Comparison of regression curves for detection of differential item functioning

Adéla Hladká & Patrícia Martinková

ISCB ČR, 12.9.2019

Institute of Computer Science of the Czech Academy of Sciences
Faculty of Mathematics and Physics, Charles University

Overview

Introduction

Differential item and distractor functioning

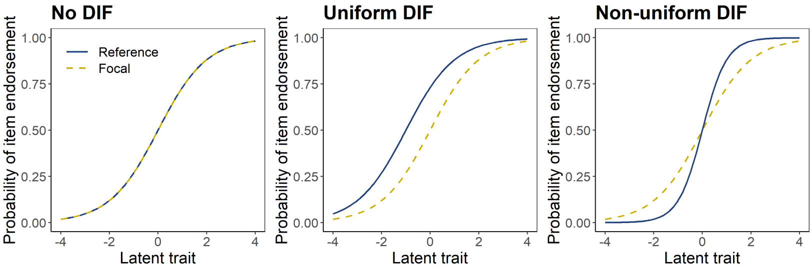
Definition of DIF

= respondents with the same latent trait but from different social groups have different probabilities to endorse an item

Latent trait = knowledge, health outcome, attitudes, etc.

Social group = gender, race, age, etc.

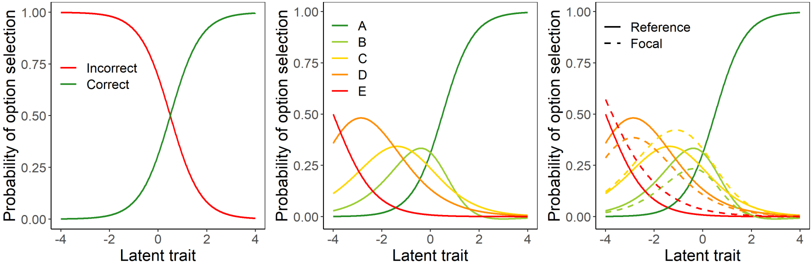
– reference (majority) and focal (minority)



Differential item and distractor functioning

Definition of DDF

= respondents with the same latent trait but from different social groups have different probabilities of option selection



Examples of DIF items

Pain *"How often did pain prevent you from walking more than 1 mile?"*
(reported more often by older patient[1])

"How often did pain prevent you from standing for more than 1 hour?"
(reported more often by older patients[1])

Depression *"I felt like crying"*
(endorsed more often by females[2])

Anger *"I was angry when people were unfair"*
(endorsed more often by older patients[2])

"I was angry when I did something stupid"
(endorsed more often by older patients[2])

---

[1] Amtmann, D. et al. (2010). Development of a PROMIS® item bank to measure pain interference. *Pain, 150*(1), 173-182.

[2] Pilkonis, P. A., et al. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. *Assessment, 18*(3), 263-283.

Examples of DIF items

Education *"Growth of long bones"*

- *A) occurs in growth cartilage*
- *B) is hormone-controlled*
- *C) usually ends at about 10-13 years of age, in boys earlier than in girls*
- *D) usually ends around 16-19 years of age, in girls earlier than in boys*

(more often correctly answered by males[3])

"Runner is to marathon as"

- *A) envoy to embassy*
- *B) martyr to massacre*
- *C) oarsman to regatta*
- *D) referee to tournament*
- *E) horse to stable*

(more often correctly answered by white students[4])

---

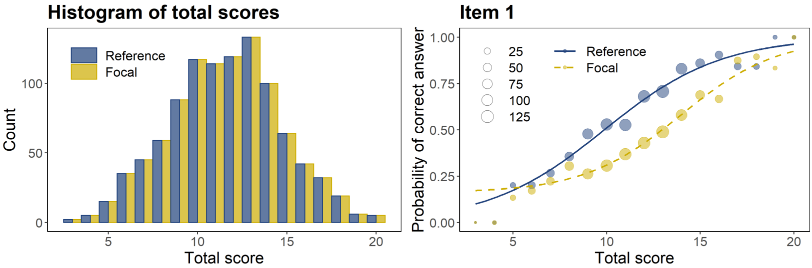
[3] Martinková, P., Hladká, A., Leupen, S., Štěpánek, L, & Králíčková, M. (2019). Submitted.

[4] Cramp, A., & McDougall, J. (2018). *Doing Theory on Education: Using Popular Culture to Explore Key Debates*. Routledge.

Why is DIF/DDF detection important?

Routine for checking item fairness in large-scale assessment[5]

- Difference in total scores does not imply DIF
- DIF can be present without differences in total score!



DIF is not necessarily threat to fairness and validity

[5] Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), rm2.

More general problem description

Two measurements on two populations (reference and focal)

$$\mathrm{E}(Y_R|X_R) = \mathrm{P}(Y_R = 1|X_R) = m_R(X_R),$$
$$\mathrm{E}(Y_F|X_F) = \mathrm{P}(Y_F = 1|X_F) = m_F(X_F),$$

$Y_R \in \{0,1\}$, $Y_F \in \{0,1\}$ (endorsement of the item)
$\mathrm{E}|Y_R| < \infty$, $\mathrm{E}|Y_F| < \infty$, X_R, X_F (standardized) total score of the test

More general problem description

Two measurements on two populations (reference and focal)

$$\mathrm{E}(Y_R|X_R) = \mathrm{P}(Y_R = 1|X_R) = m_R(X_R),$$
$$\mathrm{E}(Y_F|X_F) = \mathrm{P}(Y_F = 1|X_F) = m_F(X_F),$$

$Y_R \in \{0, 1\}$, $Y_F \in \{0, 1\}$ (endorsement of the item)
$\mathrm{E}|Y_R| < \infty$, $\mathrm{E}|Y_F| < \infty$, X_R, X_F (standardized) total score of the test

We want to test
$H_0 : m_R \equiv m_F$ vs. $H_1 : m_R \not\equiv m_F$

More general problem description

Two measurements on two populations (reference and focal)

$$\mathrm{E}(Y_R|X_R) = \mathrm{P}(Y_R = 1|X_R) = m_R(X_R),$$
$$\mathrm{E}(Y_F|X_F) = \mathrm{P}(Y_F = 1|X_F) = m_F(X_F),$$

$Y_R \in \{0,1\}$, $Y_F \in \{0,1\}$ (endorsement of the item)
$\mathrm{E}|Y_R| < \infty$, $\mathrm{E}|Y_F| < \infty$, X_R, X_F (standardized) total score of the test

We want to test
$H_0 : m_R \equiv m_F$ vs. $H_1 : m_R \not\equiv m_F$

Two main goals:

1. Estimation of m_R and m_F
2. Comparison of m_R and m_F

DIF detection methods overview

Most often used methods:

- Mantel-Haenszel test[6]
    - Odds ratio across all ability levels for a specific item
- Logistic regression method[7]
    - Effect of ability, group membership and their interaction
- SIBTEST[8]
    - Similar to MH test, uses a regression correction
- IRT models
    - Wide range of models
    - Estimate of ability as a random effect of respondent

[6] Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719-748.

