveness.

It has long been a puzzle why women do worse than men in university entrance exams, despite doing better in school. One reason might be greater risk aversion on the part of women which reduces their performance in multiple choice exams. Overall we find that while female students do act in a more risk averse manner, the impact of this is relatively limited in terms of performance and the prevalence of women in the group admitted to university.

Thus, we need to look elsewhere to quantitatively match the above puzzle. A hypothesis worth exploring is the extent to which this arises from women putting in less effort in high stakes exams because they have less at stake. To the extent that women are less likely to work, or more likely to take time off, their gains from expending effort would fall and this could explain their poorer performance in high stakes exams.

We also find that negative marking has a considerable impact on the effectiveness of the

exam: a penalty of -1 is similar to doubling the number of questions. Moreover, it does so with a minimal impact on gender bias.

References

Julie R Agnew, Lisa R Anderson, Jeffrey R Gerlach, and Lisa R Szykman. Who chooses annuities? an experimental investigation of the role of gender, framing, and defaults. *The American Economic Review*, 98(2):418–422, 2008.

Alireza Ahmadi and Nathan A Thompson. Issues affecting item response theory fit in language assessment: A study of differential item functioning in the iranian national university entrance exam. *Journal of Language Teaching & Research*, 3(3), 2012.

Eva L Baker, Paul E Barton, Linda Darling-Hammond, Edward Haertel, Helen F Ladd, Robert L Linn, Diane Ravitch, Richard Rothstein, Richard J Shavelson, and Lorrie A Shepard. *Problems with the use of student test scores to evaluate teachers*, volume 278. Economic Policy Institute Washington, DC, 2010.

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

William E Becker and Carol Johnston. The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record*, 75(4):348–357, 1999.

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

Yoella Bereby-Meyer, Joachim Meyer, and Oded M Flascher. Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making*, 15(4):313–327, 2002.

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

J Eric Bickel. Scoring rules and decision analysis education. *Decision Analysis*, 7(4):346–357, 2010.

David Budescu and Maya Bar-Hillel. To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement*, 30(4):277–291, 1993.

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

Gary Charness and Uri Gneezy. Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, 83(1):50–58, 2012.

Rachel Croson and Uri Gneezy. Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–474, 2009.

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

Avraham Ebenstein, Victor Lavy, and Sefi Roth. The long-run economic consequences of high-stakes examinations: evidence from transitory variation in pollution. *American Economic Journal: Applied Economics*, 8(4):36–65, 2016.

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

Catherine C Eckel and Philip J Grossman. Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, 68(1):1–17, 2008b.

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

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

Norman Frederiksen. The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39(3):193, 1984.

Patricia Funk, Helena Perrone, et al. Gender differences in academic performance: The role of negative marking in multiple-choice exams. Technical report, CEPR Discussion Papers, 2016.

David W Gerbing and James C Anderson. An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of marketing research*, 25(2): 186–192, 1988.

Christian Gourieroux and Alain Monfort. *Simulation-based econometric methods*. Oxford University Press, 1997.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

Klaus D Kubinger, Stefana Holocher-Ertl, Manuel Reif, Christine Hohensinn, and Martina Frebort. On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment*, 18(1):111–115, 2010.

Richard Lynn. Sex differences on the differential aptitude test in british and american adolescents. *Educational Psychology*, 12(2):101–102, 1992.

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

Gerhard Riener and Valentin Wagner. Shying away from demanding tasks? experimental evidence on gender differences in answering multiple-choice questions. *Economics of Education Review*, 2017.

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

A Appendix: Tables and Figures

Table A.1: Summary Statistics

	Female			Male		
Variable	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
Variable	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
ÖSS-SÖZ score	3,984	110.064	15.414	4,928	108.654	16.673
Normalized High School GPA	3,984	48.663	8.342	4,928	46.122	7.866
Raw Turkish Score	3,984	20.926	9.372	4,928	18.326	9.729
Raw Social Science Score	3,984	14.479	8.890	4,928	15.446	10.037
Raw Math Score	3,984	0.856	3.007	4,928	0.898	2.831
Raw Science Score	3,984	0.084	1.221	4,928	0.152	1.292
Education level of Dad						
Primary or less	3,984	0.522		4,928	0.586	
Middle/High School	3,984	0.308		4,928	0.268	
2-year higher education	3,984	0.026		4,928	0.020	
College/Master/PhD	3,984	0.047		4,928	0.043	
Missing	3,984	0.097		4,928	0.082	
Education level of Mom						
Primary or less	3,984	0.767		4,928	0.818	
Middle/High School	3,984	0.143		4,928	0.115	
2-year higher education	3,984	0.008		4,928	0.007	
College/Master/PhD	3,984	0.012		4,928	0.011	
Missing	3,984	0.070		4,928	0.048	
Prep School Expenditure						
No prep school	3,977	0.337		4,909	0.337	
Scholarship	3,977	0.010		4,909	0.010	
<1000 TL	3,977	0.205		4,909	0.213	
1000-2000 TL	3,977	0.075		4,909	0.058	
>2000 TL	3,977	0.016		4,909	0.014	
Missing	3,977	0.356		4,909	0.368	
Income Level						
<250 TL	3,913	0.427		4,857	0.485	
250-500 TL	3,913	0.409		4,857	0.372	
500-750 TL	3,913	0.104		4,857	0.088	
750-1000 TL	3,913	0.033		4,857	0.030	

