Improving the Peer-Review Process with Model-Based Estimates of Inter-Rater Reliability and Detection of Rating Bias: From Teacher Selection to Journal Submissions and Grant Applications

Patricia Martinkova\textsuperscript{1,2} Dan Goldhaber\textsuperscript{3} Elena Erosheva\textsuperscript{3} \& Carole Lee\textsuperscript{3}

\textsuperscript{1}Institute of Computer Science, Czech Academy of Sciences
\textsuperscript{2}College of Education, Charles University, Prague
\textsuperscript{3}University of Washington, Seattle, USA

PEERE International Conference, 7-9 March 2018, Rome
Acknowledgements

- Fulbright-Masaryk Fellowship (University of Washington)
- IES grants R305C130030, R305A060018
- Czech Science Foundation grant GJ15-15856Y
- PEERE, TD Cost Action TD1306
  (STSM to University of Washington, Nov 2017)

- Data processing support by Malcolm Wolf and Adela Drabinova
Outline

1. Introduction: Peer-Review in Teacher Hiring
2. Rating bias
3. Model Based Inter-Rater Reliability
4. Implications for Other Types of Peer-Review
5. Conclusion
Introduction: Teacher Selection Process

Applicants to classroom job openings in Spokane Public Schools during years (2008/09 - 2012/13)

- Responded to job posting (2,669)
- Screened with 21-pt rubric (2,433)
- Screened with 54-pt rubric (1,177)
- Advanced to interview (709)
- Hired by SPS (374)

=Roughly 100 applicants
Introduction: Ratings as Source of Error

54-Pt Screening Rubric:

- Certificate and Education
- Training
- Experience
- Classroom Management
- Flexibility
- Instructional Skills
- Interpersonal Skills
- Cultural Competency
- Preferred Qualifications
- (Quality of Recom. Letters)
Data structure

- 3474 filled forms
- 1090 applicants
- 137 raters
- 54 job locations (schools)

Applicant status

- Internal applicant (2322 forms)
  - was previously employed as a teacher in the district or
  - had completed their student teaching in the district
- External applicant (1152 forms)
- 51 applicants external for some and internal for other ratings
Data structure

- 3474 filled forms
- 1090 applicants
- 137 raters
- 54 job locations (schools)

Applicant status

- Internal applicant (2322 forms)
  - was previously employed as a teacher in the district or
  - had completed their student teaching in the district
- External applicant (1152 forms)
- 51 applicants external for some and internal for other ratings
Ratings of a single applicant

Mean and range of ratings

Applicants ranked by averaged total score
Ratings of two applicants

Mean and range of ratings

Applicants ranked by averaged total score
1. Introduction  
2. Rating bias  
3. Model-based Inter-Rater Reliability  
4. Implications for peer-review  
5. Conclusion

Ratings of all applicants

Mean and range of ratings

Applicants ranked by averaged total score
Ratings of all applicants by Internal/External Status

Mean and range of ratings

Applicants ranked by averaged total score

- Ext
- Int
- Int/Ext
Rating distributions

- About 3pt higher ratings for internal applicants
Rating distributions

- Higher ratings for internal applicants across all subcomponents
- More skewed distribution for internal applicants
Testing for bias with respect to applicant status

Model controlling for quality measures, accounting for data structure

\[ Y_{ijk} = \mu + \omega_i \beta_0 + X_i \beta + A_i + B_j + S_k + A S_{ik} + e_{ijk} \]

- Applicant internal/external status \( \omega_i \)
- Applicant quality measures \( X_i \) (e.g. experience, licensure test scores, teacher value added estimates)
- Applicant latent quality \( A_i \sim N(0, \sigma_A^2) \)
- Rater severity/leniency \( B_j \sim N(0, \sigma_B^2) \)
- School severity/leniency \( S_k \sim N(0, \sigma_S^2) \)
- Applicant-school matching effect (interaction) \( A S_{ik} \sim N(0, \sigma_{AS}^2) \)
Testing for bias with respect to applicant status

