College entry exams: A dynamic discrete choice model

José Raimundo Carvalho, Federal University of Ceara, Brazil
Thierry Magnac, Toulouse School of Economics, France

May 2009
Economic questions

Formal *entry exams* at Federal Universities in Brazil (age 17-18)

Simple mechanism for the choice of a single major ex-ante and a 2-stage exam. The first stage is common to all majors, the second stage is more specialized.

The *matching* literature, relative to the *college admission* problem, points out that this mechanism does not have valuable properties:

- it is not *fair*, or *stable*, since at least one pair of (major, student) might be made better off by changing places
- it is *strategic*: students disguise their true preferences at the moment of choice trading them off with the success probability

**Argument:** The selection list is too short. Only one major can be chosen ex-ante while the selection list is large when using a mechanism with nice properties such as Gale Shapley deferred acceptance mechanism (medical schools in the US).

**Question:** What is the welfare loss of this mechanism? What are the distributive effects between heterogeneous students?
Literature(s)


Dynamic models: the question of agent’s information/ expectations (Manski, 1992).


This paper: a dynamic mixed discrete and continuous model to allow the quantification of losses due to suboptimality of the scheme under full information assumptions (and preliminary parametric assumptions!). Full observation of performances (grades at the 2 stages).
Outline of the Talk

- The game, its timing and what we observe
- Characterization of the optimal solution
- Identification of success probabilities & preferences
- Estimation: a parametric setting
- Counterfactuals
The entry exam: timing

Several majors: $d = 1, .., D$ among which medicine, engineering, Portuguese, ... at the Federal University of Cearà (the best in this state), northeast Brazil.

- **Step 0**: A national exam, about a year before, delivering a grade which is going to matter in the further two-stages ($m_0$): $m_0$ is observed.

- **Step 1.1**: Choice of major $d$ and choice of effort $y$: $d$ is observed and effort $y$ is unobserved.

- **Step 1.2**: First stage common exam delivering a grade $m_1$. Students are ranked within a major according to a weighted average of initial grade $m_0$ and first stage grade. The number of students who pass is equal (roughly) to four times the number of positions $n_d$ in the major.

- **Step 2**: Second stage exam specific to the major delivering a grade $m_2$. Students are ranked according to a weighted sum of $m_0$, $m_1$ and $m_2$ and the first $n_d$ students pass.
Applications, positions and success probabilities

<table>
<thead>
<tr>
<th>Groups of majors</th>
<th>Applications</th>
<th>% Pass 1st stage</th>
<th>% Pass 2nd stage</th>
<th>Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accountancy</td>
<td>1,374</td>
<td>40%</td>
<td>13%</td>
<td>185</td>
</tr>
<tr>
<td>Administration</td>
<td>2,474</td>
<td>29%</td>
<td>8%</td>
<td>200</td>
</tr>
<tr>
<td>Agrosciences</td>
<td>2,996</td>
<td>41%</td>
<td>12%</td>
<td>300</td>
</tr>
<tr>
<td>Economics</td>
<td>1,516</td>
<td>37%</td>
<td>11%</td>
<td>160</td>
</tr>
<tr>
<td>Engineering</td>
<td>2,648</td>
<td>40%</td>
<td>14%</td>
<td>360</td>
</tr>
<tr>
<td>Humanities</td>
<td>4,897</td>
<td>17%</td>
<td>9%</td>
<td>430</td>
</tr>
<tr>
<td>Law</td>
<td>3,625</td>
<td>20%</td>
<td>5%</td>
<td>180</td>
</tr>
<tr>
<td>Mathematics</td>
<td>2,425</td>
<td>37%</td>
<td>11%</td>
<td>269</td>
</tr>
<tr>
<td>Medicine</td>
<td>4,024</td>
<td>23%</td>
<td>6%</td>
<td>230</td>
</tr>
<tr>
<td>Other</td>
<td>2,778</td>
<td>21%</td>
<td>6%</td>
<td>165</td>
</tr>
<tr>
<td>Pharmacy, Dentist &amp; Other</td>
<td>5,312</td>
<td>24%</td>
<td>6%</td>
<td>320</td>
</tr>
<tr>
<td>Physics &amp; Chemistry</td>
<td>1,734</td>
<td>58%</td>
<td>20%</td>
<td>349</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>5,574</td>
<td>26%</td>
<td>7%</td>
<td>385</td>
</tr>
</tbody>
</table>

Source: UFC Vestibular 2004
Observations and measurements

The grades obtained at the different stages are functions of unobserved talent \( \varepsilon \) and unobserved effort \( y \) such that:

\[
\begin{align*}
  m_0 &= \varepsilon + \eta_0 \\
  m_1 &= \varepsilon + y + \eta_1 \\
  m_2 &= \varepsilon + y + \eta_2
\end{align*}
\]

where \( \eta_i \) is supposed to be an independent "measurement" error on grades.

