Lesson 6: Reliability

Patrícia Martinková

Department of Statistical Modelling
Institute of Computer Science, Czech Academy of Sciences

NMST 570, December 12, 2017

Contents

Classical test theory

In behavioral research we are typically interested in the **true score** T but have available only the **observed score** X which is contaminated by some (uncorrelated) **measurement error** e, such that $X = T + e$.

Examples:

- Admission tests: we are interested in **applicant's knowledge or ability** T, but have available only the test score X
- Grading of essays: We are interested in **essay's quality** T but we have available only the grader's evaluation X
- Questionnaires on satisfaction: main interest is **respondent satisfaction**, but available are only his/her responses on the questionnaire.

The observed score might vary if we chose different items or different graders.

Classical test theory

Natural questions:

- How much information about the true score is indeed contained in the measurement?
- What is the strength of the relationship between true and observed score?

Reliability theory

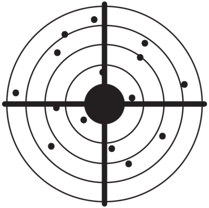
- Reliability is defined as squared correlation of the true and observed score $\rho_X = \text{corr}^2(T, X) = \rho^2_{T,X}$
- $\rho_X \in \langle 0, 1 \rangle$
- equivalently, reliability can be reexpressed as the ratio of the true score variance to total observed variance $\rho_X = \frac{\text{var}(T)}{\text{var}(X)} = \frac{\sigma^2_T}{\sigma^2_X}$

Implications of low reliability

- less accurate estimates of the true score
- wider (less precise) confidence intervals
- need of higher number of subjects to demonstrate differences between groups (keeping the same test power)
- attenuation of correlations, bound of criterion validity

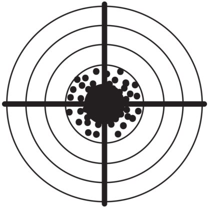
$$\rho_{X,Y} = \rho_{T_X,T_Y}\sqrt{\rho_X\rho_Y} \leq \rho_X$$

Graphical interpretation



Low reliability thus low validity

High reliability but low validity



High reliability and high validity

- center of the target represents the value we want to measure
- shots represent independent measurements on one object
- reliability represented by variability of the shots
- validity represented by overall shots' closeness to the center

Observations

- high reliability does not ensure high validity
- validity is bounded by reliability

Reliability guidelines

- Conventional requirement $\rho_X \geq .8$, but see Lee (2012)
  - $\geq .9$ for intelligence tests
  - $\geq .7$ for personality tests
  - $\sim .6$ for essay marking
- In case of low reliability we should think of instrument revision
  - adding items
  - deleting items
  - in case of graders: training, precise instructions

Importance of proper estimation of reliability

- Overestimation may imply adopting unreliable instrument
- Underestimation may imply (costly) revision of instrument
- Misunderstanding of reliability can imply deletion of important items and lowering validity

Estimation procedures

The true score T is not observed, thus we can't estimate reliability from its definition ($\rho^2_{T,X}$ nor σ^2_T/σ^2_X)

Parallel measurements

- equally precise measurements of the same true score:
- $X_1 = T + e_1, \quad X_2 = T + e_2, \quad \mathrm{var}\,(e_1) = \mathrm{var}\,(e_2) = \sigma^2_e$
- the reliability of both measurements is the same ρ
- if the errors are uncorrelated, then **correlation between the measurements is equal to** their (common) **reliability**
  $\rho_{X_1,X_2} = \frac{\mathrm{cov}\,(T+e_1,T+e_2)}{\sqrt{\mathrm{var}\,(T+e_1)\mathrm{var}\,(T+e_2)}} = \frac{\sigma^2_T}{\sigma^2_X} = \rho$

The methods differ in how they make use of multiple measurements.

