Lesson 10: Differential Item Functioning - part 2

Patrícia Martinková

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

Institute for Research and Development of Education
Faculty of Education, Charles University, Prague

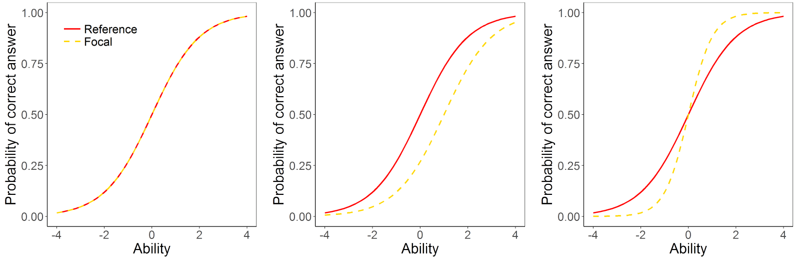
NMST570, December 11, 2018

Outline

Differential Item Functioning - Review

Differential Item Functioning (DIF)
Two subjects with the same underlying ability but from different groups have different probability to answer question correctly

- Two groups referred to as reference and focal (usually minority)
- Two types of DIF - uniform and non-uniform



Obrázek: A. No DIF. B. Uniform DIF. C. Non-uniform DIF

Examples of DIF items - Review

- Childhood illnesses (Drabinová & Martinková, 2017)
- Area of a cellar
- Tipping example (Martiniello et al., 2012)
- Spelling test - girger
- SAT item oarsman::regatta

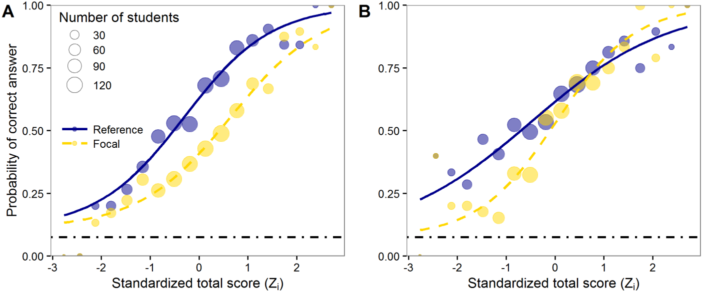
DIF and fairness:

- Existence of another dimension (secondary latent trait) besides the primary latent variable tested on the particular item
- Secondary latent trait causing DIF
  - Unrelated to content being tested
    - Unfair item, should be reworded or removed
  - Related to content being tested
    - Item may be kept, DIF may inform future teaching
- Content experts must decide on item fairness

DIF vs. difference in total scores - Review

Comparing total scores can lead to incorrect conclusions about fairness:

- Case study 1: Homeostasis Concept Inventory
  - Significant difference between males and females in total score
  - No HCI item detected as DIF
- Case study 2: Simulated dataset based on GMAT
  - Identical distributions of total score
  - Item 1 exhibits uniform DIF, Item 2 non-uniform DIF



Martinková et al. (2017)

DIF detection methods - Review

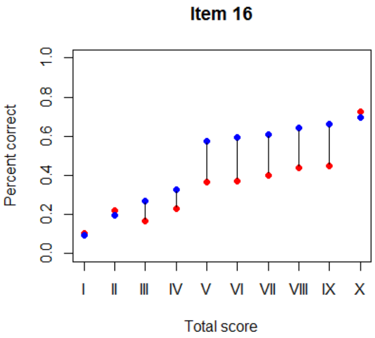
- Delta plot (Angoff & Ford (1973))
  - compares proportions of correct answers in the two groups
  - displays non-linear transformation of proportions (using quantiles)
- Mantel-Haenszel test
  - Test of independence of two binary variables: item score and group membership.
  - X^2 test, but incorporating also ability score
  - Looking at contingency tabels for each level of total score, adding up
- Logistic regression

$$\mathrm{P}(Y_{ij} = 1 | X_i, G_i) = \frac{e^{\beta_{0j} + \beta_{1j} X_i + \beta_{2j} G_i + \beta_{3j} X_i G_i}}{1 + e^{\beta_{0j} + \beta_{1j} X_i + \beta_{2j} G_i + \beta_{3j} X_i G_i}}$$