Thresholds:

Stage 1 success in major \( d \) : \( m_1 > t_1(d, m_0) \),
Stage 2 success in major \( d \) : \( m_2 > t_2(d, m_0, m_1) \).
<table>
<thead>
<tr>
<th>Subgroup</th>
<th>10th percentile</th>
<th>Min</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>First stage</td>
<td>Pass</td>
<td>All</td>
</tr>
<tr>
<td>Agrosciences</td>
<td>71.1</td>
<td>91.2</td>
<td>100.1</td>
<td>106.9</td>
</tr>
<tr>
<td>Other</td>
<td>66.1</td>
<td>102.1</td>
<td>104.8</td>
<td>102.0</td>
</tr>
<tr>
<td>Physics &amp; Chemistry</td>
<td>76.8</td>
<td>33.0</td>
<td>50.0</td>
<td>115.2</td>
</tr>
<tr>
<td>Humanities</td>
<td>67.9</td>
<td>96.3</td>
<td>99.2</td>
<td>104.2</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>68.9</td>
<td>101.0</td>
<td>102.0</td>
<td>109.4</td>
</tr>
<tr>
<td>Accountancy</td>
<td>80.5</td>
<td>120.5</td>
<td>122.9</td>
<td>120.3</td>
</tr>
<tr>
<td>Economics</td>
<td>71.8</td>
<td>113.3</td>
<td>121.1</td>
<td>110.9</td>
</tr>
<tr>
<td>Administration</td>
<td>68.6</td>
<td>108.5</td>
<td>121.0</td>
<td>108.7</td>
</tr>
<tr>
<td>Mathematics</td>
<td>75.8</td>
<td>70.3</td>
<td>73.0</td>
<td>122.1</td>
</tr>
<tr>
<td>Engineering</td>
<td>84.3</td>
<td>130.2</td>
<td>137.6</td>
<td>133.7</td>
</tr>
<tr>
<td>Pharmacy, Dentist &amp; Other</td>
<td>73.8</td>
<td>142.0</td>
<td>143.8</td>
<td>123.0</td>
</tr>
<tr>
<td>Law</td>
<td>77.4</td>
<td>165.5</td>
<td>168.0</td>
<td>139.5</td>
</tr>
<tr>
<td>Medicine</td>
<td>89.6</td>
<td>182.0</td>
<td>188.9</td>
<td>169.0</td>
</tr>
</tbody>
</table>

Table 5. Summary statistics of first stage grades in the samples of (1) all, (2) pass after first stage (3) definite pass after second stage
(The order of subgroups is given by the median of the first stage grades in the pass sample, column 5)
Solving the dynamic model backward

**Step 2:** history $h_1$ of past grades and other relevant information:

$$V_2(h_1) = \Pr_{\eta_2}\{m_2 > t_2(d, m_0, m_1) \mid h_1\}.u_d + \Pr_{\eta_2}\{m_2 < t_2(d, m_0, m_1) \mid h_1\}.v_f.$$  

where:

- $u_d$ is the *utility* of choosing major $d$ (including the discounted value of future wages)
- $v_f$ is the *outside option* utility.