(continued on next page)

	Female			Male		
Variable	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
1000-1500 TL	3,913	0.015		4,857	0.014	
1500-2000 TL	3,913	0.007		4,857	0.005	
>2000 TL	3,913	0.005		4,857	0.005	
Time Spent in Math Preparation						
<100 hours	3,984	0.117		4,928	0.117	
100-200 hours	3,984	0.078		4,928	0.068	
>200 hours	3,984	0.022		4,928	0.014	
Time Spent in Science Preparation						
<100 hours	3,984	0.079		4,928	0.072	
100-200 hours	3,984	0.017		4,928	0.013	
>200 hours	3,984	0.004		4,928	0.004	
Time Spent in Turkish Preparation						
<100 hours	3,984	0.075		4,928	0.083	
100-200 hours	3,984	0.104		4,928	0.103	
>200 hours	3,984	0.046		4,928	0.039	
Time Spent in Social Sci. Preparation						
<100 hours	3,984	0.078		4,928	0.085	
100-200 hours	3,984	0.100		4,928	0.093	
>200 hours	3,984	0.064		4,928	0.065	

Table A.2: Estimates of Risk Aversion Cutoff

	Female 1^{st} time takers	**Male 1^{st} time takers**
(0,90)	0.2145	0.2151
	(0.0030)	(0.0018)
[90,100)	0.2304	0.2269
	(0.0011)	(0.0008)
[100,110)	0.2387	0.2356
	(0.0008)	(0.0007)
[110,120)	0.2533	0.2493
	(0.0014)	(0.0011)
[120,130)	0.2662	0.2585
	(0.0027)	(0.0017)
[130,140)	0.2744	0.2622
	(0.0040)	(0.0025)
[140,∞)	0.2706	0.2638
	(0.0047)	(0.0031)

Standard errors are reported in parentheses.

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

	<90	90-100	100-110	110-120	120-130	130-140	> 140
Social Science Track	-0.18	2.56	2.92	2.16	2.44	2.59	1.23

Table A.4: Estimates of Ability Distribution Parameters

	Female		Male	
Social Science Test				
	μ	σ	μ	σ
(0,90)	-1.13	0.82	-1.37	0.95
	(0.24)	(0.23)	(0.17)	(0.12)
[90,100)	-0.65	0.71	-0.63	0.80
	(0.04)	(0.04)	(0.04)	(0.03)
[100,110)	-0.08	0.60	0.07	0.71
	(0.02)	(0.02)	(0.02)	(0.02)
[110,120)	0.49	0.56	0.69	0.66
	(0.02)	(0.03)	(0.02)	(0.02)
[120,130)	1.02	0.48	1.27	0.57
	(0.03)	(0.04)	(0.03)	(0.03)
[130,140)	1.49	0.44	1.72	0.58
	(0.04)	(0.07)	(0.05)	(0.04)
[140,∞)	1.92	0.43	2.22	0.34
	(0.06)	(0.08)	(0.05)	(0.08)
Turkish Test				
	Female		**Male**	
	μ	σ	μ	σ
(0,90)	-0.36	0.69	-0.85	0.78
	(0.15)	(0.12)	(0.11)	(0.09)
[90,100)	0.08	0.58	-0.15	0.67
	(0.03)	(0.02)	(0.03)	(0.02)
[100,110)	0.57	0.54	0.39	0.61
	(0.02)	(0.02)	(0.02)	(0.02)
[110,120)	1.10	0.53	0.92	0.57
	(0.02)	(0.02)	(0.02)	(0.02)
[120,130)	1.67	0.46	1.43	0.52
	(0.03)	(0.03)	(0.03)	(0.03)
[130,140)	2.23	0.42	1.98	0.56
	(0.05)	(0.04)	(0.05)	(0.05)
[140,∞)	2.91	0.62	2.62	0.56
	(0.09)	(0.08)	(0.07)	(0.09)

Standard errors are reported in parentheses.

