Proceedings of the Federated Conference on Computer Science and Information Systems pp. 189–194
DOI: 10.15439/2017F380
ISSN 2300-5963 ACSIS, Vol. 11

Semi-real-time analyses of item characteristics for medical school admission tests

Patrícia Martinková
Institute of Computer Science
Czech Academy of Sciences
Pod Vodárenskou věží 2, Praha 8
martinkova@cs.cas.cz

Lubomír Štěpánek
Institute of Biophysics and Informatics
First Faculty of Medicine, Charles University
Salmovská 1, Praha 2
lubomir.stepanek@lf1.cuni.cz

Adéla Drabinová
Institute of Computer Science
Czech Academy of Sciences
Pod Vodárenskou věží 2, Praha 8
adela.drabinova@gmail.com

Jakub Houdek
Institute of Computer Science
Czech Academy of Sciences
Pod Vodárenskou věží 2, Praha 8
houdek.james@gmail.com

Martin Vejražka
Institute of Medical Biochemistry and Laboratory Diagnostics
First Faculty of Medicine, Charles University
U Nemocnice 2, Praha 2
martin.vejrazka@lf1.cuni.cz

Čestmír Štuka
Institute of Biophysics and Informatics
First Faculty of Medicine, Charles University
Salmovská 1, Praha 2
cestmir.stuka@lf1.cuni.cz

***Abstract*—University admission exams belong to so-called high-stakes tests, i. e. tests with important consequences for the exam taker. Given the importance of the admission process for the applicant and the institution, routine evaluation of the admission tests and their items is desirable.**

In this work, we introduce a quick and efficient methodology and on-line tool for semi-real-time evaluation of admission exams and their items based on classical test theory (CTT) and item response theory (IRT) models. We generalize some of the traditional item analysis concepts to tailor them for specific purposes of the admission test.

On example of medical school admission test we demonstrate how R-based web application may simplify admissions evaluation work-flow and may guarantee quick accessibility of the psychometric measures. We conclude that the presented tool is convenient for analysis of any admission or educational test in general.

I. Introduction

ADEQUATE selection of students to higher education is a crucial point for both the applicant and the institution, because the quality of students influences the school's reputation and vice versa. In some countries, standardized tests have been used for decades in admission process and examination of test and item properties according to field standards [1] is a routine task [2].

In the Czech Republic, medical schools traditionally organize in-house admissions and prepare their own admission tests. Total score achieved in admission tests is usually the main criterion for the admission decision. Yet, at Czech universities, the scope of analyses checking test and item properties varies among individual institutions. While some schools publish validation studies of their exams [3], [4], [5], [6], [7], others may perform psychometric analyses as internal reports or the test and item analysis is missing. While monographs containing the methodology of test analysis have been published in Czech language [8], [9], [10], the use of robust psychometric measures in test development is still limited.

Test and item analysis can be carried out in a variety of widely available general statistical analysis software, such as R [11], SPSS [12], STATA [13], SAS [14], and others. For item analysis based on CTT, software Iteman [15] or descriptive statistics available in Rogo [16] may be used. For analyses within IRT framework, there are several commercially available packages including Winsteps [17], IRTPRO [18], and ConQuest [19]; for other psychometric software, see also [20]. While commercially available psychometric software provides graphically convenient environment for the end user, its use may be limited due to financial costs. It is usually also impossible to tailor the provided calculations to the needs of the user, for example to adopt the existing methods for multiple true-false format of the items or to take into account the ratio of admitted students.

In this work, we present a web application ShinyItemAnalysis [21] for psychometric analysis of admission tests and their items, available online at

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

which covers broad range of psychometric methods and offers training data examples while also allowing the users to upload

and analyse their own data and to generate analysis report. We further focus on generalization of traditional item characteristics and their incorporation into the application to allow for analyses tailored to the needs of specific admission test. We conclude by arguing that psychometric analysis should be a routine part of test development in order to gather proofs of reliability and validity of the measurement. With example of medical school admission test we demonstrate how ShinyItemAnalysis may simplify the workflow of admission tests.