[7] Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement, 27*(4), 361-370.

[8] Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika, 58(2), 159-194.

DIF detection methods overview

Most often used methods:

Type of DIF	Uniform	Non-uniform	Other
Mantel-Haenszel	✓	✗	✗
Logistic regression	✓	✓	✗
SIBTEST	✓	✗	✗
IRT models	✓	✓	✓

DIF detection methods overview

Most often used methods:

Type of DIF	Uniform	Non-uniform	Other
Mantel-Haenszel	✓	✗	✗
Logistic regression	✓	✓	✗
SIBTEST	✓	✗	✗
IRT models	✓	✓	✓

Other properties	Score-based	Small samples	Easy to fit
Mantel-Haenszel	✓	✓	✓
Logistic regression	✓	✓	✓
SIBTEST	✓	✓	✓
IRT models	✗	✗	✗

Research methods

Nonlinear regression for DIF detection

- Extension of logistic regression method for DIF detection[7,9]
- Introducing guessing and inattention parameters
- Allows for testing difference in these parameters
- Also called 4PL non-IRT model

[7] Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement, 27*(4), 361-370.

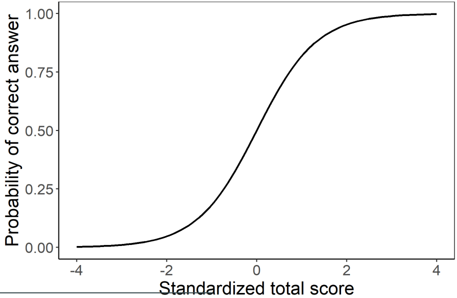
[9] Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498-517.

Nonlinear regression for DIF detection

$$\mathrm{P}(Y_{pi} = 1 | X_p, G_p) = \frac{e^{a_i \quad (X_p - b_i \quad)}}{1 + e^{a_i \quad (X_p - b_i \quad)}}$$

= probability that person p endorses an item i

X_p (standardized) total score, G_p group membership[9]



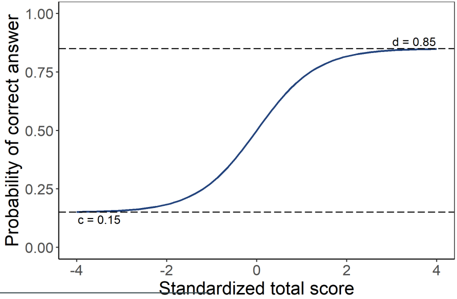
[9] Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498-517.

Nonlinear regression for DIF detection

$$\mathrm{P}(Y_{pi} = 1 | X_p, G_p) = c_i \quad + (d_i \quad - c_i \quad) \frac{e^{a_i \quad (X_p - b_i \quad)}}{1 + e^{a_i \quad (X_p - b_i \quad)}}$$

= probability that person p endorses an item i

X_p (standardized) total score, G_p group membership[9]



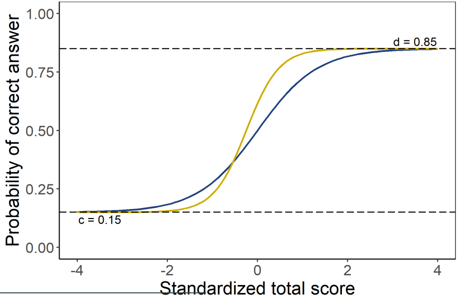
[9] Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498-517.

Nonlinear regression for DIF detection

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

= probability that person p endorses an item i

X_p (standardized) total score, G_p group membership[9]



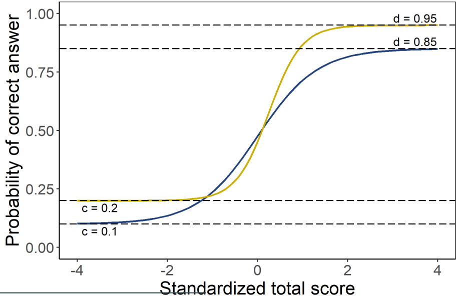
[9] Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498-517.

Nonlinear regression for DIF detection

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

= probability that person p endorses an item i

X_p (standardized) total score, G_p group membership[9]



[9] Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498-517.

Parametric approaches for DIF/DDF detection

- Extension of logistic regression for ordinal and nominal data
- Wide range of models including:
  - Cumulative logit model
  - Adjacent category logit model
  - Multinomial model

Cumulative logit model

For $K+1$ ordinal outcome

$$\mathrm{P}(Y_{ip} \geq k | X_p, G_p) = \frac{e^{a_{iG_p}(X_p - b_{iG_p k})}}{1 + e^{a_{iG_p}(X_p - b_{iG_p k})}},$$

Category probability for $k = 0, \ldots, K-1$

$$\mathrm{P}(Y_{ip} = k | X_p, G_p) = \mathrm{P}(Y_{ip} \geq k | X_p, G_p) - \mathrm{P}(Y_{ip} \geq k+1 | X_p, G_p)$$

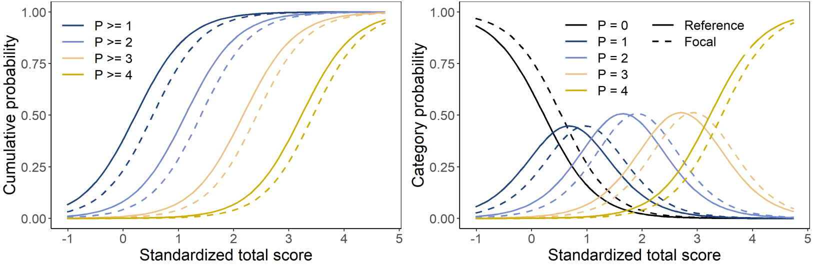
where $a_{iG_p}(X_p - b_{iG_p 0}) = 0$

X_p (standardized) total score, G_p group membership

Proxy to a graded response IRT model[10]

[10] Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika 34*(Suppl 1).

Cumulative logit model



Adjacent category logit model

For $K+1$ ordinal outcome

$$\log \frac{\mathrm{P}(Y_{ip} = k|X_p, G_p)}{\mathrm{P}(Y_{ip} = k-1|X_p, G_p)} = a_{iG_p}(X_p - b_{iG_pk})$$

Category probability for $k = 0, \ldots, K$

$$\mathrm{P}(Y_{ip} = k|X_p, G_p) = \frac{e^{\sum_{l=0}^{k} a_{iG_p}(X_p - b_{iG_pl})}}{\sum_{j=0}^{K} e^{\sum_{l=0}^{j} a_{iG_p}(X_p - b_{iG_pl})}},$$

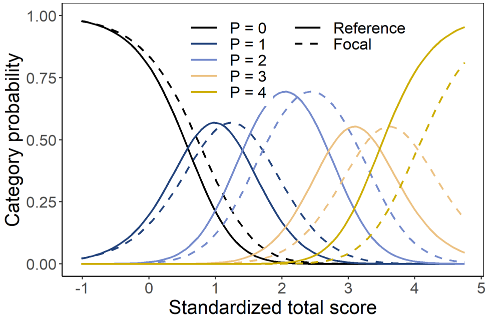
where $a_{iG_p}(X_p - b_{iG_p0}) = 0$

X_p (standardized) total score, G_p group membership

Proxy to a rating scale IRT model[11]