Estimation procedures

Use of multiple administrations

Methods employ correlation coefficient btw. observed total scores

- Test-retest method (coefficient of stability)
- Alternate test forms (coefficient of equivalence)

Use of composite measurements

Methods employ correlation coefficient btw. observed partial total scores

- Split-half coefficient
- Average split-half
- Cronbach's aplha (coefficient of internal consistency)

Test-Retest

- Assumes independent test administrations
  - No memory
  - No improvement between administrations

- Optimal interval 6-12 weeks

Parallel Forms

- Assumes trully paralel forms
    - Equally difficult
    - Parallel items and content
- Assumes the same conditions

Composite measurements

- Goal is to provide multiple converging pieces of information
- E.g. educational tests, scales, questionnaires, ...

What is the relationship between reliability of composite measurement $X = \sum_{j=1}^{m} X_j$ and reliability of its components?

Spearman-Brown prophecy formula (1910)

Assume $X_1, \ldots, X_m$ parallel measurements (with uncorrelated errors and uncorrelated with true scores). Then reliability of each X_i is the same ρ and the composite reliability is

$$\rho_X = \frac{m \cdot \rho}{1 + (m-1)\rho}$$

Remark: Adding parallel items increases reliability of total score.

Split-half coefficient

- correlation between two subscores corrected for test length
- test is split into two parts, two subscores Y_1, Y_2 are computed
- $\rho_{SH} = \frac{2\rho_{Y_1,Y_2}}{1+\rho_{Y_1,Y_2}}$
- assumes that the two subtests are parallel
- depends on how the split was carried out (even/odd, random,...)
  - even-numbered / odd-numbered
  - with intention to create two halves that are as similar as possible
  - in a random fashion
- we may also compute the mean of all possible split-half coefficients
  - Average split-half

Cronbach's alpha

- based on idea of splitting the test into individual items

$$\alpha = \frac{m}{m-1}\frac{\sum\sum_{j\neq k}\operatorname{cov}(X_j, X_k)}{\operatorname{var}(X)} = \frac{m}{m-1}\left(1 - \frac{\sigma^2_{X_1} + \cdots + \sigma^2_{X_m}}{\sigma^2_X}\right)$$

- popular estimator, provides simple and unique estimation
- equals to composite reliability σ^2_T/σ^2_X in case of parallel (or at least T-equivalent) items and uncorrelated errors
- in general case and uncorrelated errors, alpha is lower bound to reliability $\alpha \leq \rho_X$ (Novick & Lewis, 1967) and can be viewed as **index of internal consistency**
- in case of correlated errors, alpha can be lower or greater than reliability

Cronbach's alpha - limitations

Cronbach's alpha is a good estimator of reliability for

- parallel (or at least T-equivalent) items and and
- uncorrelated errors

Corrections needed for:

- Correlated errors
  - Example: Reading test, group of items associated with one text.
  - Corrections for correlated errors (Rae, 2006)
- Multidimensional measurement
  - Example: Math test, items measuring arithmetic skills but also reading skills etc.
  - Factor-analysis based estimation of reliability (Raykov & Maurcoulides, 2011)
- More sources of error (multilevel models, G-index)
- Other than normal distribution of item responses (what happens in case of binary items?)

Beyond Cronbach's alpha

How to define and estimate internal consistency in case of binary items?

Martinková P, & Zvára K. Reliability in the Rasch Model. Kybernetika, 43(3), pp. 315-26, 2007. http://www.kybernetika.cz/content/2007/3/315/paper.pdf

Martinková P, & Vlčková K. Hodnocení reliability znalostních a psychologických testů. (Estimation of Reliability of Educational and Psychological Measurements. In Czech.) *Informační bulletin České statistické společnosti*, 4, pp. 1-15, 2014. http://www.statspol.cz/cs/wp-content/uploads/IB_4_2014.pdf