Model controlling for quality measures, accounting for data structure

\[ Y_{ijk} = \mu + \omega_i \beta_0 + X_i \beta + A_i + B_j + S_k + AS_{ik} + e_{ijk} \]

- Applicant internal/external status \( \omega_i \)
- Applicant quality measures \( X_i \) (e.g. experience, licensure test scores, teacher value added estimates)
- Applicant latent quality \( A_i \sim N(0, \sigma^2_A) \)
- Rater severity/leniency \( B_j \sim N(0, \sigma^2_B) \)
- School severity/leniency \( S_k \sim N(0, \sigma^2_S) \)
- Applicant-school matching effect (interaction) \( AS_{ik} \sim N(0, \sigma^2_{AS}) \)
Testing for bias with respect to applicant status

Model controlling for quality measures, accounting for data structure

\[ Y_{ijk} = \mu + \omega_i \beta_0 + X_i \beta + A_i + B_j + S_k + A_S_{ik} + e_{ijk} \]

- Applicant internal/external status \( \omega_i \)
- Applicant quality measures \( X_i \)
  (e.g. experience, licensure test scores, teacher value added estimates)
- Applicant latent quality \( A_i \sim N(0, \sigma^2_A) \)
- Rater severity/leniency \( B_j \sim N(0, \sigma^2_B) \)
- School severity/leniency \( S_k \sim N(0, \sigma^2_S) \)
- Applicant-school matching effect (interaction) \( A_S_{ik} \sim N(0, \sigma^2_{A_S}) \)
## Rating Bias by Internal/External Status

<table>
<thead>
<tr>
<th>Model</th>
<th>Internal Only</th>
<th>Experience Only</th>
<th>WESTB</th>
<th>VA Math Only</th>
<th>VA Read Only</th>
<th>Both VA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>N = 3474</td>
<td>N = 3473</td>
<td></td>
<td>N = 1411</td>
<td>N = 303</td>
<td>N = 267</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Intercept | 36.03 (0.48)*** | 35.57 (0.50)*** | 36.23 (0.60)*** | 37.34 (1.32)*** | 36.96 (1.11)*** | 36.74 (1.37)*** |
| Internal  | 3.08 (0.31)***  | 3.16 (0.31)***  | 2.84 (0.50)***  | 3.97 (1.29)**   | 4.15 (1.11)***  | 4.80 (1.35)***  |
| Experience| 0.11 (0.03)     |               |       |              |              |         |
| WESTB    |               | 0.11 (0.35)    | 0.40 (0.33)    | 0.09 (0.27)     |               |         |
| Value Added |           |     |       |              |              |         |
| Math     | 3.90 (2.00)    | 5.62 (2.46)*   | 3.29 (2.27)    | -3.10 (3.04)   |               |         |
| Reading  |               |               |       |              |              |         |

Notes:
- Models include random effects of applicant, rater, school and applicant-school interaction.
- **Experience** in years
- **WESTB** - scores on state licensure tests.
- **Value Added** - teacher value added estimates based on changes of student performance on achievement tests.
## Rating Bias by Internal/External Status

<table>
<thead>
<tr>
<th>Model</th>
<th>Internal Only</th>
<th>Experience Only</th>
<th>WESTB</th>
<th>VA Math Only</th>
<th>VA Read Only</th>
<th>Both VA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>N = 3474</td>
<td>N = 3473</td>
<td>N = 1411</td>
<td>N = 303</td>
<td>N = 336</td>
<td>N = 267</td>
</tr>
<tr>
<td>Intercept</td>
<td>36.03 (0.48)**</td>
<td>35.57 (0.50)**</td>
<td>36.23 (0.60)**</td>
<td>37.34 (1.32)**</td>
<td>36.96 (1.11)**</td>
<td>36.74 (1.37)**</td>
</tr>
<tr>
<td>Internal</td>
<td>3.08 (0.31)**</td>
<td>3.16 (0.31)**</td>
<td>2.84 (0.50)**</td>
<td>3.97 (1.29)**</td>
<td>4.15 (1.11)**</td>
<td>4.80 (1.35)**</td>
</tr>
<tr>
<td>Experience</td>
<td>0.11 (0.03)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WESTB</td>
<td></td>
<td></td>
<td>Writing</td>
<td>0.11 (0.35)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reading</td>
<td>0.40 (0.33)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Math</td>
<td>0.09 (0.27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value Added</td>
<td></td>
<td></td>
<td>Math</td>
<td>3.90 (2.00)</td>
<td>5.62 (2.46)*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reading</td>
<td>3.29 (2.27)</td>
<td>-3.10 (3.04)</td>
<td></td>
</tr>
</tbody>
</table>