[11] Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Adjacent category logit model



Multinomial model

For $K+1$ nominal outcome

$$\mathrm{P}(Y_{pi} = k | X_p, G_p) = \frac{e^{\alpha_{iG_pk}(X_p - \beta_{iG_pk})}}{\sum_{l=0}^{K} e^{\alpha_{iG_pl}(X_p - \beta_{iG_pl})}},$$

$=$ probability of option selection k by person p on item i

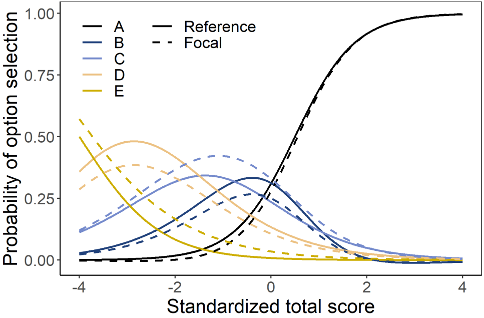
where $k = 0, \ldots, K$ and $\alpha_{iG_p0}(X_p - \beta_{iG_p0}) = 0$

X_p (standardized) total score, G_p group membership

Proxy to Bock's nominal model[12]

[12] Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Multinomial model



Nonparametric approaches for DIF detection

- Estimation of 3PL-4PL IRT and non-IRT models is challenging
- And requires large sample size in both groups (≥ 500)
- Parametric model does not necessarily correspond to reality

Nonparametric approaches for DIF detection

- Estimation of 3PL-4PL IRT and non-IRT models is challenging
- And requires large sample size in both groups ($\geq$ 500)
- Parametric model does not necessarily correspond to reality

- Need for method which detects DIF caused by various sources

Kernel smoothing estimate of ICC

Nearest-neighbor estimate[13, 14]

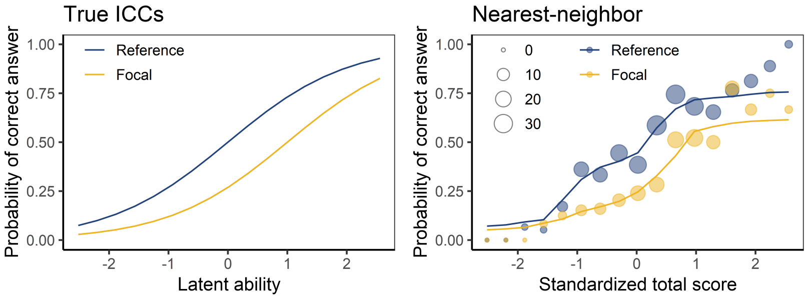
$$\hat{m}_R(x) = \sum_{p=1}^{n_R} Y_{Rp} W_{Rp}(x),$$

$$W_{Rp}(x) = \frac{K\left(\frac{\hat{F}_R(X_{Rp}) - \hat{F}_R(x)}{h}\right)}{\sum_{k=1}^{n_R} K\left(\frac{\hat{F}_R(X_{Rk}) - \hat{F}_R(x)}{h}\right)}$$

- K symmetric kernel function
- $\hat{F}_R(x)$ empirical distribution function of $X_{R1}, \ldots, X_{Rn_R}$
- h bandwidth
- n_R number of respondents in the reference group

[13] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications, 9*(1), 141-142.
[14] Srihera, R., & Stute, W. (2010). Nonparametric comparison of regression functions. *Journal of Multivariate Analysis, 101*(9), 2039–2059

Kernel smoothing estimate



Test statistic

Test statistic: [14]

$$\hat{T} = \frac{1}{n_R n_F} \sum_{i=1}^{n_R} \sum_{j=1}^{n_F} W\left(\frac{X_{Ri} + X_{Fj}}{2}\right) \left[\hat{m}_R\left(\frac{X_{Ri} + X_{Fj}}{2}\right) - \hat{m}_F\left(\frac{X_{Ri} + X_{Fj}}{2}\right)\right]$$

- Can be shown that $\hat{T}$ is normally distributed
- Which weight function W to use?

[14] Srihera, R., & Stute, W. (2010). Nonparametric comparison of regression functions. *Journal of Multivariate Analysis, 101*(9), 2039–2059

Weight function

1. Fixed weight function[14]

$$W_1(x) = 1, \quad \forall x$$

[14] Srihera, R., & Stute, W. (2010). Nonparametric comparison of regression functions. *Journal of Multivariate Analysis, 101*(9), 2039–2059

[15] Hladká, A., & Martinková, P. (2019). Nonparametric comparison of regression curves for DIF detection. In progress.

Weight function

1. Fixed weight function[14]

$$W_1(x) = 1, \quad \forall x$$

2. Optimal weight function[9,15]
(in the sense of maximizing power of the test)

$$W_O(x) = \frac{m_R(x) - m_F(x)}{(1-\lambda)m_R(x)(1-m_R(x))\frac{e(x)}{f_R(x)} + \lambda m_F(x)(1-m_F(x))\frac{e(x)}{f_F(x)}}$$

$\lambda = \lim \frac{n_R}{n_R+n_F}$

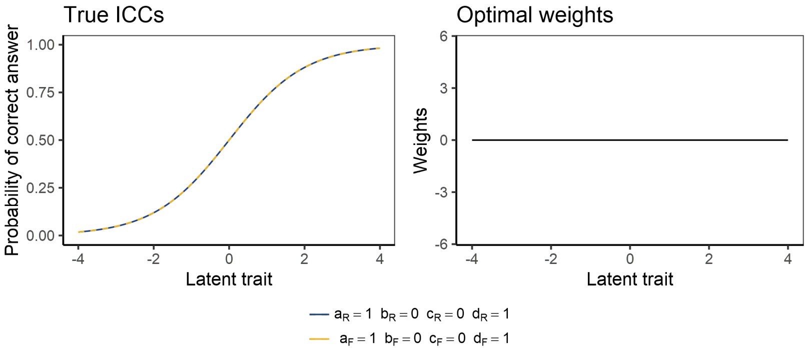
$f_R(x), f_F(x)$ pdf of X_R and X_F, $e(x)$ pdf of $\frac{X_R+X_F}{2}$

[14] Srihera, R., & Stute, W. (2010). Nonparametric comparison of regression functions. *Journal of Multivariate Analysis, 101*(9), 2039–2059

[15] Hladká, A., & Martinková, P. (2019). Nonparametric comparison of regression curves for DIF detection. In progress.

Weight function

For 4PL IRT model with normally distributed latent trait[15]



[15] Hladká, A., & Martinková, P. (2019). Nonparametric comparison of regression curves for DIF detection. In progress.

