Hierarchical Models for Assessing Inter-Rater Reliability in Teacher Hiring

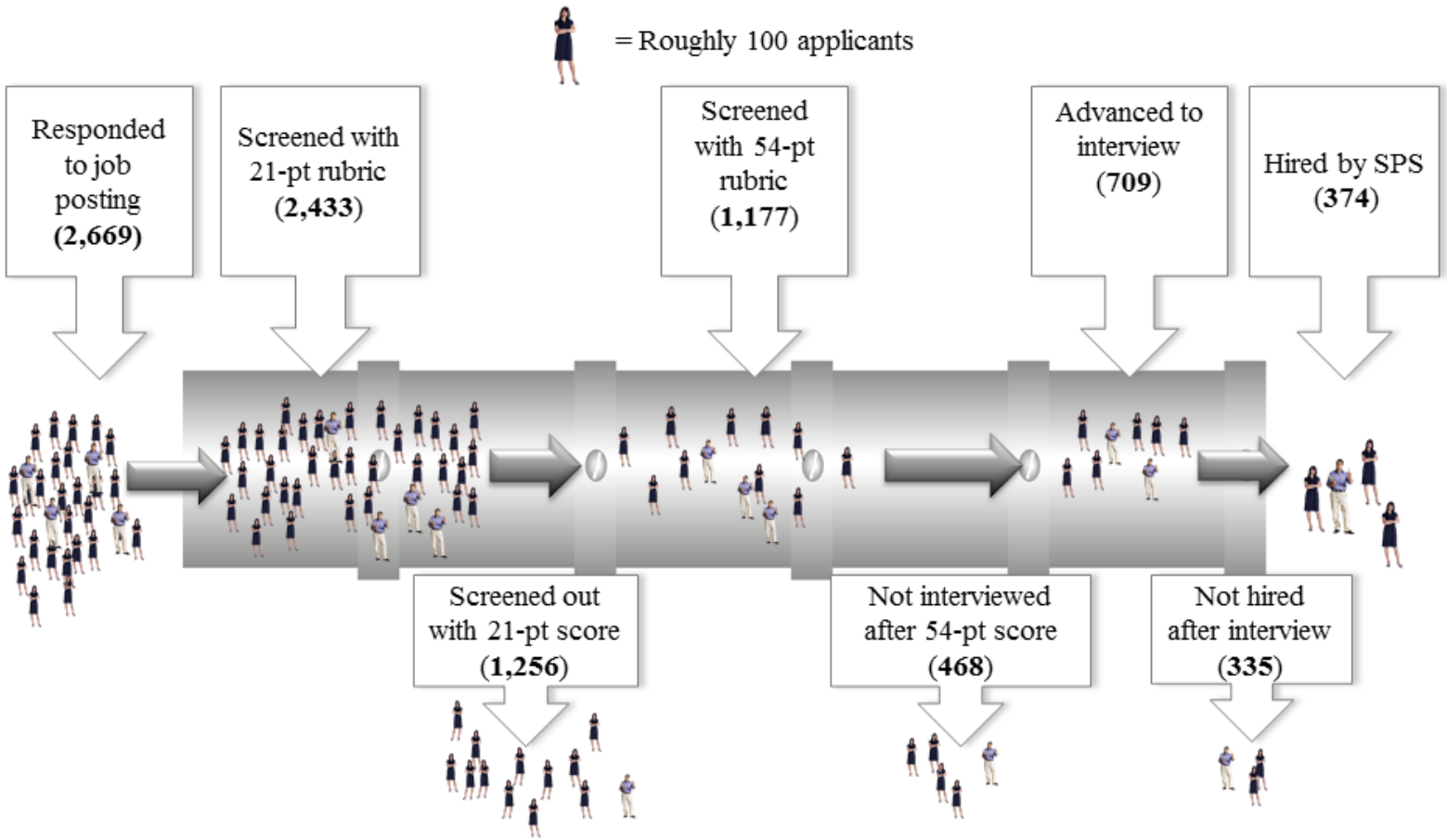
Patricia Martinkova[1,2] & Dan Goldhaber[3]

[1]Institute of Computer Science, Czech Academy of Sciences
[2]Dept. of Statistics & CSSS, University of Washington
[3]CEDR, University of Washington, Bothell

