ShinyItemAnalysis for Teaching Psychometrics and to Enforce Routine Analysis of Educational Tests

by Patrícia Martinková, Adéla Drabinová

Abstract This work introduces **ShinyItemAnalysis**, an R package and an online shiny application for psychometric analysis of educational tests and items. **ShinyItemAnalysis** covers a broad range of psychometric methods and offers data examples, model equations, parameter estimates, interpretation of results, together with a selected R code, and is therefore suitable for teaching psychometric concepts with R. Furthermore, the application aspires to be an easy-to-use tool for analysis of educational tests by allowing the users to upload and analyze their own data and to automatically generate analysis reports in PDF or HTML. We argue that psychometric analysis should be a routine part of test development in order to gather proofs of reliability and validity of the measurement, and we demonstrate how **ShinyItemAnalysis** may help enforce this goal.

Introduction

Assessments that are used to measure students' ability or knowledge need to produce valid, reliable and fair scores (Brennan 2006; Downing and Haladyna 2011; AERA, APA and NCME 2014). While many R packages have been developed to cover general psychometric concepts (e. g. **psych** (Revelle, 2018), **ltm** (Rizopoulos, 2006)) or specific psychometric topics (e. g. **difR** (Magis et al., 2010), **lavaan** (Rosseel, 2012), see also *Psychometrics*), stakeholders in this area are often non-programmers and thus, may find it difficult to overcome the initial burden of an R-based environment. Commercially available software provides an alternative but high prices and limited methodology may be an issue. Nevertheless, it is of high importance to enforce routine psychometric analysis in development and validation of educational tests of various types worldwide. Freely available software with user-friendly interface and interactive features may help this enforcement even in regions or scientific areas where understanding and usage of psychometric concepts is underdeveloped.

In this work we introduce **ShinyItemAnalysis** (Martinková et al., 2018a) - an R package and an online application based on **shiny** (Chang et al., 2018) which was initially created to support teaching of psychometric concepts and test development, and subsequently used to enforce routine validation of admission tests to Czech universities (Martinková et al., 2017). Its current mission is to support routine validation of educational and psychological measurements worldwide.

We first briefly explain the methodology and concepts in a step-by-step way, from the classical test theory (CTT) to the item response theory (IRT) models, including methods to detect differential item functioning (DIF). We then describe the implementation of **ShinyItemAnalysis** with practical examples coming from development and validation of the Homeostasis Concept Inventory (HCI, McFarland et al., 2017). We conclude with discussion of features helpful for teaching psychometrics, as well as features important for generation of PDF and HTML reports, enforcing routine analysis of admission and other educational tests.

Psychometric models for analysis of items and tests

Classical test theory

Traditional item analysis uses counts, proportions and correlations to evaluate properties of test items. Difficulty of items is estimated as the percentage of students who answered correctly to that item. Discrimination is usually described as the difference between the percent correct in the upper and lower third of the students tested (upper-lower index, ULI). As a rule of thumb, ULI should not be lower than 0.2, except for very easy or very difficult items (Ebel, 1954). ULI can be customized by determining the number of groups and by changing which groups should be compared: this is especially helpful for admission tests where a certain proportion of students (e. g. one fifth) are usually admitted (Martinková et al., 2017).

Other traditional statistics for a description of item discrimination include the point-biserial correlation, which is the Pearson correlation between responses to a particular item and scores on the total test. This correlation (R) is denoted RIT index if an item score (I) is correlated with the total score (T) of the whole test, or RIR if an item score (I) is correlated with the sum of the rest of the items (R).

In addition, difficulty and discrimination may be calculated for each response of the multiple choice item to evaluate distractors and diagnose possible issues, such as confusing wording. Respondents are divided into N groups (usually 3, 4 or 5) by their total score. Subsequently, the percentage of students in each group which selected a given answer (correct answer or distractor) is calculated and may be displayed in a distractor plot.

To gain empirical proofs of the construct validity of the whole instrument, correlation structure needs to be examined. Empirical proofs of validity may be provided by correlation with a criterion variable. For example, a correlation with subsequent university success or university GPA may be used to demonstrate predictive validity of admission scores.

To gain proofs of test reliability, internal consistency of items can be examined using Cronbach's alpha (Cronbach 1951, see also Zinbarg et al. 2005 for more discussion). Cronbach's alpha of test without a given item can be used to determine items not internally consistent with the rest of the test.

Regression models for description of item properties

Various regression models may be fitted to describe item properties in more detail. With binary data, logistic regression may be used to model the probability of a correct answer as a function of the total score by an S-shaped logistic curve. Parameter $b_0 \in R$ describes horizontal location of the fitted curve, parameter $b_1 \in R$ describes its slope.

$$\mathrm{P}\left(Y = 1|X, b_0, b_1\right) = \frac{\exp\left(b_0 + b_1 X\right)}{1 + \exp\left(b_0 + b_1 X\right)}. \tag{1}$$

A standardized total score may be used instead of a total score as an independent variable to model the properties of a given item. In such a case, the estimated curve remains the same, only the interpretation of item properties now focuses on improvement of 1 standard deviation rather than 1 point improvement. It is also helpful to re-parametrize the model using new parameters $a \in R$ and $b \in R$. Item difficulty parameter b is given by location of the inflection point, and item discrimination parameter a is given by the slope at the inflection point:

$$\mathrm{P}(Y = 1|Z, a, b) = \frac{\exp(a(Z - b))}{1 + \exp(a(Z - b))}. \tag{2}$$

Further, non-linear regression models allow us to account for guessing by providing non-zero left asymptote $c \in [0, 1]$ and inattention by providing right asymptote $d \in [0, 1]$:

$$\mathrm{P}\left(Y = 1|Z, a, b, c, d\right) = c + (d - c)\,\frac{\exp\left(a\left(Z - b\right)\right)}{1 + \exp\left(a\left(Z - b\right)\right)}. \tag{3}$$

Other regression models allow for further extensions or different data types: Ordinal regression allows for modelling Likert-scale and partial-credit items. To model responses to all given options (correct response and all distractors) in multiple-choice questions, a multinomial regression model can be used. Further complexities of the measurement data and item functioning may be accounted for by incorporating student characteristics with multiple regression models or clustering with hierarchical models.

A logistic regression model (1) for item description, and its re-parametrizations and extensions as illustrated in equations (2) and (3) may serve as a helpful introductory step for explaining and building IRT models which can be conceptualized as (logistic/non-linear/ordinal/multinomial) mixed effect regression models (Rijmen et al., 2003).