II. Research Methodology

The item analysis and evaluation of its psychometric properties should be a routine part in test development cycle [22], [23]. At First Faculty of Medicine, Charles University, the current test development cycle consists of the following phases:

(i) item writing;
(ii) item revision performed by domain experts;
(iii) test composition based on prespecified knowledge domains;
(iv) test revision;
(v) test printing distribution to admission applicants;
(vi) test administration (written examination);
(vii) automatic and anonymized test scoring using a scanner (output of which is a vector of student total scores as well as a flat-file dataset consisting of responses of all applicants to each item);
(viii) automatic evaluation of test and item psychometric properties;
(ix) feedback to item creators

In case when the evaluation detects a suspicious item e.g. with a very high or low difficulty or with very low discrimination, such item is eliminated and not used in test scoring. The item is sent back to item writer and is further reformulated or eliminated from the item bank.

The methodology of evaluation of admission tests and their items described below is particularly involved in phases (viii) and (ix) of the workflow above. In optimal case, the evaluation should be done during a pretest. This is currently not the case due to security reasons, and it is thus even more important to perform the analysis in a semi-real time.

A. Psychometric measures used in test and item evaluation

To evaluate the test, we use mix of CTT and IRT measures. Summary statistics are provided for the total score, together with a histogram and standard scores. Correlation heatmap displays dependencies between test items and internal structure of the test. Cronbach's alpha [24] is provided as a measure of internal consistency. Traditional item analysis further displays item difficulty and discrimination as well as properties of each individual distractor: Difficulty is defined as ratio of students who answered correctly to the item. Discrimination is defined as difference of percent correct in upper and lower third of students (Upper-Lower Index, ULI) and by Pearson correlation (R) between item and Total score (index RIT). By rule of thumb, discrimination should not be lower than 0.2, except for very easy or very difficult items. To analyse properties of individual distractors, respondents are divided into three groups by their total score; having the equinumerous divison of students' scores, ULI could be computed after that. Subsequently, we display percentage of students in each group who selected given answer (correct answer or distractor) in Distractor plot. The correct answer should be more often selected by strong students than by students with lower total score. The distractor should work in opposite direction, i. e. the ratio of students who picked distractors should be decreasing with total score. Items with negative or very low discrimination should be revised or discarded [23], ineffective distractors should be reconsidered as well.

Regression models [25] and so called IRT models [26] are used to give more precise description of item properties. Instead of displaying proportions, the regression models fit a smooth line with given parameters. These parameters are then used to describe difficulty and discrimination of the item and probability of guessing. In IRT models, student abilities are estimated simultaneously with item parameters and each item may influence ability estimates differently depending on its discrimination power. As a more detailed analysis, regression and IRT models may be used to detect a situation when item functions differently for different groups, e.g. males vs. females, or majorities vs. minorities. This so called *differential item functioning* [27] is a potential threat to item fairness and test validity and should therefore be tested routinely [28].

B. Generalized ULI index and Distractors Plot

Because the test used at First Faculty of Medicine is composed of multiple true-false items, as part of the CTT measures, we have developed graphical representation, Distractors plot, allowing quick visual check of item properties (see also [8], [10], [29]). This visualization represents properties of all correct answers and all distractors at once. Since ratio of admitted students is usually around $1/5$, we may consider employing quintiles instead of terciles in the above defined index ULI. More generally, any q-quantiles may be considered. Formalization of this generalization is provided below.

Let's suppose we have a flat-file dataset (created at phase (vii)) for a given exam test, where each row consists of one of applicant's answers and each column corresponds to one of the test item questions. Dimension of the flat-file dataset is thus equal to the number of all the applicants (vertical dimension) times number of all the test items (horizontal dimension). All items included within the test are multiple true-false, thus each cell consists of combination of answers the student selected (item response pattern).

