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[1,2] Dan Goldhaber[3] Elena Erosheva[3] & Carole Lee[3]

[1]Institute of Computer Science, Czech Academy of Sciences
[2]College of Education, Charles University, Prague
[3]University of Washington, Seattle, USA

PEERE International Conference, 7-9 March 2018, Rome

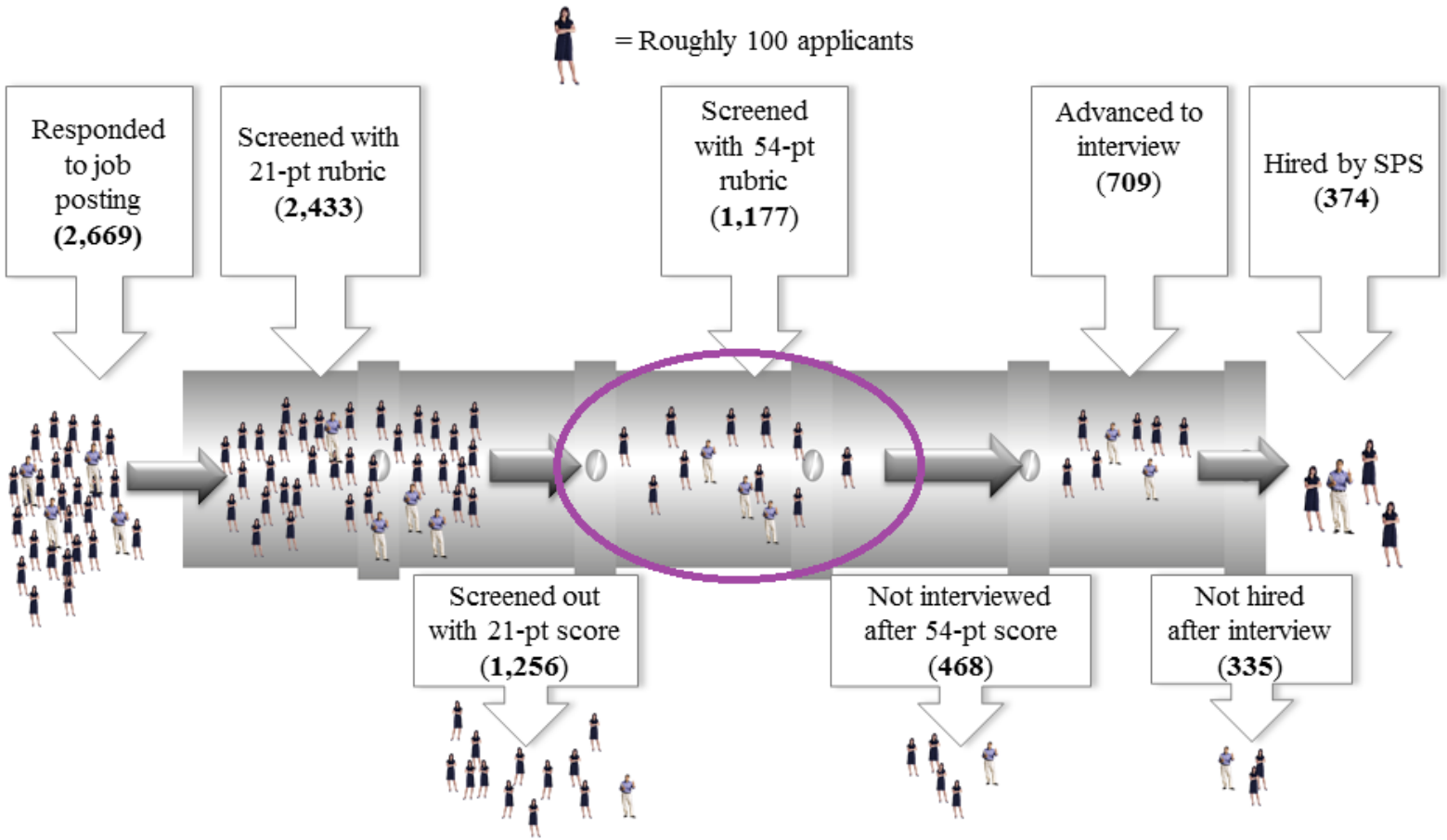
Acknowledgements

- 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

Introduction: Teacher Selection Process

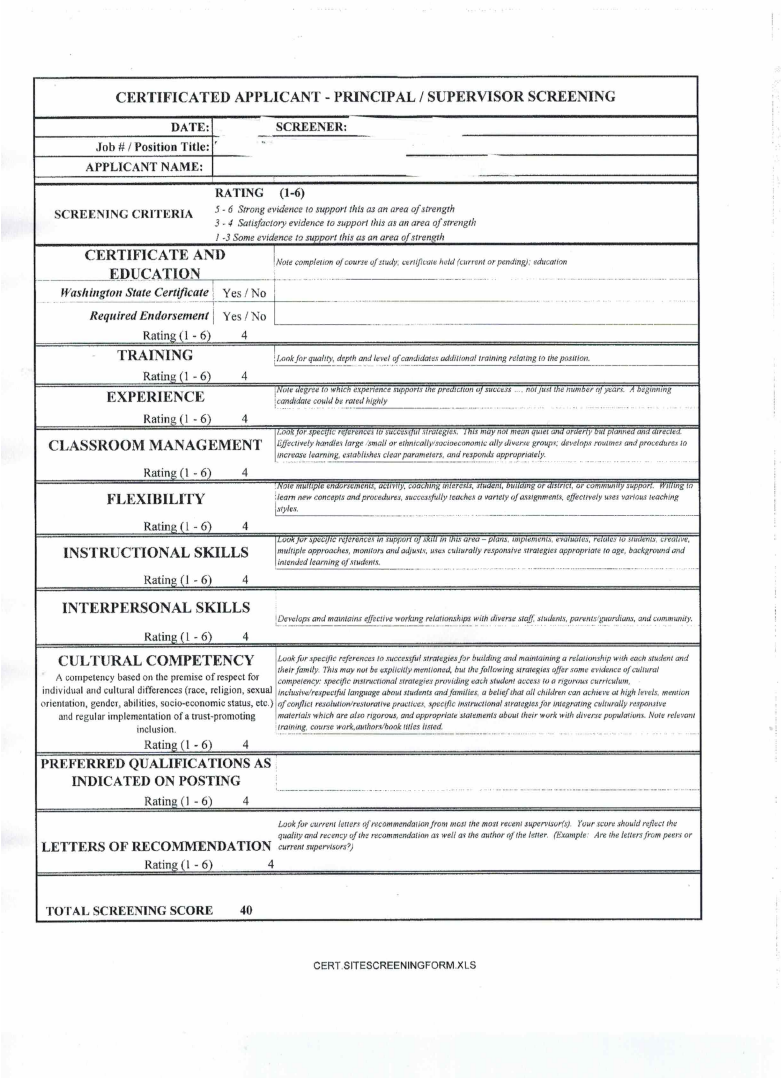


Applicants to classroom job openings in Spokane Public Schools during years (2008/09 - 2012/13)

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)



