Hit or Miss? Test Taking Behavior in Multiple Choice Exams

Ş. Pelin Akyol [1] James Key [2] Kala Krishna [3]

[1] Bilkent University

[2] University of Western Australia

[3] Pennsylvania State University

October 12, 2015

Motivation

- Multiple choice tests are widely used
  - University entrance exams (Turkey, Greece, Japan, Korea, China,...)
  - The SAT and GRE
- Disadvantage: Random guessing is possible
  - Apply penalty for incorrect answers to prevent random guessing
  - Decision to guess/not depends on knowledge and risk aversion.
- Does the exam format grant certain groups an advantage? Fair?

Literature

- Literature: Women skip more often
  - Reduced form: Ben-Shakhar and Sinai (1991)
  - Experimental: Baldiga (2013), Espinosa and Gardeazabal (2010)
  - Proper grading rules: Bernardo (1998), Burgos (2004), Espinosa and Gardeazabal (2005)
  - Semi Structural: Pekkarinen (2014) (Rasch model), Tannenbaum (2012)

Education System

- ÖSS Exam - held annually
  - Paper based multiple choice exam
  - Most important determinant of university admission weights
- Four sections: math, science, social science and Turkish
  - 45 questions in each part
- Expectation of 0 if guess randomly
  - 5 answers
  - +1 point for correct, -0.25 for incorrect
- Students can skip the question, giving 0 points
- Attitudes to risk will impact outcomes

The Data

- Sample of students taking 2002 University Entrance Exam
  - Scores in each section
  - Background information
- Focus on social science track, 1^{st} time takers (8917 students)
- Two sections of interest: social science and Turkish

Gender Gap

- There is a gender gap Scores
- Only 9% of these students gain university entrance
- Males are over-represented in the top 9%
  - 9.4% of males are in this top group
  - Compare to 8.5% of females
- A model where students form beliefs regarding the chance of success when answering a question

The Model

- Students generate beliefs regarding answers
- The questions are attempted independently
- For each answer $n \in \{1, ..., 5\}$, the student draws a signal x_n
- The correct answer draws from a Pareto distribution with shape parameter α and scale parameter A
- Incorrect answers draw from a Pareto distribution with shape parameter β and scale parameter B
- Know parameters, but not which distribution they are drawing signal from
- Based on signals, they form beliefs regarding which answer is correct answer

The Distributions

Assumption

The scale parameters of the distributions are equal: $A = B > 0$. That is, the minimum signal with positive support is the same for both the incorrect answers and the correct answer.

- Student can never be absolutely certain of the answer (either correct or incorrect)
- Simplifies the state space of student types
- Interpretation of the parameters more intuitive

Proposition

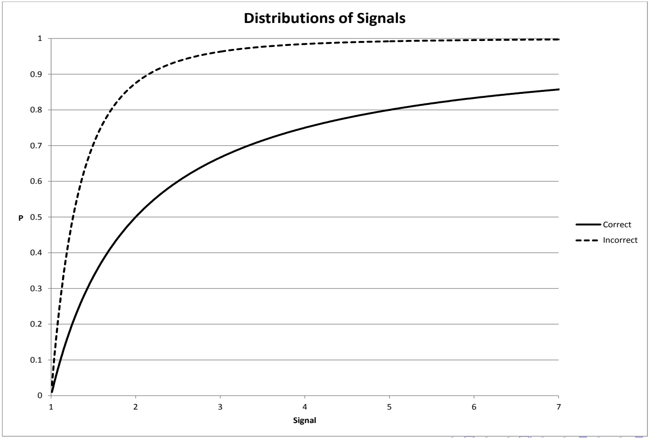
The outcome of the model is independent of the size of A

Proposition

The outcome of the model depends only on the ratio β/α

Student Ability

- Without loss of generality, $A = 1$, $\alpha = 1$, so that β is ability.
- Distributions of signals for a student with $\beta = 3$, approximately median