Figure A.1: Distributions of Social Science and Turkish Ability

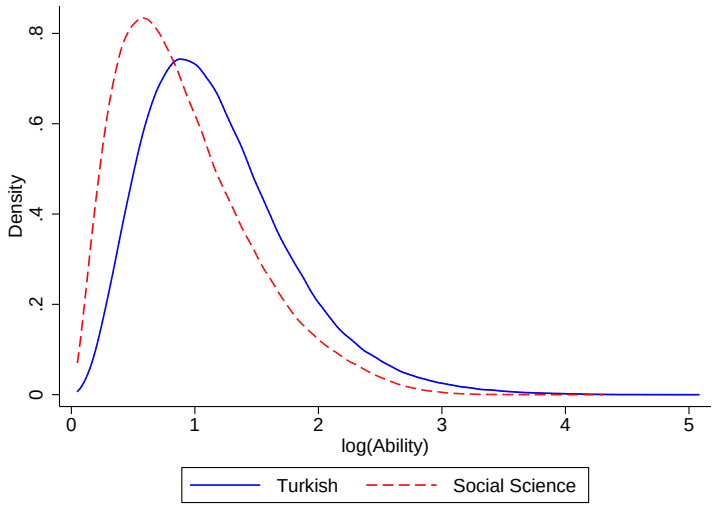
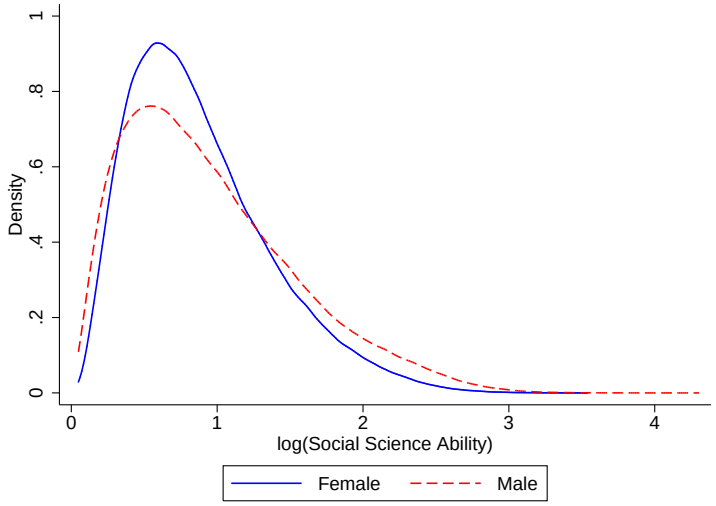


Figure A.2: Distributions of Ability by Gender

(a) Social Science



(b) Turkish

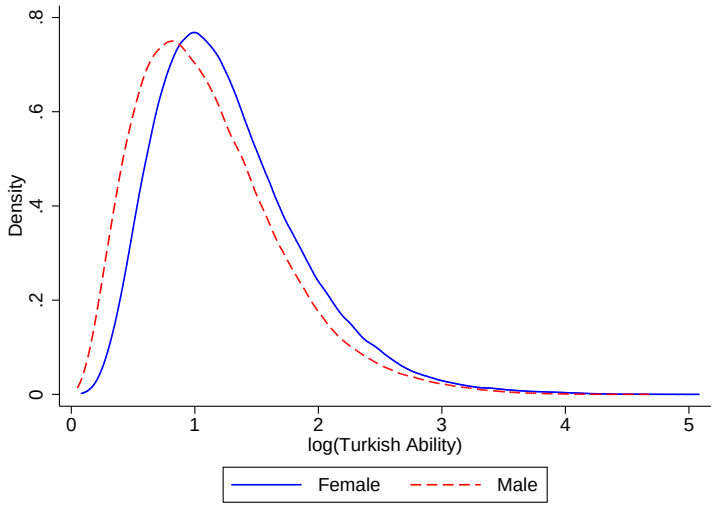


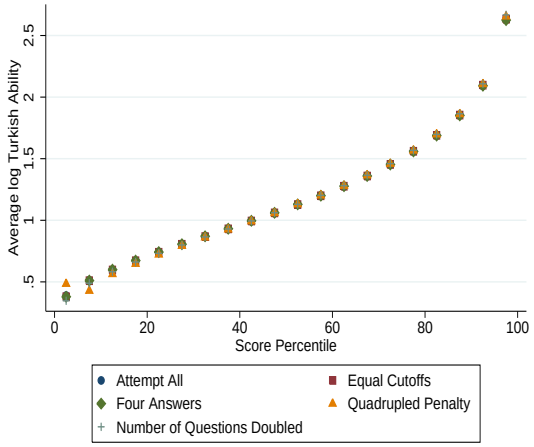
Table A.5: Estimates of Correlation between logs of Turkish and Social Science Ability