CERTIFICATED APPLICANT - PRINCIPAL / SUPERVISOR SCREENING

DATE:	SCREENER:	
Job # / Position Title:		
APPLICANT NAME:		
SCREENING CRITERIA	RATING (1-6) 5 - 6 Strong evidence to support this as an area of strength 3 - 4 Satisfactory evidence to support this as an area of strength 1 -3 Some evidence to support this as an area of strength	
CERTIFICATE AND EDUCATION		Note completion of course of study, certificate held (current or pending); education
Washington State Certificate	Yes / No	
Required Endorsement	Yes / No	
Rating (1 - 6)	4	
TRAINING		Look for quality, depth and level of candidates additional training relating to the position.
Rating (1 - 6)	4	
EXPERIENCE		Note degree to which experience supports the prediction of success .... not just the number of years. A beginning candidate could be rated highly
Rating (1 - 6)	4	
CLASSROOM MANAGEMENT		Look for specific references to successful strategies. This may not mean quiet and orderly but planned and directed. Effectively handles large /small or ethnically/racioeconomic ally diverse groups; develops routines and procedures to increase learning, establishes clear parameters, and responds appropriately.
Rating (1 - 6)	4	
FLEXIBILITY		Note multiple endorsements, activity, coaching interests, student, building or district, or community support. Willing to learn new concepts and procedures, successfully teaches a variety of assignments, effectively uses various teaching styles.
Rating (1 - 6)	4	
INSTRUCTIONAL SKILLS		Look for specific references in support of skill in this area – plans, implements, evaluates, relates to students, creative, multiple approaches, monitors and adjusts, uses culturally responsive strategies appropriate to age, background and intended learning of students.
Rating (1 - 6)	4	
INTERPERSONAL SKILLS		Develops and maintains effective working relationships with diverse staff, students, parents/guardians, and community.
Rating (1 - 6)	4	
CULTURAL COMPETENCY A competency based on the premise of respect for individual and cultural differences (race, religion, sexual orientation, gender, abilities, socio-economic status, etc.) and regular implementation of a trust-promoting inclusion.		Look for specific references to successful strategies for building and maintaining a relationship with each student and their family. This may not be explicitly mentioned, but the following strategies offer some evidence of cultural competency: specific instructional strategies providing each student access to a rigorous curriculum, inclusive/respectful language about students and families, a belief that all children can achieve at high levels, mention of conflict resolution/restorative practices, specific instructional strategies for integrating culturally responsive materials which are also rigorous, and appropriate statements about their work with diverse populations. Note relevant training, course work,authors/book titles listed.
Rating (1 - 6)	4	
PREFERRED QUALIFICATIONS AS INDICATED ON POSTING		
Rating (1 - 6)	4	
LETTERS OF RECOMMENDATION		Look for current letters of recommendation from most the most recent supervisor(s). Your score should reflect the quality and recency of the recommendation as well as the author of the letter. (Example: Are the letters from peers or current supervisors?)
Rating (1 - 6)	4	
TOTAL SCREENING SCORE	40	