Item response theory models

IRT models assume respondent abilities being unobserved/latent scores θ which need to be estimated together with item parameters. 4-parameter logistic (4PL) IRT model corresponds to regression model (3) above

$$\mathrm{P}\left(Y = 1|\theta, a, b, c, d\right) = c + (d - c)\,\frac{\exp\left(a\left(\theta - b\right)\right)}{1 + \exp\left(a\left(\theta - b\right)\right)}. \tag{4}$$

Similarly, other submodels of 4PL model (4) may be obtained by fixing appropriate parameters. E. g. the 3PL IRT model is obtained by fixing inattention to $d = 1$. The 2PL IRT model is obtained by further fixing pseudo-guessing to $c = 0$, and the Rasch model by fixing discrimination $a = 1$ in addition.

Other IRT models allow for further extensions or different data types: Modelling Likert-scale and

partial-credit items can be done by modelling cumulative responses in a graded response model (GRM, Samejima, 1969). Alternatively, ordinal items may be analyzed by modelling adjacent categories logits using a generalized partial credit model (GPCM, Muraki, 1992), or its restricted version - partial credit model (PCM, Masters, 1982), and rating scale model (RSM, Andrich, 1978). To model distractor properties in multiple-choice items, Bock's nominal response model (NRM, Bock, 1972) is an IRT analogy of a multinomial regression model. This model is also a generalization of GPCM/PCM/RSM ordinal models. Many other IRT models have been used in the past to model item properties, including models accounting for multidimensional latent traits, hierarchical structure or response times (van der Linden, 2017).

A wide variety of estimation procedures has been proposed in last decades. Joint maximum likelihood estimation treats both ability and item parameters as unknown but fixed. Conditional maximum likelihood estimation takes an advantage of the fact that in exponential family models (such as in the Rasch model), total score is a sufficient statistics for an ability estimate and the ratio of correct answers is a sufficient statistics for a difficulty parameter. Finally, in a marginal maximum likelihood estimation used by the **mirt** (Chalmers, 2018) and **ltm** (Rizopoulos, 2006) package as well as in **ShinyItemAnalysis**, parameter θ is assumed to be a random variable following certain distribution (often standard normal) and is integrated out (see for example, Johnson, 2007). An EM algorithm with a fixed quadrature is used in latent scores and item parameters estimation. Besides MLE approaches, Bayesian methods with the Markov chain Monte Carlo are a good alternative, especially for multidimensional IRT models.

Differential item functioning

DIF occurs when respondents from different groups (e. g. such as defined by gender or ethnicity) with the same underlying true ability have a different probability of answering the item correctly. Differential distractor functioning (DDF) is a phenomenon when different distractors, or incorrect option choices, attract various groups with the same ability differentially. If an item functions differently for two groups, it is potentially unfair, thus detection of DIF and DDF should be routine analysis when developing and validating educational and psychological tests (Martinková et al., 2017).

Presence of DIF can be tested by many methods including Delta plot (Angoff and Ford, 1973), Mantel-Haenszel test based on contingency tables that are calculated for each level of a total score (Mantel and Haenszel, 1959), logistic regression models accounting for group membership (Swaminathan and Rogers, 1990), nonlinear regression (Drabinová and Martinková, 2017), and IRT based tests (Lord, 1980; Raju, 1988, 1990).

Implementation

The **ShinyItemAnalysis** package can be used either locally or online. The package uses several other R packages to provide a wide palette of psychometric tools to analyze data (see Table 1). The main function is called `startShinyItemAnalysis()`. It launches an interactive shiny application (Figure 1) which is further described below. Furthermore, function `gDiscrim()` calculates generalized coefficient ULI comparing the ratio of correct answers in predefined upper and lower groups of students (Martinková et al., 2017). Function `ItemAnalysis()` provides a complete traditional item analysis table with summary statistics and various difficulty and discrimination indices for all items. Function `DDplot()` plots difficulty and selected discrimination indices of the items ordered by their difficulty. Function `DistractorAnalysis()` calculates the proportions of choosing individual distractors for groups of respondents defined by their totals score. Graphical representation of distractor analysis is provided via function `plotDistractorAnalysis()`. Other functions include item - person maps for IRT models, `ggWrightMap()`, and plots for DIF analysis using IRT methods, `plotDIFirt()`, and logistic regression models, `plotDIFlogistic()`. These functions may be applied directly on data from an R console as shown in the provided R code. The package also includes training datasets Medical 100, Medical 100 graded (Martinková et al., 2017), and HCI (McFarland et al., 2017).

Examples

Running the application

The **ShinyItemAnalysis** application may be launched in R by calling `startShinyItemAnalysis()`, or more conveniently, directly from `https://shiny.cs.cas.cz/ShinyItemAnalysis`. The intro page (Figure 1) includes general information about the application. Various tools are included in separate tabs which correspond to separate sections.

R package	Citation	Title
corrplot	(Wei and Simko, 2017)	Visualization of a correlation matrix
cowplot	(Wilke, 2018)	Streamlined plot theme and plot annotations for 'ggplot2'
CTT	(Willse, 2018)	Classical test theory functions
data.table	(Dowle and Srinivasan, 2018)	Extension of `data.frame`
deltaPlotR	(Magis and Facon, 2014)	Identification of dichotomous differential item functioning using Angoff's delta plot method
difNLR	(Drabinová et al., 2018)	DIF and DDF detection by non-linear regression models
difR	(Magis et al., 2010)	Collection of methods to detect dichotomous differential item functioning
DT	(Xie et al., 2018)	A wrapper of the JavaScript library 'DataTables'
ggdendro	(de Vries and Ripley, 2016)	Create dendrograms and tree diagrams using 'ggplot2'
ggplot2	(Wickham, 2016)	Create elegant data visualisations using the grammar of graphics
gridExtra	(Auguie, 2017)	Miscellaneous functions for 'grid' graphics
knitr	(Xie, 2018)	A general-purpose package for dynamic report generation in R
latticeExtra	(Sarkar and Andrews, 2016)	Extra graphical utilities based on lattice
ltm	(Rizopoulos, 2006)	Latent trait models under IRT
mirt	(Chalmers, 2018)	Multidimensional item response theory
moments	(Komsta and Novomestky, 2015)	Moments, cumulants, skewness, kurtosis and related tests
msm	(Jackson, 2011)	Multi-state Markov and hidden Markov models in continuous time
nnet	(Venables and Ripley, 2002)	Feed-forward neural networks and multinomial log-linear models
plotly	(Sievert, 2018)	Create interactive web graphics via 'plotly.js'
psych	(Revelle, 2018)	Procedures for psychological, psychometric, and personality research
psychometric	(Fletcher, 2010)	Applied psychometric theory
reshape2	(Wickham, 2007)	Flexibly reshape data: A reboot of the reshape package
rmarkdown	(Allaire et al., 2018)	Dynamic documents for R
shiny	(Chang et al., 2018)	Web application framework for R
shinyBS	(Bailey, 2015)	Twitter bootstrap components for shiny
shinydashboard	(Chang and Borges Ribeiro, 2018)	Create dashboards with 'shiny'
shinyjs	(Attali, 2018)	Easily improve the user experience of your shiny apps in seconds
stringr	(Wickham, 2018)	Simple, consistent wrappers for common string operations
xtable	(Dahl et al., 2018)	Export tables to LaTeX or HTML

Table 1: R packages used for developing **ShinyItemAnalysis**.