To Answer or Not

- Students draw signals, $\{x_1, x_2, x_3, x_4, x_5\}$, for each answer.
- Form beliefs
- Student knows which answer is most likely to be correct and the probability
- But should the student choose that answer? Or should they skip it?
- Risk preferences: cutoff c
    - If chance of success is greater than c, attempt
    - Otherwise, skip

To Answer or Not

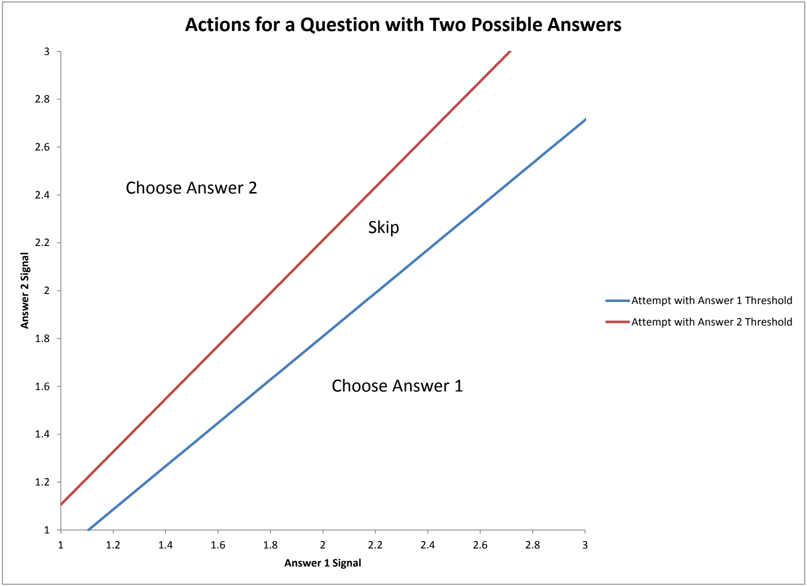
- Let $m = \arg\max_{i \in \{1,\ldots,5\}} x_i$, the answer with the highest signal, the one most likely to be correct
- Through Bayes' rule, answer m is correct with probability:

$$\frac{x_m^{\beta-\alpha}}{x_1^{\beta-\alpha} + x_2^{\beta-\alpha} + x_3^{\beta-\alpha} + x_4^{\beta-\alpha} + x_5^{\beta-\alpha}} \tag{1}$$

  where $\beta - \alpha > 0$
- The student possesses a cutoff $c \geq 0.2$, and will skip the question whenever the above equation is less than c
- Whenever there is no answer with a great enough chance of being correct, they skip the question
- Otherwise they attempt the question, choosing answer m

Probs

Three Possible Outcomes



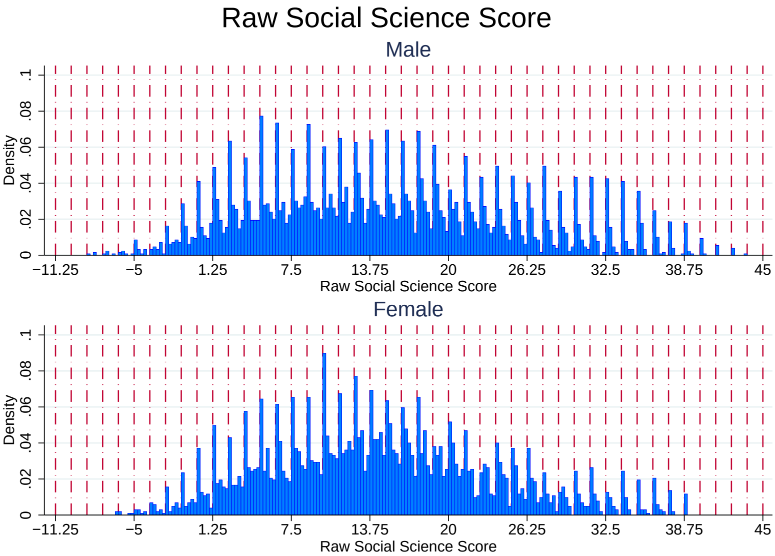
Scores

The 45 questions in each section are attempted independently, so we can find the probability that the student obtains each possible raw score, e.g. the probability to obtain a score of 34.75 in section K

