Various generalizations of the Upper-Lower Index and their critical values

Psychoco 2019

Lubomír Štěpánek[1]

Patrícia Martinková[2, 3]

[1]Institute of Biophysics and Informatics
First Faculty of Medicine
Charles University, Prague

FACULTY
OF EDUCATION
Charles University

[2]Faculty of Education
Charles University, Prague

[3]Institute of Computer Science
Czech Academy of Sciences

Content

Concept of item discrimination

- one of the item parameters characterizing item properties
  - besides item difficulty, pseudo-guessing, inattention
- describes how accurately an item can discriminate between individuals with high vs. with low ability
- item discrimination can be evaluated using both Classical Test Theory (CTT) and Item-Response Theory (IRT)

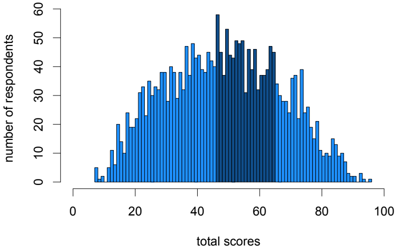
Indicators of item discrimination

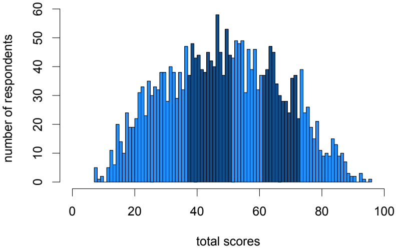
- based on (Pearson's) correlation
  - item-test correlation (RIT)
  - item-rest correlation (RIR)
  - tetrachoric coefficient of correlation, χ^2 statistics
- based on difference of proportions of correct answers
  - Kelley's discrimination index (classical Upper-Lower Index, ULI)
  - its variants and generalizations (generalized Upper-Lower Index, gULI)

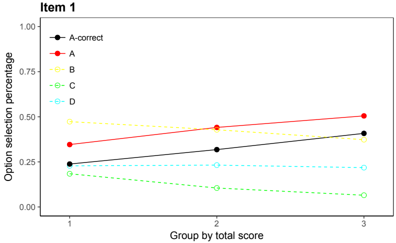
Motivation

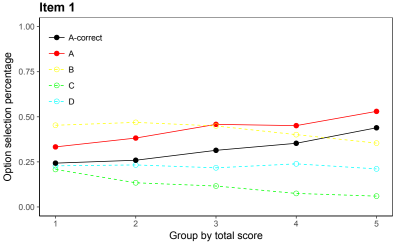
- traditional methods
  - easily explainable to practitioners
  - easily calculated
- methods based on IRT
  - sophisticated models using p-values, information criterions etc.

Motivation (con't)









Aims of the study

- to do literature review
- to look into properties of generalizations of ULI (gULI)
- to find reasoning behind suggested critical values

Kelley's item discrimination index

- let's assume a test with n test-takers and k binary items
- using ordered total scores of the n test-takers, the group is divided into three parts considering only highest m and lowest m of the test-takers
- assuming that p_U, or p_L is a proportion of correct answers in the highest, or lowest m test-takers, then[1]

$$\mathrm{ULI} \equiv p_U - p_L$$

- Kelley suggests to choose m such that $\frac{m}{n} \doteq 0.27$, but did not offer any critical values for the calculated ULI

[1] T. L. Kelley. "The selection of upper and lower groups for the validation of test items." In: *Journal of Educational Psychology* 30.1 (1939), pp. 17–24. DOI: 10.1037/h0057123. URL: https://doi.org/10.1037/h0057123.

Kelley's optimal splitting of a group of test takers

- firstly let's assume there are n test-takers with scores sorted increasingly, $(x_{(1)}, x_{(2)}, \dots, x_{(n)})^T$ such that $\forall i \in \{1, 2, \dots, n\}$ is $x_{(i)} \in \mathbb{R}$ and the scores follow $\mathcal{N}(0, 1^2)$
- Kelley searched for $q = \frac{m}{n}$ in order to maximize t statistics of two sample t-test
- assuming that μ_U and μ_L are mean scores of upper and lower m test-takers, σ is a pooled standard deviation of test scores,