First of all, q-quantiles are calculated for applicants' total scores (see also [30]), where $q \in \{2, 3, 4, \ldots\}$; q-quantiles are values that partition sorted vector of applicants' total scores into q subsets of (nearly) equal size. For example, if $q = 3$ we got terciles dividing the range of the scores vector into three subsets, for $q = 5$ obtained quintiles split the vector into five nearly equal-size subsets.

Let n be a number of all the applicants taking the test, m be a number of all the test items and $x = (x_1, x_2, \ldots, x_n)^T$ be a vector of applicants' total scores, i.e. number of items they answered correctly, where $0 \leq x_j \leq m$. Let Q_i be the i-th q-quantile for applicants' total scores, where $i \in \{1, \ldots, q-1\}$, then[1]

$$Q_i = \lceil (j - (n-1)p)x_{(j)} + ((n-1)p + 1 - j)x_{(j+1)} \rceil,$$

where $p = i/q$, $j = \lfloor (n-1)p + 1 \rfloor$ and $x_{(j)}$ is the j-th smallest value in the vector of applicants' scores $x = (x_1, x_2, \ldots, x_n)^T$. Formally, let's define $Q_0 = 0$ and let $Q_q = m$ be equal to the number of the test items. Then an applicant with a total score equal to x_j belongs to k-th subset if and only if

$$Q_{k-1} \leq x_j < Q_k,$$

where $k \in \{1, \ldots, q\}$.

As a second step, let $u_{k,t}^{\{q\}}$ be a proportion of applicants belonging to the k-th subset, who answered the item t correctly, to all applicants belonging to the k-th subset, where $k \in \{1, \ldots, q\}$, $t \in \{1, 2, \ldots, m\}$ and where $q \in \{2, 3, 4, \ldots\}$ is fixed. Let $s_{j,t} = 1$, if the j-th applicant answered the item t correctly, and $s_{j,t} = 0$ otherwise; and let $\mathcal{M} = \{j : j \in \{1, 2, \ldots, n\} \wedge Q_{k-1} \leq x_j < Q_k\}$ be the set of all applicants belonging to the k-th subset whose boundaries are Q_{k-1} and Q_k, i. e. $k-1$-th and k-th q-quantile. Then,

$$u_{k,t}^{\{q\}} = \frac{\sum_{j \in M} s_{j,t}}{|M|}.$$

for each $k \in \{1, \ldots, q\}$ and each $t \in \{1, \ldots, m\}$ and fixed q.

Furthermore, let's suppose $u_{k,t,w}^{\{q\}}$ be a proportion of applicants belonging to the k-th q-quantile who answered the item t by checking the option w, to all applicants belonging to the k-th quantille, where $q \in \{2, 3, 4, \ldots\}$ is fixed and $k \in \{1, \ldots, q\}$, $t \in \{1, \ldots, m\}$ and $w \in \{A, B, C, D\}$ in our settings. Let $c_{j,t,w} = 1$, if the j-th applicant answered the item t by checking the option w, and $c_{j,t,w} = 0$ otherwise. Then,

$$u_{k,t,w}^{\{q\}} = \frac{\sum_{j \in M} c_{j,t,w}}{|M|}$$

for each $k \in \{1, 2, \ldots, q-1, q\}$, each $t \in \{1, 2, \ldots, m\}$ and each $w \in \{A, B, C, D\}$ and fixed q.

In case we fix t and choose an appropriate q (common choice is $q = 3$) we are able to get a q-tuple in the form of $[u_{k,t}^{\{q\}}]_{k=1}^{q}$ and exactly w q-tuples in the form of $[u_{k,t,w}^{\{q\}}]_{k=1}^{q}$ which can further be used to illustrate properties of the item t and all its distractors and correct answers (as an example, see Fig. 1)

To depict attractiveness of individual answers and their combination, proportion of selected item response pattern may be depicted as formalised below for a test in which all items are multiple true-false with four options A, B, C, D. In such a case, there is exactly $2^{|\{A,B,C,D\}|} = 2^4 = 16$ possible ways how to answer the item question (16 item response patterns).

