Computer-Adaptive Testing: Item Analysis and Statistics for Effective Testing

Ivana Cechova, Jiri Neubauer and Marek Sedlacik
University of Defence, Brno, Czech Republic
ivanacechova@unob.cz
jiri.neubauer@unob.cz
marek.sedlacik@unob.cz

Abstract: Every year, hundreds of secondary school students take university entrance exams, and their results determine entry into universities or possible alternatives, such as employment. In the same way, every year university teachers face the following questions: How is it possible to cope with the number of students? And how can entrance testing be as effective as possible? One possible solution is computerized testing, which creates new opportunities as well as challenges for the production and administration of test forms. This paper presents on-going research at the Faculty of Economics and Management of the University of Defence in Brno on test item analysis as well as students' success during entrance exams. Item analysis is a procedure to increase the reliability and validity of a test by separately evaluating each test item to determine whether or not that item discriminates in the same way that the overall test is intended to discriminate. There are many ways to conduct item analysis (Tuckman 1972, Michael 1981, Nelson 2008) and ITEMAN is a useful tool to analyse objective data. For this reason, the authors of this article focused on test item selection via the ITEMAN test/scale analysis program, which provides item statistics, test/scale statistics, frequency distribution with a histogram, and item response analysis. The authors' ultimate goal is to find out how successful the students are in their entrance tests, which consist of the Learning Potential Test and the English Language Test. Careful item analysis and entrance test composition, together with entrance exam analysis, are factors that might predict academic achievement in tertiary education.

Keywords: computer-adaptive testing (CAT), test assessment, item construction, item analysis, ITEMAN

1. Introduction

Computer technologies have opened up new possibilities not only for optimizing the administration of tests, but also – and especially – for test development and assessment. Computer Assisted Testing (CAT) allows for a redesign of psychological and educational tests for effective and efficient administration by interactive computers; its objective is to select, for each examinee, a set of test questions that measures that person on the given trait effectively and efficiently (Van der Linden and Glas 2000). According to Thompson and Weiss, CAT is a sophisticated method of delivering examinations, and has nearly 40 years of technical research supporting it (Thompson and Weiss 2011).

CAT has many positive aspects that can be used to improve the assessment process as well as to overcome many of the prevailing problems in the field of traditional testing. Many advantages of CAT have been indicated in recent research (e.g. Johnson and Weiss 1981, Cudeck 1985, Koch et al. 1990, Rudner 1990, Kingsbury and Houser 1999, Weiss 2004, Vanova et al. 2012). Roever states that the biggest logistical advantage of a CAT test is its flexibility in time and space (Roever 2001). In general, CAT greatly increases the flexibility of test management. Some other benefits include the following:

- CAT reduces the number of excessively easy or difficult items;
- CAT reduces item exposure and subsequent security risks;
- CAT provides a valid and reliable measurement of students' competence.
- Tests are given "on demand" and scores are available immediately
- Tests are individually paced so that an examinee does not have to wait for others to finish before going on to the next section. Self-paced administration also offers extra time for examinees who need it, potentially reducing one source of test anxiety;
- Tests are automatically tailored to the proficiency level of the individual examinees;
- CAT offers a number of options for timing and formatting, and therefore has the potential to accommodate a wider range of item types (Vanova et al. 2012, Rudner 2014).

Despite the advantages listed above, CAT has several limitations, and can raise some technical problems, such as:

- CAT requires a facility with a large number of computers and the examinees must be computer-literate.

- CAT is not appropriate for all subjects and skills.
- Test administration procedures are different and this could cause problems for some examinees.

Overall, however, CAT represents a significant improvement over traditional assessment methods used at many schools and universities by providing more accurate scores for all students, a more detailed picture of where students excel or need additional support for teachers, and more accurate ways to evaluate students' achievement. CAT and computerized assessment allow both teachers and students to get test results immediately, and faster results mean that teachers can use the information from optional interim assessments throughout the school year to modify instruction and better meet the unique needs of their students.

2. Entrance exams at the faculty of economics and management

Predictions of academic success have been a contentious issue in educational research for a long time. This is one reason why entrance exams have been so important for all universities and the University of Defence (UoD) is not an exception. The University of Defence (UoD) provides education of both Czech Army specialists and civilian students within accredited bachelor, master and doctoral study programmes. All these programmes have two basic forms of study – full-time and combined, which are authorized by Act no. 111/1998 Coll., on universities § 44. The UoD strives to accommodate the interests of both military and civilian study candidates, who wish to complement their existing education in accordance with the rising demands of qualifications and respond to the change of professional orientation or the needs of requalification. In this, the UoD is reflecting European and worldwide life-long learning trends. Students of both forms have to fulfil the same requirements, although students of the combined form have to combine their study with a regular job and everyday duties.

