Flexible Analysis of Inter-Rater Reliability
As It Applies to Teacher Selection Instruments

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

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

Robust, September 12, 2016
Outline

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

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

Notes:

- This is just the intraclass correlation coefficient
- $R \in [0, 1]$, 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

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

Notes:
- This is just the intraclass correlation coefficient
- $R \in [0, 1]$, 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

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

Notes:

- This is just the intraclass correlation coefficient
- \( R \in [0, 1] \), 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

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

Notes:

- This is just the intraclass correlation coefficient
- $R \in [0, 1]$, 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

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

Notes:
- This is just the intraclass correlation coefficient
- $R \in [0, 1]$, 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)
  \[ \text{cor}(A_1 + e_1, A_2 + e_2) = \text{cor}(A_1, A_2) \sqrt{R_1 R_2} \]
- Higher standard error of measurement
- Wider confidence intervals
- Less powerful hypotheses tests

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

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

**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)
  \[ \text{cor}(A_1 + e_1, A_2 + e_2) = \text{cor}(A_1, A_2) \sqrt{R_1 R_2} \]
- higher standard error of measurement
- wider confidence intervals
- less powerful hypotheses tests

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

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

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)

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

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

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)

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

- \textbf{assessee effect} \( A_i \sim N(0, \sigma_A^2) \), \textbf{rater effect} \( B_j \sim N(0, \sigma_B^2) \),
- \textbf{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
- Aggregate (average of J raters) has higher IRR: \( R_n = \frac{\sigma_A^2}{\sigma_A^2/\text{J} + \sigma_B^2/\text{J} + \sigma_e^2/\text{J}} \)
Inter-Rater Reliability (Assessee–Rater Model)

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

- assessee effect \( A_i \sim N(0, \sigma_A^2) \), rater effect \( B_j \sim N(0, \sigma_B^2) \), 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

- 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 \( A_i \sim N(0, \sigma_A^2) \), rater effect \( B_j \sim N(0, \sigma_B^2) \),
  error \( e_{ij} \sim N(0, \sigma_e^2) \)
- **Inter-Rater Reliability**: 
  \[ R = \text{cor}(Y_{ij}, Y_{ij'}) = ICC = \frac{\sigma_A^2}{\sigma_Y^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_e^2} \]
  \( R \in [0, 1] \), low values mean a lot of measurement error
- 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:**

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

**IRR within school:**

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

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

- IRR within school:

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

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

- IRR within school:

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

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

- IRR within school:

\[ R_{within} = \text{cor}(Y_{ijk}, Y_{ij'k'}) = \frac{\sigma^2_A + \sigma^2_S + \sigma^2_{AS}}{\sigma^2_A + \sigma^2_B + \sigma^2_S + \sigma^2_{AS} + \sigma^2_{AB} + \sigma^2_{BS} + \sigma^2_e} \]
IRR estimation and inference

More flexible estimation using linear random-effect models
- Estimation w/ restricted maximum likelihood using \texttt{lmer} in \texttt{lme4} in R
- Model selection using AIC, BIC, likelihood ratio tests
- Confidence intervals w/ MCMC using \texttt{brms} (or bootstrap: \texttt{bootMer})

```r
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 \texttt{lmer} in \texttt{lme4} in R
- Model selection using AIC, BIC, likelihood ratio tests
- Confidence intervals w/ MCMC using \texttt{brms} (or bootstrap: \texttt{bootMer})