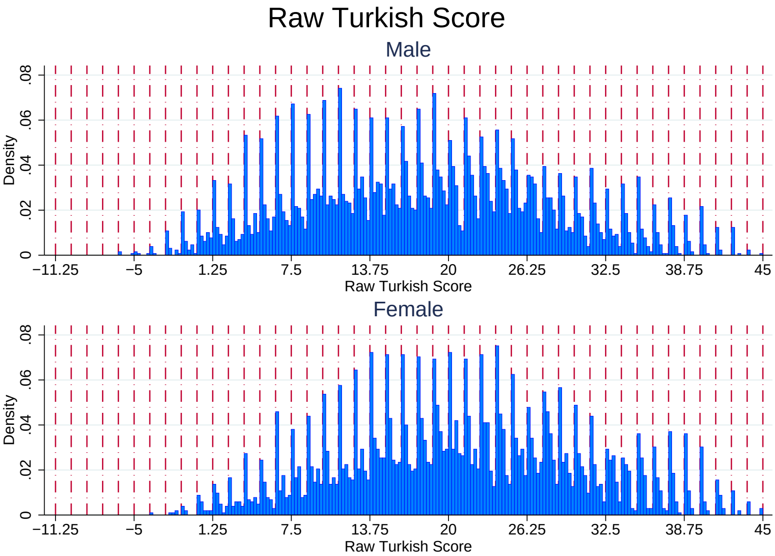
- 220 possible scores
- From -11.25 to 45
- Certain scores, for example 44.75, are impossible
- There can be multiple ways to obtain certain scores
    - 40: (40 correct, 5 skips) or (41 correct, 4 incorrect)

Exam Patterns

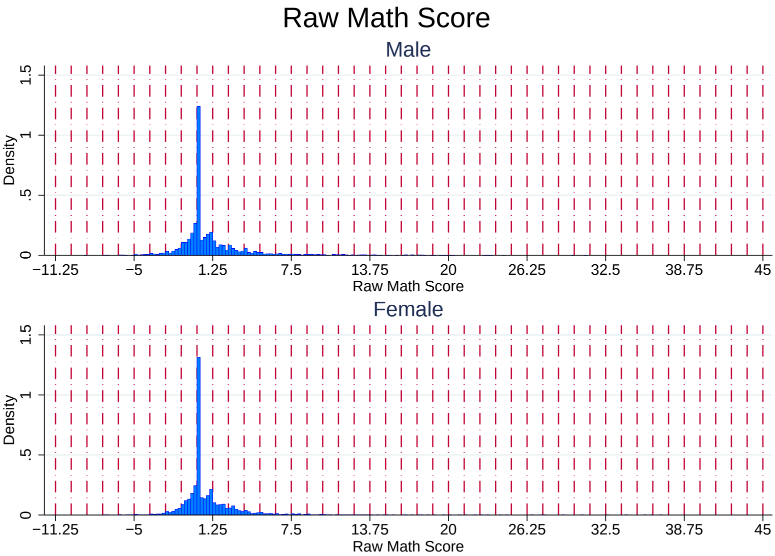
- The following graphs show the score distributions of social science track students
- Social science and Turkish sections
- First time takers, female and male students
- The score distributions exhibit interesting patterns


Raw Social Science Score
Male
Density
0
.02
.04
.06
.08
.1
−11.25
−5
1.25
7.5
13.75
20
26.25
32.5
38.75
45
Raw Social Science Score
Female
Density
0
.02
.04
.06
.08
.1
−11.25
−5
1.25
7.5
13.75
20
26.25
32.5
38.75
45
Raw Social Science Score