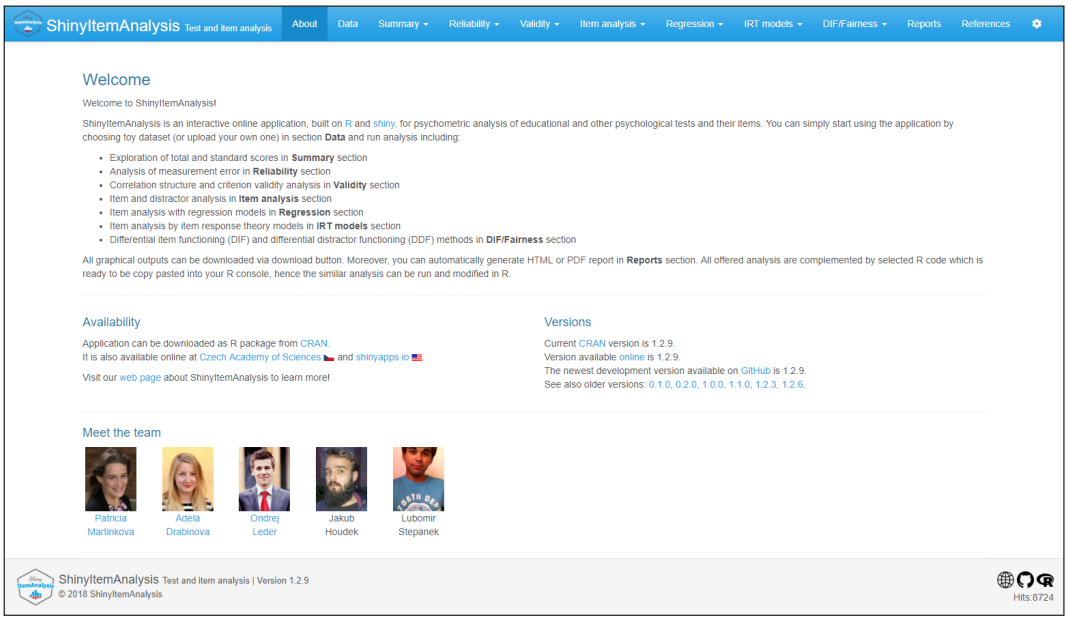

Figure 1: Intro page.

Data selection and upload

Data selection is available in section **Data**. Six training datasets may be uploaded using the **Select dataset** button: Training datasets Medical 100, Medical 100 graded (Martinková et al., 2017) and HCI (McFarland et al., 2017) from the **ShinyItemAnalysis** package, and datasets GMAT (Martinková et al., 2017), GMAT2 and MSAT-B (Drabinová and Martinková, 2017) from the **difNLR** package (Drabinová

et al., 2018).

Besides the provided toy datasets, users' own data may be uploaded as csv files and previewed in this section. To replicate examples involving the HCI dataset (McFarland et al., 2017), csv files are provided for upload in Supplemental Materials.

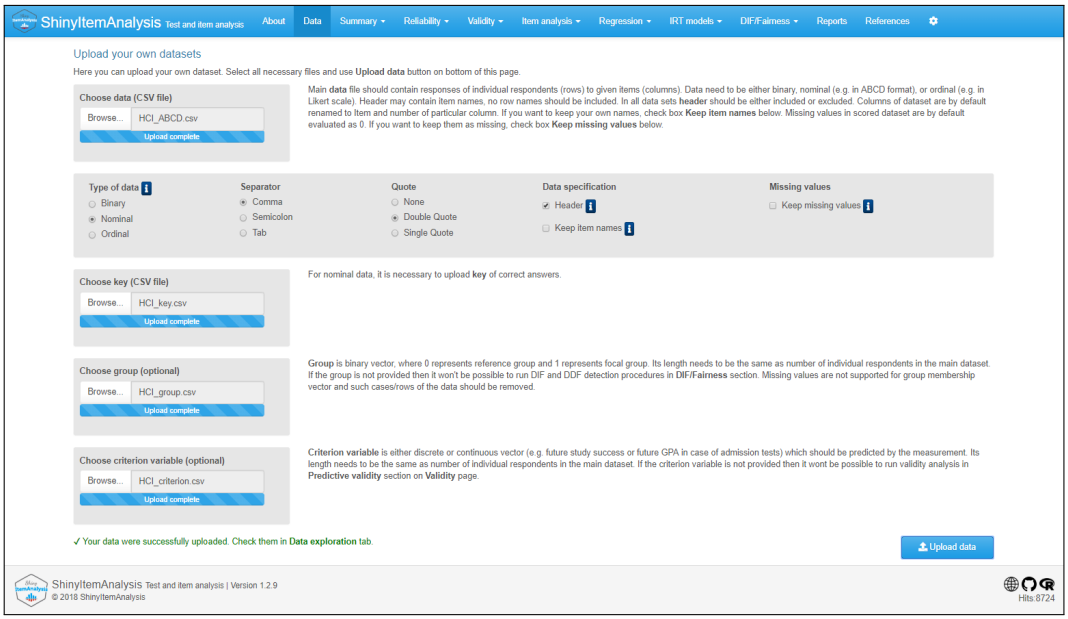

Figure 2: Page to select or upload data.

Item analysis step-by-step

Further sections of the **ShinyItemAnalysis** application allow for step-by-step test and item analysis. The first four sections are devoted to traditional test and item analyses in a framework of classical test theory. Further sections are devoted to regression models, to IRT models and to DIF methods. A separate section is devoted to report generation and references are provided in the final section. The individual sections are described below in more detail.

Section **Summary** provides for histogram and summary statistics of the total scores as well as for various standard scores (Z scores, T scores), together with percentile and success rate for each level of the total score. Section **Reliability** offers internal consistency estimates with Cronbach's alpha.

Section **Validity** provides correlation structure (Figure 3, left) and checks of criterion validity (Figure 3, right). A correlation heat map displays selected type of correlation estimates between items. Polychoric correlation is the default correlation used for binary data. The plot can be reordered using hierarchical clustering while highlighting a specified number of clusters. Clusters of correlated items need to be checked for content and other similarities to see if they are intended. Criterion validity is analyzed with respect to selected variables (e. g. subsequent study success, major, or total score on a related test) and may be analyzed for the test as a whole or for individual items.