Let

$$\begin{aligned}\mathcal{O} = \{&\emptyset,\\ &A, B, C, D,\\ &AB, AC, AD, BC, BD, CD,\\ &ABC, ABD, ACD, BCD,\\ &ABCD\}\end{aligned}$$

be the set of all possible item response patterns. For each item t, we can calculate proportion $v_{o,t}$ of number of applicants who checked item response pattern o of item t such that $o \in \mathcal{O}$ to number of all applicants[2]. Let $\mathcal{V} = \{j : j \in \{1, 2, \ldots, n\} \wedge j\text{-th applicant who chose item response pattern } o \in \mathcal{O} \text{ of item } t\}$ be the set of all applicants who chose item response pattern $o \in \mathcal{O}$ when answering item t. Then,

$$v_{o,t} = \frac{|\mathcal{V}|}{n}$$

for each $t \in \{1, \ldots, m\}$ and each $o \in \mathcal{O}$.

In case we fix t, i. e. if we choose one item t, we are able depict a 16-tuple in the form of $[v_{o,t}]_{o \in \mathcal{O}}$ which shows attractiveness of each item response pattern for item t (for example, see Fig. 2).

Finally, for each item t, we can calculate *difficulty* and *discrimination* measures. Difficulty diffc_t of the item t is defined using the proportion of applicants who correctly answered the item question t to all applicants,

$$\text{diffc}_t = 1 - \frac{\sum_{j \in \{1, \ldots, n\}} s_{j,t}}{n},$$

for each $t \in \{1, \ldots, m\}$ and where

$$s_{j,t} = \begin{cases} 1, & \text{if } j\text{-th applicant answered item } t \text{ correctly} \\ 0, & \text{otherwise.} \end{cases}$$

Intuitively, difficulty of an item is proportional to number of incorrect answers recorded for the item question. (Note: often, difficulty is defined as proportion of examinees who answered the item correctly, thus describing rather item easiness, see [23].)

Discrimination which is in CTT often described by upper-lower index (ULI, difference in proportion of correct answers in upper and lower third of students, i. e. using 3-quantiles) is here defined using quintiles (5-quantiles).

In general, discrimination $\text{discr}_t^{\{q\}}(l_1, l_2)$ is a difference between two proportions:

$$\text{discr}_t^{\{q\}}(l_1, l_2) = u_{l_2,t}^{\{q\}} - u_{l_1,t}^{\{q\}},$$

where

[1]Function $\lfloor x \rfloor$, floor of x, returns the greatest integer less than or equal to x, and function $\lceil x \rceil$, ceiling of x, returns the least integer greater than or equal to x.

[2]No option checked, i. e. $o = \emptyset$ is also possible item response pattern.

- $u_{l_1,t}^{\{q\}}$ is a proportion of applicants belonging to the l_1-th group, who answered the item t correctly, to all applicants belonging to the l_1-th group, where $l_1 \in \{1, \ldots, q\}$
- a proportion $u_{l_2,t}^{\{q\}}$ of applicants belonging to the l_2-th group, who answered the item t correctly, to all applicants belonging to the l_2-th group, where $l_2 \in \{1, \ldots, q\}$ and where $l_1 < l_2$ for a fixed q.

Intuitively, discrimination of an item describes, how well does the item discriminate between two groups of applicants which are defined by their total scores.

In our particular case, we use discrimination measure depicting differences between first and fifth quintile ($q = 5$) groups as a general discrimination measure analogous to traditional index ULI (where $q = 3$). We are even more interested in item discrimination between fourth and fifth quintile, because we usually admit only upper fifth of the students, thus this (the fourth quintile) is the cut-off where we want to discriminate the best. As an example, see difficulties and discrimination measures depicted in Fig. 3.