$$t = \frac{\mu_U - \mu_L}{\frac{\sigma}{\sqrt{m}}}$$

- given the assumed $\mathcal{N}(0, 1^2)$ distribution, therefore and equal size we get $\mu_U = -\mu_L$ and

$$t = \frac{2\mu_U \cdot \sqrt{m}}{\sigma} \propto \mu_U \cdot \sqrt{m}$$

Kelley's optimal splitting of a group of test takers

- since mean of the truncated normal distribution is $\mu_U = \frac{z_q}{q}$ where z_q is q-th quantile and $q = \frac{m}{n}$ we get

$$t \propto \mu_U \cdot \sqrt{m} \propto \frac{z_q}{q} \cdot \sqrt{m} \propto \frac{z_q}{q} \cdot \sqrt{qn} \propto \frac{z_q}{\sqrt{q}} \cdot \sqrt{n} \propto \frac{z_q}{\sqrt{q}}$$

- the first derivative with respect to x where x and $-x$ are points of division in a unit normal distribution is

$$\frac{\mathrm{d}t}{\mathrm{d}x} \propto \frac{z}{\sqrt{q}} \left(-x + \frac{z}{2q} \right)$$

- setting the derivation equal to zero we get

$$q = \frac{z}{2x}$$

  and using tables for the truncated normal distribution we could find $q \doteq 0.270$

Generalized Upper-Lower Index for binary items by Brennan

- Brennan generalized the ULI (here assigned ULI_B) by relaxing the assumption of equal sizes of the upper/lower group of test-takers[2]
- let's assume a test with n test-takers and k binary items
- using ordered total scores of the n test-takers, the group is divided into three parts considering only highest n_U and lowest n_L of the test-takers
- assuming that m_U, or m_L is a number of correct answers in the highest n_U, or lowest n_L test-takers, then

$$\mathrm{ULI}_B \equiv \frac{m_U}{n_U} - \frac{m_L}{n_L}$$

[2] Robert L. Brennan. "A Generalized Upper-Lower Item Discrimination Index". In: *Educational and Psychological Measurement* 32.2 (July 1972), pp. 289–303. DOI: 10.1177/001316447203200206. URL: https://doi.org/10.1177/001316447203200206

Critical values for Upper-Lower Index by Brennan

- Brennan also offered critical values for the ULI_B searching for the minimal values of ULI_B for which z statistic of z-test for two proportions becomes significant
- under null hypothesis $H_0 : p = p_U = p_L$ where p_U and p_L is a proportion of correct answers in both upper and lower part

$$z = \frac{\left(\frac{m_U}{n_U} - \frac{m_L}{n_L}\right) - 0}{\sqrt{\frac{n_U + n_L}{n_U n_L} p(1-p)}}$$

follows standard normal distribution

- for $n_U = n_L = 30$ and $p = p_U = p_L = 0.5$ such that $\mathbb{E}\left(\frac{m_U}{n_U}\right) = p_U$ and $\mathbb{E}\left(\frac{m_L}{n_L}\right) = p_L$ he got the well-known critical value for the ULI (or ULI_B) equal to 0.3 (approx.)

Generalized Upper-Lower Index for polytomous items by Metsämuuronen

- Metsämuuronen generalized the ULI in order to make it applicable to polytomous items

$$\mathrm{GDI}_m = \frac{\sum_{i=1}^{m} O_i^U - \sum_{i=1}^{m} O_i^L}{m(\max(O_i) - \min(O_i))}$$

where O_i^U and O_i^L refer to the score of the i-th highest and lowest test-taker given m highest and lowest observations[3]

[3] Jari Metsämuuronen. "Generalized Discrimination Index". en. In: (2018). DOI: 10.13140/rg.2.2.30933.88804/1. URL: http://rgdoi.net/10.13140/RG.2.2.30933.88804/1

Simulated optimal splitting of a group of test takers

- based on Kelley, let's assume a test with n test-takers and k binary items
- using ordered total scores of the n test-takers, the group is divided into three parts considering only highest m and lowest m of the test-takers
- assuming that p_U, or p_L is a proportion of correct answers in the highest, or lowest m test-takers, then