JSM 2015, August 10

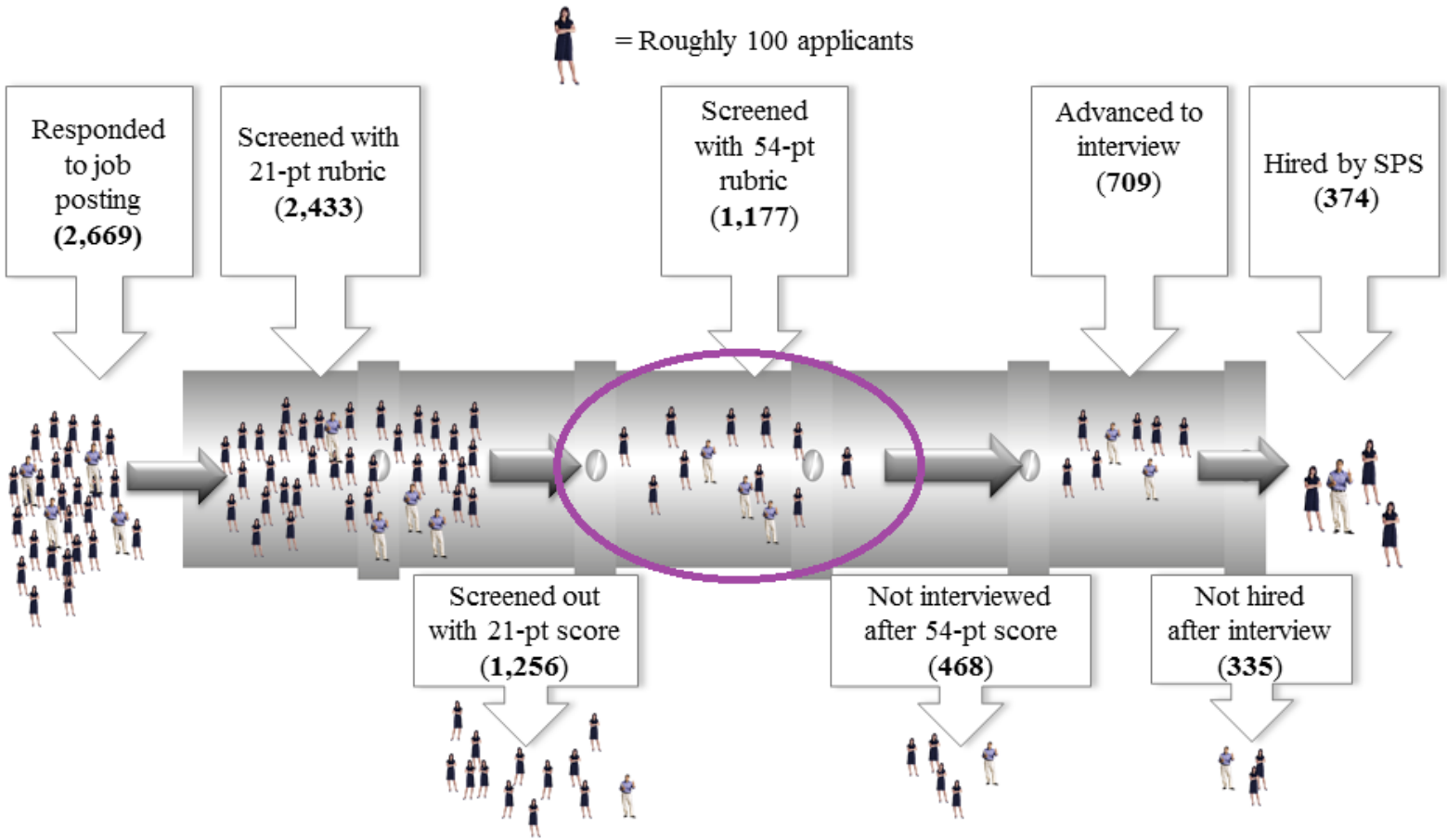
Outline

Motivation: Teacher Hiring Data



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

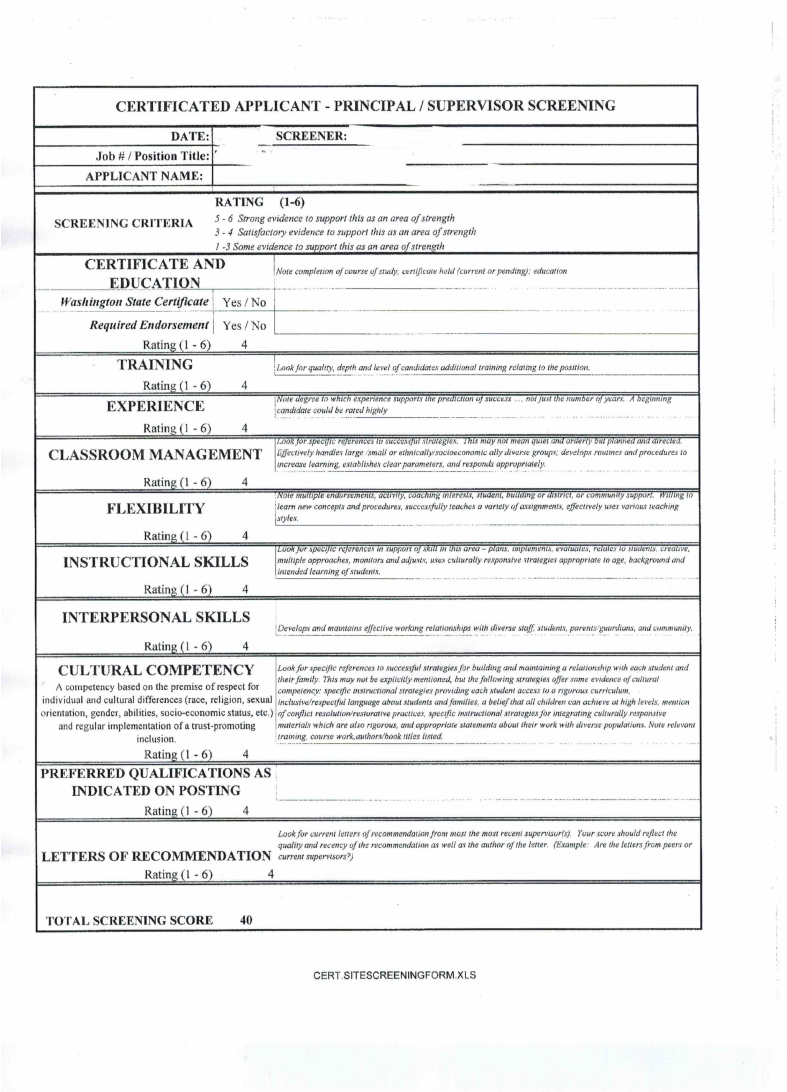
Motivation: Teacher Hiring Data



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

Teacher Hiring Data: 54-Pt Screening Rubric