Notes:
- Models include random effects of applicant, rater, school and applicant-school interaction.
- Experience in years
- WESTB - scores on state licensure tests.
- Value Added - teacher value added estimates based on changes of student performance on achievement tests.
## Rating Bias by Internal/External Status

<table>
<thead>
<tr>
<th></th>
<th>Model A Internal Only N = 3474</th>
<th>Model B Experience Only N = 3473</th>
<th>Model C WESTB N = 1411</th>
<th>Model D1 VA Math Only N = 303</th>
<th>Model D2 VA Read Only N = 336</th>
<th>Model D Both VA N = 267</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>36.03 (0.48)***</td>
<td>35.57 (0.50)***</td>
<td>36.23 (0.60)***</td>
<td>37.34 (1.32)***</td>
<td>36.96 (1.11)***</td>
<td>36.74 (1.37)***</td>
</tr>
<tr>
<td>Internal</td>
<td>3.08 (0.31)***</td>
<td>3.16 (0.31)***</td>
<td>2.84 (0.50)***</td>
<td>3.97 (1.29)**</td>
<td>4.15 (1.11)***</td>
<td>4.80 (1.35)***</td>
</tr>
<tr>
<td>Experience</td>
<td></td>
<td>0.11 (0.03)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WESTB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writing</td>
<td>0.11 (0.35)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>0.40 (0.33)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td>0.09 (0.27)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value Added</td>
<td></td>
<td></td>
<td></td>
<td>3.90 (2.00)</td>
<td>5.62 (2.46)*</td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td></td>
<td></td>
<td></td>
<td>3.29 (2.27)</td>
<td>-3.10 (3.04)</td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes:
- Models include random effects of applicant, rater, school and applicant-school interaction.
- **Experience** in years
- **WESTB** - scores on state licensure tests.
- **Value Added** - teacher value added estimates based on changes of student performance on achievement tests.
# Rating Bias by Internal/External Status

<table>
<thead>
<tr>
<th>Model</th>
<th>Internal Only N = 3474</th>
<th>Experience Only N = 3473</th>
<th>WESTB N = 1411</th>
<th>VA Math Only N = 303</th>
<th>VA Read Only N = 336</th>
<th>Both VA N = 267</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept</strong></td>
<td>36.03 (0.48)*****</td>
<td>35.57 (0.50)*****</td>
<td>36.23 (0.60)*****</td>
<td>37.34 (1.32)*****</td>
<td>36.96 (1.11)*****</td>
<td>36.74 (1.37)*****</td>
</tr>
<tr>
<td><strong>Internal</strong></td>
<td>3.08 (0.31)*****</td>
<td>3.16 (0.31)*****</td>
<td>2.84 (0.50)*****</td>
<td>3.97 (1.29)**</td>
<td>4.15 (1.11)*****</td>
<td>4.80 (1.35)*****</td>
</tr>
<tr>
<td><strong>Experience</strong></td>
<td>0.11 (0.03)</td>
<td>0.11 (0.35)</td>
<td>0.40 (0.33)</td>
<td>0.09 (0.27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>WESTB</strong></td>
<td>0.11 (0.35)</td>
<td>0.40 (0.33)</td>
<td>0.09 (0.27)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Value Added</strong></td>
<td>3.90 (2.00)</td>
<td>3.29 (2.27)</td>
<td>5.62 (2.46)*</td>
<td>-3.10 (3.04)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes:
- Models include random effects of applicant, rater, school and applicant-school interaction.
- **Experience** in years
- **WESTB** - scores on state licensure tests.
- **Value Added** - teacher value added estimates based on changes of student performance on achievement tests.
## Rating Bias by Internal/External Status