  - Probability of correct answer of student i to item j
  - X_i total score, G_i group
  - Test of submodel using F test, X^2, LR test, BIC/AIC

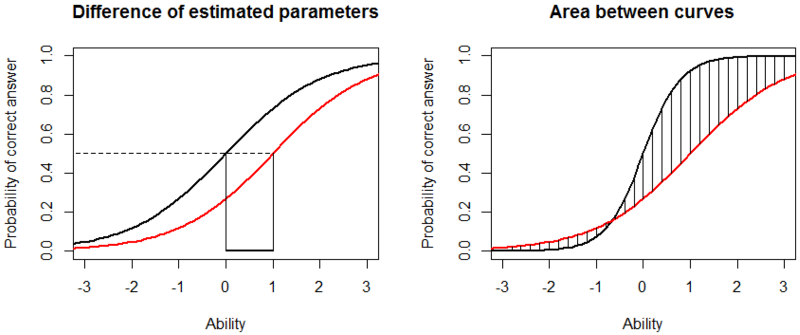
Standardization and SIBTEST

- Weighted average of the differences of success rates (at different levels of the test score) between focal and reference group



IRT-based Methods for DIF Detection

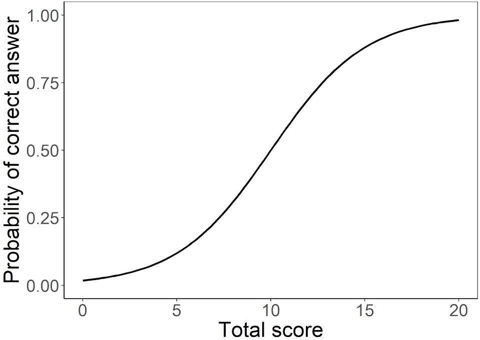
- Lord's Wald statistic: Difference between parameters
- Raju: Area between the curves (difference or absolute difference)
- Likelihood ratio test



Generalized logistic regression for DIF detection

$$\mathsf{P}(Y_{pi} = 1 | X_p, G_p) = \frac{e^{\alpha_i \quad (X_p - \beta_i \quad)}}{1 + e^{\alpha_i \quad (X_p - \beta_i \quad)}}$$

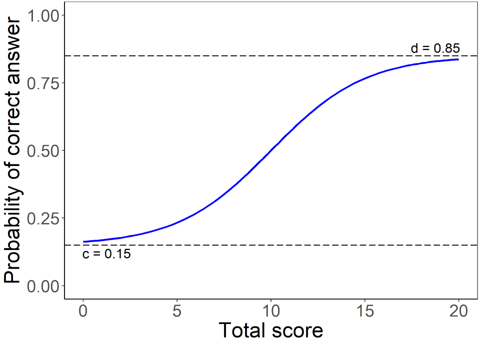
= probability of correct answer by person p on item i
X_p total score, G_p group membership



Generalized logistic regression for DIF detection

$$\mathsf{P}(Y_{pi} = 1 | X_p, G_p) = c_i + (d_i - c_i) \frac{e^{\alpha_i (X_p - \beta_i)}}{1 + e^{\alpha_i (X_p - \beta_i)}}$$

= probability of correct answer by person p on item i
X_p total score, G_p group membership

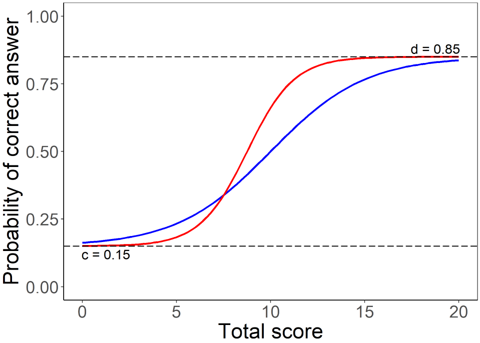


Generalized logistic regression for DIF detection

$$\mathrm{P}(Y_{pi} = 1 | X_p, G_p) = c_i + (d_i - c_i) \frac{e^{\alpha_{iG_p}(X_p - \beta_{iG_p})}}{1 + e^{\alpha_{iG_p}(X_p - \beta_{iG_p})}}$$

