

# Computational aspects of psychometrics taught with R and Shiny

---

Patrícia Martinková<sup>1,2</sup>

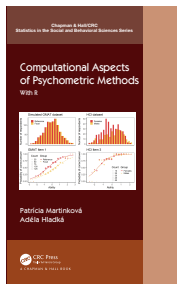
useR!2021, July 5-9

<sup>1</sup> Institute of Computer Science, Czech Academy of Sciences

<sup>2</sup> Faculty of Education, Charles University

# Outline

1. Introduction: Teaching psychometrics
2. ShinyItemAnalysis
3. Real and simulated datasets
  - Reliability and measurement error
  - Differential item functioning
4. Book in preparation
5. Discussion and conclusion



# Psychometrics

- Psychometrics deals with the advancement of quantitative measurement practices in psychology, education, health, and many other fields
- Psychometric Society <https://www.psychometricsociety.org/>
- Covers a number of statistical methods that are useful for the behavioral and social sciences, such as:
  - estimation of reliability to deal with the omnipresence of measurement error
  - detailed description of item functioning encompassed in item response theory (IRT) models
- Number of existing R packages, see CRAN task View <https://CRAN.R-project.org/view=Psychometrics>

# Teaching psychometrics

- Graduate course at University of Washington (2015)
- Graduate courses at Charles University, Prague
  - NMST570 Selected topics in psychometrics
  - NMST571 Seminar in psychometrics
- Pre-conference workshops, seminars
- Heterogeneous groups of students/participants
  - Students of psychology, education, ... and statistics
  - Researchers, practitioners from test companies
- Participants of various levels of R proficiency
- Participants of various levels of statistical focus and proficiency

# Teaching psychometrics with R and ShinyItemAnalysis

## Goals:

- Explain psychometric models and methods
  - in context of statistics and data science
- Illustrate important computational aspects
  - Real and simulated data from various fields
- Provide toolbox of R functions and packages
  - Similarities/differences across different packages
- Make procedures and concepts better available
  - Interactive application of the ShinyItemAnalysis package

# ShinyItemAnalysis

Software for psychometric analysis of educational tests, psychological assessments, health-related and other types of multi-item measurements

- R package
  - Version 1.3.7 on [▶ CRAN](#), newest version on [▶ GitHub](#)
- Interactive shiny application
  - Accessible locally from R with `startShinyItemAnalysis()`
  - Online at ICS server and shinyapps.io

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

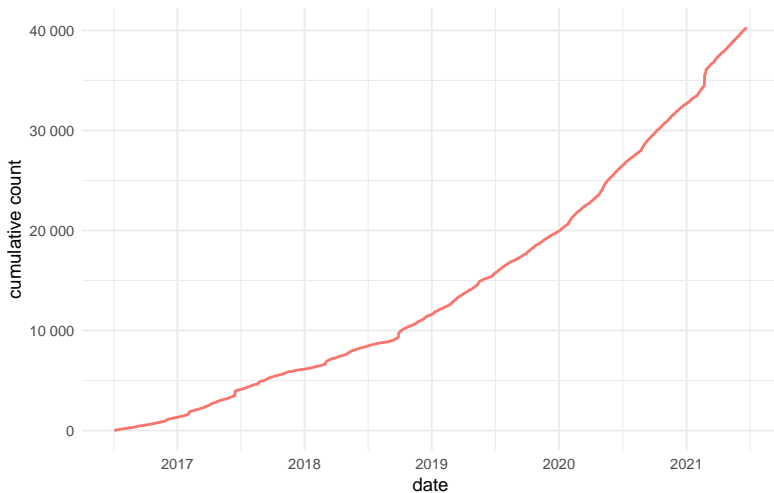
<https://cemp.shinyapps.io/ShinyItemAnalysis/>

---

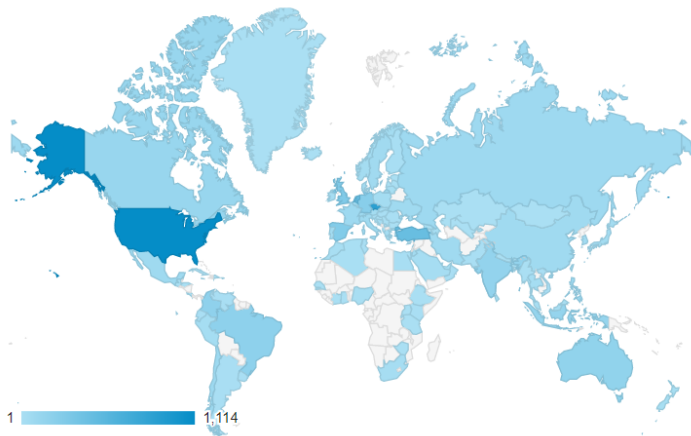
Martinková and Drabinová (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*. 10(2), 503–515.

