Flexible Analysis of Inter-Rater Reliability

As It Applies to Teacher Selection Instruments

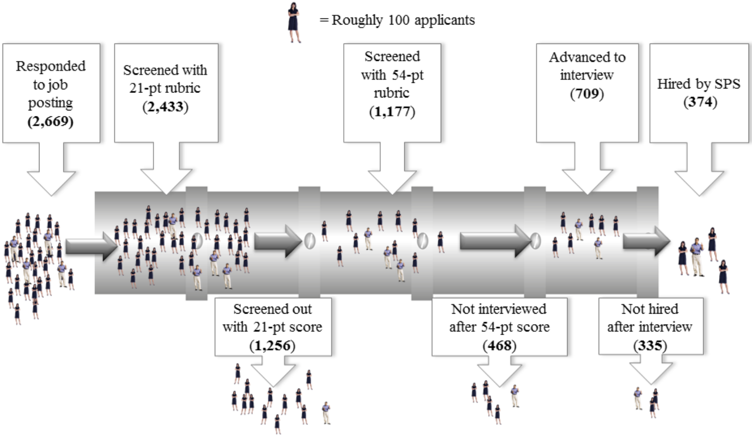
Patricia Martinkova[1], Dan Goldhaber[2] & Elena Erosheva[3]

[1]Institute of Computer Science, Czech Academy of Sciences
[2]Center for Education Data & Research, University of Washington, Bothell
[3]Dept. of Statistics, School of Social Work & CSSS, University of Washington

Robust, September 12, 2016

Outline

Motivation: Teacher Selection Process

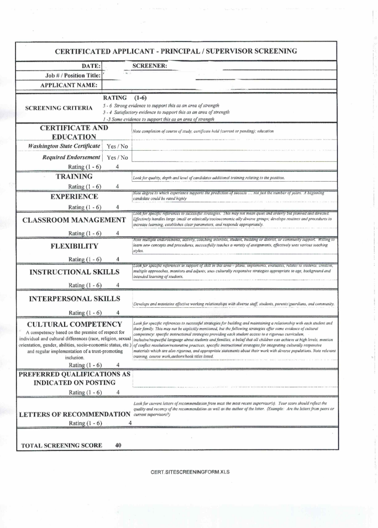


Applicants to classroom job openings in Spokane Public Schools during years (2008/09 - 2012/13)

Motivation: 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/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

Motivation: Questions

1. **Do we select the best applicants?**
   Do admission ratings predict subsequent teacher quality?
   - Goldhaber et al.

2. **Can we do better?**
   What causes error in ratings? How to eliminate the error?
   - Martinkova et al.

Motivation: Questions

1. **Do we select the best applicants?**

   Do admission ratings predict subsequent teacher quality?

   - Goldhaber et al.

2. **Can we do better?**

   What causes error in ratings? How to eliminate the error?

   - Martinkova et al.

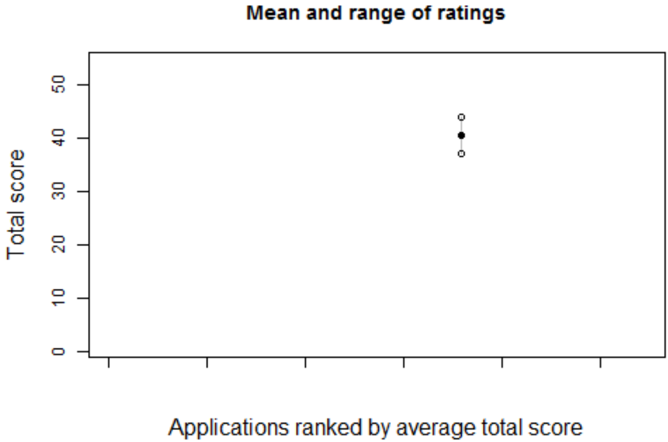
Motivation: Questions

1. **Do we select the best applicants?**
   Do admission ratings predict subsequent teacher quality?
   - Goldhaber et al.
2. **Can we do better?**
   What causes error in ratings? How to eliminate the error?
   - Martinkova et al.

Motivation: Questions

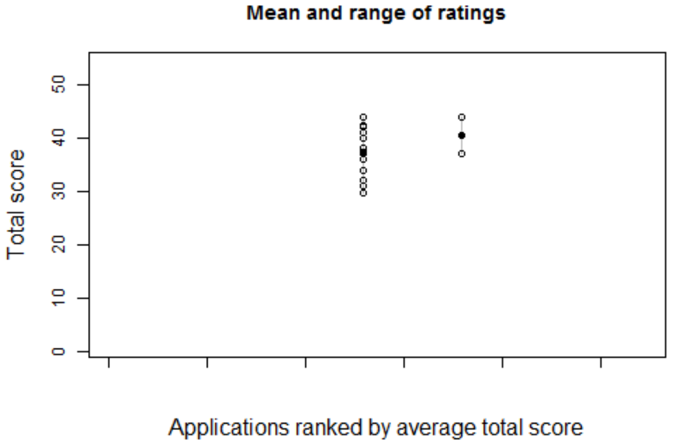
1. **Do we select the best applicants?**
   Do admission ratings predict subsequent teacher quality?
   - Goldhaber et al.
2. **Can we do better?**
   What causes error in ratings? How to eliminate the error?
   - Martinkova et al.