	Female 1^{st} time takers	**Male 1^{st} time takers**
(0,90)	1.000	0.788
	n/a	(0.021)
[90,100)	0.916	0.847
	(0.003)	(0.004)
[100,110)	0.949	0.889
	(0.001)	(0.002)
[110,120)	0.919	0.927
	(0.002)	(0.001)
[120,130)	0.991	0.894
	(0.001)	(0.004)
[130,140)	1.000	0.950
	n/a	(0.003)
[140,∞)	0.863	0.854
	(0.018)	(0.029)

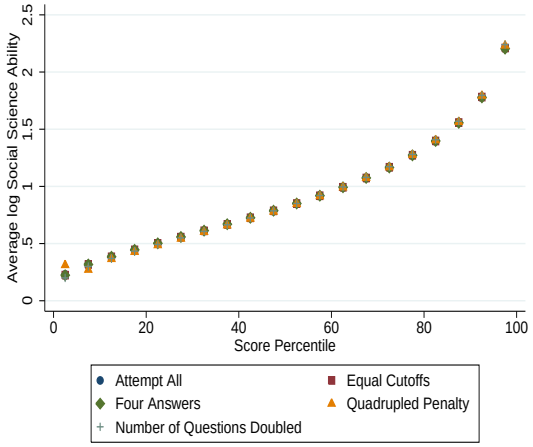
Standard errors are reported in parentheses.

Figure A.3: Counterfactual Ability Distributions

(a) Turkish Ability



(b) Social Science Ability



B Online Appendix

Table B.1: Test Weights

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

B.1 Predicted Social Science Track (ÖSS-SÖZ) Score

The risk taking behavior of students is likely to depend on their score as 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 and very similar scores may have very different utilities if they are on two opposite sides of a very desirable program's cutoff. However, we cannot use students' actual exam scores as a proxy for ability as these are endogenous objects that are 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)[60], 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 track (ÖSS-SÖZ) scores for each student.[61]

[60] 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 normalized GPA for each student, which is 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.

[61] Regression results are presented in Table B.2.

Table B.2: Regression to Predict ÖSS-SÖZ Score

Variable	Male	Female
Normalized High School GPA	0.760***	0.643***
	(0.013)	(0.013)
Income Level (base: <250 TL)		
250-500 TL	0.359	0.373
	(0.369)	(0.373)
500-750 TL	-0.116	-0.361
	(0.622)	(0.595)
750-1000 TL	-0.191	-0.589
	(1.003)	(0.985)
1000-1500 TL	0.444	-0.016
	(1.463)	(1.443)
1500-2000 TL	4.522*	0.129
	(2.541)	(2.146)
>2000 TL	4.178	5.668**
	(2.568)	(2.485)
Education level of Mom (base: Primary or less)		
Middle/High School	0.387	-0.310
	(0.566)	(0.522)
2-year higher education	-1.978	3.023
	(2.073)	(1.874)
College/Master/PhD	-0.314	2.369
	(1.786)	(1.829)
Missing	-0.094	-0.591
	(0.997)	(1.058)
Education level of Dad (base: Primary or less)		
Middle/High School	1.292***	1.059***
	(0.405)	(0.399)
2-year higher education	0.166	-0.124
	(1.197)	(1.070)
College/Master/PhD	2.287**	0.942
	(0.927)	(0.946)
Missing	-0.643	0.756
	(0.785)	(0.923)
Time Spent in Math Preparation		
<100 hours	2.504***	3.206***
	(0.948)	(1.081)
100-200 hours	3.876***	4.742***
	(1.111)	(1.197)
>200 hours	5.573***	1.071
	(1.872)	(1.718)
Time Spent in Science Preparation		
<100 hours	1.981**	0.918

(continued on next page)

Variable	Male	Female
	(0.783)	(0.760)
100-200 hours	-0.477	-1.736
	(1.597)	(1.430)
>200 hours	-5.832**	2.519
	(2.933)	(2.897)
Time Spent in Turkish Preparation		
<100 hours	-0.113	-1.416
	(1.202)	(1.437)
100-200 hours	-0.539	-1.740
	(1.183)	(1.345)
>200 hours	-1.090	-0.094
	(1.467)	(1.552)
Time Spent in Social Science Preparation		
<100 hours	-1.079	1.954*
	(1.015)	(1.176)
100-200 hours	-0.430	2.141**
	(1.016)	(1.091)
>200 hours	1.037	1.622
	(1.094)	(1.170)
Prep School Expenditure (base: No prep school)		
Scholarship	4.386**	5.521***
	(1.744)	(1.691)
<1000 TL	5.548***	3.947***
	(0.599)	(0.613)
1000-2000 TL	5.167***	5.080***
	(0.899)	(0.836)
>2000 TL	6.247***	5.371***
	(1.655)	(1.572)
Missing	-0.512	-0.219
	(0.387)	(0.395)
Constant	28.128***	37.628***
	(1.400)	(1.434)
School Type Control	Yes	Yes
Observations	4,823	3,894
R2	0.550	0.566