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



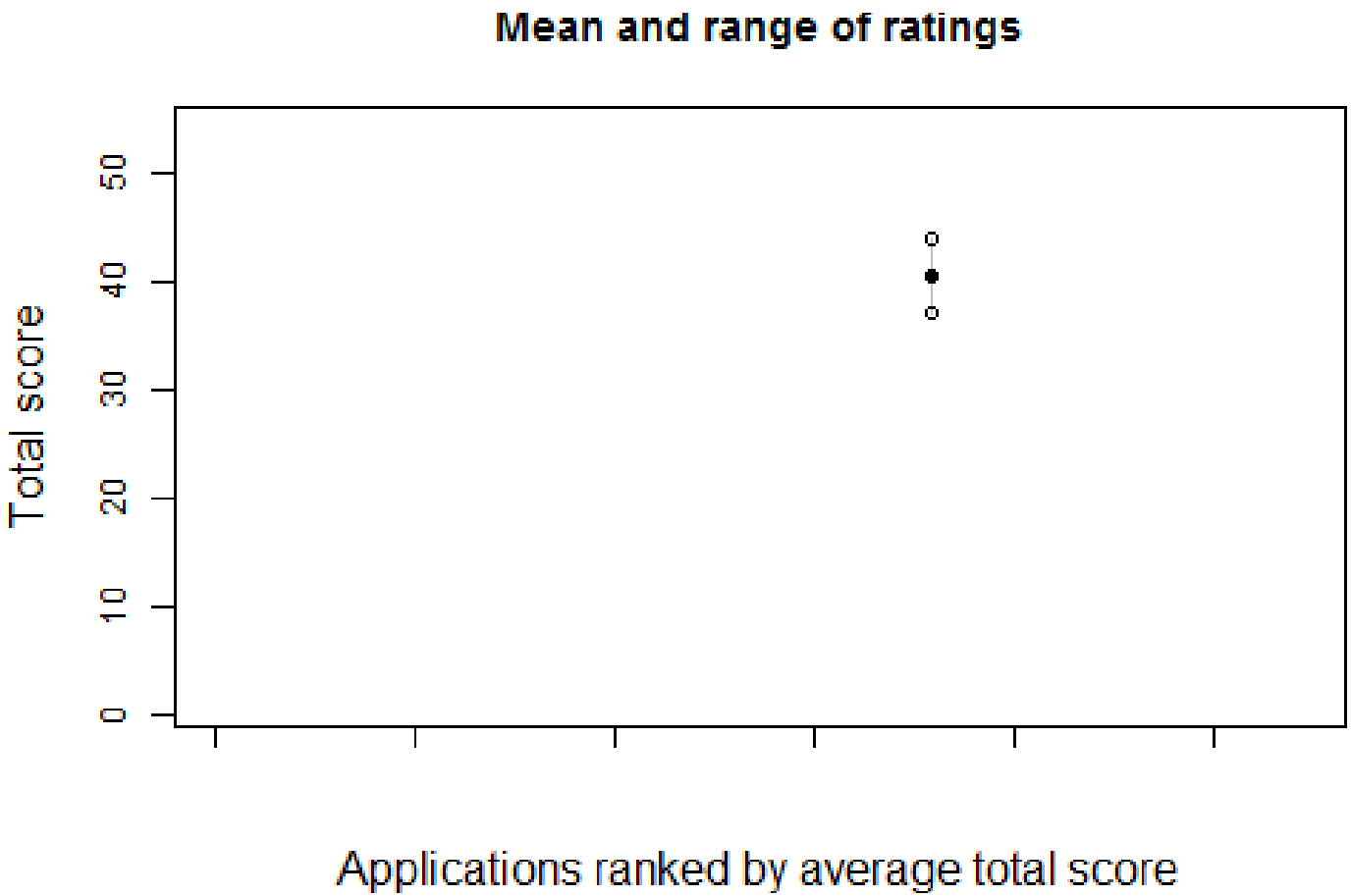
CERTIFICATED APPLICANT - PRINCIPAL / SUPERVISOR SCREENING

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/socioeconomic 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

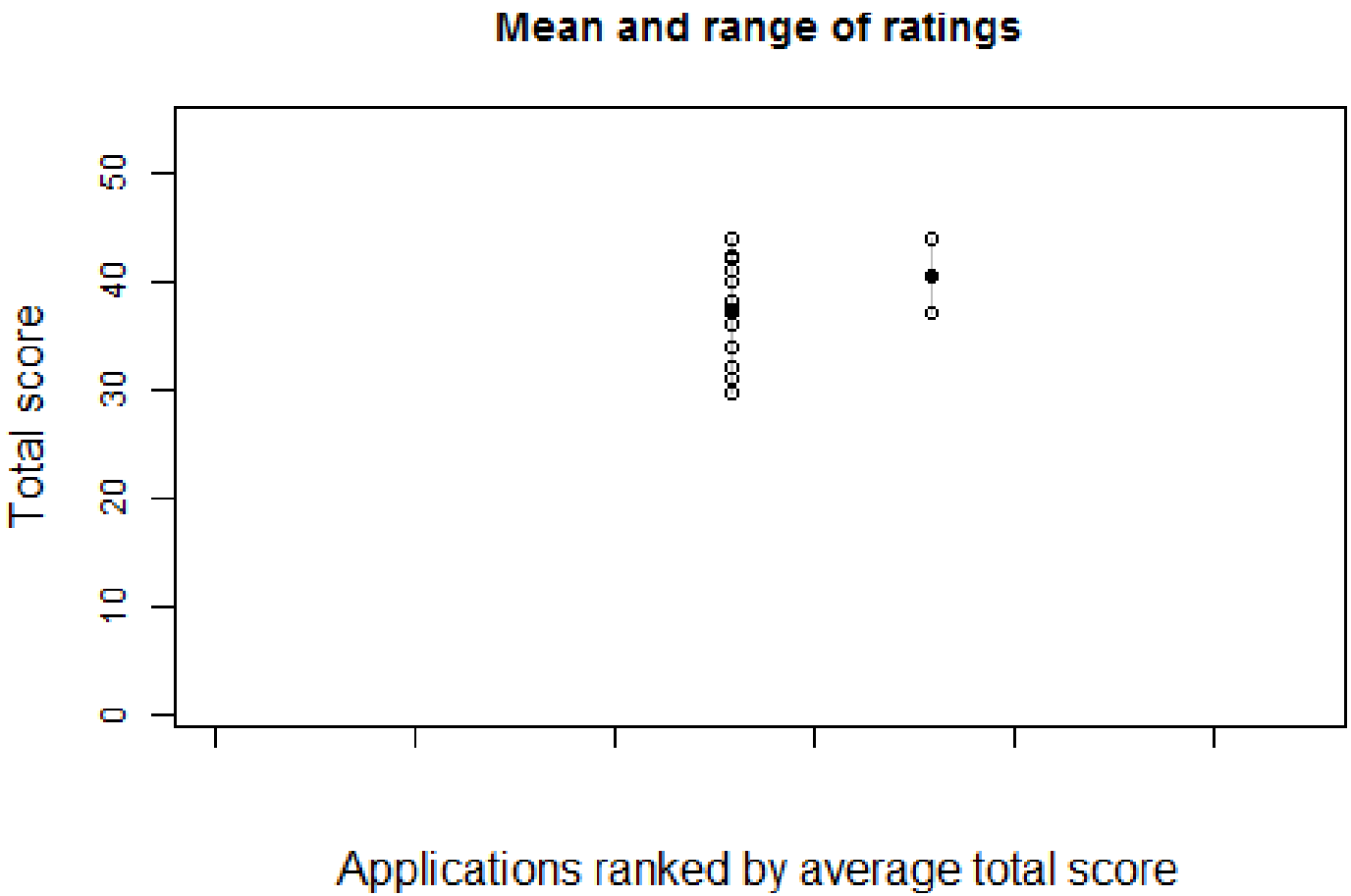
Aim of the screening rubric: To predict teacher quality