```r
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)
```
For all subcomponents, the applicant qualities are school specific.
Some subcomponents are less reliable than others.
Outline

1. Introduction
2. Hierarchical models for IRR
3. Moderators of IRR
4. Implications for validity
5. Conclusion
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

\[
Y_{ijk} = \mu + \omega_i \beta_1 + (1 - \omega_i) A_{0i} + \omega_i A_{1i} \\
+ (1 - \omega_i) B_{0j} + \omega_i B_{1j} \\
+ (1 - \omega_i) S_{0k} + \omega_i S_{1k} \\
+ A S_{ik} + A B_{ij} + B S_{jk} + e_{ijk}
\]

- \( \omega_i = 1 \) for internal and 0 for external applicants
- group fixed effect \( \beta_1 \)
- \( A_{0i} \sim N(0, \sigma_{A0}^2) \) 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 \( S_{1k} \sim N(0, \sigma_{S1}^2) \)
Assessees-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

\[
Y_{ijk} = \mu + \omega_i\beta_1 + (1 - \omega_i)A_{0i} + \omega_iA_{1i} + (1 - \omega_i)B_{0j} + \omega_iB_{1j} + (1 - \omega_i)S_{0k} + \omega_iS_{1k} + AS_{ik} + AB_{ij} + BS_{jk} + e_{ijk}
\]

- \(\omega_i = 1\) for internal and 0 for external applicants
- group fixed effect \(\beta_1\)
- \(A_{0i} \sim N(0, \sigma_{A0}^2)\) 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 \(S_{1k} \sim N(0, \sigma_{S1}^2)\)
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

\[
Y_{ijk} = \mu + \omega_i \beta_1 + (1 - \omega_i) A_{0i} + \omega_i A_{1i} \\
\quad + (1 - \omega_i) B_{0j} + \omega_i B_{1j} \\
\quad + (1 - \omega_i) S_{0k} + \omega_i S_{1k} \\
\quad + A S_{ik} + A B_{ij} + B S_{jk} + e_{ijk}
\]

- \(\omega_i = 1\) for internal and 0 for external applicants
- group fixed effect \(\beta_1\)
- \(A_{0i} \sim N(0, \sigma_{A0}^2)\) 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 \(S_{1k} \sim N(0, \sigma_{S1}^2)\)
Modeler of IRR: Internal vs. External status (Model 3)

\[ \text{model} \leftarrow \text{lmer}(\text{rating} \sim 1 + \text{internal} + \right. \\
+ (0+\text{internal}|\text{Apl}) + (0+\text{internal}|\text{Rtr}) + (0+\text{internal}|\text{Sch}) + \\
+ (1|\text{Apl}:\text{Sch}) + (1|\text{PID}:\text{rater}) + (1|\text{rater:school}), \\
+ \text{data=screening}) \]

Within-school IRR:

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

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

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

- external applicant:

\[
R_0 = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma^2_{A0} + \sigma^2_{S0} + \sigma^2_{AS}}{\sigma^2_{A0} + \sigma^2_{B0} + \sigma^2_{S0} + \sigma^2_{AS} + \sigma^2_{AB} + \sigma^2_{BS} + \sigma^2_e}
\]
Moderator of IRR: Internal vs. External status (Model 3)

```r
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:**
  
  \[
  R_1 = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma_{A1}^2 + \sigma_{S1}^2 + \sigma_{AS}^2}{\sigma_{A1}^2 + \sigma_{B1}^2 + \sigma_{S1}^2 + \sigma_{AS}^2 + \sigma_{AB}^2 + \sigma_{BS}^2 + \sigma_e^2}
  \]

- **External applicant:**
  
  \[
  R_0 = \text{cor}(Y_{ijk}, Y_{ij'k}) = \frac{\sigma_{A0}^2 + \sigma_{S0}^2 + \sigma_{AS}^2}{\sigma_{A0}^2 + \sigma_{B0}^2 + \sigma_{S0}^2 + \sigma_{AS}^2 + \sigma_{AB}^2 + \sigma_{BS}^2 + \sigma_e^2}
  \]
## Model 3: Variance decomposition, IRR