Weight function

3. Natural estimate of optimal weights[15]

$$\hat{W}_O(x) = \frac{\hat{m}_R(x) - \hat{m}_F(x)}{(1-\hat{\lambda})\hat{m}_R(x)(1-\hat{m}_R(x))\frac{\hat{e}(x)}{\hat{f}_R(x)} + \hat{\lambda}\hat{m}_F(x)(1-\hat{m}_F(x))\frac{\hat{e}(x)}{\hat{f}_F(x)}}$$

- Using kernel smoothing estimates $\hat{m}_R(x)$ and $\hat{m}_F(x)$
- Test statistic is no longer normally distributed
- Asymptotic distribution not known

[15] Hladká, A., & Martinková, P. (2019). Nonparametric comparison of regression curves for DIF detection. In progress.

Wild bootstrap

Wild bootstrap[15, 16, 17]

1. **Perform DIF detection:**
   - Estimate m_R and m_F with $\hat{m}_R$ and $\hat{m}_F$
   - Estimate W_O with $\hat{W}_O$
   - Calculate $\hat{T}$ using $\hat{W}_O$
2. **Estimate under H_0:**

   $(\hat{y}_p)_{p=1}^N$ fitted values
   $(\hat{e}_p)_{p=1}^N$ residuals

[15] Hladká, A., & Martinková, P. (2019). Nonparametric comparison of regression curves for DIF detection. In progress.

[16] Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics, 14*(4), 1261-1295.

[17] Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics, 21*(1), 255-285.

Wild bootstrap

3. **Bootstrapped samples,** for $b = 1, \ldots B$:

   3A. **Create samples:**

   $$y^*_{pb} = \hat{y}_p + v_{pb}\hat{e}_p, \text{ where}$$

   $$v_{pb} = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}), \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}) \end{cases}$$

   3B. **Estimates:**

   - Mean functions m^*_{Rb} and m^*_{Fb}
   - Optimal weight function W^*_{Ob}

   3B. **Perform DIF detection:**

   - Calculate $\hat{T}^*_b$

4. **Compare $\hat{T}^*_b$ with $\hat{T}$**

Other topics

Most methods for DIF detection

- Test for DIF in one item after another
- This may cause two issues
  1. Potential bias if DIF items are present
  2. Inflated Type I error rates due to multiple comparisons

Other topics

Most methods for DIF detection

- Test for DIF in one item after another
- This may cause two issues
  1. Potential bias if DIF items are present
  2. Inflated Type I error rates due to multiple comparisons

These drawbacks can be addressed by two controlling procedures:

1. Item purification
2. Adjustments for multiple comparisons

Other topics

Most methods for DIF detection

- Test for DIF in one item after another
- This may cause two issues
  1. Potential bias if DIF items are present
  2. Inflated Type I error rates due to multiple comparisons

These drawbacks can be addressed by two controlling procedures:

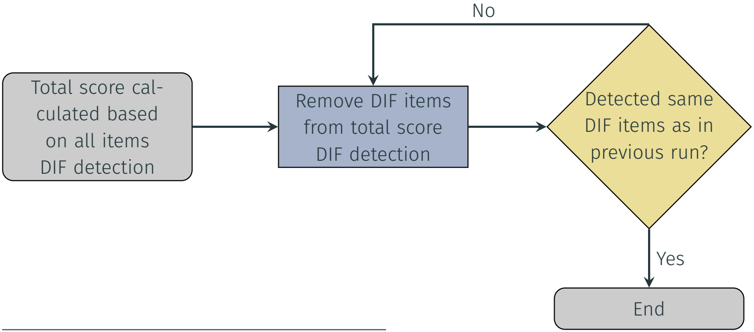
1. Item purification
2. Adjustments for multiple comparisons

- Conceptually different with different purposes
- Share the same objective - improvement of DIF detection

Item purification

Item purification[18]

= iterative removal of items flagged as DIF from the matching criterion (e.g., total score)



[18] Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*(3), 253-260.

Multiple comparison corrections

Multiple comparison corrections

- also called adjustments of p-values
- easy to implement
- non-iterative procedures that improve the accuracy of DIF detection[19]

Holm's procedure controls family-wise error[20]

Benjamini-Hochberg (BH) procedure controls false discovery rate[21]

---

[19] Kim, J., & Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement, 73*(3), 458-470.

[20] Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 65-70.

[21] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological), 57*(1), 289-300.

Multiple comparison corrections

Example

Item	Order	p-value	DIF	Holm's Boundary	Holm's DIF	BH Boundary	BH DIF
5	1	0.001	✓	0.005	✓	0.005	✓
10	2	0.004	✓	0.006	✓	0.010	✓
9	3	0.011	✓	0.006	X	0.015	✓
8	4	0.018	✓	0.007	X	0.020	✓
3	5	0.021	✓	0.008	X	0.025	✓
6	6	0.031	✓	0.010	X	0.030	X
2	7	0.039	✓	0.013	X	0.035	X
4	8	0.243	X	0.017	X	0.040	X
7	9	0.362	X	0.025	X	0.045	X
1	10	0.783	X	0.050	X	0.050	X

Simulation studies

Simulation study 1: Nonlinear regression

Aims[9]

- Investigation of properties of 3PL non-IRT model (nonlinear regression)
- Comparison to commonly used methods

---

[9] Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498-517.

Simulation study 1: Nonlinear regression

Aims[9]

- Investigation of properties of 3PL non-IRT model (nonlinear regression)
- Comparison to commonly used methods

DIF detection methods:

- Mantel-Haenszel test
- Logistic regression
- Lord's test (3PL IRT model)
- Nonlinear regression (3PL non-IRT model)

Design factors:

- Sample size (5)
- DIF type (2)
- DIF proportion (2 + 1)
- DIF size (3)

In total 4 detection approaches

In total $5 \times 2 \times 2 \times 3 + 5 = 65$ designs

[9] Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498-517.

Simulation study 1: Results

- Lower rate of convergence failures compared to 3PL IRT model
- Good control of type I error
- Sufficient power

[9] Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498-517.

Simulation study 1: Results

- Lower rate of convergence failures compared to 3PL IRT model
- Good control of type I error
- Sufficient power

- Possibility to account for guessing
- Possibility to detect DIF caused by various guessing

[9] Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498-517.