Ratings of Single Applicant (2008/09 - 2012/13)



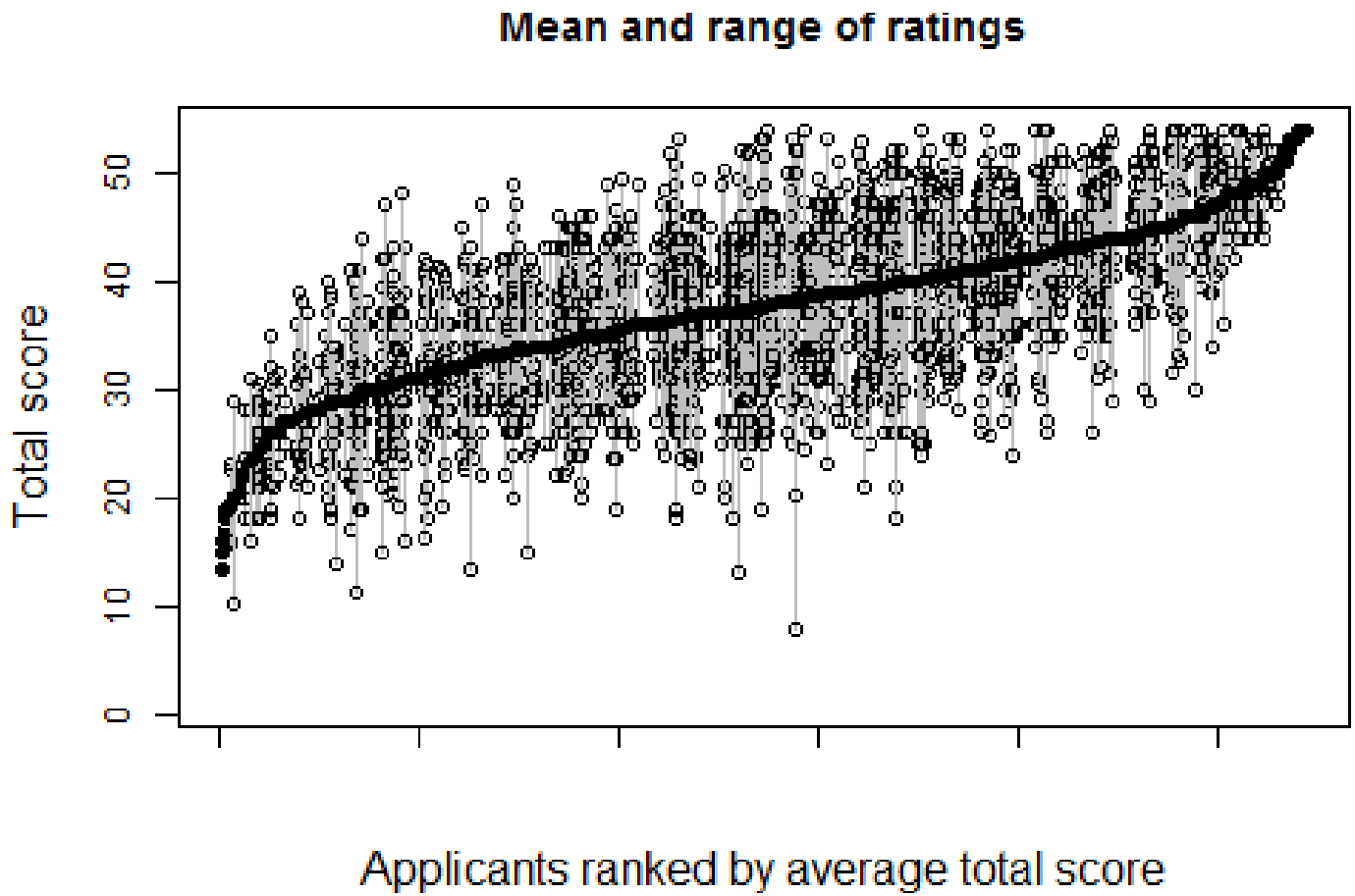
Are the ratings consistent?

Ratings of Single Applicant (2008/09 - 2012/13)



Are the ratings consistent?