<https://doi.org/10.32614/RJ-2018-074>

# R package ShinyItemAnalysis downloads from CRAN



# ShinyItemAnalysis online app is used worldwide!





# Interactive application

Home Data Scores ▾ Validity ▾ Reliability ▾ Item analysis ▾ Regression ▾ IRT models ▾ DIF/Fairness ▾ Reports  

## Welcome

Welcome to ShinyItemAnalysis!

ShinyItemAnalysis is an interactive online application for the psychometric analysis of educational tests, psychological assessments, health-related and other types of multi-item measurements, or ratings from multiple raters, built on [R](#) and [shiny](#). You can easily start using the application with the default toy dataset. You may also select from a number of other toy datasets or upload your own in the **Data** section. Offered methods include:

- Exploration of total and standard scores in the **Summary** section
- Analysis of measurement error in the **Reliability** section
- Correlation structure and criterion validity analysis in the **Validity** section
- Item and distractor analysis in the **Item analysis** section
- Item analysis with regression models in the **Regression** section
- Item analysis by item response theory models in the **IRT models** section
- Detection of differential item functioning in the **DIF/Fairness** section

All graphical outputs and selected tables can be downloaded via the download button. Moreover, you can automatically generate a HTML or PDF report in the **Reports** section. All offered analyses are complemented by selected R codes which are ready to be copied and pasted into your R console, therefore a similar analysis can be run and modified in R.

Visit the [www.ShinyItemAnalysis.org](http://www.ShinyItemAnalysis.org) webpage to learn more about ShinyItemAnalysis!

# ShinyItemAnalysis: Newest developments

- New features of the interactive application
  - New toy data, new data types allowed for one's own upload
  - Validity: New corrplot, dendrograms, factor analysis
  - Reliability: Inter-rater reliability in restricted samples
  - Traditional item analysis: Item criterion validity
  - Regression models: Models for polytomous data
  - IRT models: reorganized
  - DIF: polytomous data, uploaded matching criterion
  
- Interactive training sections with exercises
- All plots interactive, created with `plotly`
- Downloadable plots, tables and reports
- Sample R code

# Toy datasets

- Number of toy datasets, upload of one's own data is possible

## Upload your own datasets

Here you can upload your own dataset. Select all necessary files and use the **Upload data** button on bottom of this page.

**Choose data (CSV file)**

Browse... HCI\_ABCD.csv

Upload complete

The main **data** file should contain the responses of individual respondents (rows) to given items (columns). Data need to be either binary, nominal (e.g. in ABCD format), or ordinal (e.g. in Likert scale). The header may contain item names, however, no row names should be included. In all data sets, the **header** should be either included or excluded. Columns of dataset are by default renamed to the item and number of a particular column. If you want to keep your own names, check the box **Keep item names** below. Missing values in scored dataset are by default evaluated as 0. If you want to keep them as missing, check the box **Keep missing values** below.

<p><b>Type of data</b> <b>i</b></p> <p><input type="radio"/> Binary</p> <p><input checked="" type="radio"/> Nominal</p> <p><input type="radio"/> Ordinal</p>	<p><b>Separator</b></p> <p><input type="radio"/> Comma</p> <p><input checked="" type="radio"/> Semicolon</p> <p><input type="radio"/> Tab</p>	<p><b>Quote</b></p> <p><input type="radio"/> None</p> <p><input checked="" type="radio"/> Double Quote</p> <p><input type="radio"/> Single Quote</p>	<p><b>Data specification</b></p> <p><input checked="" type="checkbox"/> Header <b>i</b></p> <p><input checked="" type="checkbox"/> Keep item names <b>i</b></p> <p><b>Missing values</b></p> <p><input type="checkbox"/> Keep missing values <b>i</b></p>
--	---	--	---

**Choose key (CSV file)**

Browse... HCI\_key.csv

Upload complete

For nominal data, it is necessary to upload **key** of correct answers.