Simulation study 2: Nonparametric methods

Aims[15]

- Investigation of properties of nonparametric method

Design factors:

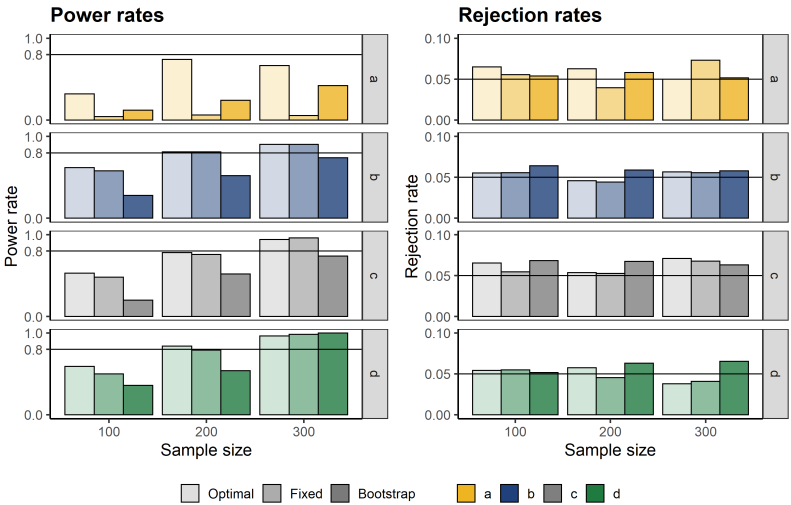
- 20 items (1 DIF, 19 non-DIF)
- 4PL IRT model with DIF caused parameters *a*, *b*, *c*, or *d*
- Sample sizes N = 100, 200, and 300

Simulation setting:

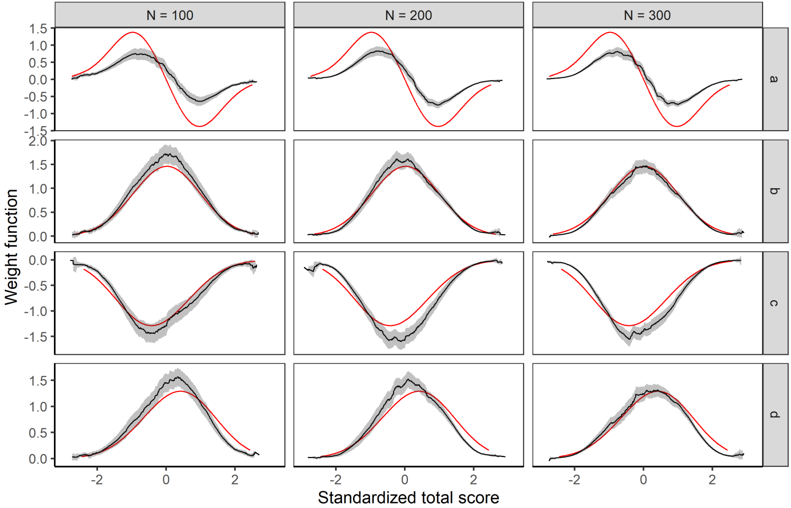
- Epanechnikov kernel $K(u) = \frac{3}{4}(1 - u^2), |u| \leq 1, h \sim n^{-\frac{7}{24}}$
- Using optimal weights W_0, fixed weights W_1, and natural estimate $\hat{W}_0$ with bootstrap
- 100 simulation runs

[15] Hladká, A., & Martinková, P. (2019). Nonparametric comparison of regression curves for DIF detection. In progress.

Simulation study 2: Very first results



Simulation study 2: Estimates of weights



Simulation study 3: Item purification and corrections

Research questions:[22]

Q1. Are the DIF detection methods able to control for Type I error with sufficient power even without any controlling procedure?

Q2. Which of the studied controlling procedures are significantly superior over others?

Q3. What factors have significant impact on Type I error and power rates?

[22] Hladká, A., Martinková, P., & Magis, D. (2019). Issues and practice in detection of differential item functioning: Applying item purification, correction for multiple comparisons, or combination of both? *Educational Measurement: Issues and Practice.* Under review.

Simulation study 3: Study design

DIF detection methods:

- Mantel-Haenszel test
- Logistic regression
- Lord's test (2PL IRT model)
- SIBTEST

Controlling procedures:

- None
- Item purification
- 2 corrections: Holm's and BH
- 2 mixtures

Design factors:

- Sample size (3)
- Test length (2)
- DIF type (2)
- DIF proportion (2 + 1)
- DIF size (2)
- Ability distribution (3)

In total $4 \times 6 = 24$ detection approaches

In total $3 \times 2 \times 2 \times 2 \times 2 \times 3 + 3 \times 2 \times 3 = 162$ designs[22]

[22] Hladká, A., Martinková, P., & Magis, D. (2019). Issues and practice in detection of differential item functioning: Applying item purification, correction for multiple comparisons, or combination of both? *Educational Measurement: Issues and Practice.* Under review.

Simulation study 3: Questions and answers

Research questions:

Q1. Are the DIF detection methods able to control for Type I error with sufficient power even without any controlling procedure?

Researchers' answers:

A1.
- Good control of Type I error in MH, LR, and SIBTEST
- Poor control of Type I error in Lord's test of 2PL IRT model
- MH and SIBTEST not able to detect non-uniform DIF

Simulation study 3: Questions and answers

Research questions:

Q2. Which of the studied controlling procedures are significantly superior over others?

Researchers' answers:

A2.
- No significant effect of item purification on power
- Significant increase of Type I error with item purification for all methods except MH
- Corrections caused rapid significant decrease in both Type I error and power rate
- Mixtures caused significant decrease in both Type I error and power rate
- Mixture of purification and BH correction was the most powerful after purification and none controlling procedure

Simulation study 3: Questions and answers

Research questions:

Q3. Are the DIF detection methods able to control for Type I error with sufficient power even without any controlling procedure?

Researchers' answers:

A3.
- Type I error mainly influenced by test length and sample size
- Power rate positively influenced by sample size, DIF proportion, DIF size and test length

Implementation and examples

Implementation - parametric methods

difNLR: DIF and DDF detection by non-linear regression models[23]

- R package (over 23,000 downloads)
- Version 1.3.0 on CRAN

```
install.packages("difNLR")
```

- The newest development version on GitHub

```
devtools::install_github("adelahladka/difNLR")
```

- Run it with

```
library("difNLR")
```

- Try some features online

https://shiny.cs.cas.cz/ShinyItemAnalysis/

[23] Hladká, A. & Martinková, P. (2019). difNLR: Generalized Logistic Regression Models for DIF and DDF Detection. *The R Journal*. Under review.

Implementation of parametric models

Main functions[15]

`difNLR()` DIF detection for dichotomous data based on non-linear regression model

`ddfORD()` DDF detection for ordinal data based either on adjacent category logit model or on cumulative logit model

`ddfMLR()` DDF detection for nominal data based on multinomial model

[23] Hladká, A. & Martinková, P. (2019). difNLR: Generalized Logistic Regression Models for DIF and DDF Detection. *The R Journal*. Under review.