Ratings of a single applicant (2008/09 - 2012/13)



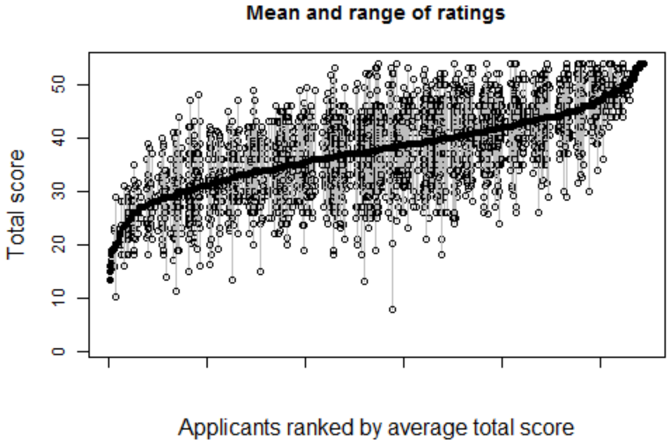
Are the ratings consistent?

Ratings of two applicants (2008/09 - 2012/13)



Are the ratings consistent?

Ratings of all applicants (2008/09 - 2012/13)



What is causing the inconsistencies in rating?

Reliability

- Consider subject with a given *true score* T_i
- Measurements Y_{ij} are imprecise: $Y_{ij} = T_i + e_{ij}$

Reliability is generally defined as

$$\mathrm{R} = \frac{\text{variance of true scores}}{\text{variance of observed scores}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}$$

Notes:

- 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
- Aggregates (average of J raters) have higher reliability: $R_n = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2/J}$

Reliability

- Consider subject with a given *true score* T_i
- Measurements Y_{ij} are imprecise: $Y_{ij} = T_i + e_{ij}$

Reliability is generally defined as

$$\mathrm{R} = \frac{\text{variance of true scores}}{\text{variance of observed scores}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}$$

Notes:

- 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
- Aggregates (average of J raters) have higher reliability: $R_n = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2/J}$

Reliability

- Consider subject with a given *true score* T_i
- Measurements Y_{ij} are imprecise: $Y_{ij} = T_i + e_{ij}$

Reliability is generally defined as

$$\mathrm{R} = \frac{\text{variance of true scores}}{\text{variance of observed scores}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}$$

Notes:

- 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
- Aggregates (average of J raters) have higher reliability: $R_n = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2/J}$

Reliability

- Consider subject with a given *true score* T_i
- Measurements Y_{ij} are imprecise: $Y_{ij} = T_i + e_{ij}$

Reliability is generally defined as

$$\mathrm{R} = \frac{\text{variance of true scores}}{\text{variance of observed scores}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}$$

Notes:

- 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
- Aggregates (average of J raters) have higher reliability: $R_n = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2/J}$

Reliability

- Consider subject with a given *true score* T_i
- Measurements Y_{ij} are imprecise: $Y_{ij} = T_i + e_{ij}$

Reliability is generally defined as

$$\mathrm{R} = \frac{\text{variance of true scores}}{\text{variance of observed scores}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}$$

Notes:

- 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
- Aggregates (average of J raters) have higher reliability: $R_n = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2/J}$

Reliability

Why it matters? Low reliability implies:

- attenuation of correlations (lower predictive power, lower validity)

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

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

How it can be estimated?

- In simple designs, R is usually estimated using mean squares
- Inference traditionally based on F statistics (McGraw & Wong, 1996)

Reliability

Why it matters? Low reliability implies:

- attenuation of correlations (lower predictive power, lower validity)

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

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

How it can be estimated?

- In simple designs, R is usually estimated using mean squares
- Inference traditionally based on F statistics (McGraw & Wong, 1996)

Reliability

Why it matters? Low reliability implies:

- attenuation of correlations (lower predictive power, lower validity)

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

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

How it can be estimated?

- In simple designs, R is usually estimated using mean squares
- Inference traditionally based on F statistics (McGraw & Wong, 1996)

Reliability

Why it matters? Low reliability implies:

- attenuation of correlations (lower predictive power, lower validity)

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

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

How it can be estimated?

- In simple designs, R is usually estimated using mean squares
- Inference traditionally based on F statistics (McGraw & Wong, 1996)

Reliability

Why it matters? Low reliability implies:

- attenuation of correlations (lower predictive power, lower validity)

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

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

How it can be estimated?

- In simple designs, R is usually estimated using mean squares
- Inference traditionally based on F statistics (McGraw & Wong, 1996)

Hiring data: Data structure

- 3986 filled forms
- 1177 applicants
  - internal and external
- 141 raters
  - various levels of experience
- 54 schools
  - 3 school types: elementary, middle, high
- 526 job openings
  - 15 types of jobs: grade teacher, math, English, science, ...