**Choose group (optional)**

Browse... HCI\_group.csv

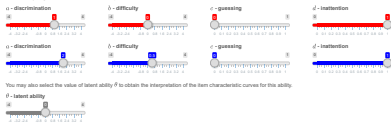
**Group** is a variable for DIF and DDF analyses. It should be a binary vector, where 0 represents the reference group and 1 represents the focal group. Its length needs to be the same as the number of individual respondents in the main dataset. Missing values are not supported for the group variable and such cases/rows of the data should be removed.

# Interactive training sections

## • Interactive training sections for IRT models and DIF

### Parameters

Select parameters  $\alpha$  (discrimination),  $\beta$  (difficulty),  $c$  (guessing), and  $d$  (leakage). By constraining  $\alpha = 1$ ,  $c = 0$ ,  $d = 1$  you get the Rasch model. With option  $c = 0$  and  $d = 1$  you get the 2PL model, and with option  $d = 1$  the 3PL model.



You may also select the value of latent ability  $\theta$  to obtain the interpretation of the item characteristic curves for this ability.



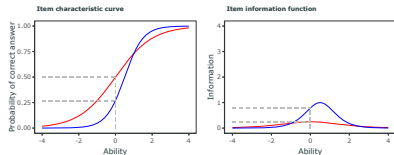
### Equations

$$P(Y = 1|X) = c + (d - c) \frac{e^{a(\theta - \beta)}}{1 + e^{a(\theta - \beta)}}$$

$$I(\theta) = \frac{e^{a(\theta - \beta)}}{\pi(\theta)(1 - \pi(\theta))} = \frac{e^a \cdot [a(\theta - \beta)]^2 \cdot (d - c)(\theta - \beta)^2}{\pi(\theta) \cdot (1 - \pi(\theta)) \cdot (d - c)^2}$$

**Interpretation:** The probability of the correct answer with the latent ability  $\theta = 0$  in the red item with parameters  $\alpha = 1$ ,  $\beta = 0$ ,  $c = 0$ , and  $d = 1$  is equal to 0.50. The information for the latent ability  $\theta = 0$  in the red item is equal to 0.25. The probability of the correct answer with the latent ability  $\theta = 0$  in the blue item with parameters  $\alpha = 2$ ,  $\beta = 0.5$ ,  $c = 0$ , and  $d = 1$  is equal to 0.27. The information for the latent ability  $\theta = 0$  in the blue item is equal to 0.79.

Note that for 1PL and 2PL models, the item information is the highest at  $\theta = \hat{\beta}$ . This is not necessarily the case for 3PL and 4PL models.



### DIF training

In this section, you can explore the group-specific model for testing differential item functioning among two groups - reference and focal.

#### Parameters

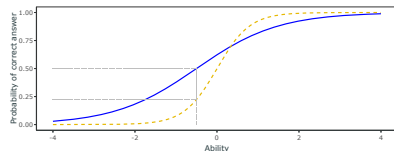
Select parameters  $\alpha$  (discrimination) and  $\beta$  (difficulty) for an item given by 2PL IRT model for **reference** and **focal** group. When the item parameters for the reference and the focal group differ, this phenomenon is termed differential item functioning.



You may also select the value of latent ability  $\theta$  to obtain the interpretation of the item characteristic curves for this ability.



**Interpretation:** In the **reference** group, a respondent with the ability  $\theta = -0.5$  has the probability of the correct answer to an item with parameters  $\alpha = 1$ ,  $\beta = -0.5$ ,  $c = 0$ , and  $d = 1$  equal to 0.50. In the **focal** group, a respondent with the ability  $\theta = -0.5$  has the probability of the correct answer to an item with parameters  $\alpha = 2.5$ ,  $\beta = 0$ ,  $c = 0$ , and  $d = 1$  equal to 0.22.