Cronbach's alpha: 2-way mixed ANOVA approach

- X_{ij} responses of n students on m items
- $X_{ij} = T_i + b_j + e_{ij}$
  - $T_i \sim \mathrm{N}(0, \sigma_T^2)$ random, student ability
  - b_j fixed, $\sum b_j = 0$, describe item difficulty
  - $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$ random error
  - total scores $X_i = mT_i + \sum_j b_j + \sum_j e_{ij}$
- reliability: $\rho_X = \frac{var(mT_i)}{\mathrm{var}\,(X_i)} = \frac{m^2\sigma_T^2}{m^2\sigma_T^2 + m\sigma_e^2} = \frac{\sigma_T^2}{\sigma_T^2 + \frac{1}{m}\sigma_e^2}$
- Cronbach's alpha:
  $\alpha = \frac{m}{m-1}\frac{\sum\sum_{j\neq k}\mathrm{cov}\,(X_{ij}, X_{ik})}{\mathrm{var}\,(X_i)} = \frac{m}{m-1}\frac{m(m-1)\sigma_T^2}{m^2\sigma_T^2 + m\sigma_e^2} = \frac{\sigma_T^2}{\sigma_T^2 + \frac{1}{m}\sigma_e^2}$
- estimate of Cronbach's alpha: $\hat{\alpha} = \frac{m}{m-1}\frac{\sum\sum_{j\neq k} s_{jk}}{\sum\sum_{j,k} s_{jk}}, \quad \text{where } s_{jk} = \frac{1}{n-1}\sum_{t=1}^{n}(X_{tj} - \bar{X}_{\bullet j})(X_{tj} - \bar{X}_{\bullet k})$

Cronbach's alpha: 2-way mixed ANOVA approach (2)

Sums of squares

- $SS_T = \sum\sum(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 \sim (m\sigma_T^2 + \sigma_e^2)\chi^2(n-1)$
- $SS_e = \sum\sum(X_{ij} - \bar{X}_{\bullet j} - \bar{X}_{i\bullet} + \bar{X}_{\bullet\bullet})^2 \sim \sigma_e^2\chi^2((n-1)(m-1))$

Expectations of Mean sums of squares

- $\mathrm{E}\, MS_T = \mathrm{E}\, SS_T/(n-1) = m\sigma_T^2 + \sigma_e^2$
- $\mathrm{E}\, MS_e = \mathrm{E}\, SS_e/((n-1)(m-1)) = \sigma_e^2$

Cronbach's alpha

$$\alpha = \frac{\sigma_T^2}{\sigma_T^2 + \frac{1}{m}\sigma_e^2} = \frac{\mathrm{E}\, MS_T - \mathrm{E}\, MS_e}{\mathrm{E}\, MS_T}$$

Cronbach's alpha estimate

$$\hat{\alpha} = \frac{m}{m-1}\frac{\sum\sum_{j\neq k} s_{jk}}{\sum\sum_{j,k} s_{jk}} = \frac{MS_T - MS_e}{MS_T} = 1 - \frac{1}{F}$$

Cronbach's alpha: 2-way mixed ANOVA approach (3)

Estimate of Cronbach's alpha can be reexpressed as

$$\hat{\alpha} = \frac{MS_T - MS_E}{MS_T} = 1 - \frac{1}{F}$$

- F statistic used to test the submodel with no subject effect $(H_0 : \sigma_T^2 = 0)$
- Interpretation: alpha close to 1 for F high, i.e. when we reject H_0, i.e. when admission test well discriminates between students
- Gives confidence intervals
- Estimate is not generally appropriate for more complicated designs

Logistic alpha

F statistic in

$$\hat{\alpha} = 1 - \frac{1}{F}$$

assumes normality of items

- How does the estimate of reliability behave for binary items?
- Would a new estimate

$$\hat{\alpha}_{log} = 1 - \frac{n-1}{X^2}$$

  based on statistic used in similar situation in logistic regression (difference of deviances $X^2 = D(B) - D(A+B)$) give better results for case of binary data?