Aims of the study

- Estimate IRR while accounting for hierarchical data structure
  - schools, job openings, etc.
  - applicant-school matching, etc.

- Test for possible moderators of IRR
  - internal/external status of the applicant
  - rater experience

  (Conway et al, 1995: A Meta-Analysis of IRR of Selection Interviews)

- Apply this “model-based IRR” to analyze implications for validity
  - how IRR affects power to predict teacher value added

Aims of the study

- Estimate IRR while accounting for hierarchical data structure
  - schools, job openings, etc.
  - applicant-school matching, etc.

- Test for possible moderators of IRR
  - internal/external status of the applicant
  - rater experience

  (Conway et al, 1995: A Meta-Analysis of IRR of Selection Interviews)

- Apply this "model-based IRR" to analyze implications for validity
  - how IRR affects power to predict teacher value added

Aims of the study

- Estimate IRR while accounting for hierarchical data structure
  - schools, job openings, etc.
  - applicant-school matching, etc.
- Test for possible moderators of IRR
  - internal/external status of the applicant
  - rater experience

  (Conway et al, 1995: A Meta-Analysis of IRR of Selection Interviews)
- Apply this “model-based IRR” to analyze implications for validity
  - how IRR affects power to predict teacher value added

Aims of the study

- Estimate IRR while accounting for hierarchical data structure
  - schools, job openings, etc.
  - applicant-school matching, etc.
- Test for possible moderators of IRR
  - internal/external status of the applicant
  - rater experience

  (Conway et al, 1995: A Meta-Analysis of IRR of Selection Interviews)
- Apply this "model-based IRR" to analyze implications for validity
  - how IRR affects power to predict teacher value added

Outline

1. Introduction
2. **Hierarchical Models for Inter-Rater Reliability**
3. Moderators of Inter-Rater Reliability
4. Implications for Predictive Power
5. Conclusion

Inter-Rater Reliability (Assessee–Rater Model)

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

- assessee effect $A_i \sim \mathrm{N}(0, \sigma_A^2)$, rater effect $B_j \sim \mathrm{N}(0, \sigma_B^2)$, error $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$
- **Inter-Rater Reliability**:

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

- $\mathrm{R} \in [0, 1]$, low values mean a lot of measurement error
- Aggregate (average of J raters) has higher IRR: $R_n = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2/J + \sigma_e^2/J}$

Inter-Rater Reliability (Assessee–Rater Model)

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

- assessee effect $A_i \sim \mathrm{N}(0, \sigma_A^2)$, rater effect $B_j \sim \mathrm{N}(0, \sigma_B^2)$, error $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$
- **Inter-Rater Reliability**:

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

- $\mathrm{R} \in [0, 1]$, low values mean a lot of measurement error
- Aggregate (average of J raters) has higher IRR: $R_n = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2/J + \sigma_e^2/J}$

Inter-Rater Reliability (Assessee–Rater Model)

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

- assessee effect $A_i \sim \mathrm{N}(0, \sigma_A^2)$, rater effect $B_j \sim \mathrm{N}(0, \sigma_B^2)$, error $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$
- **Inter-Rater Reliability**:

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

- $\mathrm{R} \in [0, 1]$, low values mean a lot of measurement error
- Aggregate (average of J raters) has higher IRR: $R_n = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2/J + \sigma_e^2/J}$

Assessee-Rater-Unit Model

$$Y_{ijk} = \mu + A_i + B_j + S_k + AS_{ik} + AR_{ij} + BS_{jk} + e_{ijk}$$

- Unit (School) level $S_k \sim N(0, \sigma_S^2)$
- Applicant-unit matching effect (interaction) $AS_{ik} \sim N(0, \sigma_{AS}^2)$
- Interactions $AB_{ik} \sim N(0, \sigma_{AB}^2)$, $BS_{ik} \sim N(0, \sigma_{BS}^2)$
- **IRR** across schools:

$$R_{across} = \mathrm{cor}(Y_{ijk}, Y_{ij'k'}) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_S^2 + \sigma_{AS}^2 + \sigma_{AB}^2 + \sigma_{BS}^2 + \sigma_e^2}$$

- **IRR** within school:

$$R_{within} = \mathrm{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma_A^2 + \sigma_S^2 + \sigma_{AS}^2}{\sigma_A^2 + \sigma_B^2 + \sigma_S^2 + \sigma_{AS}^2 + \sigma_{AB}^2 + \sigma_{BS}^2 + \sigma_e^2}$$

Assessee-Rater-Unit Model

$$Y_{ijk} = \mu + A_i + B_j + S_k + AS_{ik} + AR_{ij} + BS_{jk} + e_{ijk}$$

- Unit (School) level $S_k \sim N(0, \sigma_S^2)$
- Applicant-unit matching effect (interaction) $AS_{ik} \sim N(0, \sigma_{AS}^2)$
- Interactions $AB_{ik} \sim N(0, \sigma_{AB}^2)$, $BS_{ik} \sim N(0, \sigma_{BS}^2)$
- **IRR** across schools:

$$R_{across} = \mathrm{cor}(Y_{ijk}, Y_{ij'k'}) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_S^2 + \sigma_{AS}^2 + \sigma_{AB}^2 + \sigma_{BS}^2 + \sigma_e^2}$$