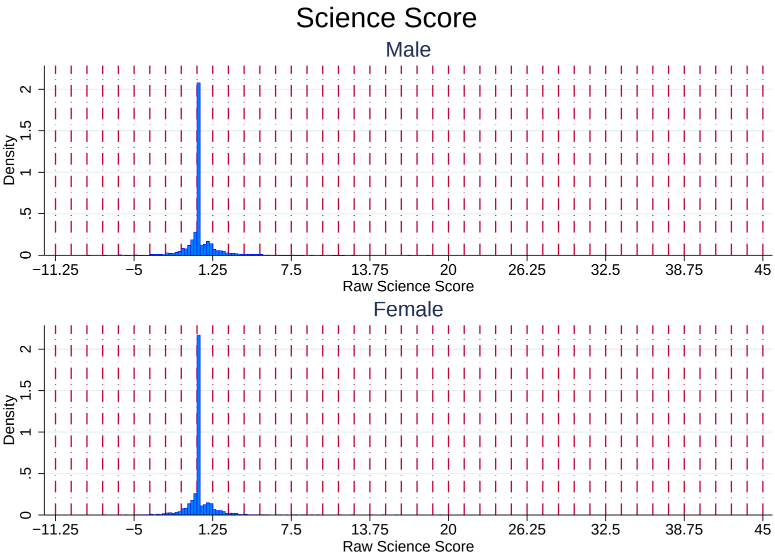
Turkish Score



Math Score



Raw Math Score
Male
Density
0
.5
1
1.5
−11.25
−5
1.25
7.5
13.75
20
26.25
32.5
38.75
45
Raw Math Score
Female
Density
0
.5
1
1.5
−11.25
−5
1.25
7.5
13.75
20
26.25
32.5
38.75
45
Raw Math Score


Science Score
Male
Density
0
.5
1
1.5
2
−11.25
−5
1.25
7.5
13.75
20
26.25
32.5
38.75
45
Raw Science Score
Female
Density
0
.5
1
1.5
2
−11.25
−5
1.25
7.5
13.75
20
26.25
32.5
38.75
45
Raw Science Score

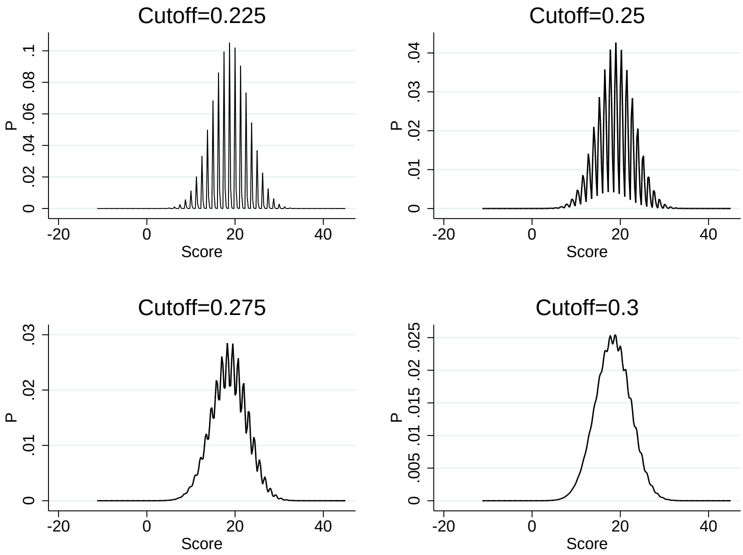
Spikes

- The social science track score distributions for Social Science and Turkish display a considerable amount of structure throughout the support
- These spikes correspond to scores which could be obtained while attempting every question
- Spikes are 1.25 apart - instead of gaining 1 point, a quarter point is lost
- This pattern implies that there is relatively little skipping behavior in these sections of the exam, for social science track students
- This pattern allows us to identify key components of the model

Identification

- Means of ability: means of section scores
- Similarly with variance/covariance of ability
- Identification of risk aversion is less obvious