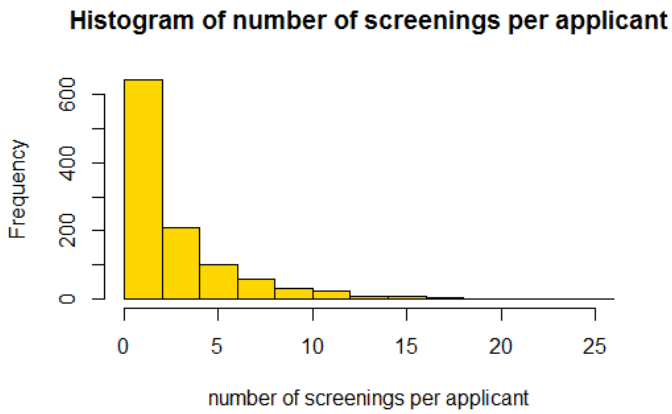
Ratings of All Applicants (2008/09 - 2012/13)

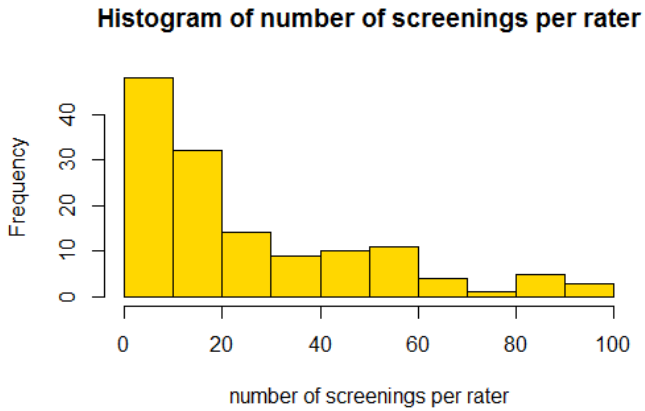


What is causing the inconsistencies in rating?

Teacher Hiring Data

- 3986 ratings (filled forms)
- 1177 applicants
  - rated 1-25 times
  - rated for 1-17 schools
  - internal and external
- by 141 raters
  - rated 1-99 times
  - rated applicants for 1-8 schools
- at 54 schools
  - elementary, middle, high
- for 526 job openings
  - 15 types of classroom jobs
  - Grade teacher, Math, English, Science, Social Studies, ...





Aims of this Study:

1. **Estimate:** Enumerate the inconsistencies
   - Inter-rater reliability (IRR)
   - Account for different schools, different job openings, ...
   - Compare IRR for subcomponents

2. **Test:** What is driving the inconsistencies in ratings?
   - School-applicant matching effect? Job-applicant matching effect?
   - Is IRR smaller in external applicants?
   - Is IRR smaller in some job types?

3. **Implications:** What is the impact of averaging ratings of more raters
   - How average of higher number of raters increases IRR
   - How higher IRR increases predictive power (measured by teacher value added)

Outline

Classical test theory model

- Classical test theory considers subject with a given *true score*
- Measurements of the true score are imprecise
- Assume simple model

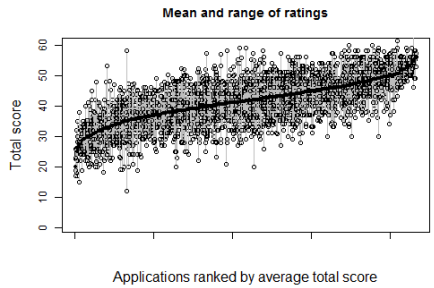
$$Y_{ij} = \mu + A_i + B_j + e_{ij} \tag{1}$$

  - Y_{ij} observed ratings
  - μ overall mean
  - $\mu + A_i \sim \mathrm{N}(\mu, \sigma_A^2)$ applicant's true score
  - $B_j \sim \mathrm{N}(0, \sigma_B^2)$ rater effect
  - $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$ random error
  - A_i, B_j and e_{ij} uncorrelated

Inter-Rater Reliability

Reliability is generally defined as

$$\text{reliability} = \frac{\text{variance of true scores}}{\text{variance of observed scores}}$$



- In model (1) $Y_{ij} = \mu + A_i + B_j + e_{ij}$
  Inter-rater reliability:

$$R = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_e^2}$$

  Note: this is just the intraclass correlation coefficient

- $\mathrm{R} \in [0, 1]$, low values mean a lot of measurement error
  - No universal heuristics, in high stakes testing $R > 0.8$ recommended

Inter-Rater Reliability: Why it Matters

Low reliability implies:

- attenuation of correlations:

$$\text{cor}(A_1 + B_1 + e_1, A_2 + B_2 + e_2) = \text{cor}(A_1, A_2)\sqrt{R_1 R_2}$$

- higher standard error of measurement
- wider confidence intervals
- less powerful hypotheses tests

Reliability of aggregates (average of J raters) is higher:

$$R_n = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2/J + \sigma_e^2/J}$$

Inter-Rater Reliability: Estimation

Traditional methods (balanced designs needed):

- correlation-based
- ANOVA-based

Our approach: More flexible estimation using hierarchical linear models

- restricted maximum likelihood using `lme4` in `R`
- parametric bootstrapping using `bootMer`
- model selection using BIC