C. *R-language web-based application for semi-real-time evaluation of admission exam tests data*

Methodology described above was implemented in an online, freely-available application and R package `ShinyItemAnalysis` [21], [31], available online at

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

The core of the application is built of source code written in language R which is a *free-as-in-beer* and *free-as-in-speech* programming language and environment for statistical computing and graphics and is widely used among statisticians, econometricians, or biologists. Code chunks written offline in R language were uploaded online using `shiny` package on server dedicated to R calculations; `shiny` package is a library written also in R which provides an online framework for R scripts.

Components of the application consist of `ui.R`, which defines graphical user interface in terms of HTML (HyperText Markup Language), CSS (Cascading Style Sheets) and a little bit of javascript, and of `server.R` covering all workhorse functions and procedures of the application. There are other components beyond the mentioned two, but these are not necessary for application running. Graphical user interface offers multi-tabelar layout as each tab displays one of several kinds of plots based on the tuples of important characteristics described in Research methodology passage.

The application is accompanied by training datasets, example R code, model equations and interpretation and is thus well suited for routine test analysis as well as for educational purposes and teaching the methods. Also, the application allows for online typesetting of TEXdocuments and downloadable .pdf and HTML reports containing tables and figures with estimates described above.

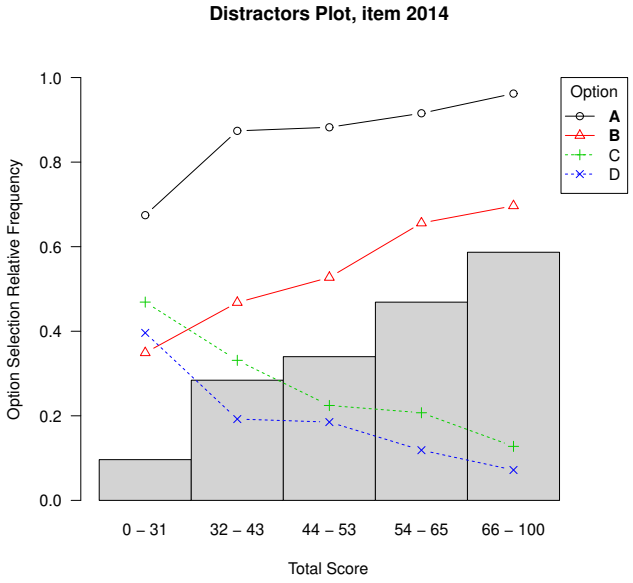

Fig. 1. Distractors plot for item 2014 based on quintiles. Bold lines (A and B) depict correct answers, as expected, percentage of applicants who chose these answers is increasing with total score. Dotted lines (C and D) represent distractors, relative frequency of their selection is decreasing with total score. Combination of correct answers (correct item response pattern) is depiced by bar graph and again is supposed to be increasing.

III. Results

Here we focus on presentation of Generalized ULI index and Distractors Plot, while we leave it upon the reader to examine the other functionalities of the *ShinyItemAnalysis* application online or in R. We present analyses and plots for medical school admission test in chemistry.

Calculation of Upper-lower index (ULI) as well as of Discrimination Plot (Fig. 1) is based on quintiles (5-quantiles) due to the fact that usually about $1/5$ of the applicants is admitted. We are thus mostly interested, how well does the item discriminate between students above and below the fourth quintile.

Detailed distribution of item response patterns is depicted in Fig. 2.

Finally, item difficulties and discrimination indices are displayed in Fig. 3.

IV. Conclusion

In this paper we have introduced `ShinyItemAnalysis` application for psychometric analysis of admission tests and their items. `ShinyItemAnalysis` provides graphical interface and web framework to open source statistical software R and thus opens up its functionality to wide audience. Application covers broad range of methods and offers data examples, model equations, parameter estimates, interpretation of results, together with selected R code, and is thus suitable for teaching psychometric concepts with R, besides, it aspires to be a simple tool for routine analysis by allowing the users to upload and analyse their own data and by generating