**Step 1:** History $h_0 = (m_0, \varepsilon)$:

$$V_1(y, d; h_0) = -c \left( y \right) + E_{\eta_1} [1\{m_1 > t_1(d, m_0)\}.V_2(h_1)] + \Pr_{\eta_1}\{m_1 < t_1(d, m_0)\}.v_f.$$  

Overall probability of success in major $d$ as:

$$P_d(y; h_0) = \Pr(\eta_1 > t_1(d, m_0) - \varepsilon - y, \eta_2 > t_2(d, m_0, \varepsilon + y + \eta_1) - \varepsilon - y).$$

**Summary:**

$$V_1(y, d; \varepsilon, m_0) = -c.y + P_d(y; \varepsilon, m_0).u_d + (1 - P_d(y; \varepsilon, m_0)).v_f.$$
Step 0:

\[ V_0(\varepsilon, m_0) = \max_{y,d} V_1(y, d; \varepsilon, m_0). \]

Normalizations of:

\[ V_1(y, d; \varepsilon, m_0) = -c.y + P_d(y; \varepsilon, m_0).u_d + (1 - P_d(y; \varepsilon, m_0)).v_f. \]

- Outside option: \( v_f = 0 \)
- Cost of effort, \( c = 1 \)

Optimality condition if \( y_d \) is the optimal effort:

\[ v_d = -y_d + P_d(y_d; \varepsilon, m_0).u_d. \]
Optimal solution

Optimality condition

\[ v_d = -y_d + P_d(y_d; \varepsilon, m_0) \cdot u_d. \]

Probability of success:

\[ \frac{d}{dy} P_d(y; \varepsilon, m_0) > 0, \quad P_d(y; \varepsilon, m_0) \text{ varies from 0 to 1 when } y \text{ varies from } -\infty \text{ to } +\infty \]

First order condition in effort (interior solution):

\[ 1 = P'_d(y; \varepsilon, m_0) u_d. \]
College entry exams: A dynamic discrete choice model

Optimal solutions

Corner solution
Optimal interior solution: \( u.p'(y) = 1 \)

Effort, \( y \)

Success density function: \( p'(y) \)

Optimal interior solution: \( u.p'(y) = 1 \)

1/\( u \)

Corner solution

Econometric of Demand, Brown University, May 4th 2009
Characterization

Denote $\tilde{y}_d$ the interior solution. The optimal effort $y_d$ is given by:

$$
\begin{align*}
    y_d &= \tilde{y}_d & \text{if } -\tilde{y}_d + P_d(\tilde{y}_d; \varepsilon, m_0) \cdot u_d > P_d(0; \varepsilon, m_0) \cdot u_d, u_d > 0, \ max_{y \geq 0} P'_d(y; \varepsilon, m_0) \geq 1/u_d \\
    y_d &= 0 & \text{if not}
\end{align*}
$$

and is a non-decreasing function of $u_d$.

Condition for non zero effort: $u_d > \tilde{u}_d$:

$$
\tilde{u}_d = \frac{\tilde{y}_d}{P_d(\tilde{y}_d; \varepsilon, m_0) - P_d(0; \varepsilon, m_0)} > 0.
$$

**Remark 1:** the optimal effort function is discontinuous in $u_d$ at the threshold.

**Remark 2:** The value $v_d$ is continuous and increasing in $u_d$. 
The econometric model: Identification

Grade, talent and effort distributions:

\[
\begin{align*}
    m_0 &= \varepsilon + \eta_0 & \text{observed in the population} \\
    m_1 &= \varepsilon + y + \eta_1 & \text{observed in the population} \\
    m_2 &= \varepsilon + y + \eta_2 & \text{observed in the selected population, } m_1 > t_1(d, m_0)
\end{align*}
\]

Non parametric identification of distributions of \(\varepsilon + y, \eta_1\) and \(\eta_2\) using either:

- Identification at infinity and Kotlarski’s argument.
- Finite mixture components

Non parametric identification of joint distribution of \(\varepsilon, \varepsilon + y\):

\[
E \exp(i\theta_0 m_0 + i\theta_1 m_1) = E \exp(i\theta_0 \varepsilon + i\theta_1 (\varepsilon + y)) \exp(i\theta_0 \eta_0) \exp(i\theta_1 \eta_1) = \phi_{\varepsilon, \varepsilon+y}(\theta_0, \theta_1) \phi_0(\theta_0) \phi_1(\theta_1)
\]

which identifies the characteristic function \(\phi_{\varepsilon, \varepsilon+y}(\theta_0, \theta_1)\) (under non-vanishing conditions).
Normalization of preferences

Assume that $d$ is a binary variable for a two-state model where 0 is the alternative to 1.

$$\Pr(d = 1) = \Pr(v_1 = \Psi_1(u_1) > v_0 = \Psi_0(u_0)) = \Pr(u_1 > \Psi_1^{-1} \circ \Psi_0(u_0))$$
$$= \Pr(T(u_1) > T \circ \Psi_1^{-1} \circ \Psi_0(u_0)) = \Pr(\Psi_1(T(u_1)) > \Psi_1 \circ T \circ \Psi_1^{-1} \circ \Psi_0(u_0))$$
$$= \Pr(\Psi_1(w_1) > \Psi_0(w_0))$$

where $T()$ is a monotonic function.