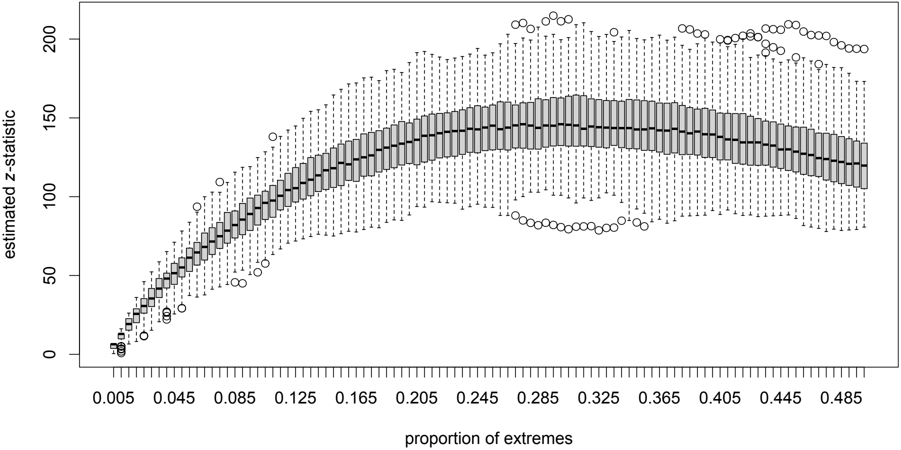
$$\mathrm{ULI} \equiv p_U - p_L$$

- let's suppose a probability of correct answer is given by $\frac{\exp(\theta)}{1+\exp(\theta)}$ where $\theta \sim \mathcal{N}(0, 1^2)$ and $p_L \sim \text{Bernoulli}\left(\frac{\exp(\theta)}{1+\exp(\theta)}\right)$ as well as $p_U \sim \text{Bernoulli}\left(\frac{\exp(\theta)}{1+\exp(\theta)}\right)$

Simulated optimal splitting of a group of test takers

```
n := 1000      // size of a sample of test-takers (or their scores);
k := 200       // number of repetitions per one given proportion i;
z := {∅}       // a list of z-statistics;
for i = 1 : ⌊n/2⌋ do
    z-vector := {∅}      // vector of z-statistics for given i;
    for j = 1 : k do
        abilities := vector of n numbers following ~ N(0, 1^2);
        scores := {∅};
        for θ in abilities do
            scores = {scores, Bernoulli(exp(θ)/(1+exp(θ)))};
        end
        z-statistic := z-statistic of test of two proportions based on lower i and upper
          i number of the vector scores;
        z-vector = {z-vector, z-statistic};
    end
    z = {z, z-vector};
end
```

Simulated optimal splitting of a group of test takers



Critical values for gULI

- based on Brennan, let's assume a test with n test-takers and k binary items
- using ordered total scores of the n test-takers, the group is divided into three parts considering only highest n_U and lowest n_L of the test-takers
- assuming that m_U, or m_L is a number of correct answers in the highest n_U, or lowest n_L test-takers, then

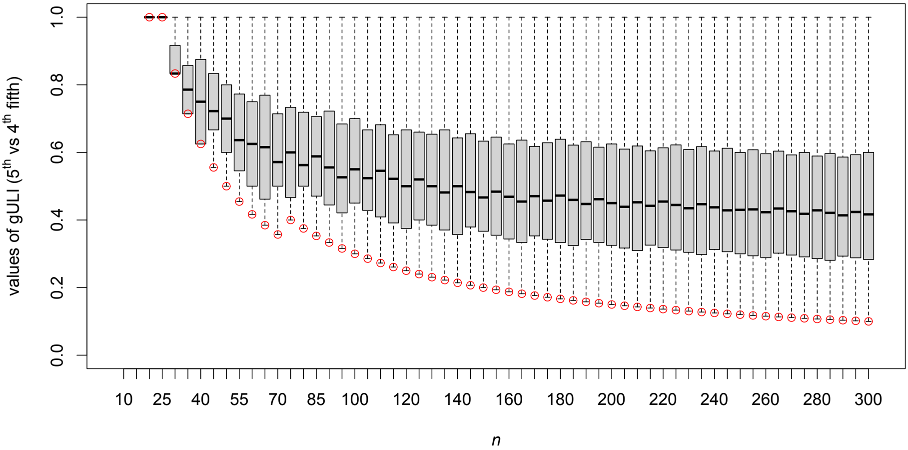
$$\mathrm{ULI}_B \equiv \frac{m_U}{n_U} - \frac{m_L}{n_L}$$