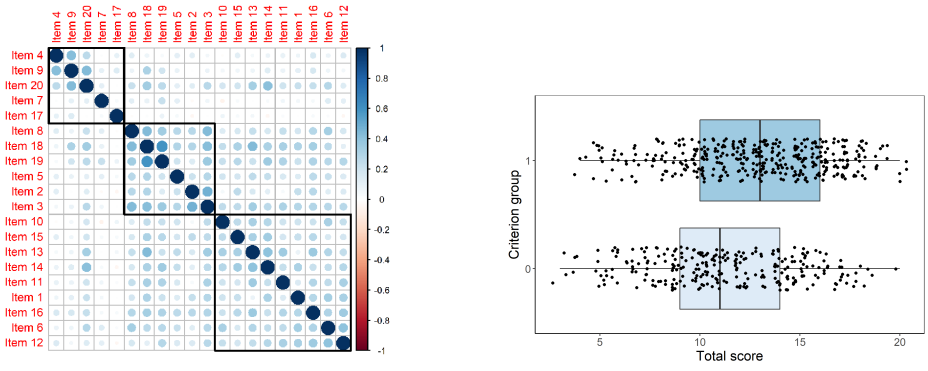

Figure 3: Validity plots for HCI data.

Section **Item analysis** offers traditional item analysis of the test as well as a more detailed distractor analysis. The so called DD plot (Figure 4) displays difficulty (red) and a selected discrimination index (blue) for all items. Items are ordered by difficulty. While lower discrimination may be expected for very easy or very hard items, all items with ULI discrimination lower than 0.2 (borderline in the plot) are worth further checks by content experts. The distractor plot (Figure 5) provides for

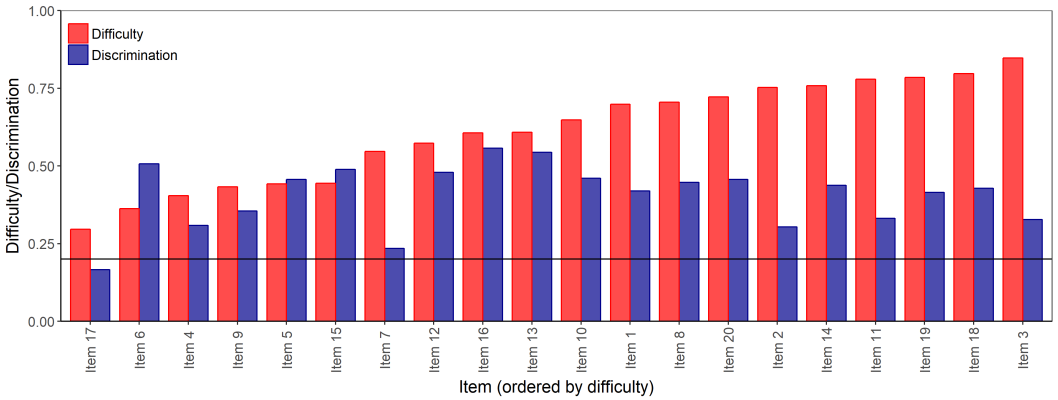

Figure 4: DD plot for HCI data.

detailed analysis of individual distractors by the respondents' total score. The correct answer should be selected more often by strong students than by students with a lower total score, i. e. the solid line in the distractor plot (see Figure 5) should be increasing. The distractors should work in an opposite direction, i. e. the dotted lines should be decreasing.

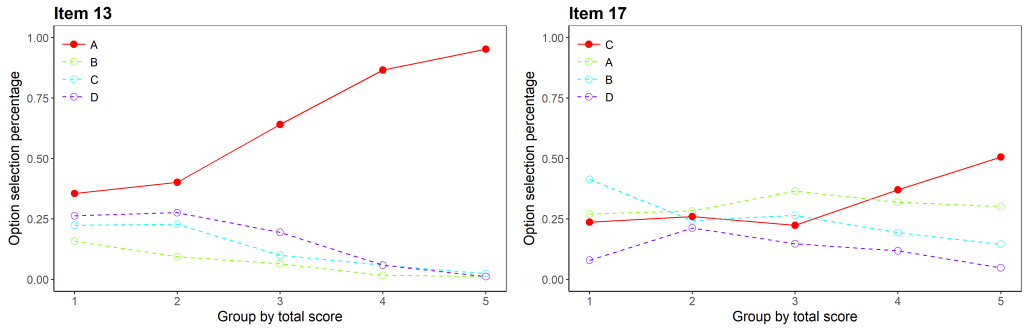

Figure 5: Distractor plots for items 13 and 17 in HCI data.

Section **Regression** allows for modelling of item properties with a logistic, non-linear or multinomial regression (see Figures 6). Probability of the selection of a given answer is modelled with respect to the total or standardized total score. Classical as well as IRT parametrization are provided for logistic and non-linear models. Model fit can be compared by Akaike's (Akaike, 1974) or Schwarz's Bayesian (Schwarz, 1978) information criteria and a likelihood-ratio test.

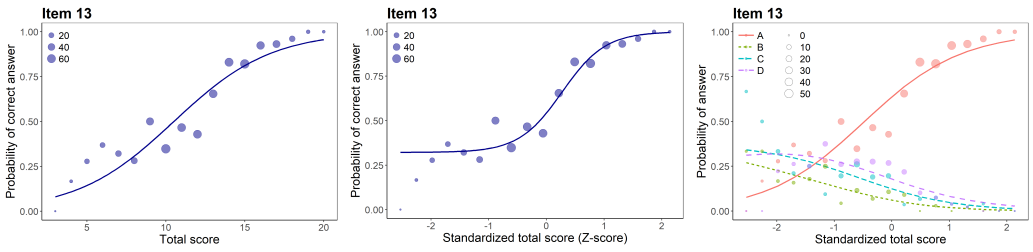

Figure 6: Regression plots for item 13 in HCI data.

Section **IRT models** provides for 1-4PL IRT models as well as Bock's nominal model, which may also be used for ordinal items. In IRT model specification, **ShinyItemAnalysis** uses default settings of the **mirt** package and for the Rasch model (Rasch, 1960) it fixes discrimination to $a = 1$, while variance of ability θ is freely estimated. Contrary to Rasch model, 1PL model allows any discrimination $a \in R$ (common to all items), while fixing variance of ability θ to 1. Similarly, other submodels of 4PL model

(4), e. g. 2PL and 3PL model, may be obtained by fixing appropriate parameters, while variance of ability θ is fixed to 1.