Example - DIF detection with difNLR() function

A Measure of Anxiety[24]

```
data(Anxiety, package = "lordif")
dim(DataOrd <- Anxiety[, ids])
[1] 766   17

head(DataOrd)
  R3 R6 R8 R9 R10 R11 R12 R13 R18 R19 R20 R21 R24 R25 R26 R29
1  1  1  2  2   1   2   2   1   2   1   2   2   2   3   2   2
2  1  1  1  1   1   1   1   1   1   1   1   1   1   2   1   1
3  1  1  1  2   1   2   2   1   2   1   1   1   1   3   2   2
4  1  1  1  2   1   1   2   1   3   1   1   1   1   1   3   1
5  1  1  1  1   1   1   1   1   1   1   1   1   1   1   1   1
6  1  1  1  1   1   2   1   1   1   1   1   1   1   2   1   1

DataBin <- sapply(DataOrd, function(x) as.numeric(x >= 2))
table(group <- Anxiety$gender)
  0   1
369 397
```

[24] PROMIS Cooperative Group. Unpublished Manual for the Patient-Reported Outcomes Measurement Information System (PROMIS) Version 1.1. October, 2008: http://www.nihpromis.org

```
(fit1 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all"))
```

```r
(fit1 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all"))
# R6:  I was concerned about my mental health
# R20: My worries overwhelmed me
# R24: Many situations made me worry
```

```
Detection of all types of differential item functioning
using generalized logistic regression model

Generalized logistic regression likelihood ratio chi-square
statistics based on 3PL model with inattention parameter

Parameters were estimated with non-linear least squares

Item purification was not applied
No p-value adjustment for multiple comparisons

    Chisq-value P-value
R3   1.8134      0.6120
R6  15.8001      0.0012 **
R8   1.4523      0.6933
R9   3.4299      0.3300
R10  4.1015      0.2507
R11  4.5327      0.2094
R12  0.6706      0.8801
R13  0.5729      0.9026
R18  1.0155      0.7975
R19  0.3552      0.9493
R20 12.5446      0.0057 **
R21  6.9948      0.0721 .
R24  8.1791      0.0425 *
R25  2.7145      0.4378
R26  0.7457      0.8624
R29  1.2394      0.7436

Sign. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Detection thresholds: 7.8147 (significance level: 0.05)

Items detected as DIF items:
 R6
 R20
 R24
```

```r
(fit1 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all"))
# R6:  I was concerned about my mental health
# R20: My worries overwhelmed me
# R24: Many situations made me worry

# coefficients
round(coef(fit1), 3)
```

```r
(fit1 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all"))
# R6:  I was concerned about my mental health
# R20: My worries overwhelmed me
# R24: Many situations made me worry

# coefficients
round(coef(fit1), 3)
```

```
        a      b     d  aDif   bDif   dDif
R3  2.799  0.851 1.000 0.000  0.000  0.000
R6  1.685  0.483 1.000 0.964  0.197  0.000
R8  1.615  0.609 1.000 0.000  0.000  0.000
R9  1.518  0.262 1.000 0.000  0.000  0.000
R10 2.787  0.816 1.000 0.000  0.000  0.000
R11 1.683  0.198 1.000 0.000  0.000  0.000
R12 2.660 -0.409 0.963 0.000  0.000  0.000
R13 1.681  0.436 1.000 0.000  0.000  0.000
R18 2.173 -0.451 0.898 0.000  0.000  0.000
R19 2.523  0.834 1.000 0.000  0.000  0.000
R20 2.403  0.705 1.000 0.189 -0.254  0.000
R21 1.256  0.688 1.000 0.000  0.000  0.000
R24 3.072 -0.172 0.977 0.397 -0.225 -0.067
R25 3.233 -0.855 0.938 0.000  0.000  0.000
R26 3.928 -0.550 0.945 0.000  0.000  0.000
R29 3.173  0.266 0.956 0.000  0.000  0.000
```

```r
(fit1 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all"))
# R6:  I was concerned about my mental health
# R20: My worries overwhelmed me
# R24: Many situations made me worry

# coefficients
round(coef(fit1), 3)

# fit measures
AIC(fit1, item = 2)
BIC(fit1, item = 2)
logLik(fit1, item = 2)
```

```r
(fit1 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all"))
# R6:  I was concerned about my mental health
# R20: My worries overwhelmed me
# R24: Many situations made me worry

# coefficients
round(coef(fit1), 3)

# fit measures
AIC(fit1, item = 2)
BIC(fit1, item = 2)
logLik(fit1, item = 2)
```

```
[1] 485.8436

[1] 513.6907

'log Lik.' -236.9218 (df=6)
```

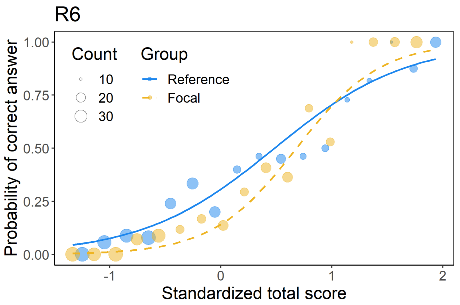
```r
(fit1 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all"))
# R6:  I was concerned about my mental health
# R20: My worries overwhelmed me
# R24: Many situations made me worry

# coefficients
round(coef(fit1), 3)

# fit measures
AIC(fit1, item = 2)
BIC(fit1, item = 2)
logLik(fit1, item = 2)

# prediction
predict(fit1, item = 2,
        match = 0, group = 0)
predict(fit1, item = 2,
        match = 0, group = 1)
```

```r
(fit1 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all"))
# R6:  I was concerned about my mental health
# R20: My worries overwhelmed me
# R24: Many situations made me worry

# coefficients
round(coef(fit1), 3)

# fit measures
AIC(fit1, item = 2)
BIC(fit1, item = 2)
logLik(fit1, item = 2)

# prediction
predict(fit1, item = 2,
        match = 0, group = 0)
predict(fit1, item = 2,
        match = 0, group = 1)
```

```
       R6
0.3071129

       R6
0.1417547
```

```r
(fit1 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all"))
# R6:  I was concerned about my mental health
# R20: My worries overwhelmed me
# R24: Many situations made me worry

# coefficients
round(coef(fit1), 3)

# fit measures
AIC(fit1, item = 2)
BIC(fit1, item = 2)
logLik(fit1, item = 2)

# prediction
predict(fit1, item = 2,
        match = 0, group = 0)
predict(fit1, item = 2,
        match = 0, group = 1)

# plotting ICC
plot(fit1, item = 2)
```

```r
(fit1 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all"))
# R6:  I was concerned about my mental health
# R20: My worries overwhelmed me
# R24: Many situations made me worry

# coefficients
round(coef(fit1), 3)

# fit measures
AIC(fit1, item = 2)
BIC(fit1, item = 2)
logLik(fit1, item = 2)

# prediction
predict(fit1, item = 2,
        match = 0, group = 0)
predict(fit1, item = 2,
        match = 0, group = 1)

# plotting ICC
plot(fit1, item = 2)
```



```r
# item purification
(fit2 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all",
                purify = TRUE))
```

```r
# item purification
(fit2 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all",
                purify = TRUE))
```

```
Detection of all types of differential item functioning
using generalized logistic regression model

Generalized logistic regression likelihood ratio chi-square
statistics based on 3PL model with inattention parameter

Parameters were estimated with non-linear least squares

Item purification was applied with 2 iterations.
No p-value adjustment for multiple comparisons

    Chisq-value P-value
R3   2.9094      0.4058
R6  12.2778      0.0065 **
R8   1.2140      0.7496
R9   4.0661      0.2544
R10  2.7692      0.4286
R11  4.5099      0.2114
R12  0.8727      0.8320
R13  0.3288      0.9545
R18  0.9653      0.8097
R19  0.0563      0.9965
R20  9.9210      0.0193 *
R21  7.4482      0.0589 .
R24  6.9028      0.0751 .
R25  2.2930      0.5139
R26  0.5606      0.9054
R29  2.0642      0.5592

Sign. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Detection thresholds: 7.8147 (significance level: 0.05)

Items detected as DIF items:
 R6
 R20
```

```r
# item purification
(fit2 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all",
                purify = TRUE))
# purification process
fit2$difPur
```

```r
# item purification
(fit2 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all",
                purify = TRUE))

# purification process
fit2$difPur
```

```
      R3 R6 R8 R9 R10 R11 R12 R13 R18
Step0  0  1  0  0   0   0   0   0   0
Step1  0  1  0  0   0   0   0   0   0
Step2  0  1  0  0   0   0   0   0   0
      R19 R20 R21 R24 R25 R26 R29
Step0   0   1   0   1   0   0   0
Step1   0   1   0   0   0   0   0
Step2   0   1   0   0   0   0   0
```

```r
# item purification
(fit2 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all",
                purify = TRUE))
# purification process
fit2$difPur
# multiple comparison correction
(fit3 <- difNLR(DataBin, group,
         p.adjust.method = "BH",
                focal.name = 1,
                model = "3PLd",
                type = "all"))
```

```r
# item purification
(fit2 <- difNLR(DataBin, group,
                focal.name = 1,
                model = "3PLd",
                type = "all",
                purify = TRUE))
# purification process
fit2$difPur
# multiple comparison correction
(fit3 <- difNLR(DataBin, group,
         p.adjust.method = "BH",
                focal.name = 1,
                model = "3PLd",
                type = "all"))
```

```
Detection of all types of differential item functioning
using generalized logistic regression model

Generalized logistic regression likelihood ratio chi-square
statistics based on 3PL model with inattention parameter

Parameters were estimated with non-linear least squares

Item purification was not applied
Multiple comparisons made with Benjamini-Hochberg
adjustment of p-values

    Chisq-value P-value Adj. P-value
R3   1.8134      0.6120  0.9493
R6  15.8001      0.0012  0.0199      *
R8   1.4523      0.6933  0.9493
R9   3.4299      0.3300  0.7542
R10  4.1015      0.2507  0.6686
R11  4.5327      0.2094  0.6686
R12  0.6706      0.8801  0.9493
R13  0.5729      0.9026  0.9493
R18  1.0155      0.7975  0.9493
R19  0.3552      0.9493  0.9493
R20 12.5446      0.0057  0.0459      *
R21  6.9948      0.0721  0.2883
R24  8.1791      0.0425  0.2264
R25  2.7145      0.4378  0.8755
R26  0.7457      0.8624  0.9493
R29  1.2394      0.7436  0.9493

Sign. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Detection thresholds: 7.8147 (significance level: 0.05)
Items detected as DIF items:
 R6
 R20
```