- **IRR** within school:

$$R_{within} = \mathrm{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma_A^2 + \sigma_S^2 + \sigma_{AS}^2}{\sigma_A^2 + \sigma_B^2 + \sigma_S^2 + \sigma_{AS}^2 + \sigma_{AB}^2 + \sigma_{BS}^2 + \sigma_e^2}$$

Assessee-Rater-Unit Model

$$Y_{ijk} = \mu + A_i + B_j + S_k + AS_{ik} + AR_{ij} + BS_{jk} + e_{ijk}$$

- Unit (School) level $S_k \sim N(0, \sigma_S^2)$
- Applicant-unit matching effect (interaction) $AS_{ik} \sim N(0, \sigma_{AS}^2)$
- Interactions $AB_{ik} \sim N(0, \sigma_{AB}^2)$, $BS_{ik} \sim N(0, \sigma_{BS}^2)$
- **IRR** across schools:

$$R_{across} = \mathrm{cor}(Y_{ijk}, Y_{ij'k'}) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_S^2 + \sigma_{AS}^2 + \sigma_{AB}^2 + \sigma_{BS}^2 + \sigma_e^2}$$

- **IRR** within school:

$$R_{within} = \mathrm{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma_A^2 + \sigma_S^2 + \sigma_{AS}^2}{\sigma_A^2 + \sigma_B^2 + \sigma_S^2 + \sigma_{AS}^2 + \sigma_{AB}^2 + \sigma_{BS}^2 + \sigma_e^2}$$

Assessee-Rater-Unit Model

$$Y_{ijk} = \mu + A_i + B_j + S_k + AS_{ik} + AR_{ij} + BS_{jk} + e_{ijk}$$

- Unit (School) level $S_k \sim N(0, \sigma^2_S)$
- Applicant-unit matching effect (interaction) $AS_{ik} \sim N(0, \sigma^2_{AS})$
- Interactions $AB_{ik} \sim N(0, \sigma^2_{AB})$, $BS_{ik} \sim N(0, \sigma^2_{BS})$
- **IRR** across schools:

$$R_{across} = \mathrm{cor}(Y_{ijk}, Y_{ij'k'}) = \frac{\sigma^2_A}{\sigma^2_A + \sigma^2_B + \sigma^2_S + \sigma^2_{AS} + \sigma^2_{AB} + \sigma^2_{BS} + \sigma^2_e}$$

- **IRR** within school:

$$R_{within} = \mathrm{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma^2_A + \sigma^2_S + \sigma^2_{AS}}{\sigma^2_A + \sigma^2_B + \sigma^2_S + \sigma^2_{AS} + \sigma^2_{AB} + \sigma^2_{BS} + \sigma^2_e}$$

IRR estimation and inference

More flexible estimation using linear random-effect models

- Estimation w/ restricted maximum likelihood using `lmer` in `lme4` in `R`
- Model selection using AIC, BIC, likelihood ratio tests
- Confidence intervals w/ MCMC using `brms` (or bootstrap: `bootMer`)

```
library(brms)
model2 <- brm(total~1+(1|Apl)+(1|Rtr)+(1|Sch)+
+(1|Apl:Sch)+(1|Rtr:Sch)+(1|Apl:Rtr), data=screening)
results <- as.matrix(model2)

IRR_across <- results[,2]/apply(results[,2:8],1,sum)

IRRa_LCL <- quantile(IRR_across, 0.025)
IRRa_UCL <- quantile(IRR_across, 0.975)
```

IRR estimation and inference

More flexible estimation using linear random-effect models

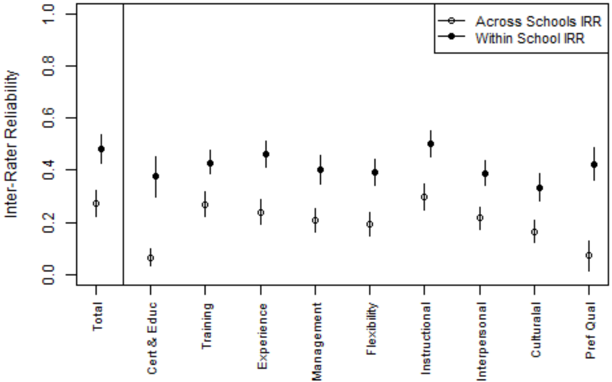
- Estimation w/ restricted maximum likelihood using `lmer` in `lme4` in `R`
- Model selection using AIC, BIC, likelihood ratio tests
- Confidence intervals w/ MCMC using `brms` (or bootstrap: `bootMer`)

```
library(brms)
model2 <- brm(total~1+(1|Apl)+(1|Rtr)+(1|Sch)+
+(1|Apl:Sch)+(1|Rtr:Sch)+(1|Apl:Rtr), data=screening)
results <- as.matrix(model2)

IRR_across <- results[,2]/apply(results[,2:8],1,sum)

IRRa_LCL <- quantile(IRR_across, 0.025)
IRRa_UCL <- quantile(IRR_across, 0.975)
```

IRR within/across Schools - Results



- For all subcomponents, the applicant qualities are school specific.
- Some subcomponents are less reliable than others.

