Biases and (Dis)agreement in Fellowship Selection Process

Insights & Strategies

2018 Wharton People Analytics Conference

Park Sinchaisri, Wharton Operations Research
Titipat Achakulvisut, Penn Neuroscience/Bioengineering
Review processes are prone to *biases*

**Domains:**
- Employment interviews
- Peer reviews in academia
Review processes are prone to **biases**

**Domains:**
- Employment interviews/Peer reviews in academia

**Existing biases of applicant's characteristics**
- Race, ethnic names, accents, appearances
- Authors from further away in networks
Review processes are prone to biases

**Domains:**
- Employment interviews/Peer reviews in academia

**Existing biases of applicant’s characteristics**
- Race, ethnic names, accents, appearances
- Authors from further away in networks

**Reviewer’s demographics**

**Nature of application**

**Multiple evaluations/rankings**
## Research Questions

| How do applicants’/reviewers’ demographics and position’s characteristics affect the evaluation? | What may influence (dis)agreements among human reviewers? Can ML help? |
Research Questions

How do applicants’/reviewers’ demographics and position’s characteristics affect the evaluation?

What may influence (dis)agreements among human reviewers? Can ML help?

Agenda


- Female with exp. citizenship bias
- Reviewer skill/happiness
- Normalized scores
- Optimal assignment
- Machine learning
Fellowship Review Process

1. Reading
2. Reading
3. Reading
4. Interview
5. Interview

2-6 reviewers
1 reviewer

5778 applicants
139 positions
Fellowship Review Process

1. Reading
2. Reading
3. Reading
4. Interview
5. Interview

2-6 reviewers
1 reviewer

5778 applicants
139 positions
244 reviewers
1204 semifinalists
Data Pre-Processing

Text Preprocessing
Features generation

Normalized Score within reviewer

\[ s_{\text{norm}} = \frac{s_i - s_{\text{min}}}{s_{\text{max}} - s_{\text{min}}} \]

OLS model
Negative Binomial model
Beta model
Probit/Logit model
## Roles of Applicants’ Characteristics

<table>
<thead>
<tr>
<th>% selected</th>
<th>Whites</th>
<th>60.31%</th>
<th>Blacks</th>
<th>51.27%</th>
<th>Hispanics</th>
<th>56.58%</th>
<th>Asians</th>
<th>54.79%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25.97%</td>
<td>50.20%</td>
<td>3.46%</td>
<td>10.81%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Roles of Applicants’ Characteristics

<table>
<thead>
<tr>
<th>% selected</th>
<th>Accept rate corrected for competition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Whites</td>
</tr>
<tr>
<td>25.97%</td>
<td>60.31%</td>
</tr>
<tr>
<td>50.20%</td>
<td>39.15%</td>
</tr>
<tr>
<td>3.46%</td>
<td></td>
</tr>
<tr>
<td>10.81%</td>
<td></td>
</tr>
</tbody>
</table>

Race of applicants do not significantly affect their scores.
# Roles of Applicants’ Characteristics

<table>
<thead>
<tr>
<th>% selected</th>
<th>Accept rate corrected for competition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Whites</td>
<td>25.97%</td>
</tr>
<tr>
<td></td>
<td>60.31%</td>
</tr>
<tr>
<td>Blacks</td>
<td>50.20%</td>
</tr>
<tr>
<td></td>
<td>53.94%</td>
</tr>
<tr>
<td>Hispanics</td>
<td>3.46%</td>
</tr>
<tr>
<td></td>
<td>36.61%</td>
</tr>
<tr>
<td>Asians</td>
<td>10.81%</td>
</tr>
<tr>
<td></td>
<td>25.04%</td>
</tr>
</tbody>
</table>

Race of applicants do not significantly affect their scores

More favorable:
- Female
- Eligible citizenship
- Work experience in public health
- Previously applied
Roles of Reviewer’s Characteristics

- Citizenship
- Gender
- Skillset
- Happiness

Fixed effects regression models
Roles of Reviewer’s Characteristics

- Citizenship
- Gender
- Skillset
- Happiness

Fixed effects regression models

Citizenship

62.7% matched
Score: +3.5%

Rank applicants of the same citizenship higher

Citizenship Bias
Roles of Reviewer’s Characteristics

- Citizenship
- Gender
- Skillset
- Happiness

Fixed effects regression models

Citizenship

<table>
<thead>
<tr>
<th>Reviewer’s</th>
<th>Applicant’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citizenship</td>
<td>Citizenship</td>
</tr>
<tr>
<td>Gender</td>
<td>Gender</td>
</tr>
<tr>
<td>Skillset</td>
<td>Skillset</td>
</tr>
<tr>
<td>Happiness</td>
<td>Happiness</td>
</tr>
</tbody>
</table>

62.7% matched
Score: +3.5%

54.6% matched
Rank: -1.5%

Rank applicants of the same citizenship higher
Citizenship Bias

Harsher in ranking applicants + selecting semifinalist when reviewing for home
Roles of **Reviewer’s**