Identification



Estimation

- The relationship between ÖSS-SÖZ score and the utility is not necessarily constant throughout the range of score:
- The degree of risk aversion may be different
  - Students with score < 105 cannot submit preference for college programs
  - $105 \leq$ score < 120 can submit preference only for 2-years college programs
  - ≥ 120 can submit preference for all 2-years and 4-years college programs
- Group students according to gender, and the range in which their predicted ÖSS-SÖZ score lies:
  - $(0, 90)$, $[90, 100)$, $[100, 110)$, $[110, 120)$, $[120, 130)$, $[130, 140)$, and $[140, \infty)$

Estimation

- For each group, for each section, estimate the following:
  - Risk aversion measure *c*, below which students will skip, common to all students in that group/score range
  - The parameters of ability distribution: β_T and β_{SS} and $\Sigma(\beta)$
- For given *c*, $\mu(\beta)$ and $\Sigma(\beta)$, simulate a number of students
- Compare the following moments to those found in the data
  - Fraction of students obtaining scores corresponding to attempting all minus fraction skipping one
  - Means and variance/covariance of scores Link

Cutoffs

	Female	**Male**
(0,90)	0.2429	0.2100
	(0.0269)	(0.0026)
[90,100)	0.2322	0.2272
	(0.0023)	(0.0019)
[100,110)	0.2396	0.2364
	(0.0009)	(0.0010)
[110,120)	0.2546	0.2480
	(0.0017)	(0.0016)
[120,130)	0.2612	0.2594
	(0.0037)	(0.0043)
[130,140)	0.2763	0.2633
	(0.0062)	(0.0036)
[140,∞)	0.2796	0.2697
	(0.0175)	(0.0076)

Standard errors are reported in parentheses.

Cutoffs

- Females tend to have higher cutoffs than males
- Consistent with males being less risk averse
- Cutoffs tend to rise as we move from low scoring students to high scoring students
- Consistent with students acting in a less risk averse manner when appropriate
  - A score below the application threshold results in no possibility of admission

Ability

- The estimation procedure also finds the distribution of ability for each group ability dist
  - We can compare ability distributions across groups
  - Turkish ability is higher than social science ability on average
  - Males have greater variance in ability
  - Males have a comparative advantage in social science

Counterfactual Experiments

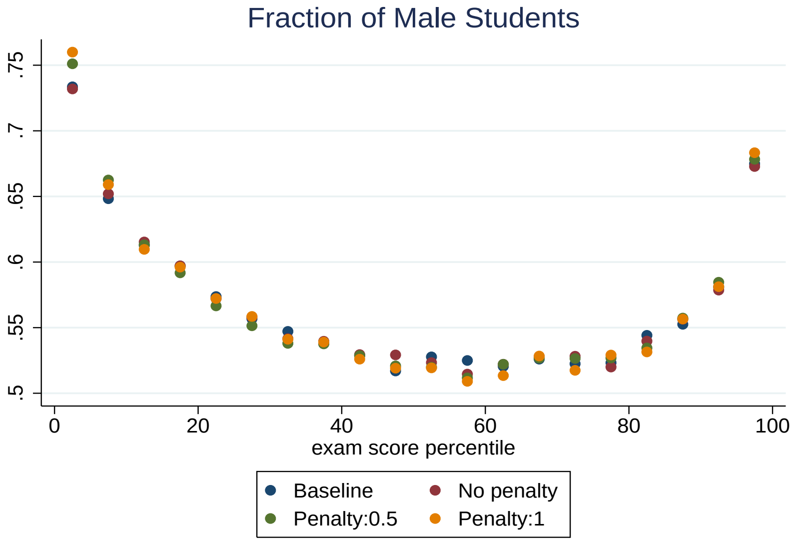
- Structural parameters of the model have been recovered.
- What would happen if we change the testing environment?
- We can conduct counterfactual experiments, to see the effect of the test regime on the relationship between ÖSS-SÖZ score percentiles and:
  1. Share of Male students
  2. Average Turkish ability
  3. Average social science ability