Outline

Assessee-Rater-Unit-Moderator Model

- 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

$$
\begin{aligned}
Y_{ijk} = \mu + \omega_i \beta_1 &+ (1-\omega_i)A_{0i} + \omega_i A_{1i} \\
&+ (1-\omega_i)B_{0j} + \omega_i B_{1j} \\
&+ (1-\omega_i)S_{0k} + \omega_i S_{1k} \\
&+ AS_{ik} + AB_{ij} + BS_{jk} + e_{ijk}
\end{aligned}
$$

  - $\omega_i = 1$ for internal and 0 for external applicants
  - group fixed effect β_1
  - $A_{0i} \sim N(0, \sigma^2_{A0})$ 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 $S_{1k} \sim N(0, \sigma^2_{S1})$

Assessee-Rater-Unit-Moderator Model

- 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

$$
\begin{aligned}
Y_{ijk} = \mu + \omega_i\beta_1 &+ (1-\omega_i)A_{0i} + \omega_i A_{1i} \\
&+ (1-\omega_i)B_{0j} + \omega_i B_{1j} \\
&+ (1-\omega_i)S_{0k} + \omega_i S_{1k} \\
&+ AS_{ik} + AB_{ij} + BS_{jk} + e_{ijk}
\end{aligned}
$$

  - $\omega_i = 1$ for internal and 0 for external applicants
  - group fixed effect β_1
  - $A_{0i} \sim N(0, \sigma^2_{A0})$ 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 $S_{1k} \sim N(0, \sigma^2_{S1})$

Assessee-Rater-Unit-Moderator Model

- 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

$$\begin{aligned} Y_{ijk} = \mu + \omega_i \beta_1 &+ (1-\omega_i)A_{0i} + \omega_i A_{1i} \\ &+ (1-\omega_i)B_{0j} + \omega_i B_{1j} \\ &+ (1-\omega_i)S_{0k} + \omega_i S_{1k} \\ &+ AS_{ik} + AB_{ij} + BS_{jk} + e_{ijk} \end{aligned}$$

  - $\omega_i = 1$ for internal and 0 for external applicants
  - group fixed effect β_1
  - $A_{0i} \sim N(0, \sigma^2_{A0})$ 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 $S_{1k} \sim N(0, \sigma^2_{S1})$

Moderator of IRR: Internal vs. External status (Model 3)

```
model <- lmer(rating ~ 1 + internal +
+ (0+internal|Apl) + (0+internal|Rtr) + (0+internal|Sch) +
+ (1|Apl:Sch) + (1|PID:rater) + (1|rater:school),
+ data=screening)
```

Within-school IRR:

- internal applicant :

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

- external applicant:

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

Moderator of IRR: Internal vs. External status (Model 3)

```
model <- lmer(rating ~ 1 + internal +
+ (0+internal|Apl) + (0+internal|Rtr) + (0+internal|Sch) +
+ (1|Apl:Sch) + (1|PID:rater) + (1|rater:school),
+ data=screening)
```

Within-school IRR:

- internal applicant :

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

- external applicant:

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

Moderator of IRR: Internal vs. External status (Model 3)

```
model <- lmer(rating ~ 1 + internal +
+ (0+internal|Apl) + (0+internal|Rtr) + (0+internal|Sch) +
+ (1|Apl:Sch) + (1|PID:rater) + (1|rater:school),
+ data=screening)
```

Within-school IRR:

- internal applicant :

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

- external applicant:

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

Model 3: Variance decomposition, IRR

Internal	b	SE(b)	Apl	Rtr	Sch	AS	RS	AR	Res.	Total	IRRw
Total	**3.35**	0.40	19%	**16%**	6%	26%	1%	0%	33%	60.61	**0.51**
Crt. Ed.	0.13	0.05	1%	9%	12%	20%	25%	0%	34%	1.12	0.33
Training	0.49	0.08	20%	9%	1%	22%	3%	2%	43%	1.65	0.43
Exper.	0.33	0.06	16%	9%	2%	28%	0%	2%	43%	1.39	0.46
Mngmnt	0.41	0.06	16%	7%	4%	20%	2%	4%	47%	1.29	0.40
Flexiblty	0.35	0.05	15%	13%	2%	21%	1%	4%	44%	1.23	0.38
Instruct.	0.47	0.06	19%	5%	6%	24%	2%	3%	41%	1.31	0.49
Interpers.	0.31	0.05	15%	11%	2%	17%	3%	8%	43%	1.14	0.35
Cultural	0.34	0.05	13%	14%	1%	17%	2%	5%	47%	1.38	0.32
Pref.Q.	0.47	0.09	7%	16%	0%	35%	3%	0%	38%	2.36	0.42
External	**b**	**SE(b)**	**Apl**	**Rtr**	**Sch**	**AS**	**RS**	**AR**	**Res.**	**Total**	**IRRw**
Total			15%	**26%**	1%	25%	1%	0%	32%	62.60	**0.41**
Crt. Ed.			18%	14%	3%	16%	20%	0%	28%	1.36	0.38
Training			17%	19%	1%	20%	3%	2%	39%	1.83	0.38
Exper.			17%	16%	1%	25%	0%	2%	39%	1.53	0.43
Mngmnt			16%	13%	3%	19%	2%	3%	45%	1.36	0.38
Flexiblty			14%	18%	1%	20%	1%	3%	43%	1.28	0.36
Instruct.			19%	12%	2%	23%	2%	3%	39%	1.37	0.45
Interpers.			16%	19%	1%	16%	2%	7%	39%	1.28	0.33
Cultural			15%	19%	0%	16%	2%	5%	43%	1.51	0.31
Pref.Q.			0%	21%	2%	35%	3%	0%	38%	2.33	0.37