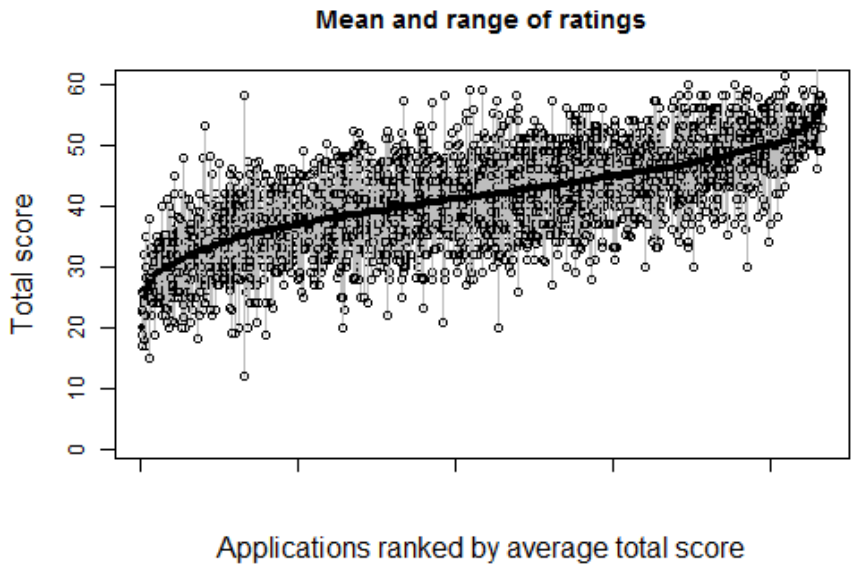
Outline

Inter-Rater Reliability across Schools

- Model 1: applicant and rater effect only

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

  - **inter-rater reliability** across schools: $R_{inter} = \frac{\sigma_A^2}{\sigma_A^2+\sigma_B^2+\sigma_e^2}$

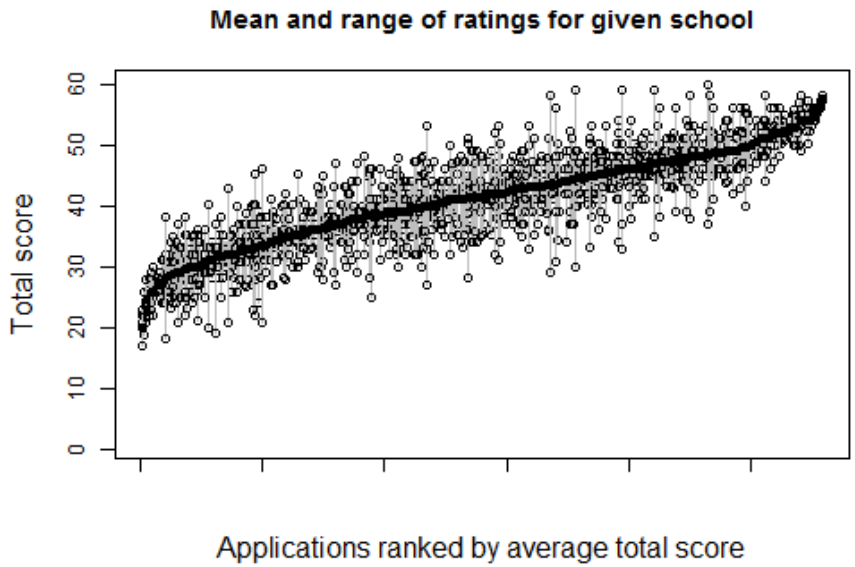


Inter-Rater Reliability within Schools

- Model 2: applicant differently suited for given school k

$$Y_{ijk} = \mu + A_i + B_j + AS_{ik} + e_{ijk} \tag{2}$$

  - **inter-rater reliability** within schools: $R_{inter} = \frac{\sigma_A^2 + \sigma_{AS}^2}{\sigma_A^2 + \sigma_{AS}^2 + \sigma_B^2 + \sigma_e^2}$