<table>
<thead>
<tr>
<th>Internal</th>
<th>b</th>
<th>SE(b)</th>
<th>Apl</th>
<th>Rtr</th>
<th>Sch</th>
<th>AS</th>
<th>RS</th>
<th>AR</th>
<th>Res.</th>
<th>Total</th>
<th>IRRw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td><strong>3.35</strong></td>
<td>0.40</td>
<td>19%</td>
<td><strong>16%</strong></td>
<td>6%</td>
<td>26%</td>
<td>1%</td>
<td>0%</td>
<td>33%</td>
<td>60.61</td>
<td><strong>0.51</strong></td>
</tr>
<tr>
<td>Crt. Ed.</td>
<td>0.13</td>
<td>0.05</td>
<td>1%</td>
<td>9%</td>
<td>12%</td>
<td>20%</td>
<td>25%</td>
<td>0%</td>
<td>34%</td>
<td>1.12</td>
<td>0.33</td>
</tr>
<tr>
<td>Training</td>
<td>0.49</td>
<td>0.08</td>
<td>20%</td>
<td>9%</td>
<td>1%</td>
<td>22%</td>
<td>3%</td>
<td>2%</td>
<td>43%</td>
<td>1.65</td>
<td>0.43</td>
</tr>
<tr>
<td>Exper.</td>
<td>0.33</td>
<td>0.06</td>
<td>16%</td>
<td>9%</td>
<td>2%</td>
<td>28%</td>
<td>0%</td>
<td>2%</td>
<td>43%</td>
<td>1.39</td>
<td>0.46</td>
</tr>
<tr>
<td>Mngmnt</td>
<td>0.41</td>
<td>0.06</td>
<td>16%</td>
<td>7%</td>
<td>4%</td>
<td>20%</td>
<td>2%</td>
<td>4%</td>
<td>47%</td>
<td>1.29</td>
<td>0.40</td>
</tr>
<tr>
<td>Flexibilty</td>
<td>0.35</td>
<td>0.05</td>
<td>15%</td>
<td>13%</td>
<td>2%</td>
<td>21%</td>
<td>1%</td>
<td>4%</td>
<td>44%</td>
<td>1.23</td>
<td>0.38</td>
</tr>
<tr>
<td>Instruct.</td>
<td>0.47</td>
<td>0.06</td>
<td>19%</td>
<td>5%</td>
<td>6%</td>
<td>24%</td>
<td>2%</td>
<td>3%</td>
<td>41%</td>
<td>1.31</td>
<td>0.49</td>
</tr>
<tr>
<td>Interpers.</td>
<td>0.31</td>
<td>0.05</td>
<td>15%</td>
<td>11%</td>
<td>2%</td>
<td>17%</td>
<td>3%</td>
<td>8%</td>
<td>43%</td>
<td>1.14</td>
<td>0.35</td>
</tr>
<tr>
<td>Cultural</td>
<td>0.34</td>
<td>0.05</td>
<td>13%</td>
<td>14%</td>
<td>1%</td>
<td>17%</td>
<td>2%</td>
<td>5%</td>
<td>47%</td>
<td>1.38</td>
<td>0.32</td>
</tr>
<tr>
<td>Pref.Q.</td>
<td>0.47</td>
<td>0.09</td>
<td>7%</td>
<td>16%</td>
<td>0%</td>
<td>35%</td>
<td>3%</td>
<td>0%</td>
<td>38%</td>
<td>2.36</td>
<td>0.42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>External</th>
<th>b</th>
<th>SE(b)</th>
<th>Apl</th>
<th>Rtr</th>
<th>Sch</th>
<th>AS</th>
<th>RS</th>
<th>AR</th>
<th>Res.</th>
<th>Total</th>
<th>IRRw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>15%</td>
<td><strong>26%</strong></td>
<td>1%</td>
<td>25%</td>
<td>1%</td>
<td>0%</td>
<td>32%</td>
<td>62.60</td>
<td><strong>0.41</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crt. Ed.</td>
<td>18%</td>
<td>14%</td>
<td>3%</td>
<td>16%</td>
<td>20%</td>
<td>0%</td>
<td>28%</td>
<td>1.36</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td>17%</td>
<td>19%</td>
<td>1%</td>
<td>20%</td>
<td>3%</td>
<td>2%</td>
<td>39%</td>
<td>1.83</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exper.</td>
<td>17%</td>
<td>16%</td>
<td>1%</td>
<td>25%</td>
<td>0%</td>
<td>2%</td>
<td>39%</td>
<td>1.53</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mngmnt</td>
<td>16%</td>
<td>13%</td>
<td>3%</td>
<td>19%</td>
<td>2%</td>
<td>3%</td>
<td>45%</td>
<td>1.36</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flexibilty</td>
<td>14%</td>
<td>18%</td>
<td>1%</td>
<td>20%</td>
<td>1%</td>
<td>3%</td>
<td>43%</td>
<td>1.28</td>
<td>0.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instruct.</td>
<td>19%</td>
<td>12%</td>
<td>2%</td>
<td>23%</td>
<td>2%</td>
<td>3%</td>
<td>39%</td>
<td>1.37</td>
<td>0.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpers.</td>
<td>16%</td>
<td>19%</td>
<td>1%</td>
<td>16%</td>
<td>2%</td>
<td>7%</td>
<td>39%</td>
<td>1.28</td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cultural</td>
<td>15%</td>
<td>19%</td>
<td>0%</td>
<td>16%</td>
<td>2%</td>
<td>5%</td>
<td>43%</td>
<td>1.51</td>
<td>0.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pref.Q.</td>
<td>0%</td>
<td>21%</td>
<td>2%</td>
<td>35%</td>
<td>3%</td>
<td>0%</td>
<td>38%</td>
<td>2.33</td>
<td>0.37</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Model comparison (BIC)