Interactive item characteristic curves (ICC), item information curves (IIC) and test information curves (TIC) are provided for all IRT models (see Figure 7). An item-person map is displayed for Rasch and 1PL models (Figure 7, bottom right). Table of item parameter estimates is completed by $S - X^2$ item fit statistics (Orlando and Thissen, 2000). Estimated latent abilities (factor scores) are also available for download. While fitting of IRT models is mainly implemented using the **mirt** package (Chalmers, 2018), sample R code is provided for work in both **mirt** and **ltm** (Rizopoulos, 2006).

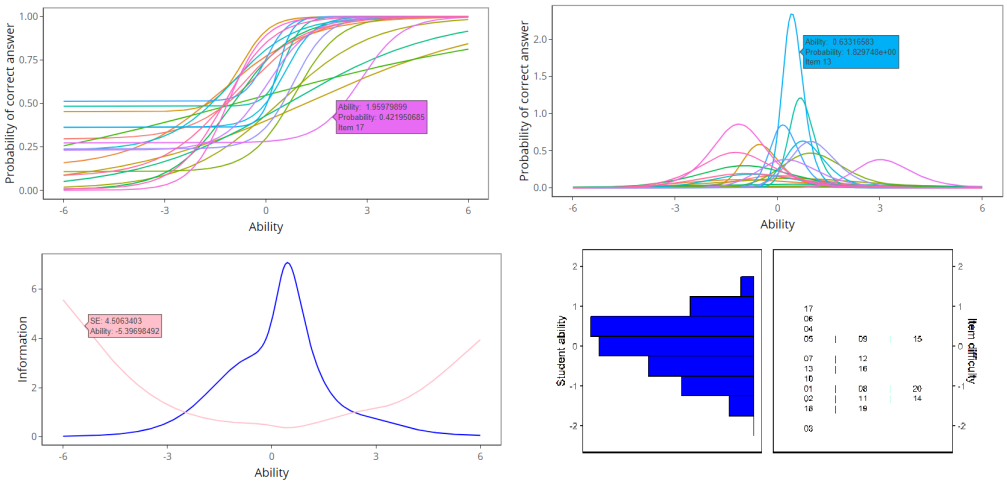

Figure 7: IRT plots for HCI data. From top left: Item characteristic curves, item information curves, test information curve with standard error of measurement, and item-person map.

Finally, section **DIF/Fairness** offers the most used tools for detection of DIF and DDF included in **deltaPlotR** (Magis and Facon, 2014), **difR** (Magis et al., 2010) and the **difNLR** package (Drabinová et al., 2018, see also Drabinová and Martinková, 2018 and Drabinová and Martinková, 2017).

Datasets GMAT and HCI provide valuable teaching examples, further discussed in Martinková et al. (2017). HCI is an example of a situation whereby the two groups differ significantly in their overall ability, yet no item is detected as DIF. Dataset GMAT was simulated to demonstrate that it is hypothetically possible that even though the distributions of the total scores for the two groups are identical, yet, there may be an item present that functions differently for each group (see Figure 8).

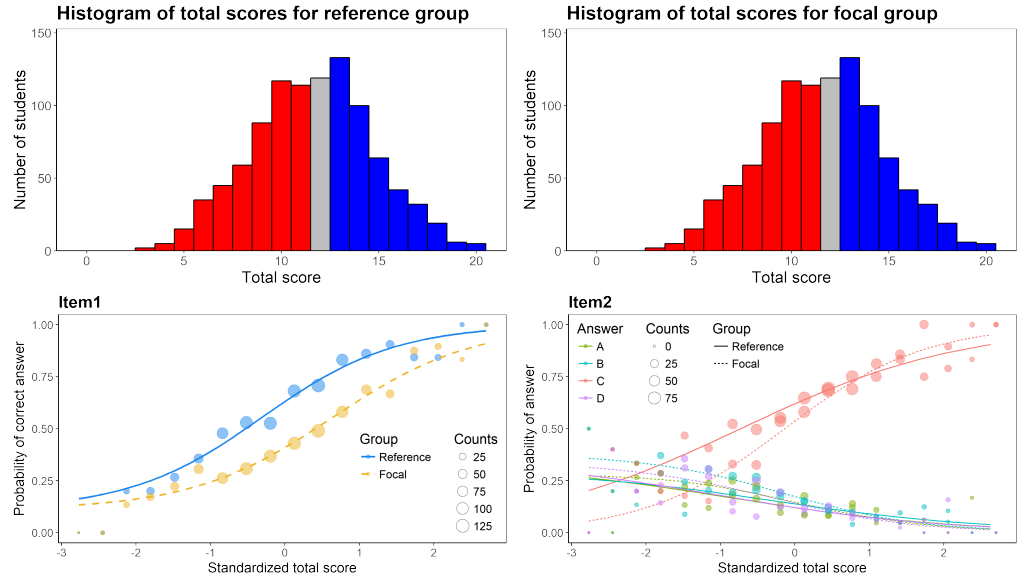

Figure 8: GMAT data simulated to show that hypothetically, two groups may have an identical distribution of total scores, yet there may be a DIF item present in the data.

Teaching with ShinyItemAnalysis

ShinyItemAnalysis is based on examples developed for a course of IRT models and psychometrics offered to graduate students at the University of Washington and at the Charles University. It has also been used at workshops for educators developing admission tests and other tests in various fields.

Besides the presence of a broad range of CTT as well as IRT methods, toy data examples, model equations, parameter estimates, and interactive interpretation of results, selected R code is also available, ready to be copy-pasted and run in R. The shiny application can thus serve as a bridge to users who do not feel secure in the R programming environment by providing examples which can be further modified or adopted to different datasets.

As an important teaching tool, an interactive training section is present, deploying item characteristic and item information curves for IRT models. For dichotomous models (Figure 9, left), the user can specify parameters a, b, c, and d of two toy items and latent ability θ of respondent. ICC is then provided interactively, displaying probability of a correct answer for any θ and highlighting this probability for selected θ. IIC compares the two items in the amount of information they provide about respondents of a given ability level.

For polytomous items (Figure 9, right), analogous interactive plots are available for GRM, (G)PCM as well as for NRM. Step functions are displayed for GRM, and category response functions are available for all three models. In addition, the expected item score is displayed for all the models.