Counterfactual Experiments

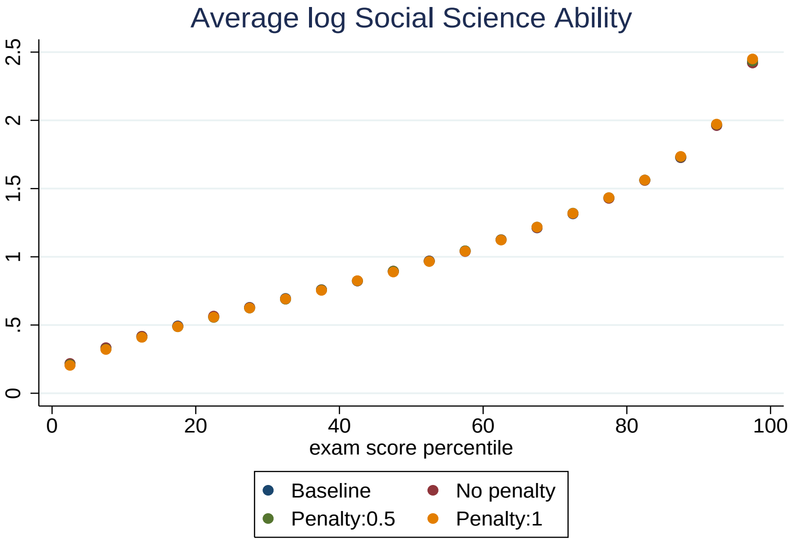
In addition to the baseline model, we consider three counterfactuals:

- No penalty
  - Students attempt every question
  - Risk aversion has no impact
- Penalty for incorrect answer is doubled
- Penalty for incorrect answer is quadrupled

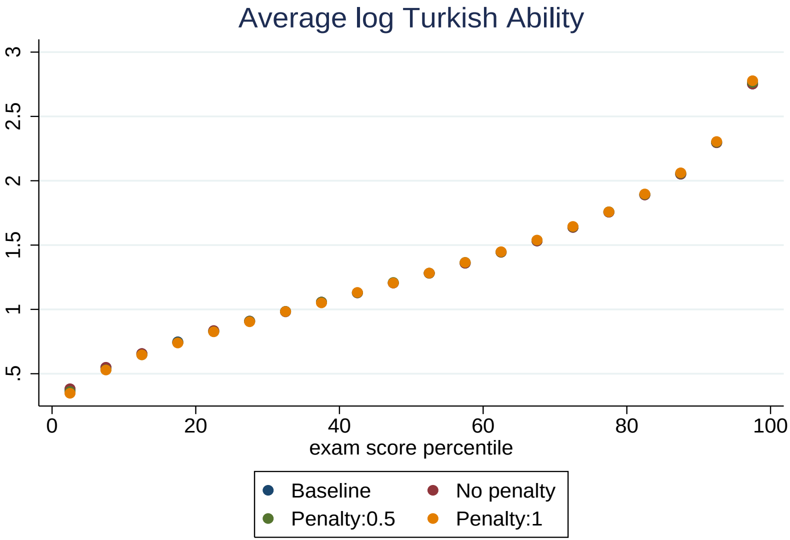
Counterfactual Results



Counterfactual Results



Counterfactual Results



Counterfactuals

- We do not observe any substantial differences
- Differences in risk aversion do not explain the gender gap
- Two reasons for this:
  - Students skip very few questions in this part of the exam
  - Given the low cutoffs, very little difference between skipping and attempting
- Could be specific to these students
- And these tests