CERT.SITESCREENINGFORM.XLS

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

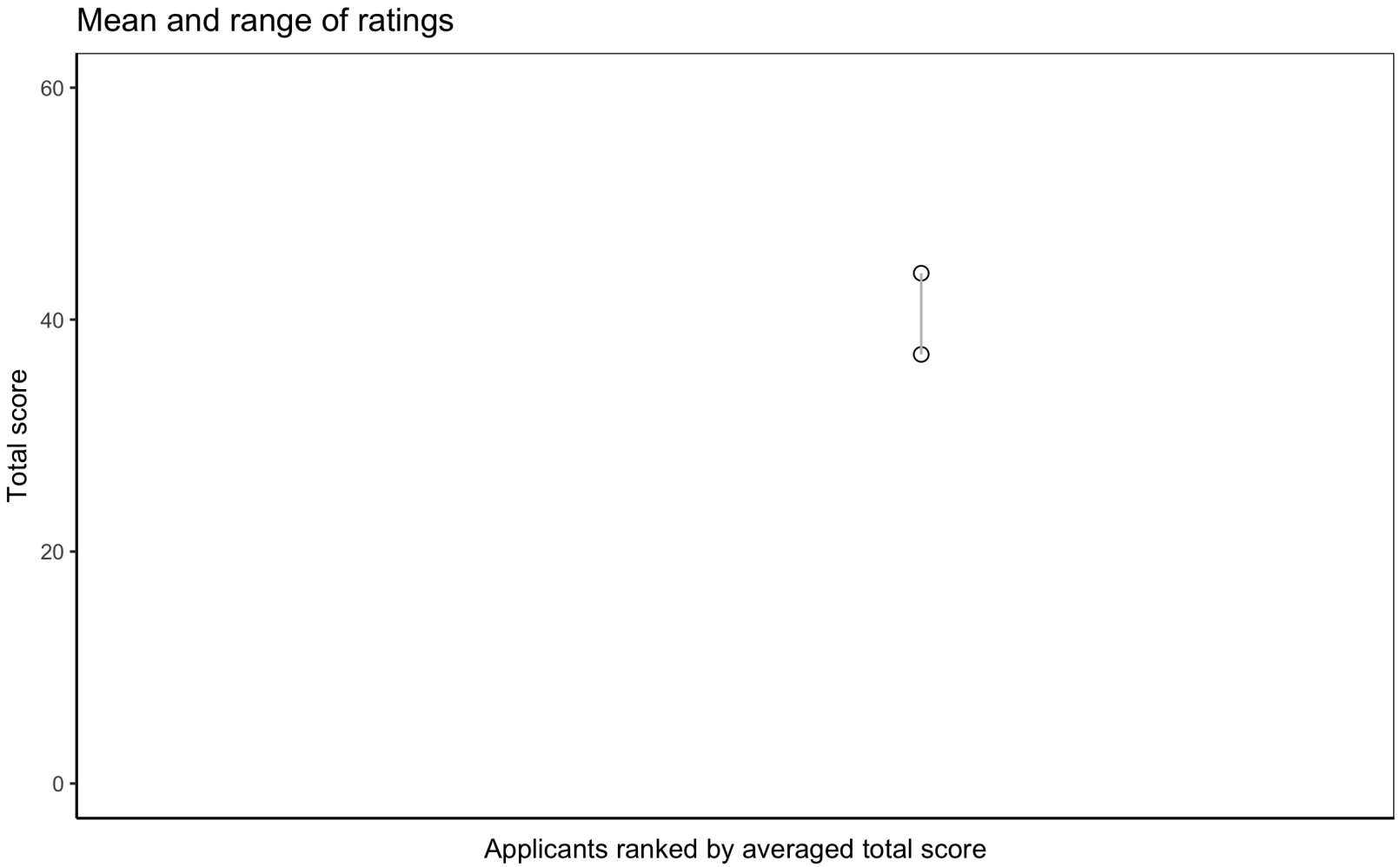
Data structure

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

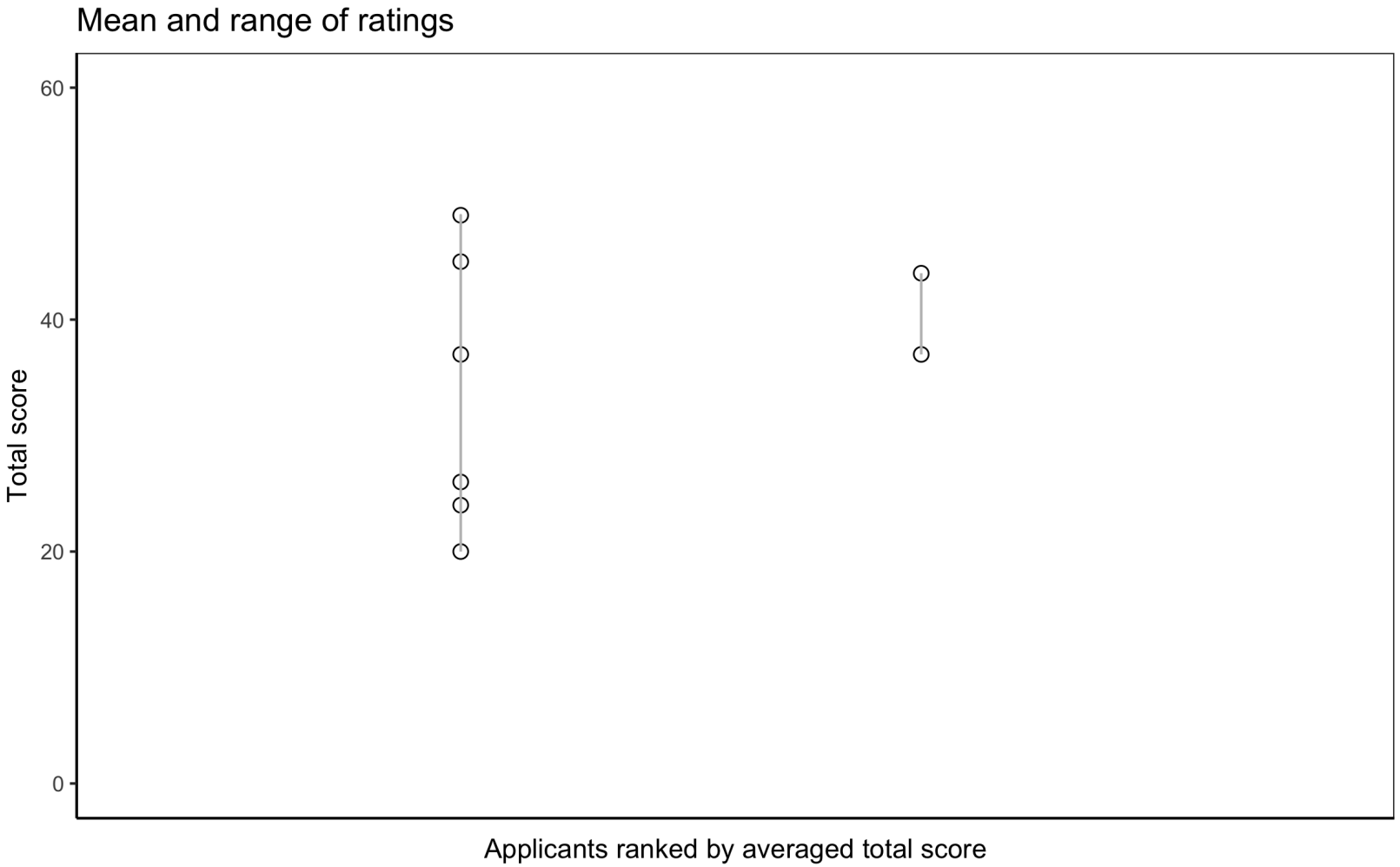
Applicant status

- 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

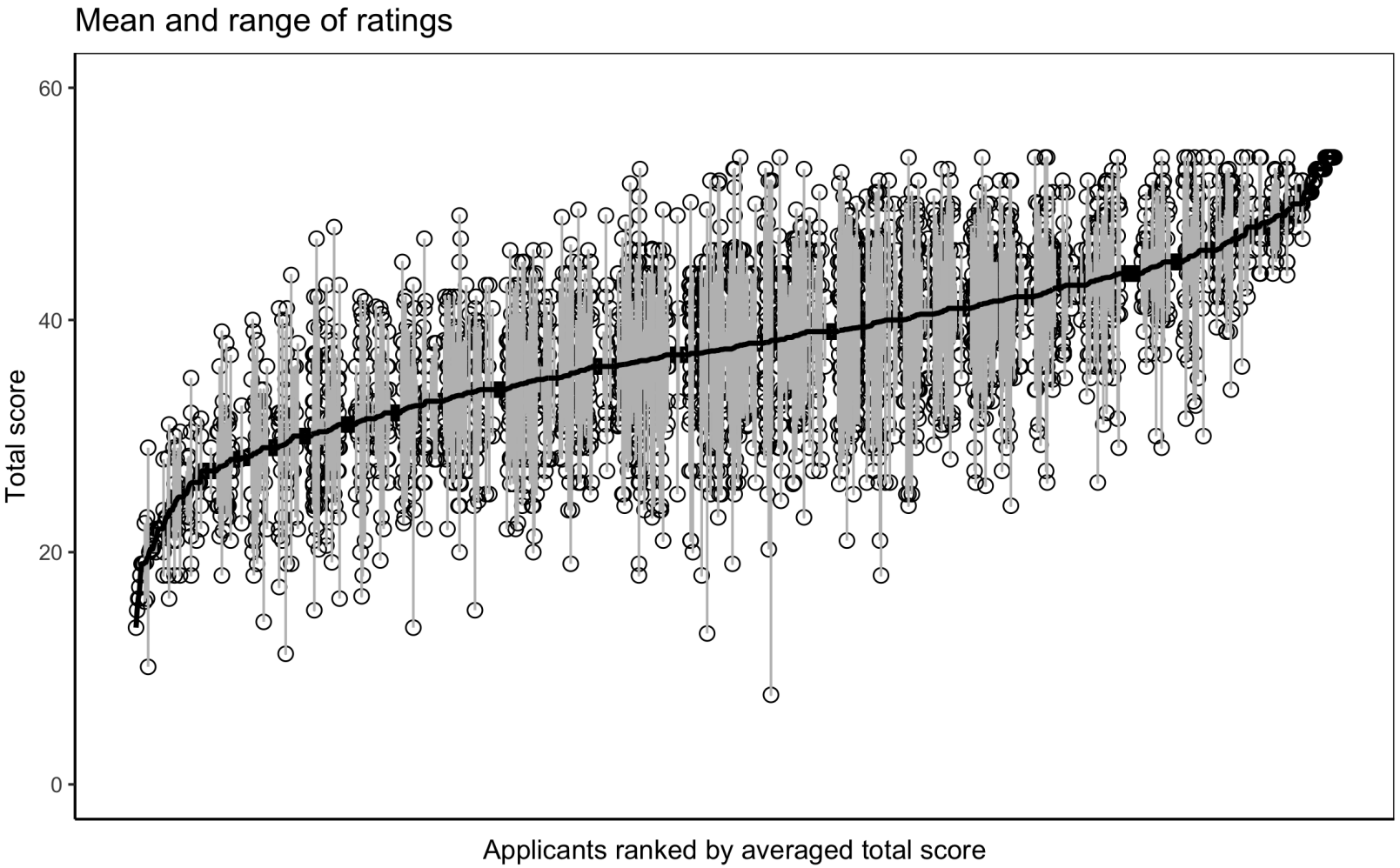
Ratings of a single applicant



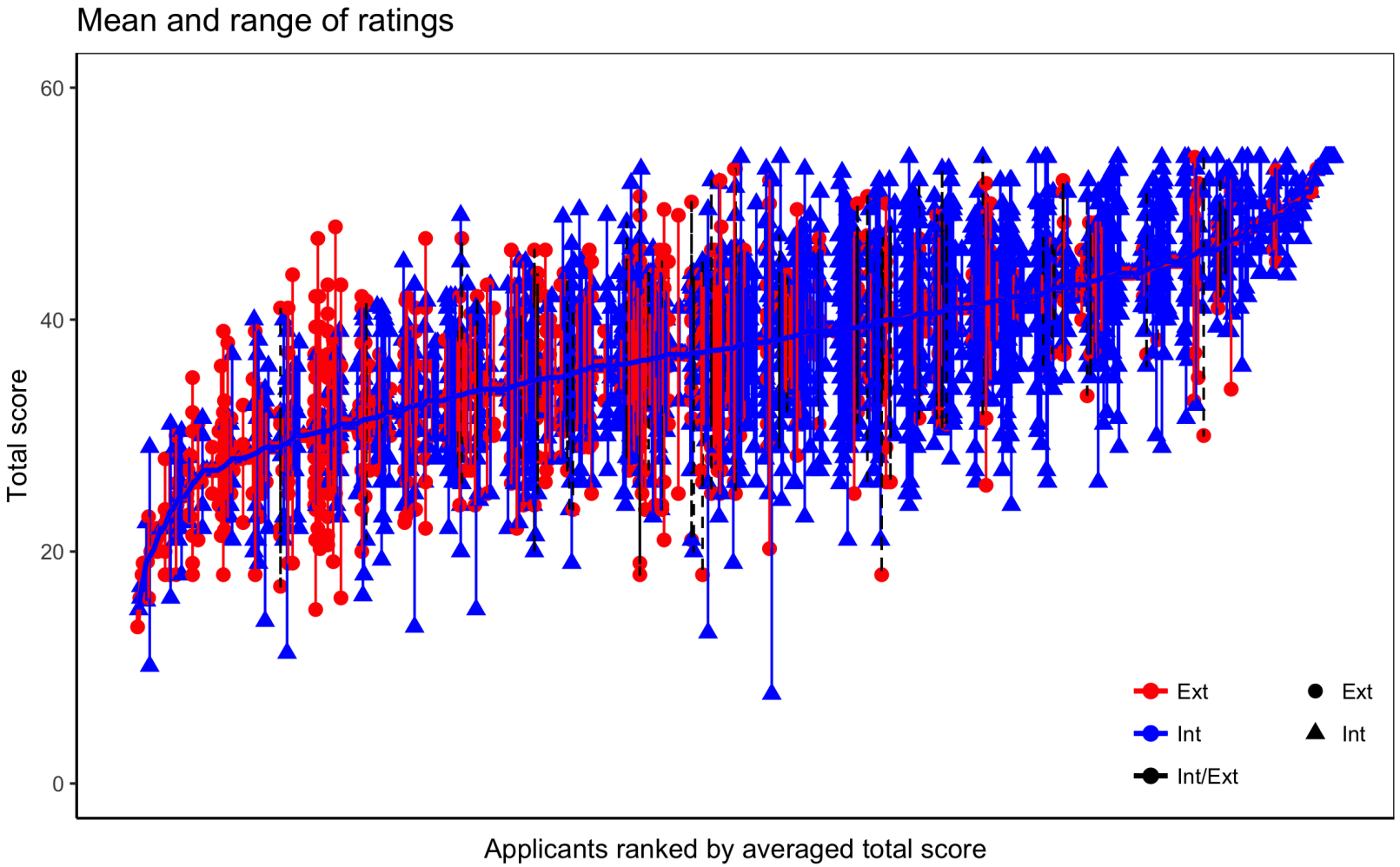
Ratings of two applicants



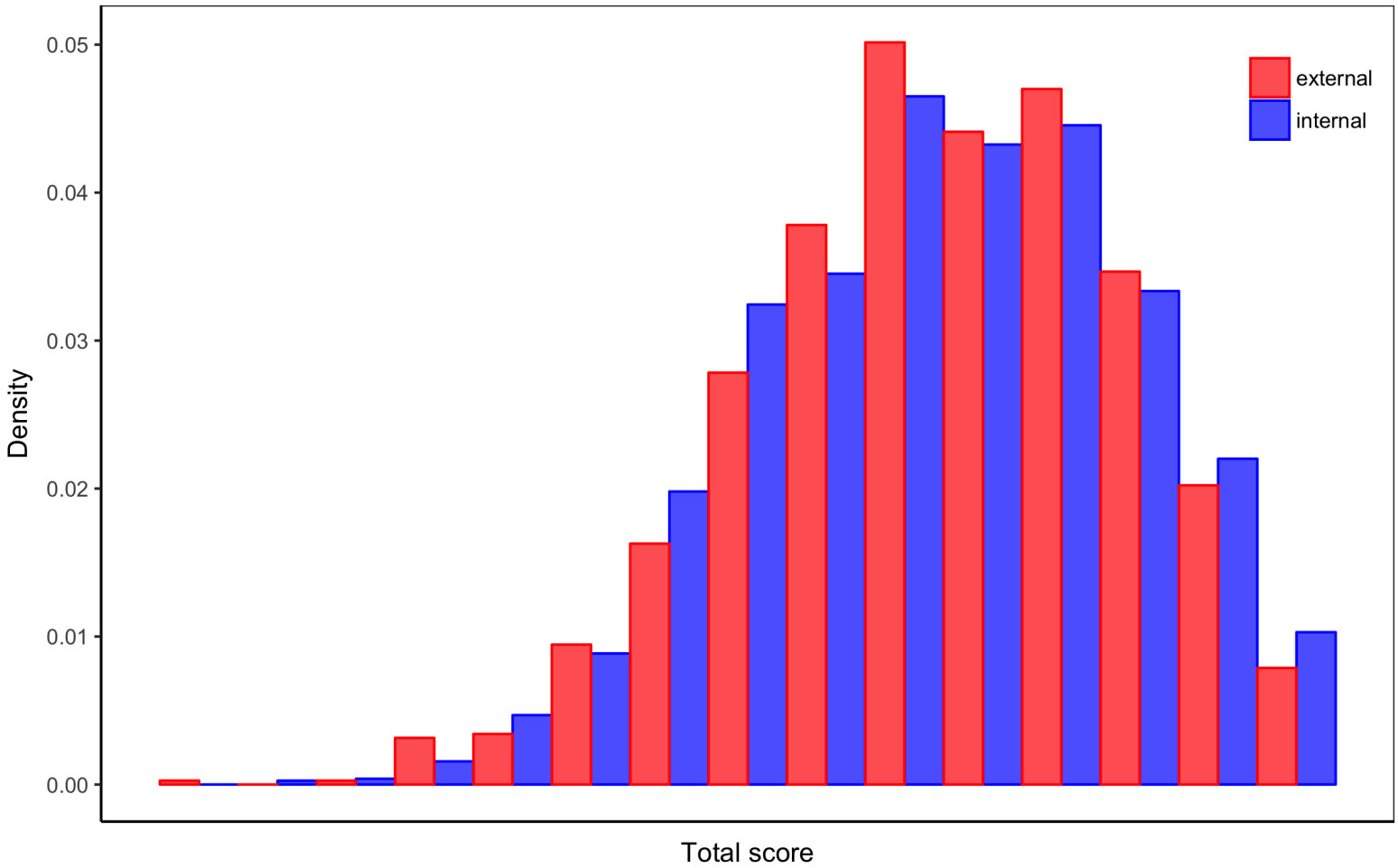
Ratings of all applicants



Ratings of all applicants by Internal/External Status

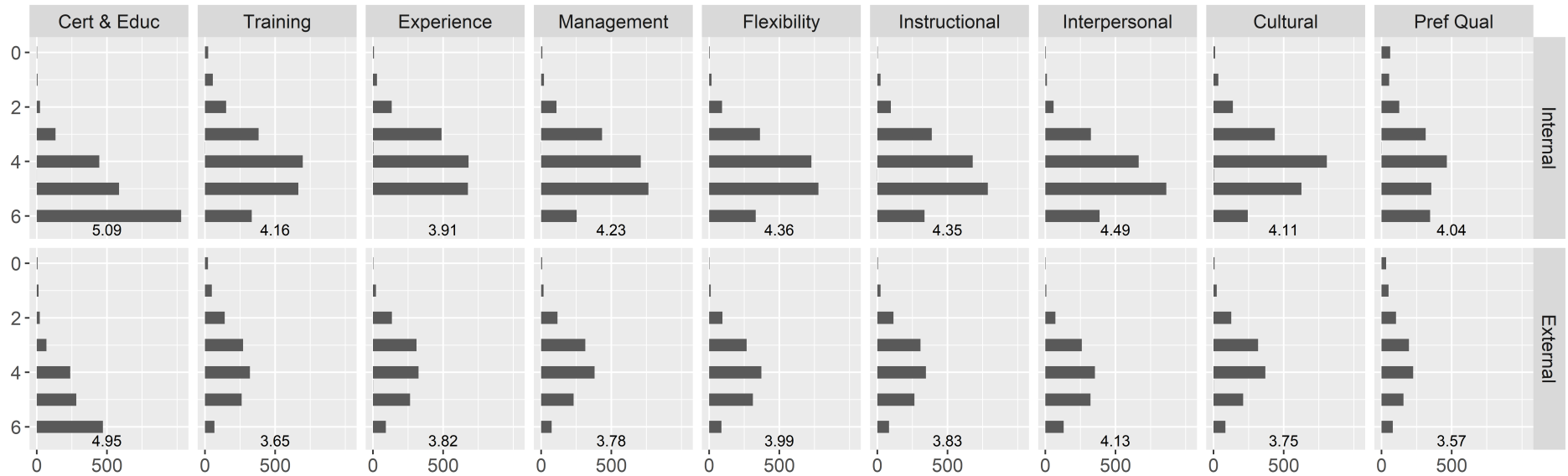


Rating distributions



- 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

$$Y_{ijk} = \mu + \omega_i \beta_0 + \boldsymbol{X}_i \boldsymbol{\beta} + A_i + B_j + S_k + AS_{ik} + e_{ijk}$$

- Applicant internal/external status ω_i
- Applicant quality measures $\boldsymbol{X}_i$
  (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

$$Y_{ijk} = \mu + \omega_i \beta_0 + \boldsymbol{X}_i \boldsymbol{\beta} + A_i + B_j + S_k + AS_{ik} + e_{ijk}$$

- Applicant internal/external status ω_i
- Applicant quality measures $\boldsymbol{X}_i$
  (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) $AS_{ik} \sim N(0, \sigma_{AS}^2)$

Testing for bias with respect to applicant status

Model controlling for quality measures, accounting for data structure

$$Y_{ijk} = \mu + \omega_i \beta_0 + \boldsymbol{X}_i \boldsymbol{\beta} + A_i + B_j + S_k + AS_{ik} + e_{ijk}$$

- Applicant internal/external status ω_i
- Applicant quality measures $\boldsymbol{X}_i$