Example - DDF detection with ddfORD( ) function

A Measure of Anxiety[24]

```
summary(DataOrd)

   R3  R6  R8  R9 R10 R11 R12 R13 R18 R19 R20 R21 R24 R25 R26 R29
1 560 508 507 448 562 440 337 482 356 560 517 508 379 237 315 488
2 117 135 143 170 119 205 239 111 200 129 128 139 212 172 204 165
3  65  80  94 107  62  83 144 103 143  60  85  91 120 210 169  86
4  18  32  18  31  18  26  36  57  45  13  27  23  40 104  67  22
5   6  11   4  10   5  12  10  13  22   4   9   5  15  43  11   5
```

[24] PROMIS Cooperative Group. Unpublished Manual for the Patient-Reported Outcomes Measurement Information System (PROMIS) Version 1.1. October, 2008: http://www.nihpromis.org

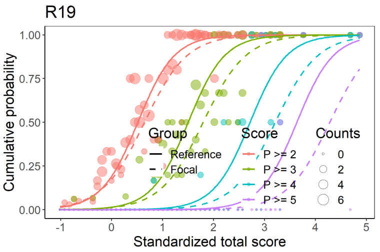
```r
# cumulative logit
(fit4 <- ddfORD(DataOrd, group,
                focal.name = 1,
         model = "cumulative"))
```

```r
# cumulative logit
(fit4 <- ddfORD(DataOrd, group,
                focal.name = 1,
         model = "cumulative"))
# R19: I found it hard to focus on anything
#      other than my anxiety
```

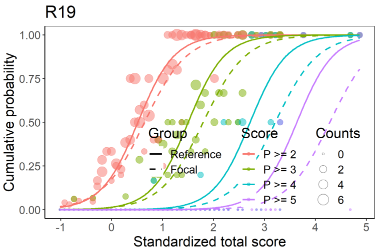
```
Detection of both types of Differential Distractor
Functioning for ordinal data using cumulative logit
regression model

Likelihood-ratio Chi-square statistics

Item purification was not applied
No p-value adjustment for multiple comparisons

    Chisq-value P-value
R3   0.1029     0.9499
R6   8.9062     0.0116 *
R8   1.6033     0.4486
R9   2.8795     0.2370
R10  3.6480     0.1614
R11  3.3894     0.1837
R12  2.5989     0.2727
R13  0.7204     0.6975
R18  1.9843     0.3708
R19  6.7181     0.0348 *
R20 15.6995     0.0004 ***
R21  4.0303     0.1333
R24  2.4008     0.3011
R25  1.2703     0.5299
R26  0.1898     0.9094
R29  0.7360     0.6921

Sign. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Items detected as DDF items:
 R6
 R19
 R20
```

```r
# cumulative logit
(fit4 <- ddfORD(DataOrd, group,
                focal.name = 1,
         model = "cumulative"))
# R19: I found it hard to focus on anything
#      other than my anxiety
# plotting cumulative probs
plot(fit4, item = 10,
     plot.type = "cumulative")
```

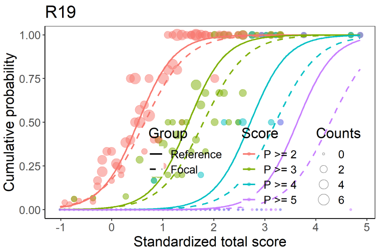
```r
# cumulative logit
(fit4 <- ddfORD(DataOrd, group,
                focal.name = 1,
           model = "cumulative"))
# R19: I found it hard to focus on anything
#      other than my anxiety

# plotting cumulative probs
plot(fit4, item = 10,
     plot.type = "cumulative")
```

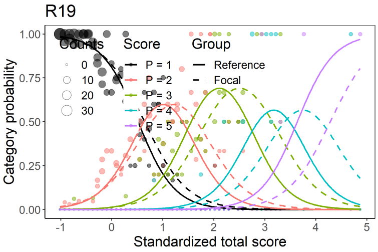


```r
# cumulative logit
(fit4 <- ddfORD(DataOrd, group,
                focal.name = 1,
       model = "cumulative"))

# R19: I found it hard to focus on anything
#      other than my anxiety

# plotting cumulative probs
plot(fit4, item = 10,
     plot.type = "cumulative")

# plotting category probs
plot(fit4, item = 10,
     plot.type = "category")
```



```
# cumulative logit
(fit4 <- ddfORD(DataOrd, group,
                focal.name = 1,
         model = "cumulative"))

# R19: I found it hard to focus on anything
#      other than my anxiety

# plotting cumulative probs
plot(fit4, item = 10,
     plot.type = "cumulative")

# plotting category probs
plot(fit4, item = 10,
     plot.type = "category")
```