Extension

- Suppose we have item level response data
- Extend model to include question difficulty
  - The correct answer draws from a Pareto distribution with shape parameter q and scale parameter 1
  - The incorrect answer draws from a Pareto distribution with shape parameter $q + s$ and scale parameter 1

where

-
  - $q_n > 0$ is the question difficulty
  - $s_m > 0$ is the student ability
- A student with ability s_m considering a question with difficulty q_n will have an effective ability

$$k_{m,n} = \frac{q_n + s_m}{q_n}$$

Extension

- Effective ability, $k_{m,n}$, is increasing in student ability, s_m, decreasing in question difficulty, q_n.
- Let $x_{m,n} \in \{Correct, Incorrect, Skip\}$ denote the outcome of student m in question n.
- Probability of each outcome can be found given (s_m, q_n, c_m), $Pr(x_{m,n}|s_m, q_n, c_m)$.
- Estimation with maximum log likelihood
- Identify difficulty of each question, ability and risk preferences of each student

Conclusions

- Rich Structure of Turkish ÖSS Exams allows us to infer how students behave during exams, and the distributions of ability for the social science and Turkish sections
- Female students are more risk averse than male students
- However, attitudes to risk are shown to have minimal impact on the ranking of students by the final allocation score
- Differences are driven primarily by ability
- Penalizing students for incorrect answers results in a more effective separation of students by ability
- Model can be extended to include question difficulty

Outcome Probabilities and Student Type

Back

β	Cutoff	Prob(S)	Prob(C)	Prob(I)	PPQ
2	0.2	0	0.405	0.595	0.257
2	0.225	0.012	0.403	0.585	0.257
2	0.25	0.085	0.386	0.529	0.254
2	0.275	0.192	0.359	0.449	0.247
2	0.3	0.303	0.328	0.370	0.235
2	0.325	0.403	0.297	0.300	0.222
3	0.2	0	0.535	0.465	0.419
3	0.225	0.003	0.534	0.463	0.419
3	0.25	0.030	0.528	0.442	0.418
3	0.275	0.081	0.515	0.404	0.414
3	0.3	0.143	0.498	0.360	0.408
3	0.325	0.208	0.478	0.315	0.399

SMM

$$\hat{c}, \hat{\mu}, \hat{\Sigma} = \hat{\theta} = \arg\min_{\theta} \left[\sum_{t=1}^{T} \left(m(o_t) - \frac{1}{S} \sum_{s=1}^{S} m(o(u_t^s, \theta)) \right) \right]' W_T^{-1} \left[\sum_{t=1}^{T} \left(m(o_t) - \frac{1}{S} \sum_{s=1}^{S} m(o(u_t^s, \theta)) \right) \right] \tag{2}$$

Back

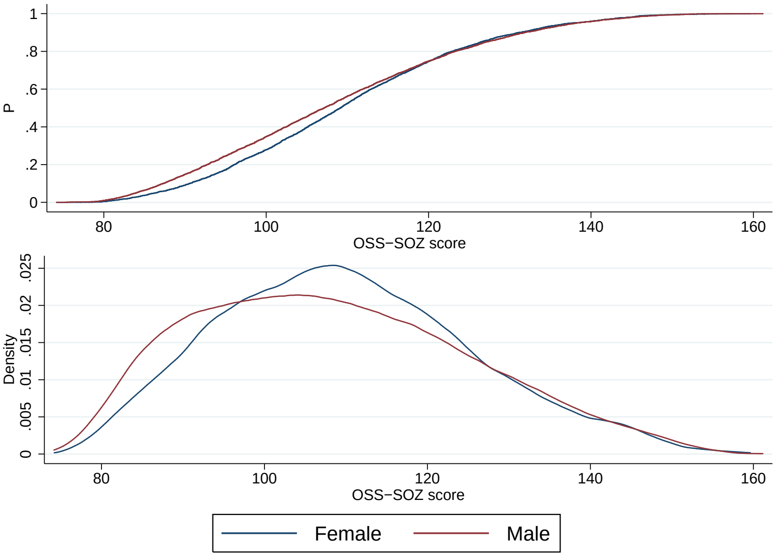
Test Weights Back

	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

- For social science track students, the math and science sections have very little weight on the total score
    - These students are told in the exam to spend more time on social science and Turkish than on science and math

Score Distribution

Back



Ability Distributions Back