The Faculty of Economics and Management provides Bachelor's degree programmes, subsequent Master's degree programmes, continuous Master's degree programmes, and PhD degree programmes. Applicants for military full-time study programmes take the following tests in their entrance exam: the Learning Potential Test (a written test); the English Language Test (also written); and a Physical Fitness Test. Applicants for the military part-time study programme, as well as all civilian applicants (both full-time and part-time) take only the Learning Potential Test (LPT). The LPT is divided into three parts; each part always contains ten questions. The first part deals with numeric thinking and logic; the second part focuses on spatial imagination and abstract thinking; and the last part concentrates on basic mathematical skills. The Learning Potential Test result is assessed on a scale between 0 and 60 points; a passing grade for this test is 30 (Sedlacik et al. 2013).

All applicants take the English language test (ELT), which examines reading comprehension, vocabulary, and grammar. The minimal entrance level should be at least A2 according to the Common European Framework of Reference for Languages, or SLP 1 (Standardized Language Profile) according to NATO STANAG 6001. The ELT is assessed on a scale between 0 and 50 points; a passing grade for this test is 25. With regard to this article, the authors will not describe Learning Potential Test and the physical fitness test, and will concentrate only on ELT analysis. Detailed information regarding the entrance exam is available at the University of Defence web site.

3. Test development, item construction and analysis

A proper test is a collection of well-combined elements. First, the construction of test items is a crucial step for the validity of a test. A good item construction process enhances the discrimination power, score variance, reliability, and evidence of validity for the intended interpretation and use of scores from the overall test (Suen and McClellan 2003). According to Sikolova et al. (2009), text and task authenticity, attractiveness and balance of distracters, length of texts, and relevance of topics in terms of the examinees' age, education, and common interests are other aspects that must be considered during test development. The ability to construct high-quality test items requires knowledge of the principles and techniques of test construction and skills in their application. Crocker states that a process of test development must include the following steps:

- Primary objective for which the test scores will be used.
- Initial test items' development, a pool of test items' construction.
- Moderation and revision of test items.
- Analysis of test items (ITEMAN, exploratory statistics).
- Test piloting.
- Reliability and validity studies.

- Guidelines for administration and scoring (Crocker and Algina 2014).

The decisions that have to be made at the beginning of the process of the development of the ELT at the UoD are strongly influenced by several external factors: these include the level of test difficulty, time allotment, and testing techniques. According to UoD regulations, the ELT must be administered within 60 minutes. Even though the requirements of the proficiency level clearly indicate the need to include all four macro-skills in the test (i.e. listening, speaking, reading, and writing), time constraints do not allow it. As a result, tests of speaking, writing, and listening, although of high importance, had to be excluded from the test (Sikolova et al. 2009).

3.1 Pre-testing

Item construction and moderation is followed by test piloting/pre-testing on a limited number of students. Since the sample population cannot be chosen from secondary school students (they might become potential applicants for the entrance exams at the UoD), a decision was made to test first-year students at the Faculty of Economics and Management, whose proficiency level in English is still at the required entrance level (Stanag SLP 1+, or B1 according to the Common European Framework of Reference for Languages).

After the questions were written and reviewed, many were pretested with a sample group similar to the population to be tested. The results enable test developers to determine:

- The difficulty of each question
- If questions are ambiguous or misleading
- If questions should be revised or eliminated
- If incorrect alternative answers should be revised or replaced.

The authors pre-tested 63 students and then analysed each item via the ITEMAN:

3.2 Item analysis

The next step in test construction is item analysis, which is a process of examining class-wide performance on individual test items. Item analysis can be defined as a process which examines student responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole ("Understanding Item Analysis Reports" 2005).

Each item on a test should be of appropriate difficulty – it should not be too difficult or too easy, as this influences the discrimination power of a test. Additionally, it must be fair to all students. However, it is impossible to decide which item is fairer, as there is no generally accepted definition of "fairness" with respect to testing (Cole and Zieky, 2001).To meet the goals of reliability and validity, testers and teachers must carefully inspect each individual test item, the test as a whole and any descriptive or preparatory materials to ensure that language, symbols, words, phrases and content generally regarded as sexist, racist or otherwise inappropriate or offensive to any subgroup of the test-taking population are eliminated.