# Interactive training sections – check your understanding

## • Interactive quizzes for IRT models and DIF

### Exercise 1

Consider the following 2PL items with parameters

Item 1:  $a = 2.5, b = -0.5$

Item 2:  $a = 1.5, b = 0$

For these items fill in the following exercises with an accuracy of up to 0.05, then click on the **Submit answers** button. If you need a hint, click on the blue button with a question mark.

- Sketch the item characteristic and information curves. ? ×
- Calculate the probability of a correct answer for latent abilities  $\theta = -2, -1, 0, 1, 2$ . ?

Item 1:  $\theta = -2$  ✓  $\theta = -1$  ×  $\theta = 0$  ×  $\theta = 1$  ×  $\theta = 2$  ×

Item 2:  $\theta = -2$  ✓  $\theta = -1$  ×  $\theta = 0$  ×  $\theta = 1$  ×  $\theta = 2$  ×

- For what level of ability  $\theta$  are the probabilities equal? ?

$\theta = ?$  ×

- Which item provides more information for weak ( $\theta = -2$ ), average ( $\theta = 0$ ) and strong ( $\theta = 2$ ) students? ?

$\theta = -2$   Item 1  Item 2 ✓

$\theta = 0$   Item 1  Item 2 ✓

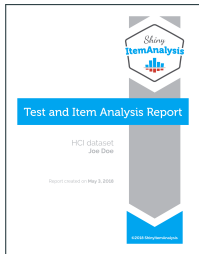
$\theta = 2$   Item 1  Item 2 ×

27% correct. Try again.

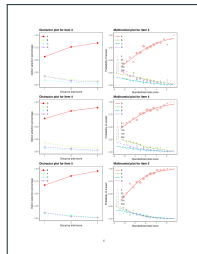
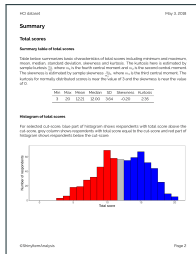
Submit answers

# Automatic report generation

- Generating PDF/HTML reports for uploaded data
- using R Markdown



<b>Contents</b>	
<b>Introduction</b>	1
<b>Summary</b>	2
Item scores	2
Histogram of total scores	2
<b>Scoring</b>	3
Summary table of standard scores	3
<b>Validity</b>	4
Cronbach's Alpha	4
Reliability: condition heat map	4
Item-to-total	4
Classification	4
<b>Tau-bounded item analysis</b>	6
Item-to-total	6
Diff. Chi-Square test	6
Tau-bounded item analysis table	6
Classification analysis	6
<b>IRT models</b>	16
Single item parameter mapping 2PL, 3PL model	16
Grading	16
Ability estimation	16
Item characteristic and information curves	16
Item characteristic and information curves	16
Item characteristic and information curves	16
<b>IRT fit indices</b>	36
Total scores by group	36
Summary table of total scores in reference and focal groups	36
Reliability of total scores by group	36
Data just method	36
Summary table	36
Delta chi	36
OIF	36
OIF: comparison using logistic regression	36
Substantive	36
OIF: decision using multinomial regression	36
Summary table	36
<b>Decision table</b>	24



# ShinyItemAnalysis: Newest developments

- New ShinyItemAnalysis package functions and functionalities
  - `startShinyItemAnalysis()` now using `rstudioapi`, runs as "Local job" in Jobs RStudio IDE pane, keeping the console available for trying sample R code
  - Testing of the online app on collection of datasets, unit tests using `testthat`
  - Refactoring the code using shiny modules, following the best practices with `golem`
  - Dealing with high number of dependencies

## Datasets demonstrating computational aspects: IRR

- Why zero inter-rater reliability estimates are plausible under restricted range
- Statistical explanation: When proposal range is restricted by perceived quality, the between-proposal variance of peer review scores  $\tau^2$  decreases.
- Interactive illustration offered in `ShinyItemAnalysis` with the AIBS dataset.
- Animation created with the `gganimate` package.