Definition of reliability in binary items

- Classical model not applicable (binary outcome can't be expressed as sum of T and independent error e)
- IRT models ussually assumed
- Reliability can be defined as (Raykov & Maurcoulides, 2011)

$$\rho_X = \frac{\mathrm{var}\left(\mathrm{E}\left(X|T\right)\right)}{\mathrm{var}\left(\mathrm{E}\left(X|T\right)\right) + \mathrm{E}\left(\mathrm{var}\left(X|T\right)\right)} = \frac{\mathrm{var}\left(\mathrm{E}\left(X|T\right)\right)}{\mathrm{var}\left(X\right)}$$

- Resulting integrals can be evaluated numerically, not explicitly
- Not equal to parallel-forms reliability, but differences negligible (Kim, 2012)
- S-B formula holds only approximately (Martinkova, Zvara 2010)

Cronbach's alpha in binary items

- Cronbach's alpha is readily applicable also for binary items
- Cronbach's alpha represents generalization of so-called Kuder-Richardson formulae (*Psychometrika*, 1937):
- $\hat{\rho}_{KR-20} = \frac{p}{p-1}\left[1 - \frac{\sum \hat{r}_k(1-\hat{r}_k)}{\hat{\sigma}_X}\right]$, where $\hat{r}_k$ is easiness of k-th item
- For test with items of common difficulties $\hat{\rho}_{KR-21} = \frac{p}{p-1}\left[1 - \frac{\hat{\mu}(p-\hat{\mu}_k)}{p\hat{\sigma}_X}\right]$, where $\hat{\mu}$ is average total score

Logistic alpha: Simulation study

Pre-defined values:

- number of students $n = 25,\ 50,\ 100,\ 500$
- number of items $m = 10,\ 20,\ 50,\ 100$
- IRT parameters (difficulty, discrimination, guessing for each item)
- 55 values of σ_T (defines true reliability)
- number of simulates $N = 1000$

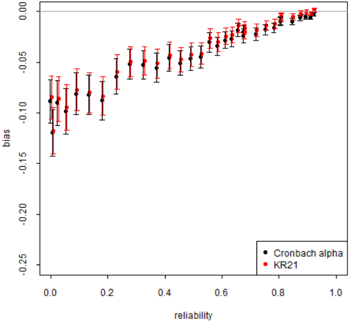
For each combination of n, m and σ_T:

- true reliability computed
- N data sets generated:
  - set of n student abilities generated $T_i \sim \mathrm{N}(0, \sigma_T^2)$
  - Y_{ij} generated from IRT model
  - estimates computed from the data

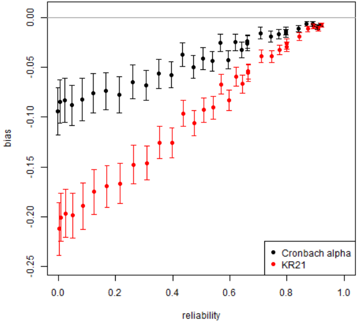
$\Rightarrow$ N estimates $\hat{\alpha}_{CR}$, KR-21 and $\hat{\alpha}_{log}$

- bias and MSE of the estimates plotted out

Simulations: Cronbach's alpha (KR-20) and KR-21



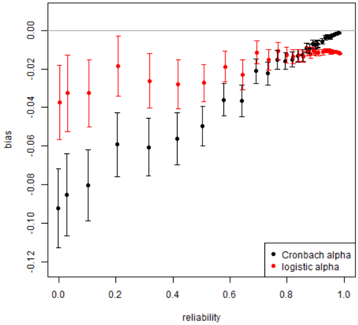
Bias and MSE of two estimators of reliability, item difficulties from $(-0.1, 0.1)$. Number of students $n = 25$, number of items $m = 10$, number of simulates $N = 1000$.



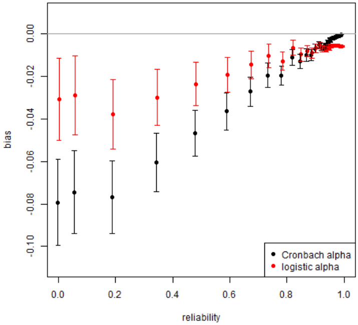
Bias and MSE of two estimators of reliability, item difficulties from $(-3, 3)$. Number of students $n = 25$, number of items $m = 10$, number of simulates $N = 1000$.