There are three common types of item analysis which provide teachers with three different types of information: the difficulty index, the discrimination index, and the analysis of response options. The ITEMAN software program is designed to provide detailed item and test analysis reports using Classical Test Theory (CTT). The program analyses item response data and provides conventional item analysis statistics for each item in order to assist in determining the extent to which items contribute to the reliability of a test and which response alternatives are functioning well for each item (User's manual for ITEMAN).

4. ITEMAN analysis

The analysed test consists of 50 items and was answered by 63 students. As can be seen in Figure 1 describing the frequency distribution of the total test score, the minimum score obtained was 25 points and the maximum was 48. The mean score value was 37.619 with the standard deviation 4.661. The value of the median was 38.

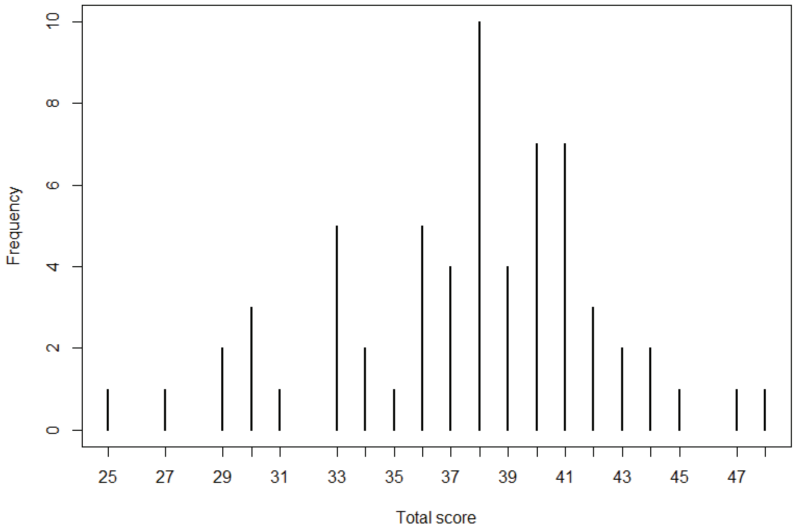


Figure 1: Frequency distribution of total test score

The ITEMAN software offers a classical item and test analysis. We will briefly describe its basic outputs, namely the **difficulty index**, the **index of endorsement**, the **discrimination index** and the **point biserial correlation coefficient** (Kerlinger and Lee 2000). In tests where answers are evaluated as either being correct or incorrect, the difficulty level can be described as follows

$$\text{Item Difficulty Index} = \frac{\text{number of people answering item correctly}}{\text{total number of people taking test}}.$$

The index of endorsement defined by the formula

$$\text{Index of Endorsement} = \frac{\text{number of people selecting the answer}}{\text{total number of people taking test}}$$

is the proportion of people selecting a particular answer (in our case answer ‘a’, ‘b’, ‘c’, or ‘d’). Most test creators agree that the best test items in terms of difficulty are those with values between 0.5 and 0.7.

After difficulty and endorsement, the next index for the item analysis is the item discrimination index. This statistic evaluates how effectively the item was able to discriminate between high scores and low scores. It is necessary to determine the high and low scoring group. The total scores are used to do this. The item discrimination index is the difference between the proportion of people in the high scoring group who answered the item correctly and the proportion of people in the low scoring group who answered this item correctly. We can write

$$\text{Discrimination Index} = \frac{N_H}{\text{number of people in high group}} - \frac{N_L}{\text{number of people in low group}},$$

where N_H is the number of people in the high scoring group that got the item correct and N_L is the number of people in the low scoring group that got the item correct. If the index is negative, the item has reverse discrimination. Good items are expected to have positive values. The higher the value is, the greater is the discrimination. For the analysed test, the low scoring group has 21 people whose score was less or equal to 36 points, the high scoring group consists of 24 people with the score greater or equal to 40 points.

The point biserial correlation coefficient (item-to-total score correlation) describes how the answers of the given item correspond to the total score obtained. One can expect that the correlation of each item should be high. An item that correlates low with the total score can be interpreted as an item that is measuring something that differs from what the other items are measuring; the item is not homogenous with other items.

The following table summarises the item analysis results of the test.