---

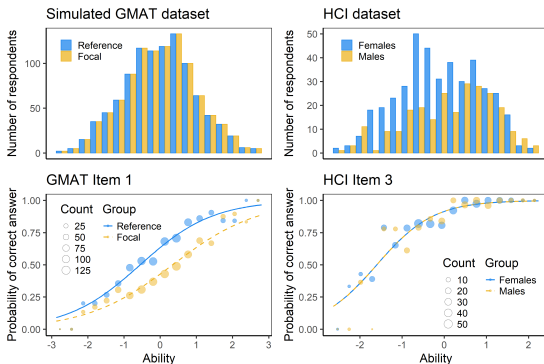
Eroshova, Martinkova, and Lee (2021). When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review. *JRSS – A*.  
doi [10.1111/rssa.12681](https://doi.org/10.1111/rssa.12681)



## IRR in restricted range: Animation

# Datasets demonstrating computational aspects: DIF

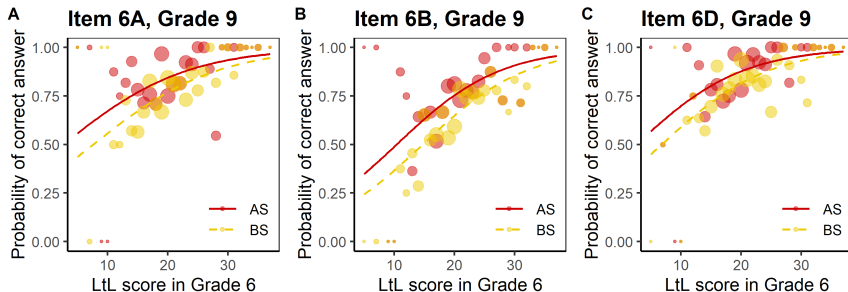
- Differential item functioning (DIF) analysis may provide deeper understanding to test functioning among groups.



Martinková et al. (2017). Checking Equity: Why DIF Analysis should be a Routine Part of Developing Conceptual Assessments. CBE-LSE, 16(2), rm2. doi 10.1187/cbe.16-10-0307

# DIF in longitudinal designs

DIF-C can provide proof of instructional sensitivity, even when differences in change are not visible in total scores.



Martinková, Hladká, and Potužníková (2020). Is academic tracking related to gains in learning competence? Using propensity score matching and differential item change functioning analysis for better understanding of tracking implications. *Learning and Instruction*, 66, 101286. doi: [10.1016/j.learninstruc.2019.101286](https://doi.org/10.1016/j.learninstruc.2019.101286)

# DIF and DIF-C analysis available in ShinyItemAnalysis

## • DIF and DIF-C analysis with difNLR package



### Observed scores

DIF analysis may come to a different conclusion than a test of group differences in total scores. Two groups may have the same distribution of total scores, yet, some items may function differently for the two groups. Also, one of the groups may have a significantly lower total score, yet, it may happen that there is no DIF item (Martinková et al., 2017). This section examines the differences in observed scores only. Explore further DIF sections to analyze differential item functioning.

In DIF analysis, the groups are compared in functioning of items with respect to respondent ability. In many methods, observed ability such as the standardized total score is used as the matching criterion. DIF can also be explored with respect to other observed scores or criteria. For example, to analyze instructional sensitivity, Martinková et al. (2020) analyzed differential item functioning in change (DIF-C) by analyzing DIF on Grade 9 item answers while matching on Grade 6 total scores of the same respondents in a longitudinal setting (see [by data learning to Learn 9](#) in the Data section).

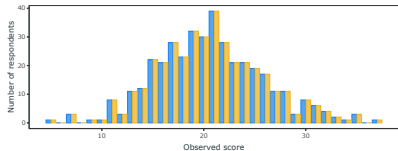
#### Observed score

Uploaded

#### Summary of uploaded variable for groups

	n	Min	Max	Mean	Median	SD	Skewness	Kurtosis
Reference group (0)	391	5.00	37.00	20.64	20.00	5.24	0.23	3.21
Focal group (1)	391	5.00	37.00	20.64	20.00	5.24	0.23	3.21

#### Histograms of uploaded variable for groups



### Summary Items

#### Generalized logistic regression

Generalized logistic regression models are extensions of a logistic regression method which account for the possibility of guessing by allowing for nonzero lower asymptote - pseudo-guessing  $\alpha_1$  (Dimitova & Martinková, 2017) or an upper asymptote lower than one - maldetection  $\alpha_2$ . Similarly to logistic regression, its extensions also provide detection of uniform and non-uniform DIF by taking the difficulty parameter  $\beta_1$  (uniform) and the discrimination parameter  $\alpha_1$  (non-uniform) differ for groups and by testing for the difference in their values. Moreover, these extensions allow for testing differences in pseudo-guessing and maldetection parameters and they can be seen as process of 3PL and 4PL RT models for DIF detection.