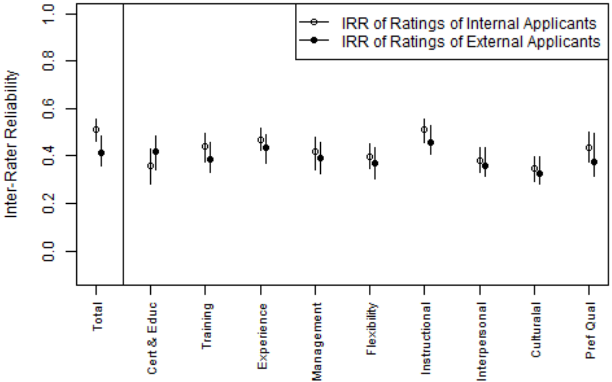
Model comparison (BIC)

Assessee-Rater-Unit-Moderator model (3) provides the best fit for all subcomponents

	model 1	model 2	model 3
Total	23,204	23,072	**22,954**
Certificate and Education	8,515	8,371	**8,336**
Training	11,050	10,981	**10,886**
Experience	10,561	10,467	**10,426**
Management	10,239	10,176	**10,093**
Flexibility	9,974	9,897	**9,838**
Instructional	10,271	10,167	**10,090**
Interpersonal	9,740	9,677	**9,643**
Cultural	10,370	10,322	**10,270**
Preferred Qualifications	9,073	8,965	**8,908**

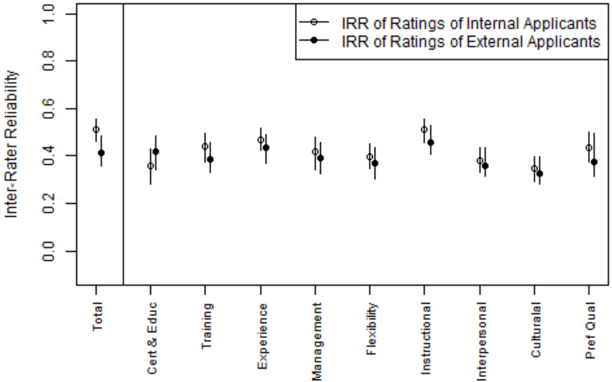
IRR for Internal and External Applicants (Model 3)

- IRR is estimated simultaneously for both groups within Model 3



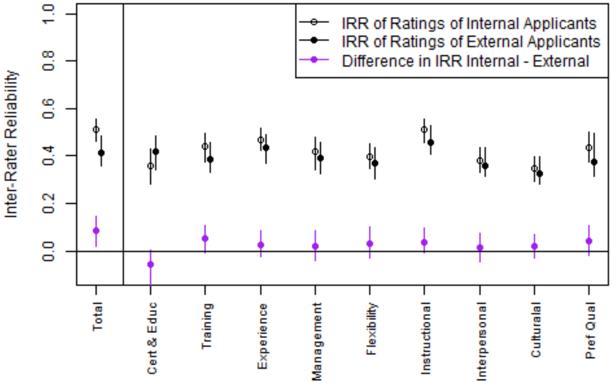
IRR for Internal and External Applicants (Model 3)

- IRR is estimated simultaneously for both groups within Model 3



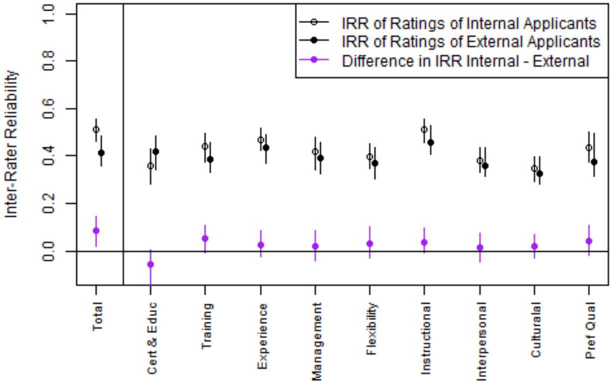
IRR for Internal and External Applicants (Model 3)