Proof of the Proposition 2.

$$\begin{aligned}
\frac{\partial c(\tau)}{\partial \tau} &= \frac{\partial \left[\frac{\exp(k\tau)-1}{\exp(k\tau)-\exp(-\tau)}\right]}{\partial \tau} \qquad (10)\\
&= \frac{k\exp(k\tau)\left[\exp(k\tau)-\exp(-\tau)\right]-\left[\exp(k\tau)-1\right]\left[k\exp(k\tau)+\exp(-\tau)\right]}{\left[\exp(k\tau)-\exp(-\tau)\right]^2}\\
&= \frac{-k\exp(k\tau-\tau)-\exp(k\tau-\tau)+k\exp(k\tau)+\exp(-\tau)}{\left[\exp(k\tau)-\exp(-\tau)\right]^2}\\
&= \frac{-(k+1)\exp(k\tau-\tau)+k\exp(k\tau)+\exp(-\tau)}{\left[\exp(k\tau)-\exp(-\tau)\right]^2} \qquad (11)
\end{aligned}$$

As the denominator of the expression is positive, it is enough to show that the nominator is positive.

We want to show that

$$k\exp(k\tau)+\exp(-\tau) > (k+1)\exp(k\tau-\tau)$$

Divide both sides of the equation by $(k+1)\exp(k\tau-\tau)$

$$\begin{aligned}
\frac{k\exp(k\tau)+\exp(-\tau)}{(k+1)\exp(k\tau-\tau)} &> 1\\
\frac{k}{k+1}\exp(\tau)+\frac{1}{k+1}\exp(-k\tau) &> 1
\end{aligned}$$

Since the exponential function is a strictly convex function, the following inequality holds

$$\frac{k}{k+1}\exp(\tau)+\frac{1}{k+1}\exp(-k\tau) > \exp(\frac{k}{k+1}\tau-\frac{1}{k+1}k\tau) = \exp(0) = 1$$

■

B.2 Other Tracks

So far we have focused on the Social Science track (ÖSS-SÖZ). Our approach can be used for subjects where there is partial knowledge such as the Turkish component in the

Turkish-Math and Language tracks. As seen in Table B.1, the Turkish-Math track (ÖSS-EA) also places high emphasis on the Turkish section of the exam, a section which is well described by the model. The model also applies to the Language section of the Language track (ÖSS-DIL), which as would be expected, accounts for a large part of the student's admission score. We do not use our model on the Math and Science tests. This is because of the limited presence of spikes which are a key part of our identification strategy. The lack of spikes we hypothesize, comes from the questions being of a different type. Science and Math problems have to be worked out and if done correctly, this eliminates all but one answer. As a result, there is a lack of partial knowledge: either the signal is fully informative, or there is no information.

There are some 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.

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.

B.3 Estimation

Estimation follows that in the main body of the paper. After separating first attempt students into predicted ÖSS score bins by gender, 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 analogous to before.

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 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.

B.4 Data

Focusing on students making their first attempt, we obtain 7972 female and 7919 male students for the Turkish Math track, and we obtain 9280 female and 3681 male students for the Language track[62].

Separating into score bins, we have sample sizes as shown in Tables B.3 and B.4.

The aggregate score patterns can be seen in Figures B.1 and B.2.[63] 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.[64]

[62]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.

[63]Gridlines are 1.25 points apart in Figure B.1 and 2.5 apart in Figure B.2.

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

Estimating the model for data showing a very different aggregate pattern also serves as a robustness check.

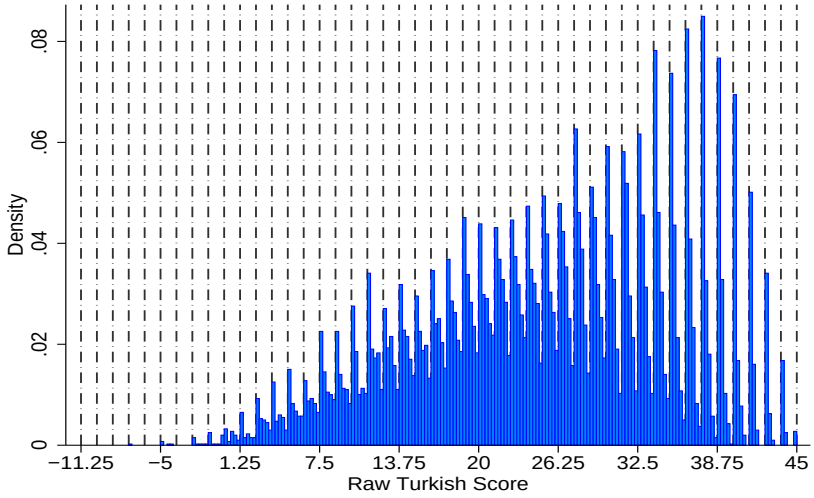
Table B.3: Number of observations vs predicted score bin, Turkish-Math track