  (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) $AS_{ik} \sim N(0, \sigma_{AS}^2)$

Rating Bias by Internal/External Status

	Model A Internal Only N = 3474	Model B Experience Only N = 3473	Model C WESTB N = 1411	Model D1 VA Math Only N = 303	Model D2 VA Read Only N = 336	Model D Both VA N = 267
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						
Writing			0.11 (0.35)			
Reading			0.40 (0.33)			
Math			0.09 (0.27)			
Value Added						
Math				3.90 (2.00)		5.62 (2.46)*
Reading					3.29 (2.27)	-3.10 (3.04)

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 perfomrance on achievement tests.

Rating Bias by Internal/External Status

	Model A Internal Only N = 3474	Model B Experience Only N = 3473	Model C WESTB N = 1411	Model D1 VA Math Only N = 303	Model D2 VA Read Only N = 336	Model D Both VA N = 267
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						
Writing			0.11 (0.35)			
Reading			0.40 (0.33)			
Math			0.09 (0.27)			
Value Added						
Math				3.90 (2.00)		5.62 (2.46)*
Reading					3.29 (2.27)	-3.10 (3.04)

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 perfomrance on achievement tests.

Rating Bias by Internal/External Status

	Model A Internal Only N = 3474	Model B Experience Only N = 3473	Model C WESTB N = 1411	Model D1 VA Math Only N = 303	Model D2 VA Read Only N = 336	Model D Both VA N = 267
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						
Writing			0.11 (0.35)			
Reading			0.40 (0.33)			
Math			0.09 (0.27)			
Value Added						
Math				3.90 (2.00)		5.62 (2.46)*
Reading					3.29 (2.27)	-3.10 (3.04)

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 perfomrance on achievement tests.