<table>
<thead>
<tr>
<th>Model</th>
<th>Internal Only</th>
<th>Experience Only</th>
<th>WESTB</th>
<th>VA Math Only</th>
<th>VA Read Only</th>
<th>Both VA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>N = 3474</td>
<td>N = 3473</td>
<td>N = 1411</td>
<td>N = 303</td>
<td>N = 336</td>
<td>N = 267</td>
</tr>
<tr>
<td>Intercept</td>
<td>36.03 (0.48)***</td>
<td>35.57 (0.50)***</td>
<td>36.23 (0.60)***</td>
<td>37.34 (1.32)***</td>
<td>36.96 (1.11)***</td>
<td>36.74 (1.37)***</td>
</tr>
<tr>
<td>Internal</td>
<td>3.08 (0.31)***</td>
<td>3.16 (0.31)***</td>
<td>2.84 (0.50)***</td>
<td>3.97 (1.29)**</td>
<td>4.15 (1.11)***</td>
<td>4.80 (1.35)***</td>
</tr>
<tr>
<td>Experience</td>
<td>0.11 (0.03)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WESTB Writing</td>
<td>0.11 (0.35)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WESTB Reading</td>
<td>0.40 (0.33)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WESTB Math</td>
<td>0.09 (0.27)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value Added Math</td>
<td>3.90 (2.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.62 (2.46)*</td>
</tr>
<tr>
<td>Value Added Reading</td>
<td>3.29 (2.27)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-3.10 (3.04)</td>
</tr>
</tbody>
</table>

Notes:
- Models include random effects of applicant, rater, school and applicant-school interaction.
- **Experience** in years
- **WESTB** - scores on state licensure tests.
- **Value Added** - teacher value added estimates based on changes of student performance on achievement tests.
# Rating Bias by Internal/External Status

<table>
<thead>
<tr>
<th>Model</th>
<th>Internal Only</th>
<th>Experience Only</th>
<th>WESTB</th>
<th>VA Math Only</th>
<th>VA Read Only</th>
<th>Both VA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>N = 3474</td>
<td>N = 3473</td>
<td>N = 1411</td>
<td>N = 303</td>
<td>N = 336</td>
<td>N = 267</td>
</tr>
<tr>
<td>Intercept</td>
<td>36.03 (0.48)**</td>
<td>35.57 (0.50)**</td>
<td>36.23 (0.60)**</td>
<td>37.34 (1.32)**</td>
<td>36.96 (1.11)**</td>
<td>36.74 (1.37)**</td>
</tr>
<tr>
<td>Internal</td>
<td>3.08 (0.31)**</td>
<td>3.16 (0.31)**</td>
<td>2.84 (0.50)**</td>
<td>3.97 (1.29)**</td>
<td>4.15 (1.11)**</td>
<td>4.80 (1.35)**</td>
</tr>
<tr>
<td>Experience</td>
<td>0.11 (0.35)</td>
<td>0.40 (0.33)</td>
<td>0.09 (0.27)</td>
<td>0.11 (0.35)</td>
<td>0.40 (0.33)</td>
<td>0.09 (0.27)</td>
</tr>
<tr>
<td>WESTB</td>
<td></td>
<td></td>
<td>Writing</td>
<td></td>
<td>Reading</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Math</td>
</tr>
<tr>
<td>Value Added</td>
<td>Math</td>
<td>3.90 (2.00)</td>
<td>5.62 (2.46)*</td>
<td>3.29 (2.27)</td>
<td>-3.10 (3.04)</td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes:
- Models include random effects of applicant, rater, school and applicant-school interaction.
- Experience in years
- WESTB - scores on state licensure tests.
- Value Added - teacher value added estimates based on changes of student performance on achievement tests.
Inter-Rater Reliability (Model 1)

\[ Y_{ij} = \mu + A_i + B_j + e_{ij} \]