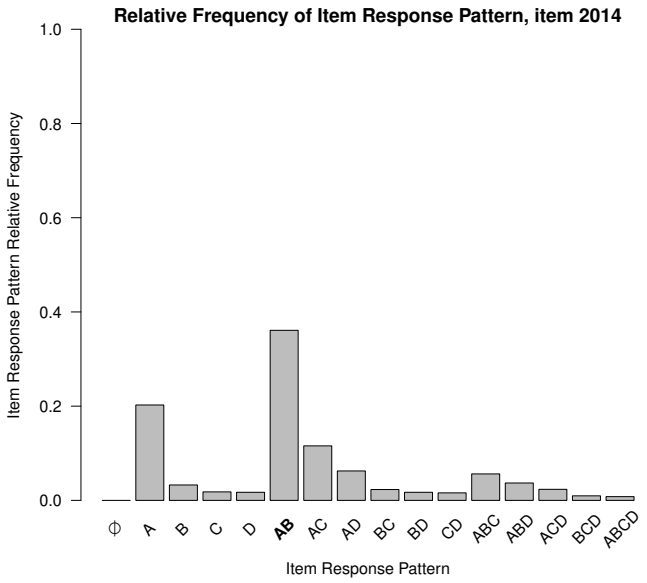

Fig. 2. Detailed distribution of item response patterns for item 2014

analysis report. We have demonstrated, how traditional Upper-Lower index may be generalized to tailor this descriptive statistics to the needs of the individual test. We conclude by arguing that psychometric analysis should be a routine part of test development in order to gather proofs of reliability and validity of the measurement. With example of admission test to medical faculty we demonstrate how `ShinyItemAnalysis` provides a simple and free tool to routinely analyse tests.

REFERENCES

[1] AERA, APA, and NCME. *Standards for educational and psychological testing*. 2014.

[2] Penny Salvatori. "Reliability and Validity of Admissions Tools Used to Select Students for the Health Professions". In: *Advances in Health Sciences Education* 6.2 (2001), pp. 159–175. ISSN: 13824996. DOI: 10.1023/A:1011489618208.

[3] Čestmír Štuka, Patrícia Martinková, Karel Zvára, et al. "The prediction and probability for successful completion in medical study based on tests and pre-admission grades". In: *The New Educational Review* 28 (2012), pp. 138–152. URL: www.educationalrev.us.edu.pl/dok/volumes/tner_2_2012.pdf.

[4] Cyril Höschl and Jiří Kožený. "Predicting academic performance of medical students: The first three years". In: *The American journal of psychiatry* 154.6 (1997), p. 86.

[5] Jiří Anděl and Karel Zvára. "Přijímací zkouška z matematiky na MFF v roce 2004". In: *Pokroky matematiky, fyziky a astronomie* 50.2 (2005), pp. 148–161. URL: http://hdl.handle.net/10338.dmlcz/141263%0A.

[6] Jiří Kožený, Lýdie Tišanská, and Cyril Höschl. "Akademická úspěšnost na střední škole: prediktor absolvování studia medicíny". In: *Československá psychologie : časopis pro psychologickou teorii a praxi* 45.1 (2001), pp. 1–6. URL: http://www.medvik.cz/link/bmc01014269.

[7] Jana Rubešová. "Souvisí úspěšnost studia na vysoké škole se středoškolským prospěchem". In: *Pedagogická orientace; Vol 19, No 3 (2009)* (2014). URL: https://journals.muni.cz/pedor/article/view/1261.

[8] Čestmír Štuka, Patrícia Martinková, Martin Vejražka, et al. *Testování při výuce medicíny. Konstrukce a analýza testů na lékařských fakultách*. 2013.

[9] Martin Chvál, Jana Straková, and Ivana Procházková. *Hodnocení výsledků vzdělávání didaktickými testy*. Česká školní inspekce, 2015. ISBN: 978-80-905632-9-2.

[10] Petr Byčkovský and Karel Zvára. *Konstrukce a analýza testů pro přijímací řízení*. Univerzita Karlova v Praze, Pedagogická fakulta, 2007. ISBN: 9788072903313. URL: https://books.google.cz/books?id=mvvjtgAACAAJ.