Rating Bias by Internal/External Status

	Model A Internal Only N = 3474	Model B Experience Only N = 3473	Model C WESTB N = 1411	Model D1 VA Math Only N = 303	Model D2 VA Read Only N = 336	Model D Both VA N = 267
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						
Writing			0.11 (0.35)			
Reading			0.40 (0.33)			
Math			0.09 (0.27)			
Value Added						
Math				3.90 (2.00)		5.62 (2.46)*
Reading					3.29 (2.27)	-3.10 (3.04)

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 perfomrance on achievement tests.

Rating Bias by Internal/External Status

	Model A Internal Only N = 3474	Model B Experience Only N = 3473	Model C WESTB N = 1411	Model D1 VA Math Only N = 303	Model D2 VA Read Only N = 336	Model D Both VA N = 267
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						
Writing			0.11 (0.35)			
Reading			0.40 (0.33)			
Math			0.09 (0.27)			
Value Added						
Math				3.90 (2.00)		5.62 (2.46)*
Reading					3.29 (2.27)	-3.10 (3.04)

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 perfomrance on achievement tests.

Rating Bias by Internal/External Status

	Model A Internal Only N = 3474	Model B Experience Only N = 3473	Model C WESTB N = 1411	Model D1 VA Math Only N = 303	Model D2 VA Read Only N = 336	Model D Both VA N = 267
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						
Writing			0.11 (0.35)			
Reading			0.40 (0.33)			
Math			0.09 (0.27)			
Value Added						
Math				3.90 (2.00)		5.62 (2.46)*
Reading					3.29 (2.27)	-3.10 (3.04)

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 perfomrance on achievement tests.