- applicant true quality \( A_i \sim N(0, \sigma_A^2) \),
- rater leniency \( B_j \sim N(0, \sigma_B^2) \),
- error \( e_{ij} \sim N(0, \sigma_e^2) \)

Inter-Rater Reliability:

\[ R = \text{cor}(Y_{ij}, Y_{ij'}) = \text{ICC} = \frac{\sigma_A^2}{\sigma_Y^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_e^2} \]

- \( R \in [0, 1] \), low values mean a lot of measurement error
Inter-Rater Reliability (Model 1)

\[ Y_{ij} = \mu + A_i + B_j + e_{ij} \]

- applicant true quality \( A_i \sim N(0, \sigma^2_A) \),
- rater leniency \( B_j \sim N(0, \sigma^2_B) \),
- error \( e_{ij} \sim N(0, \sigma^2_e) \)

Inter-Rater Reliability:

\[
R = \text{cor}(Y_{ij}, Y_{ij'}) = \text{ICC} = \frac{\sigma^2_A}{\sigma^2_Y} = \frac{\sigma^2_A}{\sigma^2_A + \sigma^2_B + \sigma^2_e}
\]

- \( R \in [0, 1] \), low values mean a lot of measurement error
Inter-Rater Reliability (Model 1)

\[ Y_{ij} = \mu + A_i + B_j + e_{ij} \]

- applicant true quality \( A_i \sim N(0, \sigma_A^2) \),
- rater leniency \( B_j \sim N(0, \sigma_B^2) \),
- error \( e_{ij} \sim N(0, \sigma_e^2) \)

**Inter-Rater Reliability:**

\[ R = \text{cor}(Y_{ij}, Y_{ij'}) = \text{ICC} = \frac{\sigma_A^2}{\sigma_Y^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_e^2} \]

- \( R \in [0, 1] \), low values mean a lot of measurement error
Within-School IRR (Model 2)

\[ Y_{ijk} = \mu + A_i + B_j + S_k + AS_{ik} + e_{ijk} \]

- School leniency \( S_k \sim N(0, \sigma^2_S) \)
- Applicant-school matching effect (interaction) \( AS_{ik} \sim N(0, \sigma^2_{AS}) \)

Within-school IRR:

\[
R = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma^2_A + \sigma^2_S + \sigma^2_{AS}}{\sigma^2_A + \sigma^2_B + \sigma^2_S + \sigma^2_{AS} + \sigma^2_e}
\]
Within-School IRR (Model 2)

\[ Y_{ijk} = \mu + A_i + B_j + S_k + AS_{ik} + e_{ijk} \]

- School leniency: \( S_k \sim N(0, \sigma_S^2) \)
- Applicant-school matching effect (interaction): \( AS_{ik} \sim N(0, \sigma_{AS}^2) \)

**Within-school IRR:**

\[ R = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma_A^2 + \sigma_S^2 + \sigma_{AS}^2}{\sigma_A^2 + \sigma_B^2 + \sigma_S^2 + \sigma_{AS}^2 + \sigma_e^2} \]
IRR for Internal vs. External Applicants (Model 3)

Q: Does IRR differ in ratings of internal vs. external applicants?

Model 3: Variance components may vary by group
  - e.g. Rater variance may higher when rating external applicants

\[ Y_{ijk} = \mu + \omega_i \beta_0 + (1 - \omega_i) A_{0i} + \omega_i A_{1i} + (1 - \omega_i) B_{0j} + \omega_i B_{1j} + (1 - \omega_i) S_{0k} + \omega_i S_{1k} + AS_{ik} + e_{ijk} \]

\[ \omega_i = 1 \text{ for internal and 0 for external applicants} \]
- \( A_{0i} \sim N(0, \sigma_{A0}^2) \) and \( A_{1i} \sim N(0, \sigma_{A1}^2) \)
- \( B_{0j} \sim N(0, \sigma_{B0}^2) \) and \( B_{1j} \sim N(0, \sigma_{B1}^2) \)
- \( S_{0k} \sim N(0, \sigma_{S0}^2) \) and \( S_{1k} \sim N(0, \sigma_{S1}^2) \)
IRR for Internal vs. External Applicants (Model 3)

Q: Does IRR differ in ratings of internal vs. external applicants?