	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 B.4: 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 B.1: Turkish Score Distribution of Turkish Math Track Students

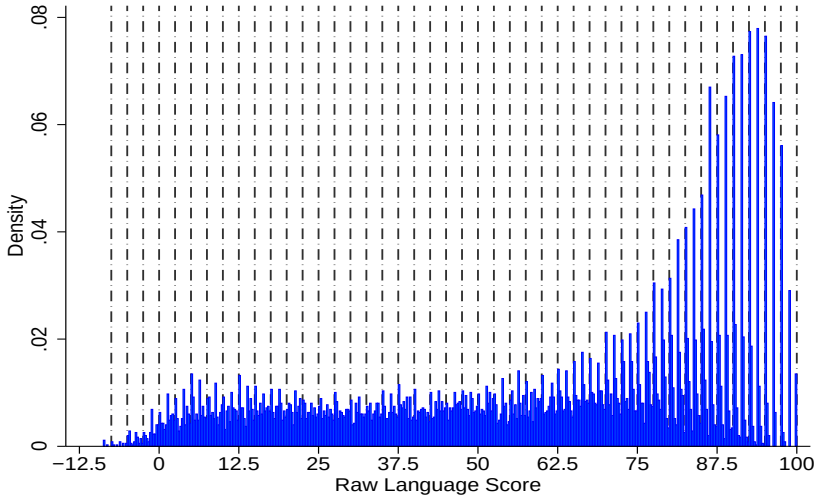


B.5 Results

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

As before, there are two important patterns. First, as seen in Figures B.3 and B.4, 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: students

Figure B.2: Language Score Distribution of Language Track Students



are less risk averse for low scores as they are below the cutoff for applying. Secondly, males tend to have lower cutoffs than females, i.e., they are less risk averse, and this difference tends to be significant only in higher score bins. 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.

Note the magnitude of and patterns in the cutoffs, and the degree of differences between male and female students, are relatively similar across tracks. In all three tracks, cutoffs rise with score and males are less risk averse.

Table B.5: 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 B.6: 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

Figure B.3: Estimates of Attempt Cutoffs: Turkish Math Track

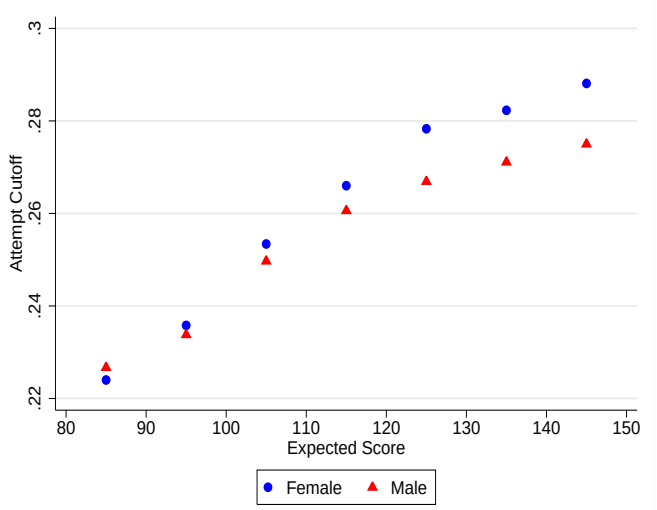


Table B.7: t-stat for test: (female cutoff-male cutoff)=0

	<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

B.6 Extension of Model to Item Response Data

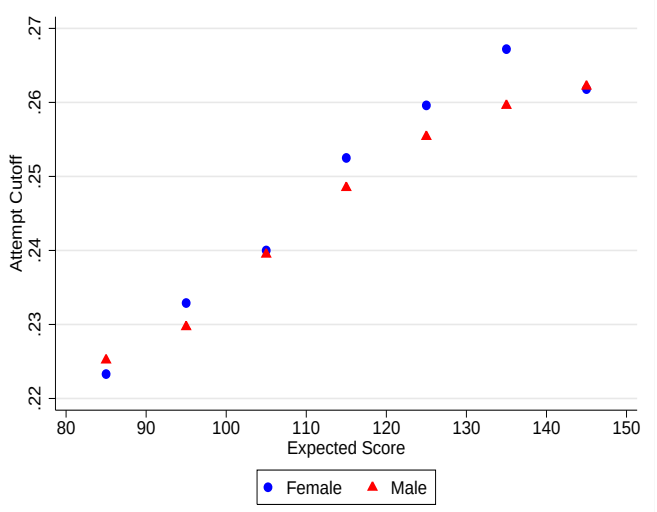
The data used to estimate the behavior of students in this paper was limited, showing only the raw scores in each section. Therefore, we have assumed that students within a group are identically and independently distributed in their test taking characteristics, and that questions have the same level of difficulty. However, with item level datasets it is possible to say more about individual students, and individual questions.