- IRR is estimated simultaneously for both groups within Model 3



IRR for Internal and External Applicants (Model 3)

- IRR is estimated simultaneously for both groups within Model 3



Outline

Increasing IRR (Generalized Prophecy Formula)

Increasing model-based IRR (model 2) by averaging ratings of J raters (J=2, 3):

$$R_J = \frac{\sigma_A^2 + \sigma_S^2 + \sigma_{AS}^2}{\sigma_A^2 + \sigma_B^2/J + \sigma_S^2 + \sigma_{AS}^2 + \sigma_{AB}^2/J + \sigma_{BS}^2/J + \sigma_e^2/J}$$

Results:

- 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

Increasing IRR (Generalized Prophecy Formula)

Increasing model-based IRR (model 2) by averaging ratings of J raters (J=2, 3):

$$R_J = \frac{\sigma_A^2 + \sigma_S^2 + \sigma_{AS}^2}{\sigma_A^2 + \sigma_B^2/J + \sigma_S^2 + \sigma_{AS}^2 + \sigma_{AB}^2/J + \sigma_{BS}^2/J + \sigma_e^2/J}$$

Results:

- 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

Increasing IRR (Generalized Prophecy Formula)

Increasing model-based IRR (model 2) by averaging ratings of J raters (J=2, 3):

$$R_J = \frac{\sigma_A^2 + \sigma_S^2 + \sigma_{AS}^2}{\sigma_A^2 + \sigma_B^2/J + \sigma_S^2 + \sigma_{AS}^2 + \sigma_{AB}^2/J + \sigma_{BS}^2/J + \sigma_e^2/J}$$

Results:

- 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

Implications for Predictive Power (Attenuation Formula)

IRR affects instrument's power to predict teacher value added (VA):

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

- A_1 applicant rating
- A_2 subsequent teacher quality (teacher value added)
- R_1, R_2 reliabilities of rating / VA estimates

Results:

- Low correlation with VA for low reliability ratings *(Cultural)*
- 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%

Implications for Predictive Power (Attenuation Formula)

IRR affects instrument's power to predict teacher value added (VA):

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

- A_1 applicant rating
- A_2 subsequent teacher quality (teacher value added)
- R_1, R_2 reliabilities of rating / VA estimates

Results:

- Low correlation with VA for low reliability ratings *(Cultural)*
- 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

Conclusions for hiring data (Questions and Answers)

- Is rating school specific?
  - Model 2: Yes, rating is school-specific.
- Are the ratings more consistent for some *groups*?
  - Model 3: Yes, (total) ratings are more consistent for 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

Conclusions for hiring data (Questions and Answers)

- Is rating school specific?
  - Model 2: Yes, rating is school-specific.
- Are the ratings more consistent for some *groups*?
  - Model 3: Yes, (total) ratings are more consistent for 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 (Methodology)

We suggest using LMM for more flexible analysis of inter-rater reliability:

- Estmation with restricted maximum likelihood (`lme4` in `R`)
- CIs with MCMC (`brms`) or parametric bootstrap (`bootMer` in `lme4`)

- Interaction terms to test for applicant-school matching effect
  (IRR within school, IRR across schools)
- Random slopes to test for differences in variance components for groups
  (different IRR for internal and external applicants)
- Model comparison using AIC, BIC, likelihood ratio tests
- MCMC/bootstraped confidence intervals for decissions about IRR

Conclusion (Methodology)

We suggest using LMM for more flexible analysis of inter-rater reliability:

- Estmation with restricted maximum likelihood (`lme4` in `R`)
- CIs with MCMC (`brms`) or parametric bootstrap (`bootMer` in `lme4`)

- Interaction terms to test for applicant-school matching effect (IRR within school, IRR across schools)
- Random slopes to test for differences in variance components for groups (different IRR for internal and external applicants)
- Model comparison using AIC, BIC, likelihood ratio tests
- MCMC/bootstraped confidence intervals for decissions about IRR

Conclusion (Methodology)

We suggest using LMM for more flexible analysis of inter-rater reliability:

- Estmation with restricted maximum likelihood (`lme4` in `R`)
- CIs with MCMC (`brms`) or parametric bootstrap (`bootMer` in `lme4`)

- Interaction terms to test for applicant-school matching effect (IRR within school, IRR across schools)
- Random slopes to test for differences in variance components for groups (different IRR for internal and external applicants)
- Model comparison using AIC, BIC, likelihood ratio tests
- MCMC/bootstraped confidence intervals for decissions about IRR

Conclusion (Methodology)

We suggest using LMM for more flexible analysis of inter-rater reliability:

- Estmation with restricted maximum likelihood (`lme4` in `R`)
- CIs with MCMC (`brms`) or parametric bootstrap (`bootMer` in `lme4`)

- Interaction terms to test for applicant-school matching effect (IRR within school, IRR across schools)
- Random slopes to test for differences in variance components for groups (different IRR for internal and external applicants)
- Model comparison using AIC, BIC, likelihood ratio tests
- MCMC/bootstraped confidence intervals for decissions about IRR

Conclusion (Methodology)

