Researching Reliability Estimates in the Context of Czech Admission Tests

Patricia Martinkova

Fulbright-Masaryk Fellow at CSSS UW
patmar@uw.edu
http://faculty.washington.edu/patmar/

BEAR seminar, Feb 11, 2014

Introduction

- study at Dept. of Prob. & Statistics, Charles University in Prague
- research at Czech Academy of Sciences

- joint work with Karel Zvara, Marie Turcicova and Katarina Vlckova

Introduction

Motivation for this study

Admission tests to Czech universities

Educational Measurement in Czech Republic

- very little standardized testing in schools
- a lot of classroom testing
- colledges often run their own admission tests (different tests for different schools of medicine, etc.)
- no (graduate) program in Educational Measurement
- research scattered, conducted under different programs

Introduction

Current increased interest for testing

- standardized high school graduation examination
- studies on validity of admission tests
- new books on test construction methodology
- effort for more sophisticated item banks, item analysis
- debates on quality of tests: require to report Cronbach's alpha?

Aims of the study

1. to research Cronbach's alpha and its assumpions
2. to research properties of newly proposed estimate *logistic alpha*

Outline

1. Introduction
2. **Reliability**
3. Cronbach's alpha
4. Reliability in case of binary items
5. Simulation study
6. Discussion and conclusion

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 : $X = T + e$

Examples:

- Admission tests: we are interested in **student's knowledge** 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

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

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

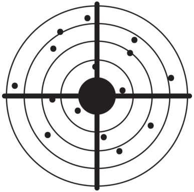
- Reliability defined as squared correlation of the true and observed score $\rho_x = \mathrm{corr}^2(T, X) = \rho^2_{T,X}$
- $\rho_X \in \langle 0, 1 \rangle$
- equivalently can be reexpressed as the ratio of the true score variance to total observed variance $\rho_X = \frac{\mathrm{var}(T)}{\mathrm{var}(X)} = \frac{\sigma^2_T}{\sigma^2_X}$
- T not observed, thus we can't estimate reliability from its definition

Implications of low reliability

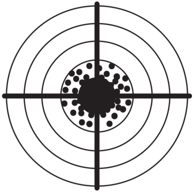
- less accurate estimates of the true score
- wider (less precise) confidence intervals
- need of higher number of subjects to demonstrate differences between groups (keeping the same test power)
- attenuation of correlations, bound of criterion validity

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

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 bound by reliability

Importance of proper estimation of reliability

- In case of low reliability we should think of instrument revision
  - adding items
  - deleting items
  - in case of graders: training, precise instructions
- Conventional requirement $\rho_X \geq .8$
- Underestimation may imply (costly) revision of instrument
- Misunderstanding of reliability can imply deletion of important items and lowering validity
- Overestimation may imply adopting unreliable instrument

Procedures for estimating reliability?

- 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:
- $X_1 = T + e_1, \quad X_2 = T + e_2, \quad \mathrm{var}(e_1) = \mathrm{var}(e_2) = \sigma_e^2$
- 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**

$$\rho_{X_1,X_2} = \frac{\mathrm{cov}(T+e_1, T+e_2)}{\sqrt{\mathrm{var}(T+e_1)\mathrm{var}(T+e_2)}} = \frac{\sigma_T^2}{\sigma_X^2} = \rho$$

Procedures for estimating reliability (1)

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

Both methods require two measurement administrations.

Composite measurements

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

Is there any 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 proper items increases reliability.

Procedures to estimate reliability(2)

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,...)
- we may also compute the mean of all possible split-half coefficients

Outline

1. Introduction
2. Reliability
3. **Cronbach's alpha**
4. Reliability in case of binary items
5. Simulation study
6. Discussion and conclusion

Cronbach's alpha

- based on idea of splitting the test into individual items

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

- 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: 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

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

Expectations of Mean sums of squares

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

Cronbach's alpha

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

Cronbach's alpha estimate

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

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

Estimate of Cronbach's alpha can be reexpressed as

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

- 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

Procedures to estimate reliability(3)

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?)