#### Method specification

Here you can specify the assumed model. In 3PL and 4PL models, the abbreviations  $\alpha_1$  or  $\alpha_2$  mean that parameters  $\alpha_1$  or  $\alpha_2$  are assumed to be the same for both groups, otherwise they are allowed to differ. With type you can specify the type of DIF to be tested by choosing the parameters in which a difference between groups should be tested. You can also select correction method for multiple comparison or item purification.

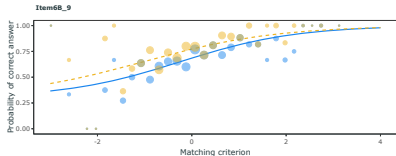
Finally, you may change the Observed score. While matching on the standardized total score is typical, the upload of other observed scores is possible in the Data section. Using a pre-test (standardized) total score allows for testing differential item functioning in change (DIF-C) to provide proofs of instructional sensitivity (Martinková et al., 2020); also see [learning to Learn 9](#) by test. For selected item you can display plot of its characteristic curves and table of its estimated parameters with standard errors.

Model: 3PLog Type:   $\alpha_1$    $\alpha_2$    $\beta_1$    $\beta_2$  Correction method: None Observed scores: Standardized uploaded Item: Item68\_9

Item purification

#### Plot with estimated DIF generalized logistic curve

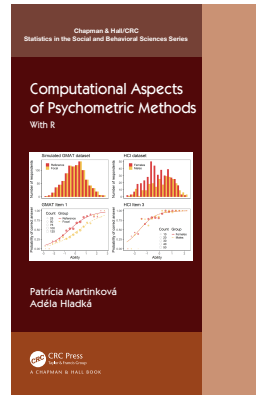
Points represent a given level of the correct answer (empirical probabilities) with respect to the observed score. Their size is determined by the count of respondents who achieved a given level of observed score with respect to the group membership.



Hladká and Martinková (2020). difNLR: Generalized logistic regression models for DIF and DDFdetection. *The R Journal*, 12(1), 300–323. doi: [10.32614/RJ-2020-014](https://doi.org/10.32614/RJ-2020-014)

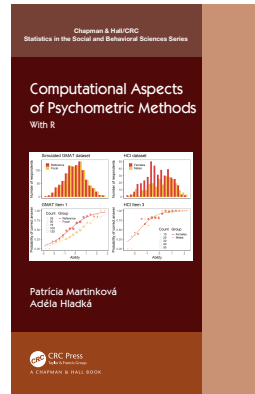
# Book planned for publication in 2022

- Deeper understanding to psychometric models and methods
- For a wide audience
- Accompanied by sample R code, practical examples and datasets
- Each chapter includes a section presenting the analysis with respective tab of the `ShinyItemAnalysis` interactive application



# Discussion

- Teaching computational aspects of psychometrics with R and ShinyItemAnalysis
- Demonstrating the power of R
- Importance of sample R code within the Shiny app
- Importance of relevant simulated and real data examples
  
- Stay tuned for the new book!





# Thank you for your attention!

`www.cs.cas.cz/martinkova`



## Acknowledgements:

- Czech Science Foundation grant 21-03658S
- Technology Agency of the Czech Republic grant TL05000008
- Computational Psychometrics Group: <https://www.cs.cas.cz/comps/>

# References

- Erosheva, E. A., Martinkova, P., & Lee, C. J. (2021). When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review. *Journal of the Royal Statistical Society – Series A*. doi: 10.1111/rssa.12681
- Hladká, A., & Martinková, P. (2020). difNLR: Generalized logistic regression models for DIF and DDF detection. *The R Journal*, 12(1), 300–323. doi: 10.32614/RJ-2020-014
- Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, 10(2). doi: 10.32614/RJ-2018-074
- 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. doi: 10.1187/cbe.16-10-0307
- Martinková, P., Hladká, A., & Potužníková, E. (2020). Is academic tracking related to gains in learning competence? Using propensity score matching and differential item change functioning analysis for better understanding of tracking implications. *Learning and Instruction*, 66, 101286. doi: j.learninstruc.2019.101286