- $\hat{\rho}_{KR-21}$ is not appropriate in case of different item difficulties

Simulations: Cronbach's and logistic alpha

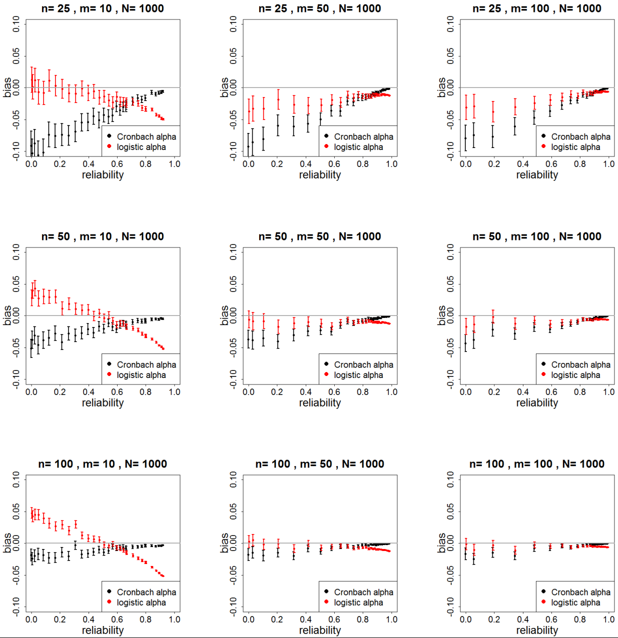


Bias and MSE of two estimators of reliability, number of students $n = 25$, number of items $m = 50$, number of simulates $N = 1000$.



Bias and MSE of two estimators of reliability, number of students $n = 25$, number of items $m = 100$, number of simulates $N = 1000$.

- $\hat{\alpha}_{log}$ has promising properties especially for high number of items

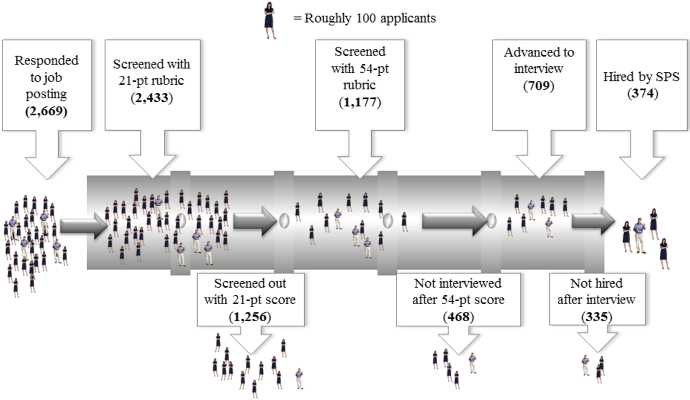

n= 25 , m= 10 , N= 1000
n= 25 , m= 50 , N= 1000
n= 25 , m= 100 , N= 1000
n= 50 , m= 10 , N= 1000
n= 50 , m= 50 , N= 1000
n= 50 , m= 100 , N= 1000
n= 100 , m= 10 , N= 1000
n= 100 , m= 50 , N= 1000
n= 100 , m= 100 , N= 1000
bias
-0.10
-0.05
0.00
0.05
0.10
reliability
0.0
0.2
0.4
0.6
0.8
1.0
Cronbach alpha
logistic alpha

Logistic alpha - conclusions

- Idea behind Logistic alpha was explained
- Logistic alpha has promising properties for some scenarios
- Cases of true reliabilities close to 1 need some adjustment
- Cases of high number of students?

More on inter-rater reliability

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:

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?