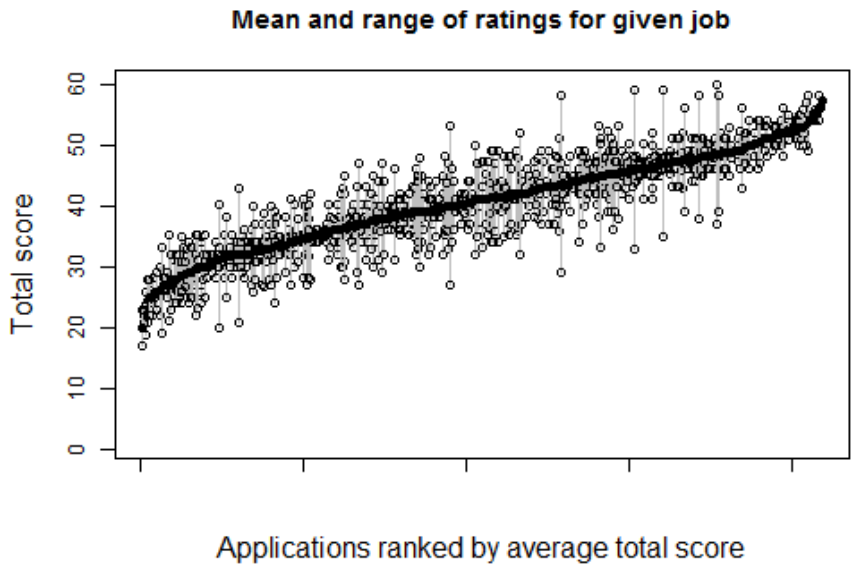


Inter-Rater Reliability within Job Openings

- Model 3: applicant differently suited for given job opening l

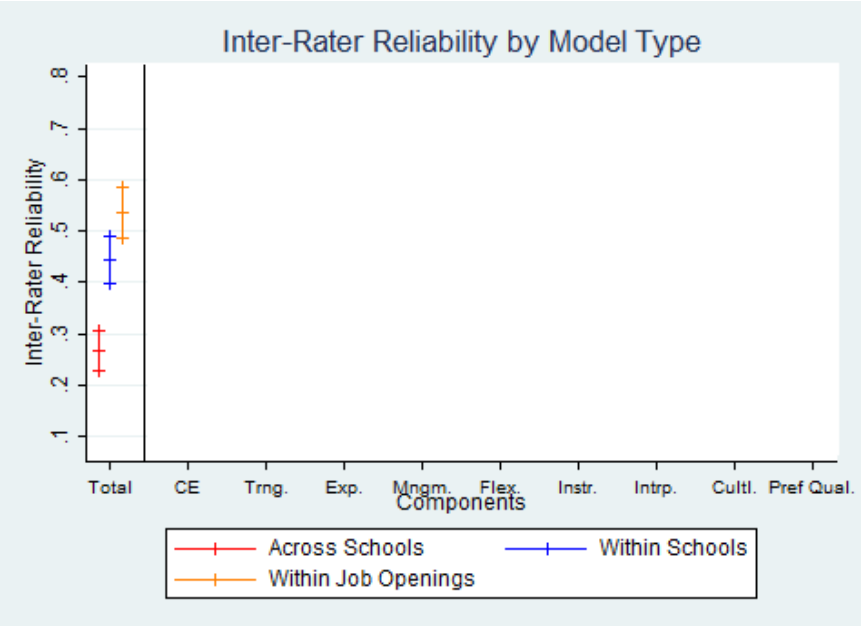
$$Y_{ijkl} = \mu + A_i + B_j + AS_{ik} + AJ_{il} + e_{ijkl} \tag{3}$$

  - **inter-rater reliability** within job openings: $R_{inter} = \frac{\sigma_A^2+\sigma_{AS}^2+\sigma_{AJ}^2}{\sigma_A^2+\sigma_{AS}^2+\sigma_{AJ}^2+\sigma_B^2+\sigma_e^2}$



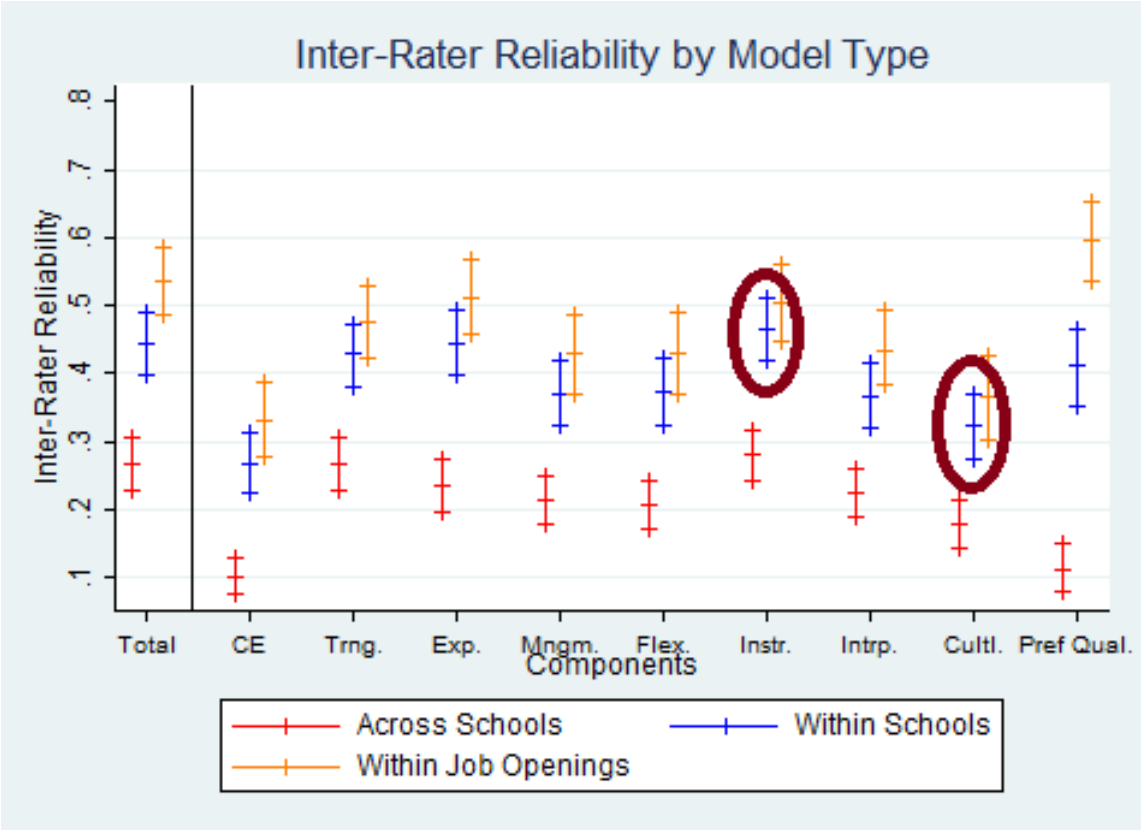
Hierarchical Models: Model Comparison

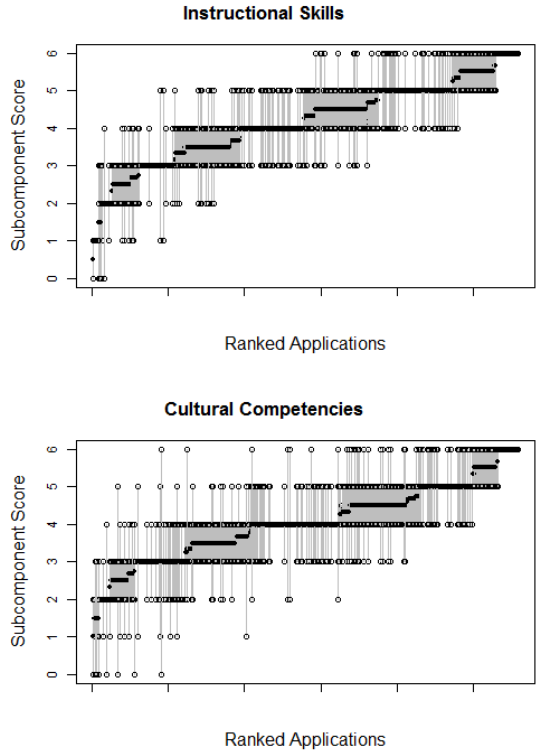
Model	Description	df	BIC
Model 1	Applicant and Rater effect only	4	20722.08
Model 2	Applicant:school interaction	5	20616.94
Model 3	Applicant:jobID interaction	6	**20592.74**



Conclusion: Applicants' qualities are school and job specific.