Inter-Rater Reliability (Model 1)

$$Y_{ij} = \mu + A_i + B_j + e_{ij}$$

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

Inter-Rater Reliability:

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

- $\mathrm{R} \in [0, 1]$, 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 $A_i \sim \mathrm{N}(0, \sigma_A^2)$,
- rater leniency $B_j \sim \mathrm{N}(0, \sigma_B^2)$,
- error $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$

Inter-Rater Reliability:

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

- $\mathrm{R} \in [0, 1]$, 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 $A_i \sim \mathrm{N}(0, \sigma_A^2)$,
- rater leniency $B_j \sim \mathrm{N}(0, \sigma_B^2)$,
- error $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$

Inter-Rater Reliability:

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

- $\mathrm{R} \in [0, 1]$, 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 leniencyl $S_k \sim N(0, \sigma_S^2)$
- Applicant-school matching effect (interaction) $AS_{ik} \sim N(0, \sigma_{AS}^2)$

Within-school IRR:

$$R = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma_A^2 + \sigma_S^2 + \sigma_{AS}^2}{\sigma_A^2 + \sigma_B^2 + \sigma_S^2 + \sigma_{AS}^2 + \sigma_e^2}$$

Within-School IRR (Model 2)

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

- School leniencyl $S_k \sim N(0, \sigma_S^2)$
- 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

$$\begin{aligned} Y_{ijk} = \mu + \omega_i\beta_0 &+ (1-\omega_i)A_{0i} + \omega_i A_{1i} \\ &+ (1-\omega_i)B_{0j} + \omega_i B_{1j} \\ &+ (1-\omega_i)S_{0k} + \omega_i S_{1k} \\ &+ AS_{ik} + e_{ijk} \end{aligned}$$

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

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

$$\begin{aligned} Y_{ijk} = \mu + \omega_i\beta_0 &+ (1-\omega_i)A_{0i} + \omega_i A_{1i} \\ &+ (1-\omega_i)B_{0j} + \omega_i B_{1j} \\ &+ (1-\omega_i)S_{0k} + \omega_i S_{1k} \\ &+ AS_{ik} + e_{ijk} \end{aligned}$$

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

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

$$\begin{aligned} Y_{ijk} = \mu + \omega_i\beta_0 &+ (1-\omega_i)A_{0i} + \omega_i A_{1i} \\ &+ (1-\omega_i)B_{0j} + \omega_i B_{1j} \\ &+ (1-\omega_i)S_{0k} + \omega_i S_{1k} \\ &+ AS_{ik} + e_{ijk} \end{aligned}$$

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

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:

$$R_0 = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma^2_{A0} + \sigma^2_{S0} + \sigma^2_{AS}}{\sigma^2_{A0} + \sigma^2_{B0} + \sigma^2_{S0} + \sigma^2_{AS} + \sigma^2_e}$$

IRR for Internal vs. External Applicants (Model 3)

Within-school IRR:

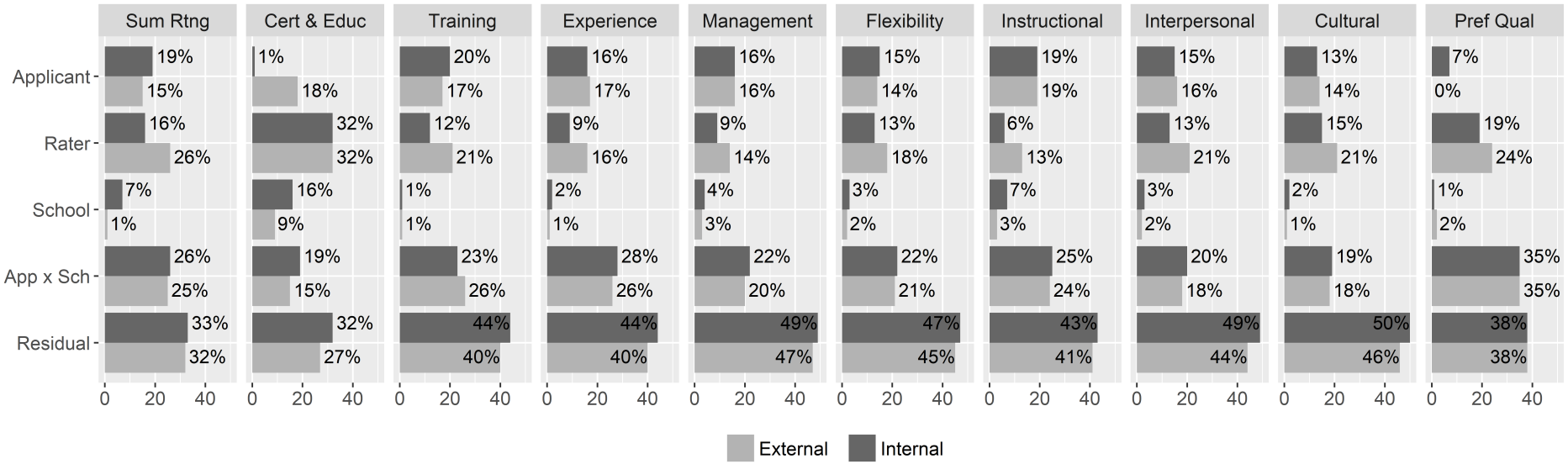
- For internal applicant :

$$R_1 = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma^2_{A1} + \sigma^2_{S1} + \sigma^2_{AS}}{\sigma^2_{A1} + \sigma^2_{B1} + \sigma^2_{S1} + \sigma^2_{AS} + \sigma^2_e}$$

- For external applicant:

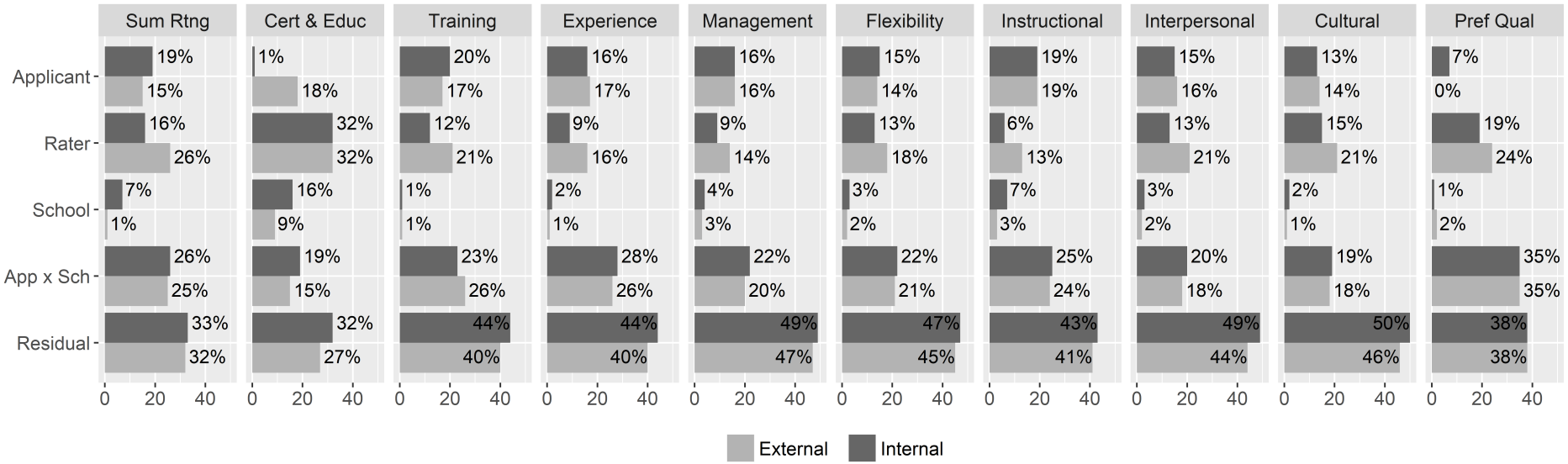
$$R_0 = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma^2_{A0} + \sigma^2_{S0} + \sigma^2_{AS}}{\sigma^2_{A0} + \sigma^2_{B0} + \sigma^2_{S0} + \sigma^2_{AS} + \sigma^2_e}$$

Results: Variance decomposition (Model 3)