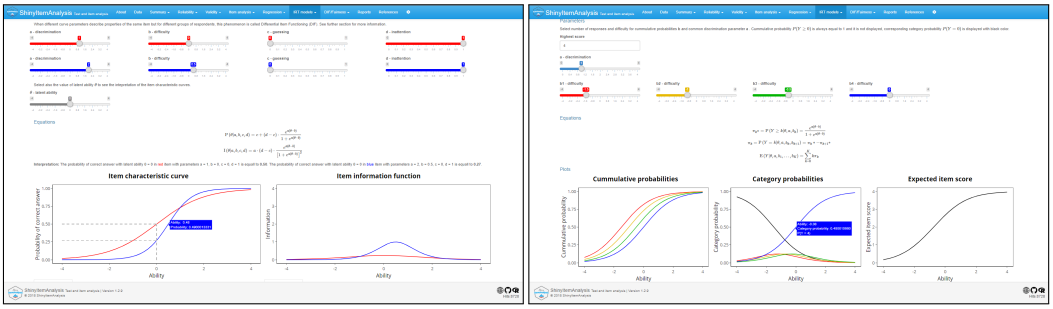

Figure 9: Interactive training IRT section.

The training sections also contain interactive exercises where students may check their understanding of the IRT models. They are asked to sketch ICC and IIC functions of items with given parameters, and to answer questions, e. g. regarding probabilities of correct answers and the information these items provide for certain ability levels.

Automatic report generation

To support routine usage of psychometric methods in test development, **ShinyItemAnalysis** offers the possibility to upload your own data for analysis as csv files, and to generate PDF or HTML reports. A sample PDF report and the corresponding csv files used for its generation are provided in Supplemental Materials.

Report generation uses **rmarkdown** templates and **knitr** for compiling (see Figure 10). LaTeX is used for creating PDF reports. The latest version of LaTeX with properly set paths is needed to generate PDF reports locally.

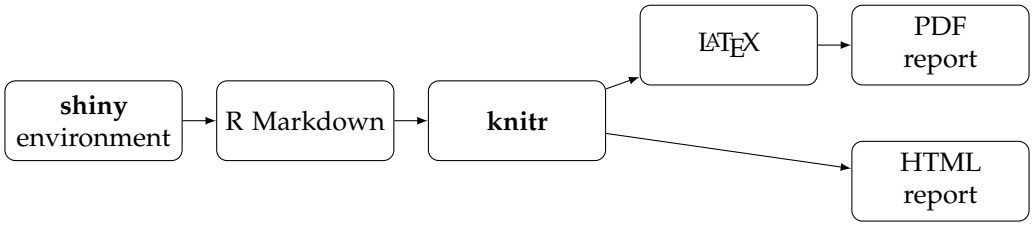

Figure 10: Report generation workflow.

Page with report setting allows user to specify a dataset name, the name of the person who generated the report, to select a method and to customize the settings (see Figure 11). The **Generate report** button at the bottom starts analyses needed for the report to be created. Subsequently, the **Download report** button initializes compiling the text, tables and figures into a PDF or an HTML file.

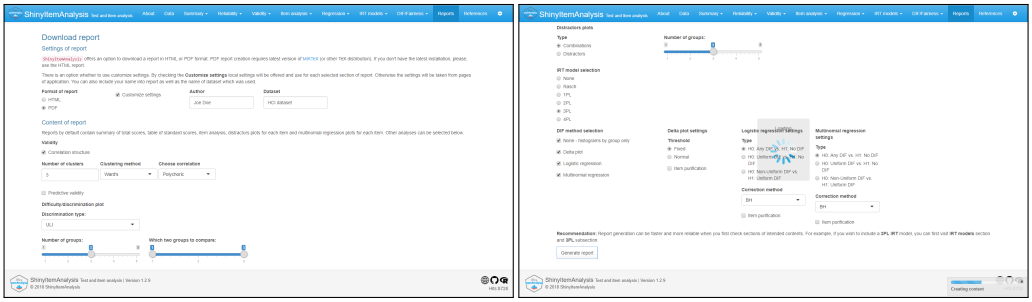

Figure 11: Report settings for the HCI data analysis.

Sample pages of a PDF report on the HCI dataset are displayed in Figure 12. Reports provide a quick overview of test characteristics and may be a helpful material for test developers, item writers and institutional stakeholders.

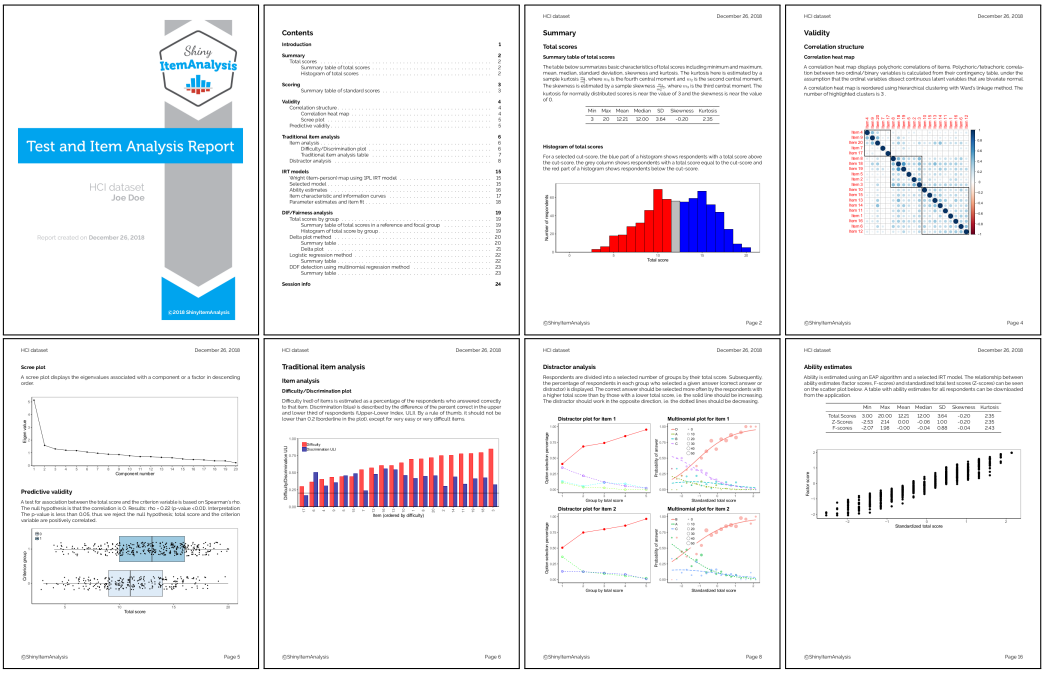

Figure 12: Selected pages of a report on the HCI data.

Discussion and conclusion

ShinyItemAnalysis is an R package (currently version 1.2.9) and an online shiny application for psychometric analysis of educational tests and items. It is suitable for teaching psychometric concepts and it aspires to be an easy-to-use tool for routine analysis of educational tests. For teaching psychometric concepts, a wide range of models and methods are provided, together with interactive plots, exercises, data examples, model equations, parameter estimates, interpretation of results, and selected R code to bring new users to R. For analysis of educational tests by educators, **ShinyItemAnalysis** interactive application allows users to upload and analyze their own data online or locally, and to automatically generate analysis reports in PDF or HTML.