= probability of correct answer by person p on item i

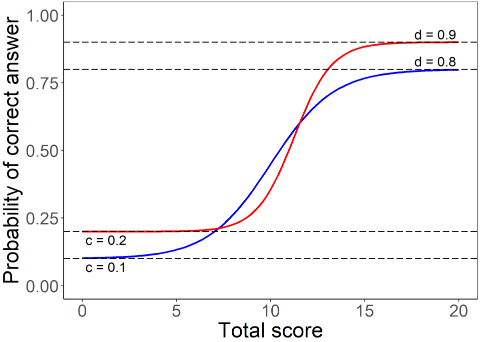
X_p total score, G_p group membership



Generalized logistic regression for DIF detection

$$\mathrm{P}(Y_{pi} = 1 | X_p, G_p) = c_{iG_p} + (d_{iG_p} - c_{iG_p}) \frac{e^{\alpha_{iG_p}(X_p - \beta_{iG_p})}}{1 + e^{\alpha_{iG_p}(X_p - \beta_{iG_p})}}$$

= probability of correct answer by person p on item i
X_p total score, G_p group membership



Technical details

We use:

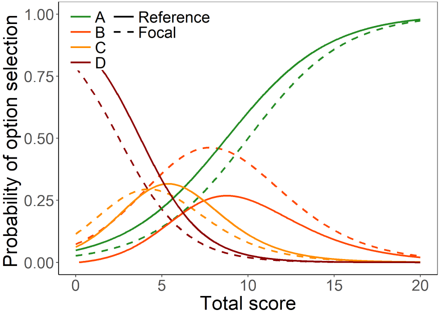
- Z-scores instead of total score
- IRT parameterization
- Non-linear least squares for parameter estimation
- DIF testing based on F or LR test
- Multiple comparison corrections

Method is implemented in R library `difNLR` (Drabinová, Martinková & Zvára, 2017)

Drabinová, Martinková & Zvára (2018): `difNLR`: Detection of Dichotomous DIF by Non-linear Regression. R package Version 1.2.2 https://CRAN.R-project.org/package=difNLR

Differential Distractor Functioning

Differential Distractor Functioning (DDF)
Two subjects with the same underlying ability but from different groups have different probability to choose given distractor in multiple-choice item

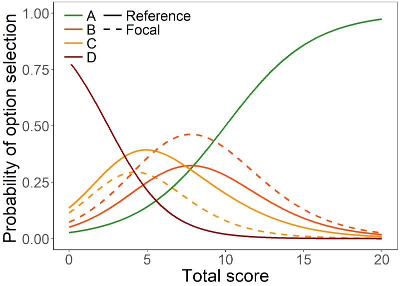


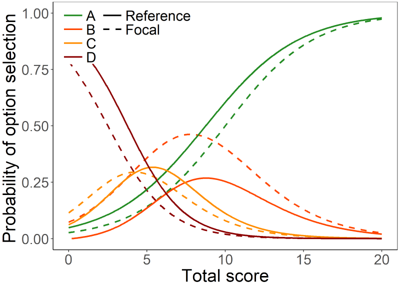
DDF with multinomial regression

$$\mathsf{P}(Y_{pi} = k|X_p, G_p) = \frac{e^{\alpha_{iG_pk}(X_p - \beta_{iG_pk})}}{1 + \sum_{l=1}^{K-1} e^{\alpha_{iG_pl}(X_p - \beta_{iG_pl})}} \qquad \text{(distractor)}$$

$$\mathsf{P}(Y_{pi} = K|X_p, G_p) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\alpha_{iG_pl}(X_p - \beta_{iG_pl})}} \qquad \text{(correct answer)}$$

= probability of option selection by person p on item i
X_p total score, G_p group membership