Table 1: Item analysis results

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Difficulty	0.84	0.54	0.84	0.92	0.92	0.71	0.78	0.92	0.57	0.56	0.65	0.17	0.73	0.25	0.92	0.71	0.94
Discrimination	0.15	0.15	0.29	0.10	0.19	0.17	0.11	0.14	0.27	0.14	0.45	-0.02	0.48	0.14	0.05	0.21	0.14
Correlation	0.24	0.24	0.33	0.16	0.16	0.29	0.15	0.20	0.34	0.26	0.45	0.03	0.48	0.17	0.04	0.12	0.17
Item	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Difficulty	0.57	0.84	0.86	0.71	0.86	0.60	0.95	0.95	0.81	0.73	0.84	0.79	0.89	0.87	0.95	0.84	0.94
Discrimination	0.23	0.25	0.02	0.49	0.34	0.54	0.14	0.01	0.03	0.08	0.38	0.30	0.11	0.15	-0.08	0.29	0.05
Correlation	0.22	0.23	0.06	0.47	0.58	0.48	0.22	0.00	0.04	0.17	0.50	0.43	0.19	0.22	-0.15	0.37	0.02
Item	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	
Difficulty	0.60	0.86	0.92	0.84	0.92	0.89	0.84	0.35	0.44	0.65	0.71	0.90	0.67	0.71	0.57	0.73	
Discrimination	0.32	0.29	0.19	0.24	0.15	0.20	0.15	0.27	0.16	0.09	0.13	0.15	0.18	0.36	0.15	0.08	
Correlation	0.24	0.44	0.30	0.15	0.35	0.35	0.11	0.30	0.13	0.07	0.16	0.30	0.13	0.32	0.23	0.23	

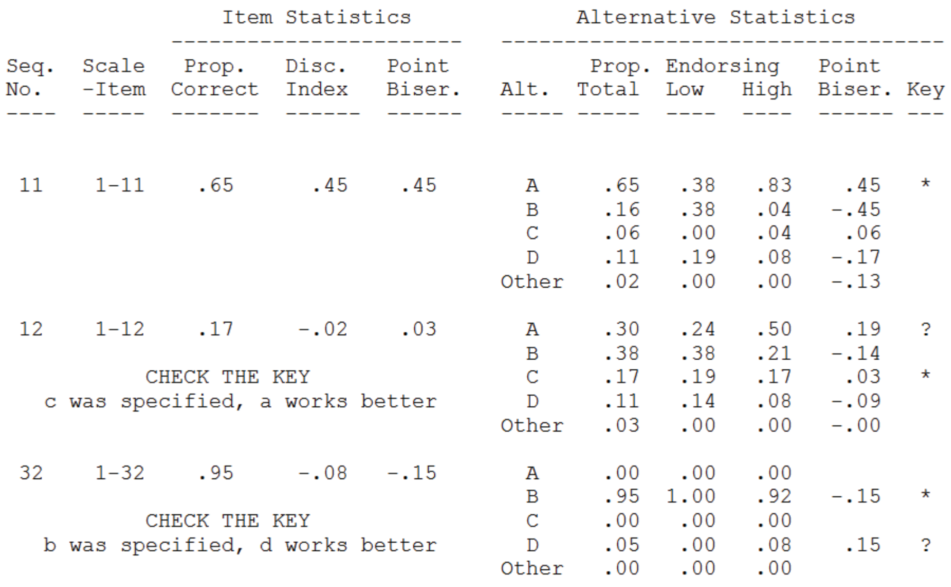
Reliability of the test is often estimated by Cronbach alpha

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} var(I_j)}{var(I_c)}\right),$$

where k denotes number items var(I_j) are the variances calculated for each item(Ij) and var(I_c) is the variance of the total score of the test. The value of these statistics is usually expected to be larger than 0.7. The estimated reliability of the analysed test is 0.659 which is slightly below this limit.

According to the result obtained (see Table 1), we can identify several items which can be considered very easy (the difficulty index is greater than 0.9 meaning that more than 90% of students answered this item correctly), namely items 4, 5, 8, 15, 17, 24, 32, 34, 37 and 39. Items 12 and 14 have a difficulty index of less than 0.3, which indicates that those items are too difficult. The discrimination index for such items is usually very small; these items do not help to discriminate between good and bad results. If we focus on the items 20, 26 and 44, we will see that the difficulty indices have acceptable values but the discrimination indices and correlation coefficients are very low. The responses to these items do not correspond to the total score obtained.