Model 3: Variance components may vary by group
  - e.g. Rater variance may higher when rating external applicants

\[ Y_{ijk} = \mu + \omega_i \beta_0 + (1 - \omega_i) A_{0i} + \omega_i A_{1i} + (1 - \omega_i) B_{0j} + \omega_i B_{1j} + (1 - \omega_i) S_{0k} + \omega_i S_{1k} + AS_{ik} + e_{ijk} \]

\[ \omega_i = 1 \text{ for internal and } 0 \text{ for external applicants} \]

- \( A_{0i} \sim N(0, \sigma_{A0}^2) \) and \( A_{1i} \sim N(0, \sigma_{A1}^2) \)
- \( B_{0j} \sim N(0, \sigma_{B0}^2) \) and \( B_{1j} \sim N(0, \sigma_{B1}^2) \)
- \( S_{0k} \sim N(0, \sigma_{S0}^2) \) and \( S_{1k} \sim N(0, \sigma_{S1}^2) \)
Q: Does IRR differ in ratings of internal vs. external applicants?

Model 3: Variance components may vary by group
- e.g. Rater variance may higher when rating external applicants

\[ Y_{ijk} = \mu + \omega_i \beta_0 + (1 - \omega_i)A_{0i} + \omega_i A_{1i} + (1 - \omega_i)B_{0j} + \omega_i B_{1j} + (1 - \omega_i)S_{0k} + \omega_i S_{1k} + AS_{ik} + e_{ijk} \]

- \( \omega_i = 1 \) for internal and 0 for external applicants
- \( A_{0i} \sim N(0, \sigma_{A0}^2) \) and \( A_{1i} \sim N(0, \sigma_{A1}^2) \)
- \( B_{0j} \sim N(0, \sigma_{B0}^2) \) and \( B_{1j} \sim N(0, \sigma_{B1}^2) \)
- \( S_{0k} \sim N(0, \sigma_{S0}^2) \) and \( S_{1k} \sim N(0, \sigma_{S1}^2) \)
Within-school IRR:

- For internal applicant:
  \[ R_1 = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma^2_{A1} + \sigma^2_{S1} + \sigma^2_{AS}}{\sigma^2_{A1} + \sigma^2_{B1} + \sigma^2_{S1} + \sigma^2_{AS} + \sigma^2_e} \]

- For external applicant:
  \[ R_0 = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma^2_{A0} + \sigma^2_{S0} + \sigma^2_{AS}}{\sigma^2_{A0} + \sigma^2_{B0} + \sigma^2_{S0} + \sigma^2_{AS} + \sigma^2_e} \]
IRR for Internal vs. External Applicants (Model 3)

Within-school IRR:
- For internal applicant:
  \[ R_1 = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma^2_{A1} + \sigma^2_{S1} + \sigma^2_{AS}}{\sigma^2_{A1} + \sigma^2_{B1} + \sigma^2_{S1} + \sigma^2_{AS} + \sigma^2_e} \]

- For external applicant:
  \[ R_0 = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma^2_{A0} + \sigma^2_{S0} + \sigma^2_{AS}}{\sigma^2_{A0} + \sigma^2_{B0} + \sigma^2_{S0} + \sigma^2_{AS} + \sigma^2_e} \]
Results: Variance decomposition (Model 3)

- High applicant-school variability
- Lower applicant variability for external applicants
- Higher rater variability for external applicants
- Lower inter-rater reliability for external applicants

Patricia Martinkova (martinkova@cs.cas.cz)
Results: Variance decomposition (Model 3)

- High applicant-school variability
- Lower applicant variability for external applicants
- Higher rater variability for external applicants
- Lower inter-rater reliability for external applicants
Results: Variance decomposition (Model 3)

- High applicant-school variability
- Lower applicant variability for external applicants
- Higher rater variability for external applicants
- Lower inter-rater reliability for external applicants
Results: Variance decomposition (Model 3)

