Hit or Miss? Test Taking Behavior in Multiple Choice Exams

Ş. Pelin Akyol ¹  James Key ²  Kala Krishna ³

¹Bilkent University
²University of Western Australia
³Pennsylvania 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: Women skip more often

- Reduced form: Ben-Shakhar and Sinai (1991)
ÖSS Exam - held annually
  - Paper based multiple choice exam
  - Most important determinant of university admission

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, 1st time takers (8917 students)
- Two sections of interest: social science and Turkish
There is a gender gap in 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 $n \in \{1, \ldots, 5\}$, the student draws a signal $x_n$.
- The correct answer draws from a Pareto distribution with shape parameter $\alpha$ and scale parameter $A$.
- Incorrect answers draw from a Pareto distribution with shape parameter $\beta$ 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 \( \beta / \alpha \)
Without loss of generality, $A = 1$, $\alpha = 1$, so that $\beta$ is ability.

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

$$x_m^{\beta - \alpha} \over x_1^{\beta - \alpha} + x_2^{\beta - \alpha} + x_3^{\beta - \alpha} + x_4^{\beta - \alpha} + x_5^{\beta - \alpha}$$

(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 $m$.
For a student with ability $\beta_K$ and cutoff $c$ for exam section $K$, there are three possible outcomes for each question:

1. Answer correctly
2. Answer incorrectly
3. Skip the question

Given $(\beta_K, c)$, we can find the possibilities of each of the three outcomes.
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)
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
Social Science Score

Raw Social Science Score

Male

Female

Akyol, Key and Krishna

ÖSS Exams
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.
Means of ability: means of section scores
Similarly with variance/covariance of ability
Identification of risk aversion is less obvious
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
- $\geq 120$ 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: $\beta_T$ and $\beta_{SS}$ 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 to Akyol, Key and Krishna ÖSS Exams]
<table>
<thead>
<tr>
<th>Range</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,90)</td>
<td>0.2429</td>
<td>0.2100</td>
</tr>
<tr>
<td></td>
<td>(0.0269)</td>
<td>(0.0026)</td>
</tr>
<tr>
<td>[90,100)</td>
<td>0.2322</td>
<td>0.2272</td>
</tr>
<tr>
<td></td>
<td>(0.0023)</td>
<td>(0.0019)</td>
</tr>
<tr>
<td>[100,110)</td>
<td>0.2396</td>
<td>0.2364</td>
</tr>
<tr>
<td></td>
<td>(0.0009)</td>
<td>(0.0010)</td>
</tr>
<tr>
<td>[110,120)</td>
<td>0.2546</td>
<td>0.2480</td>
</tr>
<tr>
<td></td>
<td>(0.0017)</td>
<td>(0.0016)</td>
</tr>
<tr>
<td>[120,130)</td>
<td>0.2612</td>
<td>0.2594</td>
</tr>
<tr>
<td></td>
<td>(0.0037)</td>
<td>(0.0043)</td>
</tr>
<tr>
<td>[130,140)</td>
<td>0.2763</td>
<td>0.2633</td>
</tr>
<tr>
<td></td>
<td>(0.0062)</td>
<td>(0.0036)</td>
</tr>
<tr>
<td>[140,∞)</td>
<td>0.2796</td>
<td>0.2697</td>
</tr>
<tr>
<td></td>
<td>(0.0175)</td>
<td>(0.0076)</td>
</tr>
</tbody>
</table>

Standard errors are reported in parentheses.
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
The estimation procedure also finds the distribution of ability for each group. 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.
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
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

Fraction of Male Students

exam score percentile

Baseline  No penalty  Penalty:0.5  Penalty:1

Akyol, Key and Krishna  ÖSS Exams
Counterfactual Results

Average log Social Science Ability

exam score percentile

Baseline  No penalty  Penalty:0.5  Penalty:1

Akyol, Key and Krishna

ÖSS Exams
Counterfactual Results

Average log Turkish Ability

exam score percentile

Baseline  No penalty
Penalty:0.5  Penalty:1

Akyol, Key and Krishna

ÖSS Exams
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
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

$$k_{m,n} = \frac{q_n + s_m}{q_n}$$
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.
<table>
<thead>
<tr>
<th>β</th>
<th>Cutoff</th>
<th>Prob(S)</th>
<th>Prob(C)</th>
<th>Prob(I)</th>
<th>PPQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.2</td>
<td>0</td>
<td>0.405</td>
<td>0.595</td>
<td>0.257</td>
</tr>
<tr>
<td>2</td>
<td>0.225</td>
<td>0.012</td>
<td>0.403</td>
<td>0.585</td>
<td>0.257</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>0.085</td>
<td>0.386</td>
<td>0.529</td>
<td>0.254</td>
</tr>
<tr>
<td>2</td>
<td>0.275</td>
<td>0.192</td>
<td>0.359</td>
<td>0.449</td>
<td>0.247</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>0.303</td>
<td>0.328</td>
<td>0.370</td>
<td>0.235</td>
</tr>
<tr>
<td>2</td>
<td>0.325</td>
<td>0.403</td>
<td>0.297</td>
<td>0.300</td>
<td>0.222</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0</td>
<td>0.535</td>
<td>0.465</td>
<td>0.419</td>
</tr>
<tr>
<td>3</td>
<td>0.225</td>
<td>0.003</td>
<td>0.534</td>
<td>0.463</td>
<td>0.419</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>0.030</td>
<td>0.528</td>
<td>0.442</td>
<td>0.418</td>
</tr>
<tr>
<td>3</td>
<td>0.275</td>
<td>0.081</td>
<td>0.515</td>
<td>0.404</td>
<td>0.414</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
<td>0.143</td>
<td>0.498</td>
<td>0.360</td>
<td>0.408</td>
</tr>
<tr>
<td>3</td>
<td>0.325</td>
<td>0.208</td>
<td>0.478</td>
<td>0.315</td>
<td>0.399</td>
</tr>
</tbody>
</table>
\[ \hat{c}, \hat{\mu}, \hat{\Sigma} = \hat{\theta} = \arg \min_\theta \left[ \sum_{t=1}^{T} \left( m(o_t) - \frac{1}{S} \sum_{s=1}^{S} m(o(u_t^s, \theta)) \right) \right]' \]

\[ W_T^{-1} \left[ \sum_{t=1}^{T} \left( m(o_t) - \frac{1}{S} \sum_{s=1}^{S} m(o(u_t^s, \theta)) \right) \right] \] (2)
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.
Ability Distributions

1\textsuperscript{st} time takers – Turkish

1\textsuperscript{st} time takers – Social Science

Female  Male

Akyol, Key and Krishna

ÖSS Exams