Functionality of the **ShinyItemAnalysis** has been validated on three groups of users. As the first group, two university professors teaching psychometrics and test development provided their written feedback on using the application and suggested edits for wording used for interpretation of results provided by the shiny application. As the second group, over 20 participants of a graduate seminar on "Selected Topics in Psychometrics" at the Charles University in 2017/2018 used **ShinyItemAnalysis** throughout the year in practical exercises. Students prepared their final projects with **ShinyItemAnalysis** applied on their own datasets and provided closer feedback on their experience. The third group consisted of more than 20 university academics from different fields who participated in a short-term course on "Test Development and Validation" in 2018 at the Charles University. During the

course, participants used the application on toy data embedded in the package. In addition, an online knowledge test was prepared in Google Docs, answered by participants, and subsequently analyzed in **ShinyItemAnalysis** during the same session. Participants provided their feedback and commented on usability of the package and shiny application. As a result, various features were improved (e. g. data upload format was extended, some cases of missing values are now handled).

Current developments of the **ShinyItemAnalysis** package comprise implementation of wider types of models, especially ordinal and multinomial models and models accounting for effect of the rater and a hierarchical structure. In reliability estimation, further sources of data variability are being implemented to provide estimation of model-based inter-rater reliability (Martinková et al., 2018b). Technical improvements include a more complex data upload or automatic testing of new versions of the application.

We argue that psychometric analysis should be a routine part of test development in order to gather proofs of reliability and validity of the measurement and we have demonstrated how **ShinyItemAnalysis** may enforce this goal. It may also serve as an example for other fields, demonstrating the ability of shiny applications to interactively present theoretical concepts and, when complemented with sample R code, to bring new users to R, or to serve as a bridge to those who have not yet discovered the beauties of R.

Acknowledgments

This work was initiated while P. Martinková was a Fulbright-Masaryk fellow with University of Washington. Research was partly supported by the Czech Science Foundation (grant number GJ15-15856Y). We gratefully thank Jenny McFarland for providing HCI data and David Magis, Hynek Cígler, Hana Ševčíková, Jon Kern and anonymous reviewers for helpful comments to previous versions of this manuscript. We also wish to acknowledge those who contributed to the **ShinyItemAnalysis** package or provided their valuable feedback by e-mail, on GitHub or through an online Google form at http://www.ShinyItemAnalysis.org/feedback.html.

Bibliography

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. URL https://doi.org/10.1109/TAC.1974.1100705. [p508]

J. Allaire, Y. Xie, J. McPherson, J. Luraschi, K. Ushey, A. Atkins, H. Wickham, J. Cheng, and W. Chang. *rmarkdown: Dynamic Documents for R*, 2018. URL https://CRAN.R-project.org/package=rmarkdown. R package version 1.10. [p506]

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). *Standards for Educational and Psychological Testing*. American Educational Research Association, 2014. [p503]

D. Andrich. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573, 1978. URL https://doi.org/10.1007/BF02293814. [p505]

W. H. Angoff and S. F. Ford. Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 10(2):95–105, 1973. URL https://doi.org/10.1002/j.2333-8504.1971.tb00812.x. [p505]

D. Attali. *shinyjs: Easily Improve the User Experience of Your shiny Apps in Seconds*, 2018. URL https://CRAN.R-project.org/package=shinyjs. R package version 1.0. [p506]

B. Auguie. *gridExtra: Miscellaneous Functions for 'Grid' Graphics*, 2017. URL https://CRAN.R-project.org/package=gridExtra. R package version 2.3. [p506]

E. Bailey. *shinyBS: Twitter Bootstrap Components for shiny*, 2015. URL https://CRAN.R-project.org/package=shinyBS. R package version 0.61. [p506]

R. D. Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51, 1972. URL https://doi.org/10.1007/BF02291411. [p505]

R. L. Brennan. *Educational Measurement*. Praeger Publishers, 2006. [p503]

P. Chalmers. *mirt: Multidimensional Item Response Theory*, 2018. URL https://CRAN.R-project.org/package=mirt. R package version 1.29. [p505, 506, 509]

W. Chang and B. Borges Ribeiro. *shinydashboard: Create Dashboards with 'shiny'*, 2018. URL https://CRAN.R-project.org/package=shinydashboard. R package version 0.7.1. [p506]

W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson. *shiny: Web Application Framework for R*, 2018. URL https://CRAN.R-project.org/package=shiny. R package version 1.2.0. [p503, 506]

L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951. URL https://doi.org/10.1007/BF02310555. [p504]

D. B. Dahl, D. Scott, C. Roosen, A. Magnusson, and J. Swinton. *xtable: Export Tables to LaTeX or HTML*, 2018. URL https://CRAN.R-project.org/package=xtable. R package version 1.8-3. [p506]

A. de Vries and B. D. Ripley. *ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'*, 2016. URL https://CRAN.R-project.org/package=ggdendro. R package version 0.1-20. [p506]

M. Dowle and A. Srinivasan. *data.table: Extension of 'data.frame'*, 2018. URL https://CRAN.R-project.org/package=data.table. R package version 1.11.8. [p506]

S. M. Downing and T. M. Haladyna, editors. *Handbook of Test Development*. Lawrence Erlbaum Associates, Inc., 2011. [p503]

A. Drabinová and P. Martinková. Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement*, 54(4):498–517, 2017. URL https://doi.org/10.1111/jedm.12158. [p505, 506, 509]

A. Drabinová and P. Martinková. difNLR: Generalized logistic regression models for DIF and DDF detection. *The R Journal*, 2018. Submitted. [p509]

A. Drabinová, P. Martinková, and K. Zvára. *difNLR: DIF and DDF Detection by Non-Linear Regression Models*, 2018. URL https://CRAN.R-project.org/package=difNLR. R package version 1.2.2. [p506, 509]

R. L. Ebel. Procedures for the analysis of classroom tests. *Educational and Psychological Measurement*, 14 (2):352–364, 1954. URL https://doi.org/10.1177/001316445401400215. [p503]

T. D. Fletcher. *psychometric: Applied Psychometric Theory*, 2010. URL https://CRAN.R-project.org/package=psychometric. R package version 2.2. [p506]

C. H. Jackson. Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, 38(8):1–29, 2011. URL https://doi.org/10.18637/jss.v038.i08. [p506]

M. S. Johnson. Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software*, 20(10):1–24, 2007. URL https://doi.org/10.18637/jss.v020.i10. [p505]

L. Komsta and F. Novomestky. *moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests*, 2015. URL https://CRAN.R-project.org/package=moments. R package version 0.14. [p506]