**Non parametric restrictions**: Median restrictions since functions $\Psi$'s are monotonic.
Parametric estimation

For simplicity, talent is supposed to be equal to the initial grade: \( m_0 = \varepsilon \).

**Normality** assumptions:
- *Measurement errors* on grades are normally distributed: corrections of truncations using the usual truncated normal equations.
- *Preferences* \( u_d \)

**Approximations:**
- *Simulations*: The GHK simulator performs well in this non-linear setting.
- *Structural model*: Approximations through a finite grid.
Data

Using 3 majors in Medicine which location differs: Barbalha (40 positions), Sobral (40) and Fortaleza (160).

3606 observations

<table>
<thead>
<tr>
<th>Major</th>
<th>Median(m0)</th>
<th>Median(m1)</th>
<th>Median(m2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barbalha</td>
<td>48</td>
<td>152.61905</td>
<td>242.35794</td>
</tr>
<tr>
<td>Sobral</td>
<td>52</td>
<td>171.7619</td>
<td>246.49241</td>
</tr>
<tr>
<td>Fortaleza</td>
<td>52</td>
<td>172.95238</td>
<td>245.09035</td>
</tr>
</tbody>
</table>

Results: The reference alternative is the last one. Preferences are very dispersed for the first major, very concentrated for the second one.

Econometric of Demand, Brown University, May 4th 2009
Counterfactuals

- Students could choose more than one major before the first stage. This is "in the direction" of the Gale-Shapley deferred acceptance algorithm although the list of choices is (potentially) infinite.

**Enlarging choices**: the choice set is composed by pairs \((d^*, d^{**})\) instead of a single choice \(d^*\) (although \((d^*, \emptyset)\) is allowed). The timing of the game remains the same, choices and investment being made before the first stage.

After the first stage, there are now three possibilities:

- \(m_1 > t_1^A(d^*, m_0)\) : the student qualifies for the second stage of major \(d^*\).
- \(m_1 < t_1^A(d^*, m_0)\) and \(m_1 > t_1^A(d^{**}, m_0)\) : the student qualifies for the second stage of major \(d^{**}\).
- \(m_1 < t_1^A(d^{**}, m_0)\) : the student fails.

**Remark**: The order \((d^*, d^{**})\) is compulsory. No possible changes after observing grade \(m_1\).
Construction

*Augmented list* of choices: 9 possible choices instead of three.

*Prediction*:
- choice probabilities
- probabilities of choice and success at stage 1
- probabilities of choice and success at stage 2

Computation of a *Nash equilibrium* through the thresholds $t_1^A(d^*, m_0)$:

The sum of probabilities of choice and success at stages 1 and 2 should be equal to the number of positions offered at each stage.
Results

Thresholds:

<table>
<thead>
<tr>
<th></th>
<th>Barbalha</th>
<th>Sobral</th>
<th>Fortaleza</th>
</tr>
</thead>
<tbody>
<tr>
<td>First stage</td>
<td>Current</td>
<td>182.0476</td>
<td>185.0476</td>
</tr>
<tr>
<td></td>
<td>Counterfactual</td>
<td>202.4889</td>
<td>186.2268</td>
</tr>
<tr>
<td>Second stage</td>
<td>Current</td>
<td>236.0816</td>
<td>237.8046</td>
</tr>
<tr>
<td></td>
<td>Counterfactual</td>
<td>264.6613</td>
<td>218.9603</td>
</tr>
</tbody>
</table>

Counterfactual and current (utilitarian) welfare:

- \( v_0 = 1353.776 \)
- \( v_1 = 1406.750 \)
Figure 9: Current and counterfactual values

Expected values

Density

Current values

Counterfactual values
Figure 10: Change from current to counterfactual
Future work

- Semiparametric estimation

- Other counterfactuals: Students could choose between the two stages and not before the first stage i.e. they condition their decisions on their first-stage grade.

- Less sophisticated expectation formation