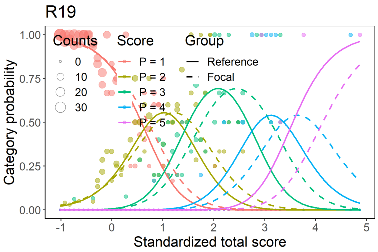
```r
# cumulative logit
(fit4 <- ddfORD(DataOrd, group,
                focal.name = 1,
         model = "cumulative"))
# R19: I found it hard to focus on anything
#      other than my anxiety
# plotting cumulative probs
plot(fit4, item = 10,
     plot.type = "cumulative")
# plotting category probs
plot(fit4, item = 10,
     plot.type = "category")


# adjacent category
(fit5 <- ddfORD(DataOrd, group,
                focal.name = 1,
         model = "adjacent"))
```

```r
# cumulative logit
(fit4 <- ddfORD(DataOrd, group,
                focal.name = 1,
         model = "cumulative"))
# R19: I found it hard to focus on anything
#      other than my anxiety
# plotting cumulative probs
plot(fit4, item = 10,
     plot.type = "cumulative")
# plotting category probs
plot(fit4, item = 10,
     plot.type = "category")


# adjacent category
(fit5 <- ddfORD(DataOrd, group,
                focal.name = 1,
         model = "adjacent"))
```

```
Detection of both types of Differential Distractor
Functioning for ordinal data using adjacent category
logit regression model

Likelihood-ratio Chi-square statistics

Item purification was not applied
No p-value adjustment for multiple comparisons

    Chisq-value P-value
R3   0.2987      0.8613
R6   5.9257      0.0517 .
R8   1.4320      0.4887
R9   1.6799      0.4317
R10  3.2452      0.1974
R11  4.4222      0.1096
R12  2.5353      0.2815
R13  0.6878      0.7090
R18  0.9893      0.6098
R19  6.3403      0.0420 *
R20 16.5813      0.0003 ***
R21  2.0704      0.3552
R24  2.2645      0.3223
R25  1.3606      0.5065
R26  0.2213      0.8953
R29  0.8419      0.6564

Sign. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Items detected as DDF items:
 R19
 R20
```

```r
# cumulative logit
(fit4 <- ddfORD(DataOrd, group,
                focal.name = 1,
        model = "cumulative"))
# R19: I found it hard to focus on anything
#      other than my anxiety
# plotting cumulative probs
plot(fit4, item = 10,
     plot.type = "cumulative")
# plotting category probs
plot(fit4, item = 10,
     plot.type = "category")


# adjacent category
(fit5 <- ddfORD(DataOrd, group,
                focal.name = 1,
        model = "adjacent"))
# plotting category probs
plot(fit5, item = 10)
```

```r
# cumulative logit
(fit4 <- ddfORD(DataOrd, group,
                focal.name = 1,
         model = "cumulative"))
# R19: I found it hard to focus on anything
#      other than my anxiety

# plotting cumulative probs
plot(fit4, item = 10,
     plot.type = "cumulative")

# plotting category probs
plot(fit4, item = 10,
     plot.type = "category")


# adjacent category
(fit5 <- ddfORD(DataOrd, group,
                focal.name = 1,
         model = "adjacent"))

# plotting category probs
plot(fit5, item = 10)
```

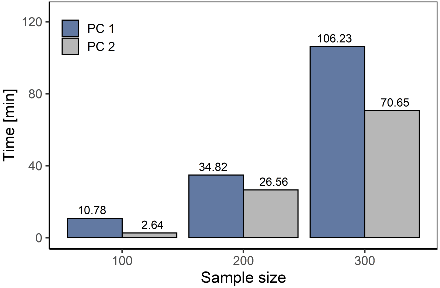


Implementation of nonparametric method

- Work in progress
- Standard R kernel estimating functions do not return kernel values
- Computationally complex
- Implementation into C++

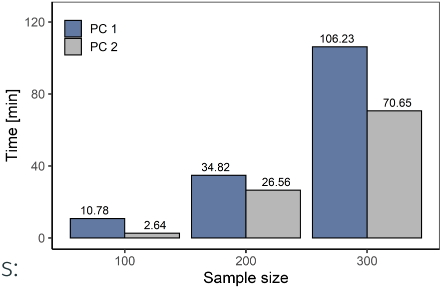
Implementation of nonparametric method

- Work in progress
- Standard R kernel estimating functions do not return kernel values
- Computationally complex
- Implementation into C++



Implementation of nonparametric method

- Work in progress
- Standard R kernel estimating functions do not return kernel values
- Computationally complex
- Implementation into C++



- Possible reasons:
  - Bootstrapping
  - Length of $\left(\frac{X_{Ri}+X_{Fj}}{2}\right)_{i=1,j=1}^{nR,nF}$ vector

Conclusion and future work

Conclusion and future work

Summary

- DIF and DDF phenomena
- Mostly used methods for their detection

Conclusion and future work

Summary

- DIF and DDF phenomena
- Mostly used methods for their detection
- New methods including
  - Nonlineaer regression (3-4PL non-IRT models)
  - Cumulative logit and adjacent category logit models
  - Multinomial model
  - Nonparametric comparison of regression curves

Conclusion and future work

Summary

- DIF and DDF phenomena
- Mostly used methods for their detection
- New methods including
  - Nonlineaer regression (3-4PL non-IRT models)
  - Cumulative logit and adjacent category logit models
  - Multinomial model
  - Nonparametric comparison of regression curves
- Simulation studies
  - Nonlineaer regression (3-4PL non-IRT models)
  - Nonparametric comparison of regression curves
  - Item purification and multiple comparison corrections

Conclusion and future work

Summary

- DIF and DDF phenomena
- Mostly used methods for their detection
- New methods including
  - Nonlineaer regression (3-4PL non-IRT models)
  - Cumulative logit and adjacent category logit models
  - Multinomial model
  - Nonparametric comparison of regression curves
- Simulation studies
  - Nonlineaer regression (3-4PL non-IRT models)
  - Nonparametric comparison of regression curves
  - Item purification and multiple comparison corrections
- Implementation of methods

Conclusion and future work

Future work

- Nonparametric comparison of regression curves
  - Complex simulation study
  - Show possible superiority when true model is not 4PL IRT
  - Implementation to C++ and R user-friendly functions

Conclusion and future work

Future work

- Nonparametric comparison of regression curves
    - Complex simulation study
    - Show possible superiority when true model is not 4PL IRT
    - Implementation to C++ and R user-friendly functions
- Dissertation

Questions and ideas are welcomed!

hladka@cs.cas.cz
www.cs.cas.cz/hladka/