In this section, we present an extended version of the model presented in Section 4 that can be used to provide estimates taking advantage of item-by-item responses. We, then, apply this model to a small sample of students taking a mock exam in a "dersane" (schools that prepare students for the exam).

B.6.1 The Extended Model

As in Section 4, students approach a question by observing signals for each answer. Signals for incorrect and correct answers are draw from Pareto distributions, with identical

Figure B.4: Estimates of Attempt Cutoffs: Language Track (English)



support but different shape parameters. The student interprets the signals rationally, and finds probabilities that each answer is correct. The student will either answer the question, choosing the answer with the highest likelihood of being correct, or skip the question. Again, the choice is determined by comparing the likelihood of success to some cutoff $c \in [0.2, 1]$.[65]

In contrast to the model presented in Section 4, here we allow questions to vary in difficulty. Some questions will be difficult, and in the context of the model, this will be a result of having similar signal distributions for the correct and incorrect answers. Some questions will be easy which means students will tend to have very different distributions for the correct and incorrect answers. As before, students are heterogeneous in their ability in the different exam sections.

We need to introduce the following parameters to extend the model: question difficulty $q > 0$ and student ability $s > 0$. Both correct and incorrect answers generate signals drawn from a Pareto distribution with scale parameter 1.[66] The distribution for the correct answer has shape parameter q, and the distribution for incorrect answers has shape parameter $q + s$.

[65] With the penalty for incorrect answers set to 0.25, and five possible answers, we cannot distinguish between $c \in [0, 0.2]$, as the student will always answer every question. With alternative structures, the range of cutoffs we can consider will be different.

[66] As previously, the choice of scale parameter/support is without loss of generality if identical support has been assumed.

As shown previously, it is the ratio of the shape parameters that determines the student's ability to distinguish the correct answer from the incorrect answers. Thus, a student with ability s considering a question with difficulty q will have an effective ability (comparable to β previously) of $k = \frac{q+s}{q} > 1$. As with β, higher values of k are associated with a higher likelihood of success. This parameterization allows for variation in question difficulty, and in particular, allows for both very hard questions, where even the top students have great difficulty, and very easy questions, where even the worst students have a high chance of selecting the correct answer. In addition, it maintains the effect of student ability: k is increasing in s, student ability, regardless of question difficulty.

B.6.2 Estimation

The model can be estimated through maximum likelihood. Let $x_{m,n} \in \{Correct, Incorrect, Skip\}$ denote the outcome of student m in question n; the data consists of question outcomes, $x_{m,n}$, for students $m = 1, ..., M$ and questions $n = 1, ..., N$. The probability of each outcome can be found, given $(k_{m,n}, c_m)$, or equivalently (s_m, q_n, c_m), where $k_{m,n}$ is the effective ability of student m in question n, and is equal to $\frac{q_n+s_m}{q_n}$, where q_n is the difficulty of the n^{th} question and s_m is the ability of the m^{th} student. The cutoff of the m^{th} student is c_m. We denote this probability as $P(x_{m,n}|s_m, q_n, c_m)$.[67] The estimates of student abilities $\{s_m\}_{m=1}^M$, the student risk aversion cutoffs $\{c_m\}_{m=1}^M$, and the question difficulties $\{q_n\}_{n=1}^N$ come from the following:[68]

$$\max_{s_m, q_n, c_m} \sum_{n=1}^{N} \sum_{m=1}^{M} \log P(x_{m,n}|s_m, q_n, c_m). \tag{12}$$

[67] While Section 4 featured a correct answer and four incorrect answers, alternative numbers of incorrect answers can be considered. It is possible to have different numbers of possible answers in the same test; one may use a CARA utility function to find cutoffs for questions with different numbers of answers which feature the same level of risk aversion.

[68] Note that it is necessary to make a normalization. We cannot identify the absolute difficulty of questions, only the relative difficulties. Without loss of generality, normalize the difficulty of the first question, $q_1 = 1$.