- High applicant-school variability
- Lower applicant variability for external applicants
- Higher rater variability for external applicants
- Lower inter-rater reliability for external applicants
Results: Variance decomposition (Model 3)

- High applicant-school variability
- Lower applicant variability for external applicants
- Higher rater variability for external applicants
- Lower inter-rater reliability for external applicants
IRR for Internal and External Applicants (Model 3)

- IRR is estimated simultaneously for both groups within Model 3
- Bootstrapped confidence intervals
IRR for Internal and External Applicants (Model 3)

- Significant difference in IRR between Internal and External applicants
Conclusion for Teacher Hiring Data

- Rating is school-specific
  - Accounting for applicant-school matching in the model is important

- Significantly lower ratings of external applicants confirmed
  - Accounting for previous experience and licensure scores
  - Accounting for subsequent teacher value added

- Significantly lower inter-rater reliability when rating external applicants
  - Similar variance decomposition in stratified data
  - Our approach allows for testing differences in variance terms and in IRR by group
Conclusion for Teacher Hiring Data

- Rating is school-specific
  - Accounting for applicant-school matching in the model is important

- Significantly lower ratings of external applicants confirmed
  - Accounting for previous experience and licensure scores
  - Accounting for subsequent teacher value added

- Significantly lower inter-rater reliability when rating external applicants
  - Similar variance decomposition in stratified data
  - Our approach allows for testing differences in variance terms and in IRR by group
Conclusion for Teacher Hiring Data

- Rating is school-specific
  - Accounting for applicant-school matching in the model is important

- Significantly lower ratings of external applicants confirmed
  - Accounting for previous experience and licensure scores
  - Accounting for subsequent teacher value added

- Significantly lower inter-rater reliability when rating external applicants
  - Similar variance decomposition in stratified data
  - Our approach allows for testing differences in variance terms and in IRR by group
Implications for Peer-Review in other areas

Model-based IRR is applicable to testing differences w/ respect to:

- assessee status (experienced, matching gender etc.)
  - more likely to matter in fellowships or grant reviews
  - not expected to matter in double-blind review

- other grouping variable
  - reviewer type (experience, research field)
  - journal, journal type
  - grant panel, grant type
  - etc.
Implications for Peer-Review in other areas

Model-based IRR is applicable to testing differences with respect to:

- assessee status (experienced, matching gender etc.)
  - more likely to matter in fellowships or grant reviews
  - not expected to matter in double-blind review

- other grouping variable
  - reviewer type (experience, research field)
  - journal, journal type
  - grant panel, grant type
  - etc.
Conclusion

- Significantly lower ratings and lower IRR showed for external applicants to teacher hiring positions.
- Model-based approach allows to:
  - account for data structure (applicant-school matching etc.)
  - test for difference in IRR between groups
- Method is applicable to grant or journal peer-review.

Thank you for your attention!

http://www.cs.cas.cz/martinkova/

martinkova@cs.cas.cz
Conclusion

- Significantly lower ratings and lower IRR showed for external applicants to teacher hiring positions.

- Model-based approach allows to
  - account for data structure (applicant-school matching etc.)
  - test for difference in IRR between groups

- Method is applicable to grant or journal peer-review.

Thank you for your attention!

http://www.cs.cas.cz/martinkova/

martinkova@cs.cas.cz
Conclusion

- Significantly lower ratings and lower IRR showed for external applicants to teacher hiring positions.

- Model-based approach allows to
  - account for data structure (applicant-school matching etc.)
  - test for difference in IRR between groups

- Method is applicable to grant or journal peer-review.

Thank you for your attention!

http://www.cs.cas.cz/martinkova/

martinkova@cs.cas.cz
Significantly lower ratings and lower IRR showed for external applicants to teacher hiring positions.

Model-based approach allows to:
- account for data structure (applicant-school matching etc.)
- test for difference in IRR between groups

Method is applicable to grant or journal peer-review.

Thank you for your attention!

http://www.cs.cas.cz/martinkova/

martinkova@cs.cas.cz