[11] R Development Core Team. "R: A Language and Environment for Statistical Computing". In: *R Foundation for Statistical Computing Vienna Austria* 0 (2016), {ISBN} 3–900051–07–. ISSN: 16000706. DOI: doi.org/10.1038/sj.hdy.6800737. arXiv: /www.R-project.org. URL: http://www.r-project.org/.

[12] IBM Corp. *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corp. 2013.

[13] StataCorp. *Stata Statistical Software: Release 14*. 2015. DOI: 10.2307/2234838.

[14] SAS Institute Inc. *SAS 9.4 Language Reference: Concepts*. Cary, NC, USA: SAS Institute Inc., 2013. ISBN: 1612905641, 9781612905648.

[15] Assessment Systems Corporation. *Iteman 4.3*. Woodbury, MN: Assessment Systems Corporation. 2013.

[16] University of Nottingham. *Rogo: eAssessment Management System*. 2016.

[17] John Michael Linacre. "Rasch dichotomous model vs. one-parameter logistic model". In: *Rasch Measurement Transactions* 19.3 (2005), p. 1032.

[18] Li Cai, Dave Thissen, and Stephen Henry Charles du Toit. *IRTPRO for Windows*. Lincolnwood, IL, 2011.

[19] M. L. Wu, R. J. Adams, and M. R. Wilson. *ConQuest: Multi-Aspect Test Software*. Camberwell, 2007.

[20] Wim J van der Linden. *Handbook of Item Response Theory, Three Volume Set*. CRC Press, 2017.

[21] Patrícia Martinková, Adéla Drabinová, Ondrej Leder, et al. *ShinyItemAnalysis: Test and Item Analysis via Shiny*. 2017. URL: https://cran.r-project.org/package=ShinyItemAnalysis.

[22] Steven Downing. *Handbook of test development*. Mahwah, N.J: L. Erlbaum, 2006. ISBN: 0805852654.

[23] Mohsen Tavakol and Reg Dennick. "Post-examination analysis of objective tests". In: *Medical Teacher* 33.6 (May 2011), pp. 447–458. DOI: 10.3109/0142159x.2011.564682. URL: https://doi.org/10.3109%2F0142159x.2011.564682.