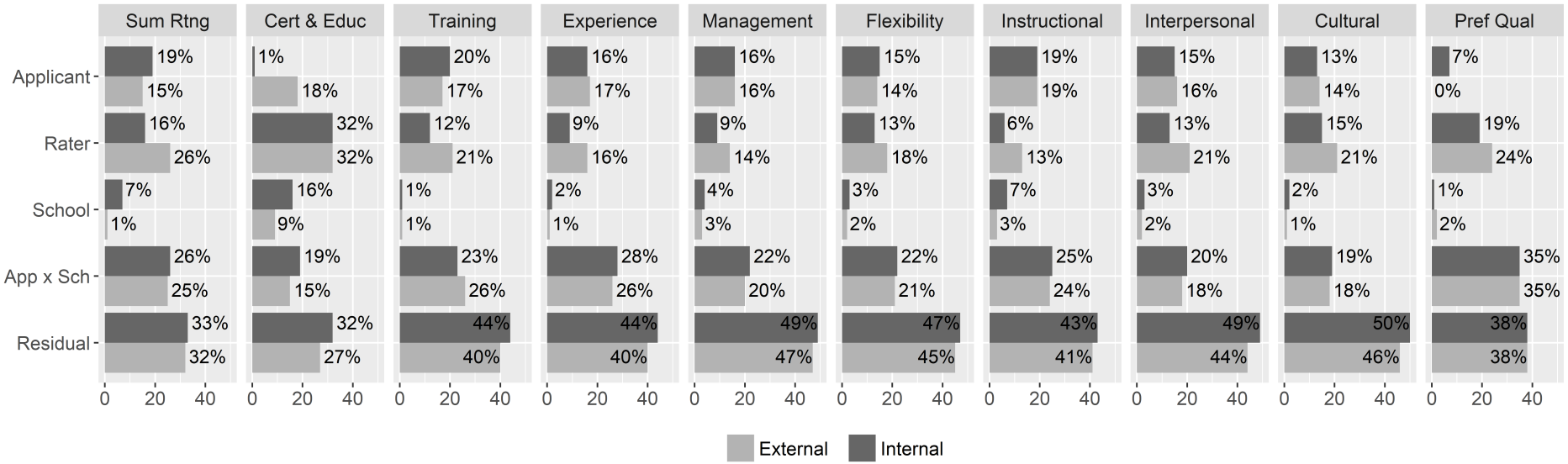
- High applicant-school variability
- Lower applicant variability for external applicants
- Higher rater variability for external applicants
- Lower inter-rater reliability for external applicants

Results: Variance decomposition (Model 3)



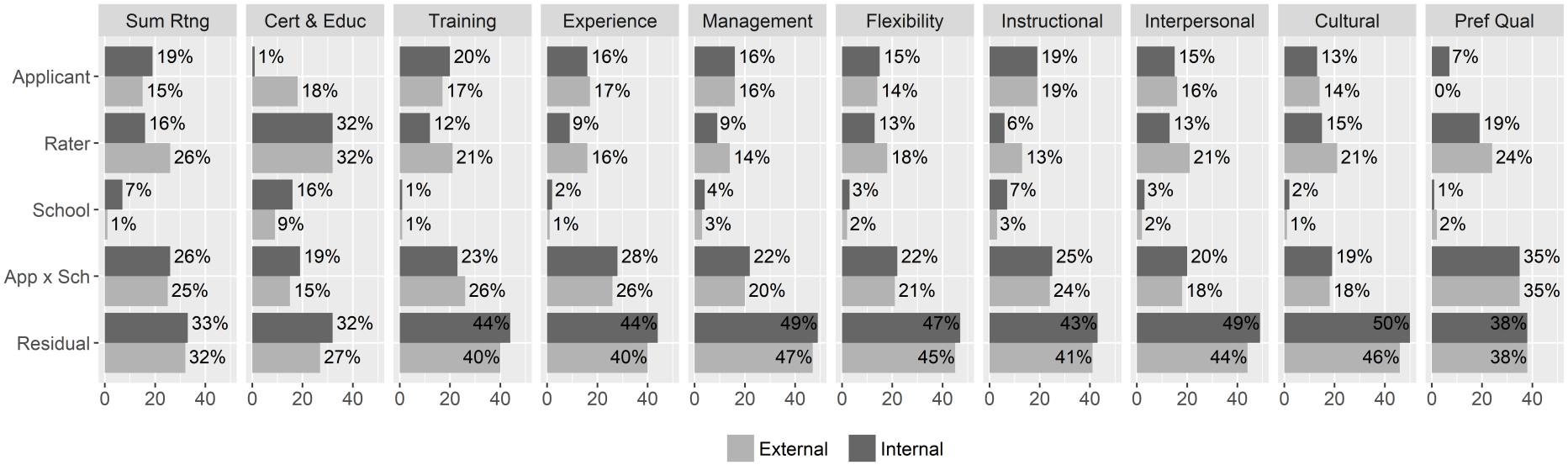
- High applicant-school variability
- Lower applicant variability for external applicants
- Higher rater variability for external applicants
- Lower inter-rater reliability for external applicants

Results: Variance decomposition (Model 3)



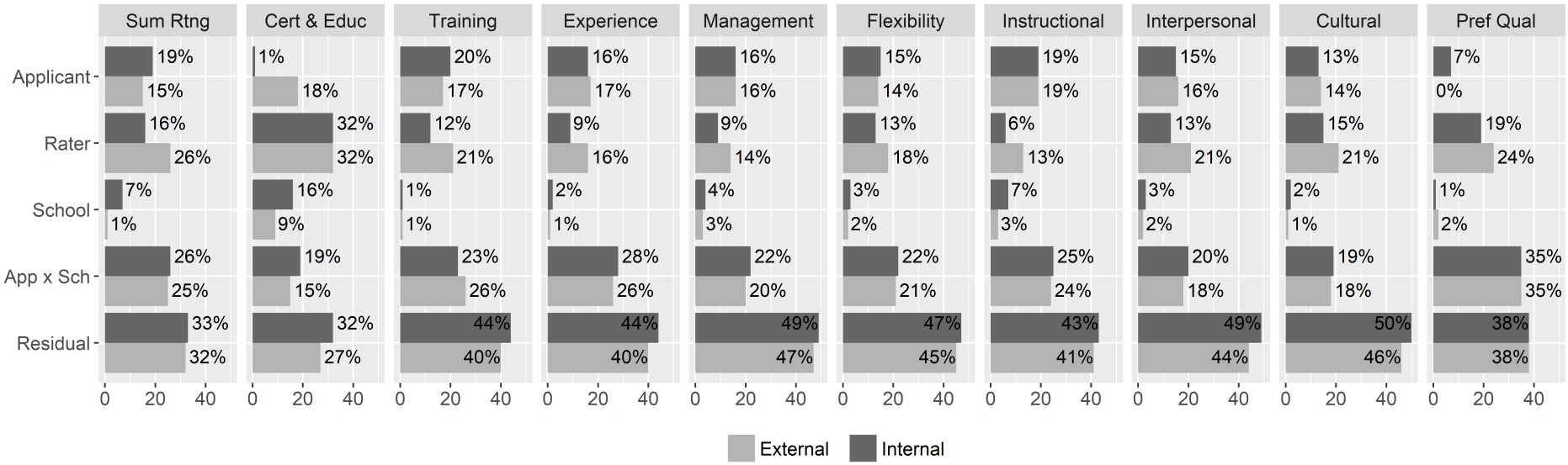
- High applicant-school variability
- Lower applicant variability for external applicants
- Higher rater variability for external applicants
- Lower inter-rater reliability for external applicants