Outline

1. Introduction
2. Reliability
3. Cronbach's alpha
4. **Reliability in case of binary items**
5. Simulation study
6. Discussion and conclusion

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 formulas (*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

Outline

1. Introduction
2. Reliability
3. Cronbach's alpha
4. Reliability in case of binary items
5. **Simulation study**
6. Discussion and conclusion

Simulation study in IRT models

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 $N = 1000$

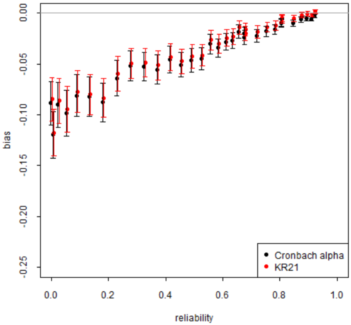
For each combination of n, m and σ_T:

- 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

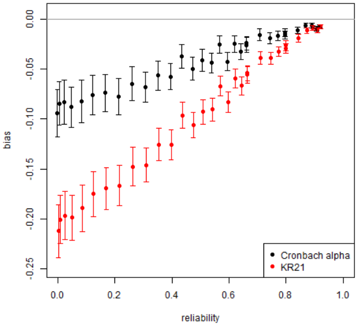
$\Rightarrow$ N estimates $\hat{\alpha}_{CR}$, KR-21 and $\hat{\alpha}_{log}$

- bias and MSE of the estimates plotted out

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



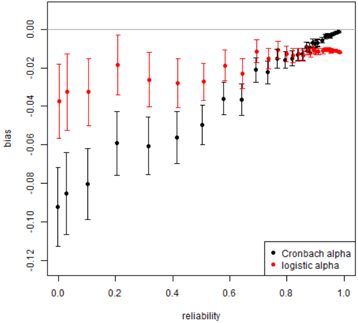
Bias and MSE of two estimators of reliability, item difficulties from $(-0.1, 0.1)$. Number of students $n = 25$, number of items $m = 10$, number of simulates $N = 1000$.



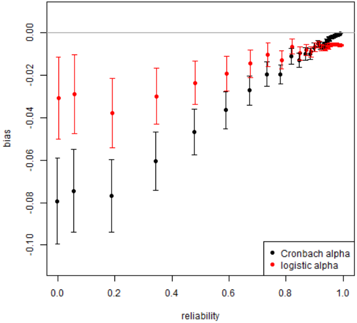
Bias and MSE of two estimators of reliability, item difficulties from $(-3, 3)$. Number of students $n = 25$, number of items $m = 10$, number of simulates $N = 1000$.

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

Simulations: Cronbach's and logistic alpha

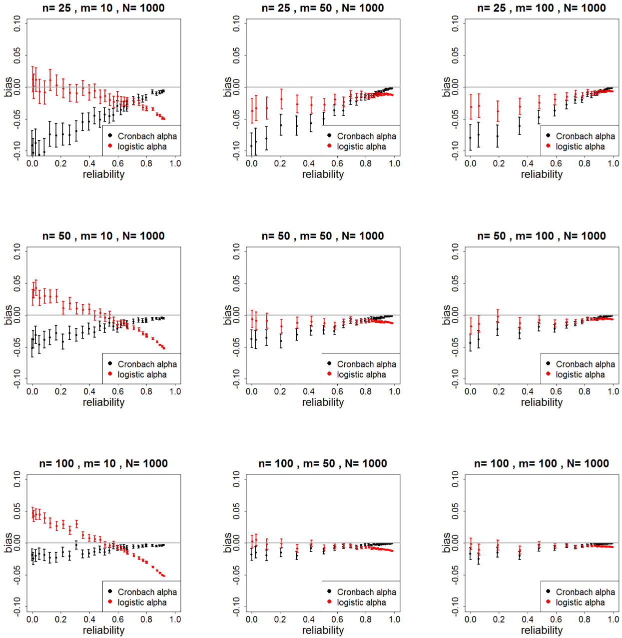


Bias and MSE of two estimators of reliability, number of students $n = 25$, number of items $m = 50$, number of simulates $N = 1000$.



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

Outline

1. Introduction
2. Reliability
3. Cronbach's alpha
4. Reliability in case of binary items
5. Simulation study
6. **Discussion and conclusion**

Discussion and Conclusion

- Estimation of reliability is important. It needs to be followed by analysis of validity.
- Cronbach's alpha is suitable only in special situations (uncorrelated errors, T-equivalent items), and shouldn't be recommended as the generally most appropriate estimator of reliability.
- New estimate of reliability for case of binary items has promising properties especially for lower true reliabilities and high number of items.
- Nevertheless, under assumptions of uncorrelated errors and T-equivalent items, Cronbach's alpha has good properties in case of binary items, too, and it is easier to compute.

Discussion and Conclusion

- psychometric research in the Czech Republic

Thank you for your attention!