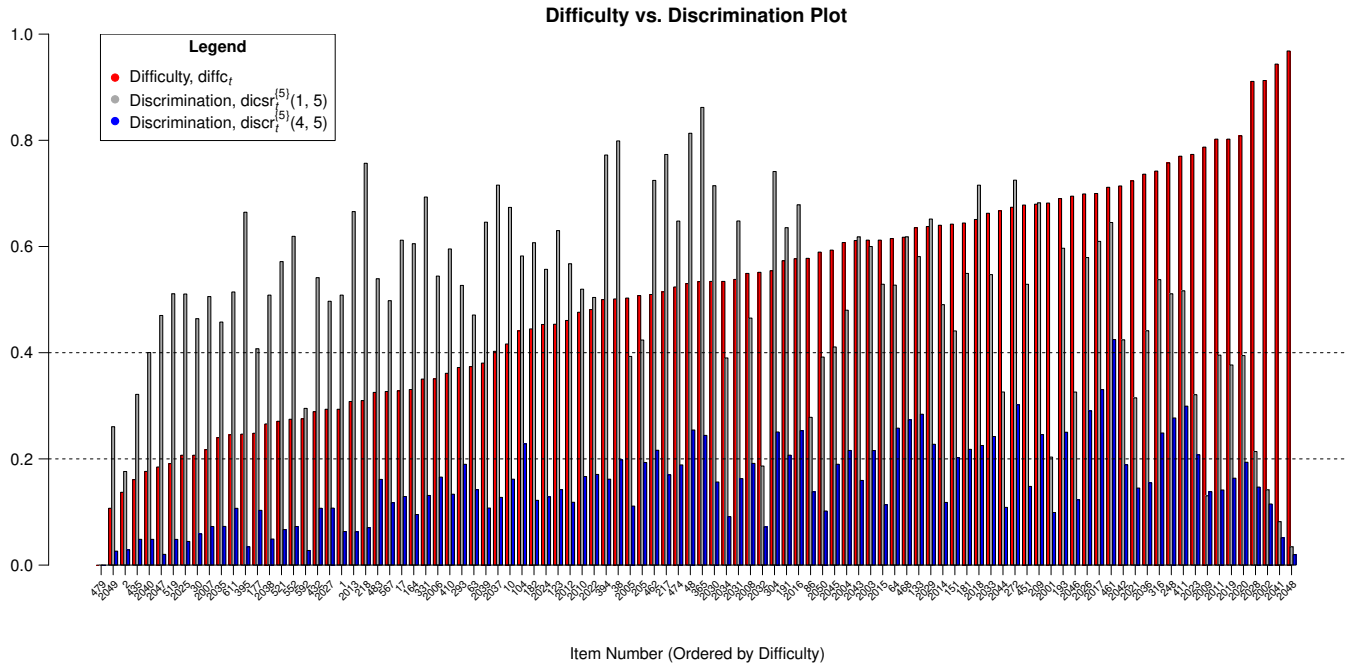

Fig. 3. Difficulty and discrimination measures of all items of medical school admission test in Chemistry. Items are ordered by their difficulty. Discrimination index is based on quintiles depicting difference in proportions of correct answers between first and fifth group ordered by total score, and between fourth and fifth group. Items with low or even negative discriminations need to be revised or discreted. Note: Item 479 was discarded due to a typo in its wording.

[24] Lee J. Cronbach. "Coefficient alpha and the internal structure of tests". In: *Psychometrika* 16.3 (Sept. 1951), pp. 297–334. DOI: 10.1007/bf02310555. URL: https://doi.org/10.1007/bf02310555.

[25] Alan Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN: 9780470463635. URL: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470463635.html.

[26] R J de Ayala. "The theory and practice of item response theory." In: (2009).

[27] Bruno Zumbo. "A handbook on the theory and methods of differential item functioning (DIF)". In: *Ottawa: National Defense Headquarters* (1999).

[28] Martinková Patrícia, Drabinová Adéla, Yuan-Ling Liaw, et al. "Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments". In: *CBE-Lifesciences Education* 16.2 (2017). Ed. by Ross Nehm, rm2. DOI: 10 . 1187 / cbe . 16 - 10 - 0307. URL: http://doi.org/10.1187/cbe.16-10-0307.

[29] Jenny L. McFarland, Rebecca M. Price, Mary Pat Wenderoth, et al. "Development and Validation of the Homeostasis Concept Inventory". In: *CBE-Lifesciences Education* 16.2 (2017). Ed. by Peggy Brickman, ar35. DOI: 10 . 1187 / cbe . 16 - 10 - 0305. URL: http://doi.org/10.1187/cbe.16-10-0305.

[30] Rob J. Hyndman and Yanan Fan. "Sample Quantiles in Statistical Packages". In: *The American Statistician* 50.4 (1996), pp. 361–365. ISSN: 00031305. URL: http://www.jstor.org/stable/2684934.

[31] Patrícia Martinková, Adéla Drabinová, and Jakub Houdek. "ShinyItemAnalysis: Analyzing admission and other educational and psychological tests". In: *Testfórum* (2017). Accepted/In press. DOI: 10 . 5817 / TF2017-9-129. URL: http://dx.doi.org/10.5817/TF2017-9-129.

[32] Lambert W. T. Schuwirth and Cees P. M. van der Vleuten. "General overview of the theories used in assessment: AMEE Guide No. 57". In: *Medical Teacher* 33.10 (Sept. 2011), pp. 783–797. DOI: 10 . 3109 / 0142159x . 2011 . 611022. URL: https://doi.org/10.3109%2F0142159x.2011.611022.