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

<table>
<thead>
<tr>
<th></th>
<th>model 1</th>
<th>model 2</th>
<th>model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>23,204</td>
<td>23,072</td>
<td>22,954</td>
</tr>
<tr>
<td>Certificate and Education</td>
<td>8,515</td>
<td>8,371</td>
<td>8,336</td>
</tr>
<tr>
<td>Training</td>
<td>11,050</td>
<td>10,981</td>
<td>10,886</td>
</tr>
<tr>
<td>Experience</td>
<td>10,561</td>
<td>10,467</td>
<td>10,426</td>
</tr>
<tr>
<td>Management</td>
<td>10,239</td>
<td>10,176</td>
<td>10,093</td>
</tr>
<tr>
<td>Flexibility</td>
<td>9,974</td>
<td>9,897</td>
<td>9,838</td>
</tr>
<tr>
<td>Instructional</td>
<td>10,271</td>
<td>10,167</td>
<td>10,090</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>9,740</td>
<td>9,677</td>
<td>9,643</td>
</tr>
<tr>
<td>Cultural</td>
<td>10,370</td>
<td>10,322</td>
<td>10,270</td>
</tr>
<tr>
<td>Preferred Qualifications</td>
<td>9,073</td>
<td>8,965</td>
<td>8,908</td>
</tr>
</tbody>
</table>
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
Increasing IRR (Generalized Prophecy Formula)

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

$$R_J = \frac{\sigma^2_A + \sigma^2_S + \sigma^2_{AS}}{\sigma^2_A + \sigma^2_B/J + \sigma^2_S + \sigma^2_{AS} + \sigma^2_{AB}/J + \sigma^2_{BS}/J + \sigma^2_e/J}$$

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):

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

- $A_1$ 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%
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

1. Introduction

2. Hierarchical Models for Inter-Rater Reliability

3. Moderators of Inter-Rater Reliability

4. Implications for Predictive Power

5. Conclusion
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:

- Estimation 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/bootstrapped confidence intervals for decisions about IRR
Conclusion (Methodology)

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

- Estimation 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/bootstrapped confidence intervals for decisions about IRR
We suggest using LMM for more flexible analysis of inter-rater reliability:

- Estimation 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/bootstrapped confidence intervals for decisions about IRR
Conclusion (Methodology)

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

- Estimation 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/bootstrapped confidence intervals for decisions about IRR
Conclusion (Methodology)

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

- Estimation 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/bootstrapped confidence intervals for decisions about IRR
Conclusion (Methodology)

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

- Estimation 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/bootstrapped confidence intervals for decisions about IRR
Conclusion (Methodology)

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

- Estimation 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/bootstrapped confidence intervals for decisions 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
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:

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