Inter-Rater Reliability of Subcomponents



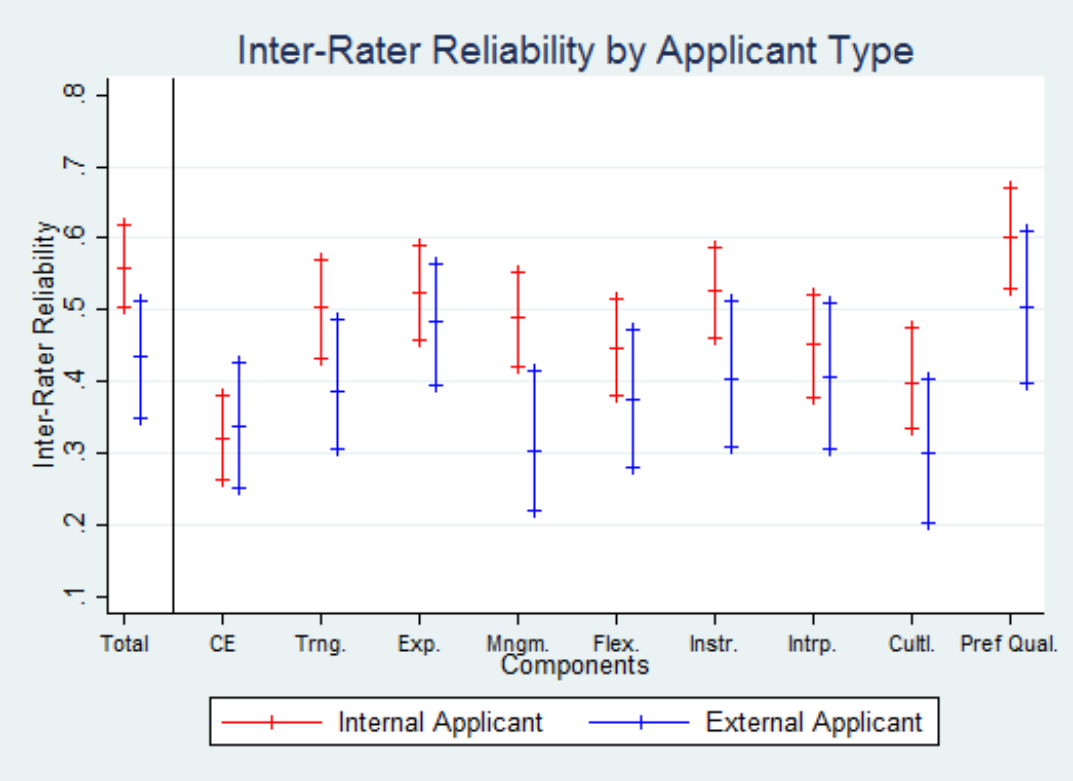


- For all subcomponents, the applicant qualities are school specific.
- For some (e.g. Pref. Qual) also job-specific.
- Some subcomponents are less reliable than others.

Difference in IRR between groups: Internal vs. External

Model 4: group effect, variance components vary by group

$$Y_{ijkl} = \mu + \omega_i \beta_{A1} + \omega_i A_{0i} + (1 - \omega_i) A_{1i} + \ldots \quad (4)$$



- These models provide better fit (BIC) for all subcomponents

Outline

Increasing IRR and Implications for Predictive Power

Increasing reliability by averaging ratings:

- IRR can be increased by averaging higher number of raters (J=2, 3)
- Two raters enough to raise IRR to 0.65 on some subcomponents *(Experience, Instructional, Pref. Qualifications)*
- Three raters enough to increase IRR to 0.80

Direct impact on predictive power of the rubric:

- Predictive power measured by correlation with teacher value added
- High reliability is necessary but not sufficient for high correlation w/ VA *(Instructional* vs. *Management)*
- Averaging ratings of two raters increases correlation of about 20%

Outline

Teacher Hiring Data: Questions and Answers

- What drives the inconsistencies in ratings?
  - Applicant's qualities are school and job specific.
- Are ratings more consistent in some *items*?
  - Ratings seem to be more consistent for some *better defined* items
  - Optimal weighting of items might be determined.
- Are the ratings more consistent in some *types of screening*?
  - Ratings in some subcomponents are more consistent in internal applicants
- How big is the impact of inconsistencies in ratings on ability of ratings to predict subsequent teacher quality?
  - Adding one rater would lead to increase about 20% in correlation with value added

Conclusion

- We suggest using HLM for more flexible estimation of inter-rater reliability
  - Restricted maximum likelihood with `lmer` in `lme4` in `R`
  - Parametric bootstrapping with `bootMer` in `lme4` in `R`
  - Model comparison using BIC
  - Interaction terms to test for applicant-school matching effect and applicant-job matching effect
    (IRR within schools, IRR within job openings)
  - Random slopes to test for differences in variance components for groups
    (different IRR for internal and external applicants)
- Possible further steps:
  - Ordinal models for subcomponents
  - Accounting for correlations between subcomponents

Thank you for your attention!

References:

- Martinkova & Goldhaber: Improving Teacher Selection: The Effect of Inter-Rater Reliability in the Screening Process. CEDR WP 2015-7. http://www.cedr.us/papers/working/CEDR WP 2015-7.pdf

- Goldhaber, Grout & Huntington-Klein: Screen Twice, Cut Once: Assessing the Predictive Validity of Teacher Selection Tools. CEDR WP 2014-9.