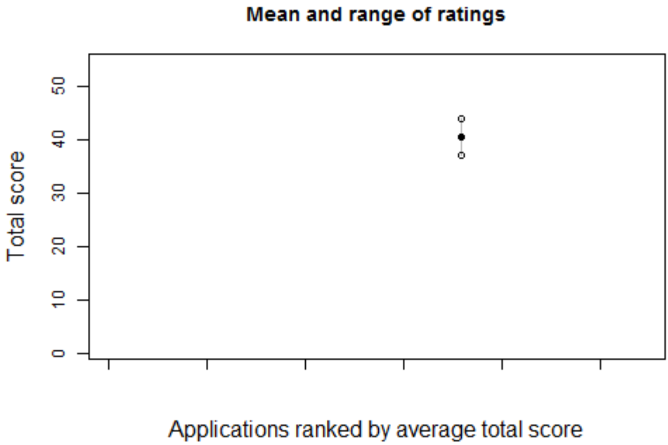
   - Goldhaber et al. 2017

2. **Can we do better?**

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

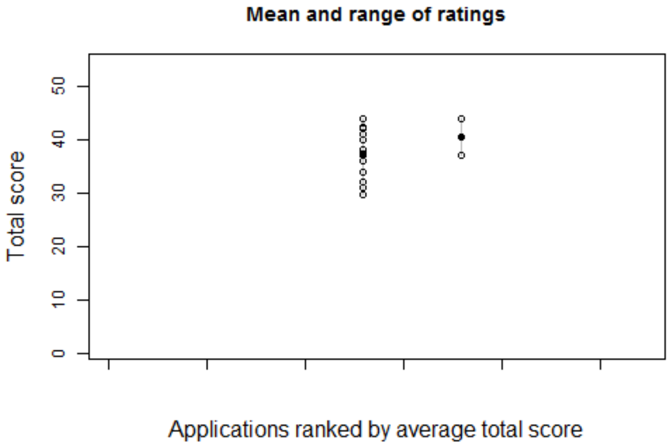
   - Martinkova et al. 2015

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



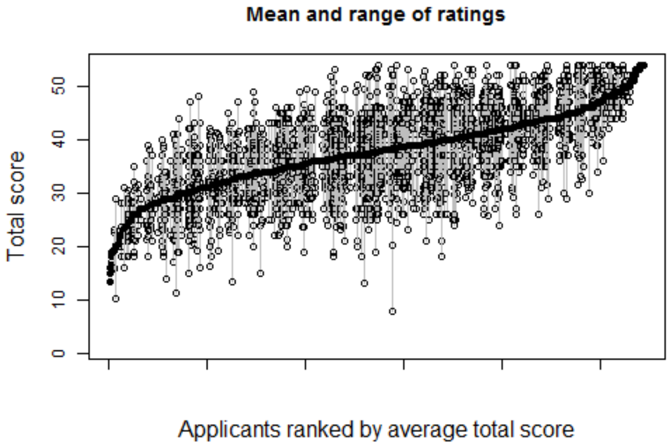
Are the ratings consistent?

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?

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, ...

Model-based reliability estimates

- 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

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

Conclusion

In this presentation, we have

- explained motivation beind reliability
- presented mostly used approaches for reliability estimation
  - test-retest
  - parallel forms
  - split-half coefficient
  - Cronbach's alpha
- presented research on alternative to Cronbach's alpha
- discussed use of model-based reliability estimates (for IRR)

Thank you for your attention!

www.cs.cas.cz/martinkova

References

- Haertel EH. (2006). Reliability. In Brennan RL (ed.), (2006). *Educational Measurement (4th edn.)*. Westport, CT: Praeger. pp. 65–110.
- Martinková P, & Vlčková K. Hodnocení reliability znalostních a psychologických testů. (Estimation of Reliability of Educational and Psychological Measurements. In Czech.) *Informační bulletin České statistické společnosti*, 4, pp. 1-15, 2014. http://www.statspol.cz/cs/wp-content/uploads/IB_4_2014.pdf
- Martinková P, & Zvára K. Reliability in the Rasch Model. Kybernetika, 43(3), pp. 315-26, 2007. http://www.kybernetika.cz/content/2007/3/315/paper.pdf
- Martinková P, & Goldhaber D. (2015). Improving Teacher Selection: The Effect of Inter-Rater Reliability in the Screening Process. *CEDR Working Paper 2015-7*. University of Washington, Seattle, WA. http://www.cedr.us/papers/working/CEDR%20WP%202015-7.pdf