**Gender**

- **Reviewer’s**
  - 26.9% male
  - Score: -7%

Male reviewers assign lower scores but select more semifinalists
Male reviewers assign lower scores but select more semifinalists.
## Roles of Reviewer’s

<table>
<thead>
<tr>
<th>Gender</th>
<th>Skillsets</th>
<th>Happiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male reviewers assign lower scores but select more semifinalists</td>
<td>Skilled reviewers are stricter</td>
<td>Disappointed reviewers tend to be less consistent/certain</td>
</tr>
<tr>
<td>26.9% male</td>
<td>55% matched</td>
<td>11 disappointed</td>
</tr>
<tr>
<td>Score: -7%</td>
<td>Chance: -11%</td>
<td>SD: +5.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(Dis)agreement among Reviewers

**Metrics:** mean + |diff| of ranks/scores, # overlap semifinalists, Spearman’s rank correlation

**Tools:** t and Wilcoxon rank sum tests to compare distributions, regressions of metrics

---

R1 vs R2

- Gender
- Citizenship
- Placement
- Skillset
- Status
(Dis)agreement among Reviewers

**Metrics:** mean + |diff| of ranks/scores, # overlap semifinalists, Spearman’s rank correlation

**Tools:** t and Wilcoxon rank sum tests to compare distributions, regressions of metrics

R1 vs R2

- Gender
- Citizenship
- Placement
- Skillset
- Status

**Same...**

- Larger rank correlation
- Smaller score differences

**Agree** on same semifinalists
(Dis)agreement among Reviewers

Metrics: mean + |diff| of ranks/scores, # overlap semifinalists, Spearman’s rank correlation
Tools: t and Wilcoxon rank sum tests to compare distributions, regressions of metrics

Gender  Citizenship  Placement  Skillset  Status

R1 vs R2

Same...

Larger rank correlation  Smaller score differences

Agree on same semifinalists  No significant differences
(Dis)agreement among Reviewers

**Metrics:** mean + |diff| of ranks/scores, # overlap semifinalists, Spearman’s rank correlation

**Tools:** t and Wilcoxon rank sum tests to compare distributions, regressions of metrics

---

R1 vs R2

<table>
<thead>
<tr>
<th>Gender</th>
<th>Citizenship</th>
<th>Placement</th>
<th>Skillset</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Larger rank correlation</td>
<td>Smaller score differences</td>
<td></td>
<td></td>
<td>Disagree more (trend)</td>
</tr>
</tbody>
</table>

**Same...**

- Agree on same semifinalists
- No significant differences
Optimal Reviewer Assignment

Use Normalized Scores

Applicant’s ≠ Reviewer’s ≠ Reviewer’s

Reviewer’s ≠ Reviewer’s

1+ review for home 1+ matched skill Assign as requested
Optimal Reviewer Assignment

Use Normalized Scores

Weights determined by other matched reviewers

Applicant’s ≠ Reviewer’s ≠ Reviewer’s ≠ Reviewer’s

1+ review for home 1+ matched skill Assign as requested
Round 3 Selection

Round 2

2 Suggested

831

1 Suggested

1231
Round 3 Selection

Round 2
- 2 Suggested: 831
- 1 Suggested: 1231

Round 3
- 2 Suggested: 682, 149
- 1 Suggested: 483, 784

82.1% Selected
39.2% Selected

Selection bias?
Round 3 Selection

<table>
<thead>
<tr>
<th>Round 2</th>
<th>2 Suggested</th>
<th>1 Suggested</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>831</td>
<td>1231</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Round 3</th>
<th>82.1% Selected</th>
<th>39.2% Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>682</td>
<td>483</td>
</tr>
<tr>
<td></td>
<td>149</td>
<td>784</td>
</tr>
</tbody>
</table>

Selection bias? No selection bias
Maximum of normalized scores predicts selection
Round 3 Selection

Selects applicants from two recommenders then by ranking of normalized scores

Maximum of normalized scores predicts selection

Selection bias? No selection bias

82.1% Selected

39.2% Selected
Data-Driven Selection in Round 3

Score Ranking
2 reviewers and normalized score

Random Forest Ensemble
Learn selection probability from 30% of data

Measure overlap between ranking model and selection in round 3
Data-Driven Selection in Round 3

Score Ranking
2 reviewers and normalized score

Random Forest Ensemble
Learn selection probability from 30% of data

R1
- 29
- 27
- 27
- 25
- 24

R2
- 29
- 27
- 27
- 25
- 24

Random selection
39.7%

Maximum Average Score
70.3%

73.4% 77.3%
Discussion and Future Research

$0/1 > [0,1)$

Features improvement

Round 3 quality checking

Review details
Conclusion

Insights

Proposed Strategies

Round 2

Round 3

Applicants
Reviewers
Interactions

ML for ranking can replace human reviewer

No selection bias

Use Normalized Score

Applicant-Reviewer
Reviewer-Reviewer

R2
2
1

R3

Conclusion