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

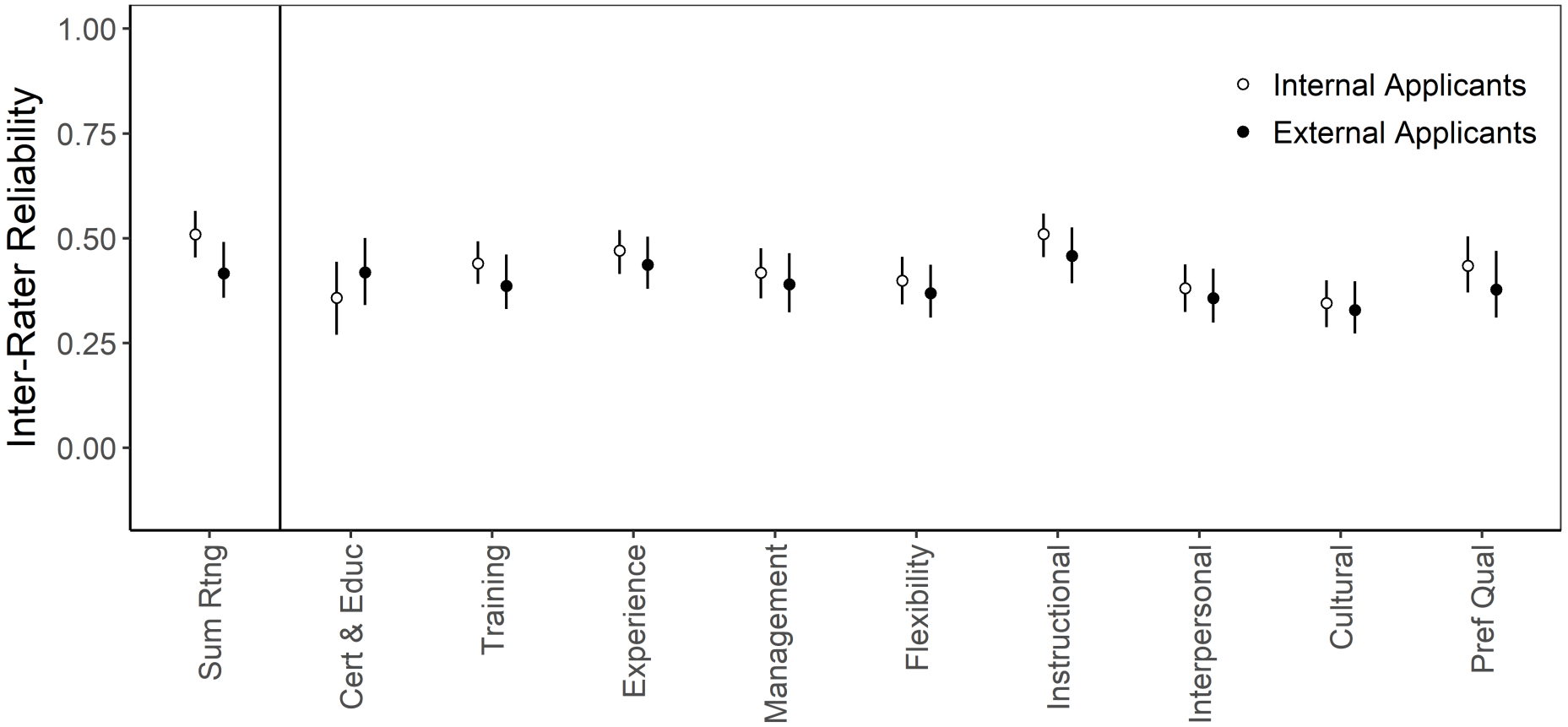
Results: Variance decomposition (Model 3)



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

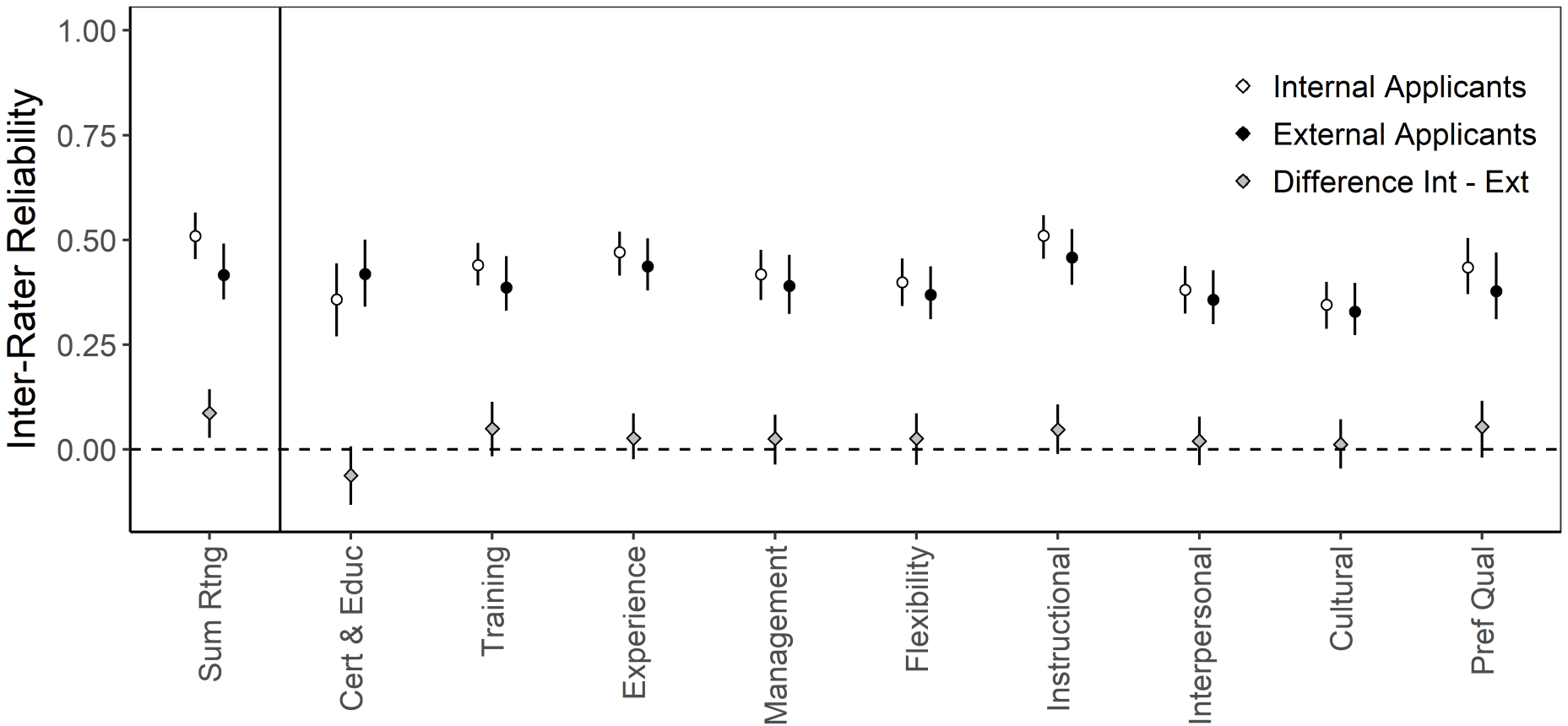
IRR for Internal and External Applicants (Model 3)

- 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*
- Singificantly 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*
- Singificantly 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*
- Singificantly 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 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.

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 aaplicable 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 aaplicable 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 aaplicable 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 aaplicable to grant or journal peer-review

Thank you for your attention!

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

martinkova@cs.cas.cz