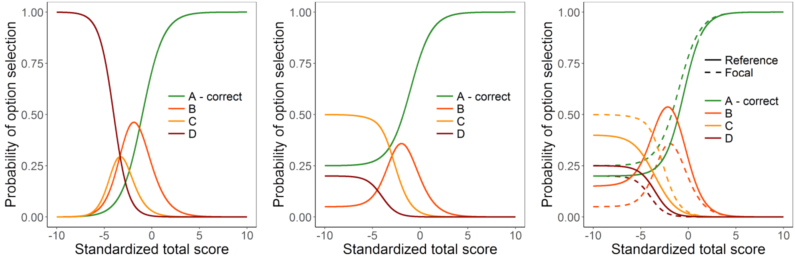
DDF for detection of differential attractiveness of distractors

Extending multinomial regression model

- To better describe attractiveness of distractors

Extending DDF model

- To account for differential attractiveness of distractors in multiple-choice items



Further Topics

Correction for multiple comparisons

- DIF analysis usually involves J multiple simultaneous statistical tests (J number of items)
- We are looking for adjusted p value, confidence level for the whole family of these tests
- Bonferroni correction, Benjamini-Hochberg, Holm, etc.

Item purification

- Iteratively removing the items currently flagged as DIF from the test scores
- Goal is to get purified sets of items, unaffected by DIF

DIF Effect size

- For very high number of respondents p values may all be significant
- Effect size measures enumerate magnitude of DIF

Monte Carlo simulation study

Goal

- To compare Non-linear regression method with other DIF detection methods

Design

- 5 levels of sample size (500+500, 500+1,000, 1,000+1,000, 1,000+2,000, 2,000+2,000)
- 20 items
- Answers generated using 3PL model
- DIF caused by difference in difficulty, discrimination and guessing parameters
- 0%, 5%, or 15% DIF proportion
- DIF size based on (weighted) area between characteristic curves

Drabinová & Martinková (2017): Detection of Differential Item Functioning with Non-Linear Regression: Non-IRT Approach Accounting for Guessing. *Journal of Educational Measurement*, 54(4), pp. 498-517, 2017. dx.doi.org/10.1111/jedm.12158

Monte Carlo simulation study

DIF detection

- Mantel-Haenszel, Logistic Regression, Lord (3PL IRT), **NLR**
- Benjamini-Hochberg multiple comparison correction

Results

- Less convergence issues than for Lord (3PL IRT)
- Good control of rejection rates in almost all scenarios
- Comparable power to other DIF detection methods
- Accounts for guessing
- Allows for testing group difference in guessing

Drabinová & Martinková (2017): Detection of Differential Item Functioning with Non-Linear Regression: Non-IRT Approach Accounting for Guessing. *Journal of Educational Measurement*, 54(4), pp. 498-517, 2017. dx.doi.org/10.1111/jedm.12158

Conclusion and vocabulary

- Differential item functioning (DIF)
- Differential distractor functioning (DDF)
- Reference and focal group
- Uniform and non-uniform DIF

DIF/DDF analysis should be used routinely in test development

- to check for fairness with respect to groups
- to inform teaching

Vocabulary cont.

DIF/DDF detection methods

- Delta-Plot
- Mantel-Haenszel test
- Standardization, SIBTEST
- Logistic regression
- Non-linear regression
- Multinomial regression (DDF)
- IRT-based methods: Lord's (Wald) test, LRT, Raju's test

Further issues in DIF detection:

- Correction for multiple comparisons
- Item purification
- DIF effect size

Simulation studies

Thank you for your attention!

www.cs.cas.cz/martinkova

References

- Martinková, Drabinová, Liaw, Sanders, McFarland & Price (2017). Checking Equity: Why DIF Analysis should be a Routine Part of Developing Conceptual Assessments. *CBE-Life Sciences Education*, 16(2), rm2. doi 10.1187/cbe.16-10-0307
- Drabinová & Martinková (2017). Detection of Differential Item Functioning with Non-Linear Regression: Non-IRT Approach Accounting for Guessing. *Journal of Educational Measurement*, 54(4), pp. 498-517, 2017. dx.doi.org/10.1111/jedm.12158
- Drabinová, Martinková & Zvára (2018): `difNLR`: Detection of Dichotomous DIF by Non-linear Regression. R package Version 1.2.2 https://CRAN.R-project.org/package=difNLR