We suggest using LMM for more flexible analysis of inter-rater reliability:

- Estmation with restricted maximum likelihood (`lme4` in `R`)
- CIs with MCMC (`brms`) or parametric bootstrap (`bootMer` in `lme4`)

- Interaction terms to test for applicant-school matching effect (IRR within school, IRR across schools)
- Random slopes to test for differences in variance components for groups
(different IRR for internal and external applicants)
- Model comparison using AIC, BIC, likelihood ratio tests
- MCMC/bootstraped confidence intervals for decissions about IRR

Conclusion (Methodology)

We suggest using LMM for more flexible analysis of inter-rater reliability:

- Estmation with restricted maximum likelihood (`lme4` in `R`)
- CIs with MCMC (`brms`) or parametric bootstrap (`bootMer` in `lme4`)

- Interaction terms to test for applicant-school matching effect
  (IRR within school, IRR across schools)
- Random slopes to test for differences in variance components for groups
  (different IRR for internal and external applicants)
- Model comparison using AIC, BIC, likelihood ratio tests
- MCMC/bootstraped confidence intervals for decissions about IRR

Conclusion (Methodology)

We suggest using LMM for more flexible analysis of inter-rater reliability:

- Estmation with restricted maximum likelihood (`lme4` in `R`)
- CIs with MCMC (`brms`) or parametric bootstrap (`bootMer` in `lme4`)

- Interaction terms to test for applicant-school matching effect (IRR within school, IRR across schools)
- Random slopes to test for differences in variance components for groups (different IRR for internal and external applicants)
- Model comparison using AIC, BIC, likelihood ratio tests
- MCMC/bootstraped confidence intervals for decissions about IRR

Discussion

Possible further steps:

- Compare with other LMM procedures (`lme`)
- Analyzing error term structure (`weights` in `lme`)
- Continuous moderator of IRR (rater experience in years)
- Ordinal models for subcomponents (`glmer`)
- Incorporating subcomponents (items) into model
- Accounting for correlations between subcomponents
- Optimal weighting of items with respect to IRR

Discussion

Possible further steps:

- Compare with other LMM procedures (`lme`)
- Analyzing error term structure (`weights` in `lme`)
- Continuous moderator of IRR (rater experience in years)
- Ordinal models for subcomponents (`glmer`)
- Incorporating subcomponents (items) into model
- Accounting for correlations between subcomponents
- Optimal weighting of items with respect to IRR

Discussion

Possible further steps:

- Compare with other LMM procedures (`lme`)
- Analyzing error term structure (`weights` in `lme`)
- Continuous moderator of IRR (rater experience in years)
- Ordinal models for subcomponents (`glmer`)
- Incorporating subcomponents (items) into model
- Accounting for correlations between subcomponents
- Optimal weighting of items with respect to IRR

Discussion

Possible further steps:

- Compare with other LMM procedures (`lme`)
- Analyzing error term structure (`weights` in `lme`)
- Continuous moderator of IRR (rater experience in years)
- Ordinal models for subcomponents (`glmer`)
- Incorporating subcomponents (items) into model
- Accounting for correlations between subcomponents
- Optimal weighting of items with respect to IRR

Discussion

Possible further steps:

- Compare with other LMM procedures (`lme`)
- Analyzing error term structure (`weights` in `lme`)
- Continuous moderator of IRR (rater experience in years)
- Ordinal models for subcomponents (`glmer`)
- Incorporating subcomponents (items) into model
- Accounting for correlations between subcomponents
- Optimal weighting of items with respect to IRR

Discussion

Possible further steps:

- Compare with other LMM procedures (`lme`)
- Analyzing error term structure (`weights` in `lme`)
- Continuous moderator of IRR (rater experience in years)
- Ordinal models for subcomponents (`glmer`)
- Incorporating subcomponents (items) into model
- Accounting for correlations between subcomponents
- Optimal weighting of items with respect to IRR

Discussion

Possible further steps:

- Compare with other LMM procedures (`lme`)
- Analyzing error term structure (`weights` in `lme`)
- Continuous moderator of IRR (rater experience in years)
- Ordinal models for subcomponents (`glmer`)
- Incorporating subcomponents (items) into model
- Accounting for correlations between subcomponents
- Optimal weighting of items with respect to IRR

Discussion

Possible further steps:

- Compare with other LMM procedures (`lme`)
- Analyzing error term structure (`weights` in `lme`)
- Continuous moderator of IRR (rater experience in years)
- Ordinal models for subcomponents (`glmer`)
- Incorporating subcomponents (items) into model
- Accounting for correlations between subcomponents
- Optimal weighting of items with respect to IRR

Thank you for your attention!

References:

- Martinkova, Goldhaber & Erosheva: Mixed-Effect Models for Assessing Inter-Rater Reliability and Its Moderators in Complex Designs. Under review, *J Educ Behav Stat*
  older working paper: CEDR WP 2015-7.

Acknowledgement:

- Czech Science Foundation grant GJ15-15856Y
- The Fulbright Commission in the Czech Republic
- IES grants R305C130030, R305A060018