B.6.3 Data: Mock Exams

The estimation procedure is applied to a mock exam held by a "dersane" (schools that prepare students for the exam). The exam consists of 120 questions: 30 for each subject, Turkish, social science, math and science. The same questions are shuffled to create two versions of the exam to prevent cheating.[69] We observe 30 students taking one version of the exam. We first apply the estimation procedure to the Turkish section of the exam, then to the social science section of the exam, then to both at the same time [70]

We recover question difficulties and student characteristics. Of particular importance is the relationship between scores and estimated abilities of students. A combination of the two abilities (Turkish and Social Science) gives us a measure of quality of the student. Figure B.5 depicts the correlation between ability ranking and score (with negative marking) as well as ability and the number of questions answered correctly.[71] As the figure shows, the three measures are highly correlated, but not perfectly so. In particular, the score ranking (one point for a correct answer, minus a quarter point for an incorrect answer) is closer to the 45 degree line than is the rank by total number of correct answers.[72] The reasons for this are discussed below.

With item response data, one can recover the ability parameters of students and question difficulties by using our model or the Rasch model. However, in addition to accommodating for skip behavior, our model has some important features. In the Rasch model, the number of questions that a student answers correctly is a sufficient statistic for ability. For example, if a student gets 1 out of 2 questions correct, we can find ability, without knowing which questions. However, in our model, it is important which questions you answer correctly,

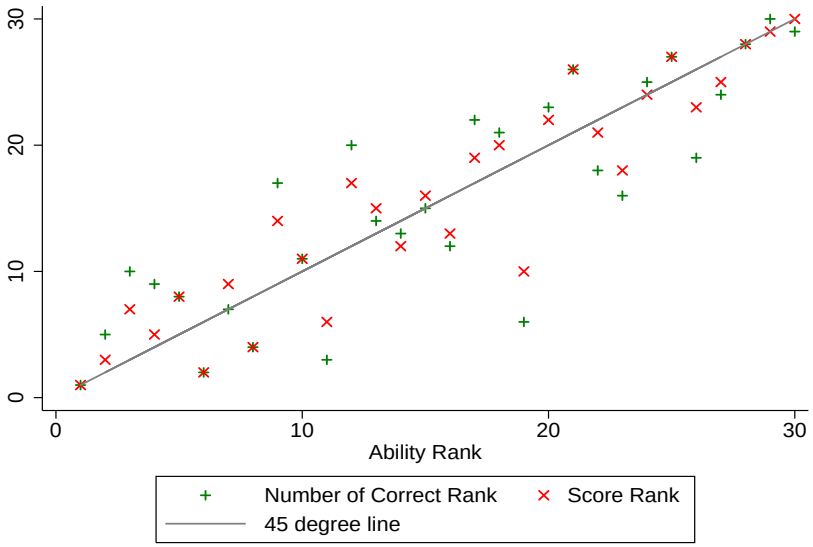
[69] We don't observe questions that correspond to each other in both versions. Therefore, we use one version of the exam.

[70] Here a student m is characterized by (s_m^T, s_m^{SS}, c_m). We normalize the first question of both exams to have a difficulty of 1 so that the difficulty of each question in each part of the exam is relative to the first question. Following this, the difficulties could be easily rescaled to be relative to the average difficulty for ease of interpretation.

[71] The number of questions answered correctly is important for the Rasch model.

[72] While the correlation between ability ranking and the score ranking is 0.929, the correlation between ability ranking and the number of correct ranking is 0.844.

Figure B.5: Ranking of students by estimated ability compared to ranking based on observables: number of correct answers and total score.



and which questions you answer incorrectly (as well as skip). Suppose that one question is very easy, and one question is very challenging. Which student would likely have the higher ability - the one who answers the easy question correctly and not the difficult question, or the one who answers the difficult question correctly but not the easy question? Keeping in mind that a student with minimal ability can still have a one in five chance of correctly answering the question (random guess), we would say that the former student would have a higher estimated ability; the latter student could not answer the easy question correctly and simply got lucky with the hard question. Our estimation procedure incorporates this intuition.

Related to the previous observation, the fraction of students who get a question correct is not sufficient for describing difficulty. *It is important to know which students get it right, and which students get it wrong.* While the fraction getting it correct is highly correlated with difficulty, more so in large samples, it is not perfect. For small samples, there is valuable information contained in the identity of the students who correctly answer.

The Rasch model also does not distinguish between incorrect answers and a skip, while our model does. Suppose that there are twenty questions and two students with the same

ability. However one is more risk averse than the other. As a result, the risk averse student skips questions and so has a lower fraction of correct questions on average. The Rasch model in this case would wrongly deem the risk averse student to have a lower ability. Risk preferences can, and do, vary across individuals. By incorporating skipping behavior, we can more accurately compare student's abilities. Our model, will account for differences in, and the impact of risk preferences.