- assuming n, $n_U = \frac{1}{5}n$, $n_L = \frac{1}{5}n$ is given in our case, there is $|\{0, 1, \ldots, n_L\} \times \{0, 1, \ldots, n_U\}|$ combinations of possible ULI_B
- we search for the lowest value of ULI_B such that z statistic of z-test of two proportions is significant given n, n_U, n_L

Critical values for gULI using the 5$^{\text{th}}$ and 4$^{\text{th}}$ fifth of scores

$n_{\min} := 10$ `// minimal size of a sample of test-takers or their scores;`
$n_{\max} := 300$ `// maximal size of a sample of test-takers or their scores;`
ULI $:= \{\emptyset\}$ `// a list of ULI's;`
for $n = n_{\min} : n_{\max}$ **do**
  $n_{\text{lower}} := 0.2n$ `// a size of lower part of test-takers' sample;`
  $n_{\text{upper}} := 0.2n$ `// a size of upper part of test-takers' sample;`
  significant_ULI $:= \{\emptyset\}$ `// a vector of significant ULI's for given` n;
  for $i = 0 : n_{\text{lower}}$ **do**
    for $j = 0 : n_{\text{upper}}$ **do**
      if $\frac{j}{n_{\text{upper}}} \geq \frac{i}{n_{\text{lower}}}$ and z-statistic for $\frac{i}{n_{\text{lower}}}$ vs. $\frac{j}{n_{\text{upper}}}$ is significant **then**
        significant_ULI $= \left\{\text{significant\_ULI}, \frac{j}{n_{\text{upper}}} - \frac{i}{n_{\text{lower}}}\right\}$;
      end
    end
  end
  ULI $= \{\text{ULI}, \text{significant\_ULI}\}$;
end

Critical values for gULI using the 5$^{\text{th}}$ and 4$^{\text{th}}$ fifth of scores



Critical values for gULI using the 5$^{\text{th}}$ and 4$^{\text{th}}$ fifth of scores

- values for gULI between the 5$^{\text{th}}$ and 4$^{\text{th}}$ fifth of scores for given n guaranteed the z statistics using the gULI would be significant, $\alpha = 0.05$, regardless of p_L, p_U

n	gULI (5$^{\text{th}}$, 4$^{\text{th}}$ fifth)
25	0.833
30	0.714
40	0.556
50	0.455
80	0.352
100	0.286
150	0.194

Going further

- deriving analytically (optimally) properties of distribution of minimal values of significant ULI's
- more IRT-based approach
  - using the approach based on searching for lowest values of discrimination coefficient a of a selected IRT model such that $H_0 : \hat{a} = 0$ is rejected
- taking account into the fact that z statistic does not respect an order or "closeness" of the upper and lower part of scores, e. g. the 5$^{\text{th}}$ and 4$^{\text{th}}$ fifth of scores vs. the 5$^{\text{th}}$ and 1$^{\text{st}}$ fifth of scores and their gULI's

Conclusion

- there are still good reasons to use traditional item analysis approaches
- Upper-Lower index could be generalized to tailor this descriptive statistics to the needs of the individual test
- critical values for generalized ULI's are not usually provided or published
- relatively simple concepts and ideas such as basic inference hypotheses tests could help to search for the critical values
- some older concepts were revisited analytically and new one were derived using simulations in R language

References

Brennan, Robert L. "A Generalized Upper-Lower Item Discrimination Index". In: *Educational and Psychological Measurement* 32.2 (July 1972), pp. 289–303. DOI: 10.1177/001316447203200206. URL: https://doi.org/10.1177/001316447203200206.

Kelley, T. L. "The selection of upper and lower groups for the validation of test items." In: *Journal of Educational Psychology* 30.1 (1939), pp. 17–24. DOI: 10.1037/h0057123. URL: https://doi.org/10.1037/h0057123.

Metsämuuronen, Jari. "Generalized Discrimination Index". en. In: (2018). DOI: 10.13140/rg.2.2.30933.88804/1. URL: http://rgdoi.net/10.13140/RG.2.2.30933.88804/1.

Thank you for your attention!

lubomir.stepanek@lf1.cuni.cz