F. M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Routledge, 1980. [p505]

D. Magis and B. Facon. deltaPlotR: An R package for differential item functioning analysis with Angoff's delta plot. *Journal of Statistical Software, Code Snippets*, 59(1):1–19, 2014. URL https://doi.org/10.18637/jss.v059.c01. [p506, 509]

D. Magis, S. Beland, F. Tuerlinckx, and P. De Boeck. A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, 42:847–862, 2010. URL https://doi.org/10.3758/BRM.42.3.847. [p503, 506, 509]

N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4):719–748, 1959. URL https://doi.org/10.1093/jnci/22.4.719. [p505]

P. Martinková, A. Drabinová, and J. Houdek. ShinyItemAnalysis: Analýza přijímacích a jiných znalostních či psychologických testů [ShinyItemAnalysis: Analyzing admission and other educational and psychological tests]. *TESTFÓRUM*, 6(9):16–35, 2017. URL https://doi.org/10.5817/TF2017-9-129. [p503]

P. Martinková, L. Štěpánek, A. Drabinová, J. Houdek, M. Vejražka, and Č. Štuka. Semi-real-time analyses of item characteristics for medical school admission tests. In *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*, pages 189–194. IEEE, 2017. URL https://doi.org/10.15439/2017F380. [p503, 505, 506]

P. Martinková, A. Drabinová, O. Leder, and J. Houdek. *ShinyItemAnalysis: Test and Item Analysis via shiny*, 2018a. URL https://CRAN.R-project.org/package=ShinyItemAnalysis. R package version 1.2.9. [p503]

P. Martinková, D. Goldhaber, and E. Erosheva. Disparities in ratings of internal and external applicants: A case for model-based inter-rater reliability. *PloS ONE*, 13(10):e0203002, 2018b. URL https://doi.org/10.1371/journal.pone.0203002. [p512]

P. Martinková, A. Drabinová, Y.-L. Liaw, E. A. Sanders, J. L. McFarland, and R. M. Price. Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE-Life Sciences Education*, 16(2):rm2, 2017. URL https://doi.org/10.1187/cbe.16-10-0307. [p505, 506, 509]

G. N. Masters. A Rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982. URL https://doi.org/10.1007/BF02296272. [p505]

J. L. McFarland, R. M. Price, M. P. Wenderoth, P. Martinková, W. Cliff, J. Michael, H. Modell, and A. Wright. Development and validation of the homeostasis concept inventory. *CBE-Life Sciences Education*, 16(2):ar35, 2017. URL https://doi.org/10.1187/cbe.16-10-0305. [p503, 505, 506, 507]

E. Muraki. A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, 1992(1), 1992. URL https://doi.org/10.1002/j.2333-8504.1992.tb01436.x. [p505]

M. Orlando and D. Thissen. Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1):50–64, 2000. URL https://doi.org/10.1177/2F01466216000241003. [p509]

N. S. Raju. The area between two item characteristic curves. *Psychometrika*, 53(4):495–502, 1988. URL https://doi.org/10.1007/BF02294403. [p505]

N. S. Raju. Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14(2):197–207, 1990. URL https://doi.org/10.1177/014662169001400208. [p505]

G. Rasch. *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*. Nielsen & Lydiche, 1960. [p508]

W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*, 2018. URL https://CRAN.R-project.org/package=psych. R package version 1.8.10. [p503, 506]

F. Rijmen, F. Tuerlinckx, P. De Boeck, and P. Kuppens. A nonlinear mixed model framework for item response theory. *Psychological methods*, 8(2):185, 2003. URL https://doi.org/10.1037/1082-989X.8.2.185. [p504]

D. Rizopoulos. ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25, 2006. URL https://doi.org/10.18637/jss.v017.i05. [p503, 505, 506, 509]

Y. Rosseel. lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36, 2012. URL https://doi.org/10.18637/jss.v048.i02. [p503]

F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34 (1):1–97, 1969. URL https://doi.org/10.1007%2FBF03372160. [p505]

D. Sarkar and F. Andrews. *latticeExtra: Extra Graphical Utilities Based on Lattice*, 2016. URL https://CRAN.R-project.org/package=latticeExtra. R package version 0.6-28. [p506]

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. URL https://doi.org/10.1214/aos/1176344136. [p508]

C. Sievert. *plotly for R*, 2018. URL https://plotly-book.cpsievert.me. [p506]

H. Swaminathan and H. J. Rogers. Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, 27(4):361–370, 1990. URL https://doi.org/10.1111/j.1745-3984.1990.tb00754.x. [p505]

W. J. van der Linden. *Handbook of Item Response Theory*. Chapman and Hall/CRC, 2017. [p505]

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, New York, 4th edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. [p506]

T. Wei and V. Simko. *corrplot: Visualization of a Correlation Matrix*, 2017. URL https://CRAN.R-project.org/package=corrplot. R package version 0.84. [p506]

H. Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007. URL https://doi.org/10.18637/jss.v021.i12. [p506]

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, 2016. URL http://ggplot2.org. [p506]

H. Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2018. URL https://CRAN.R-project.org/package=stringr. R package version 1.3.1. [p506]

C. O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2018. URL https://CRAN.R-project.org/package=cowplot. R package version 0.9.3. [p506]

J. T. Willse. *CTT: Classical Test Theory Functions*, 2018. URL https://CRAN.R-project.org/package=CTT. R package version 2.3.3. [p506]

Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2018. URL https://CRAN.R-project.org/package=knitr. R package version 1.20. [p506]

Y. Xie, J. Cheng, and X. Tan. *DT: A Wrapper of the JavaScript Library 'DataTables'*, 2018. URL https://CRAN.R-project.org/package=DT. R package version 0.5. [p506]

R. E. Zinbarg, W. Revelle, I. Yovel, and W. Li. Cronbach's α, revelle's β, and mcdonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1): 123–133, 2005. URL https://doi.org/10.1007/s11336-003-0974-7. [p504]

Patrícia Martinková
Department of Statistical Modelling, Institute of Computer Science, Czech Academy of Sciences
Pod Vodárenskou věží 271/2, Prague, 182 07
and
Institute for Research and Development of Education, Faculty of Education, Charles University
Myslíkova 7, Prague, 110 00
Czech Republic
ORCiD: 0000-0003-4754-8543
martinkova@cs.cas.cz

Adéla Drabinová
Department of Statistical Modelling, Institute of Computer Science, Czech Academy of Sciences
Pod Vodárenskou věží 271/2, Prague, 182 07
and
Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University
Sokolovská 83, Prague, 186 75
Czech Republic
ORCiD: 0000-0002-9112-1208
drabinova@cs.cas.cz