Figure 2 shows the output from ITEMAN software (version 3.6). It contains, beside other information, the values of the difficulty index (‘Prop. Correct’), the discrimination index (‘Disc. Index’), the point biserial correlation coefficient (‘Point. Biser.’) and the index of endorsement for the low and high group (‘Endorsing Low’ and ‘Endorsing High’). Three items of the test were selected as examples of ‘good’ and ‘bad’ questions. Item no. 11 represents properly a constructed question. Item no. 12 is an example of a question which is too difficult. It has a negative discrimination index and almost zero correlation coefficient. The third item, no. 32, represents a very easy question with a negative discrimination index and correlation coefficient.



```
                  Item Statistics          Alternative Statistics
              -----------------------   -----------------------------------
Seq.  Scale   Prop.    Disc.   Point           Prop. Endorsing     Point
No.   -Item   Correct  Index   Biser.   Alt.   Total  Low   High   Biser. Key
----  -----   -------  ------  ------   -----  -----  ----  ----   ------ ---


 11    1-11     .65      .45     .45      A     .65   .38   .83     .45    *
                                          B     .16   .38   .04    -.45
                                          C     .06   .00   .04     .06
                                          D     .11   .19   .08    -.17
                                        Other   .02   .00   .00    -.13

 12    1-12     .17     -.02     .03      A     .30   .24   .50     .19    ?
                                          B     .38   .38   .21    -.14
              CHECK THE KEY               C     .17   .19   .17     .03    *
   c was specified, a works better        D     .11   .14   .08    -.09
                                        Other   .03   .00   .00    -.00

 32    1-32     .95     -.08    -.15      A     .00   .00   .00
                                          B     .95  1.00   .92    -.15    *
              CHECK THE KEY               C     .00   .00   .00
   b was specified, d works better        D     .05   .00   .08     .15    ?
                                        Other   .00   .00   .00
```

Figure 2: Example of ITEMAN output (version 3.6)

Based on the item analysis (values of the difficulty and discrimination index, the correlation coefficient), it would be appropriate to modify some items in the proposed test. Such modification may cause an increase in test reliability.

5. Conclusion

Today the use of computer technology in the field of language assessment and testing has become so widespread and so inclusive that it is regarded as an inseparable part of today's education system. The item analysis shows the teachers and testers what further steps must be done to ensure the entrance test's reliability and validity. It gives them useful information dealing with item appropriateness. If an item analysis shows that an item does not work properly (Figure 2), it must be replaced or its distracters must be changed.

The ITEMAN item analysis provides teachers and testers with a user friendly tool to analyse a test. If all above mentioned steps (Chapter 3) are followed, a test is reliable and fair enough to test what it claims to be tested. CAT as well as ITEMAN are viable technologies with good potential to provide improved test analysis and measurement.

In conclusion, a testing process that provides a fair, reliable, efficient, and cost-effective assessment of applicants remains the main objective for the University of Defence educational programs. The established system of test design, moderation, pre-testing, item analysis, and administration has become a viable and transparent way of selecting candidates to the university according to their knowledge level.

References

Cole, N. S. and Zieky, M. J. (2001) "The new faces of fairness." *Journal of Educational Measurement*, Vol. 38, no 4, pp. 369-382.

Crocker, L. and Algina, J. (2014) Introduction to classical and modern test theory. [online], University of Florida, http://myweb.facstaff.wwu.edu/~graham7/crocker.pdf, accessed 19 May 2014.

Cudeck, R. (1985) "A structural comparison of conventional and adaptive versions of ASVAB." *Multivariate Behavioral Research*, Vol. 20, No. 3, pp. 305-322.

Kerlinger, F. N. and Lee, H. B. (2000) *Foundations of Behavioral Research*. 4th ed. Belmont: Cengage Learning.

Johnson, M. F. and Weiss, D. J. (1981) "Effects of immediate feedback and pacing of item presentations to testing." Research Report 81-2. Minneapolis: University of Minnesota.

Kingsbury, G. G. and Houser, R. L. (1999) "Developing computerized adaptive tests for school children." In: F. Drasgow and J. B. Olson-Buchanan (eds.), *Innovations in Computerized Assessment*, Malwah, NJ: Erlbaum, pp. 93-115.

Koch, W. R., Dodd, B. G. and Fitzpatrick, S. J. (1990). "Computerized adaptive measurement of attitudes." *Measurement and Evaluation in Counseling and Development*, Vol. 23, pp. 20-30.

Michael, A. (1981) "Multiple choice tests: Analysis", *Nutrition & Food Science*, Vol. 81, No. 6, p. 18.

Moore, E. L., Galindo, J. l. and Dodd, B. G. (2012) “Balancing flexible constraints and measurement precision in computerized adaptive testing.” *Educational and Psychological Measurement*, Vol. 72, No. 4, pp. 629-648.

Hampel, D., Myskova, K. (2013) “Software GRETL as a support of the Econometrics 2 course at FBE Mendelu.” *Efficiency and Responsibility in Education*. Prague: CULS, pp. 166-173.

Neubauer, J. (2013) “Comparison of Students’ Attitude to the Study Statistics at the Faculty of Economics and Management. *Ekonomika a Management [Economics and management]*, vol. 2013, no. 4, pp. 51-57.

Nelson, L. (2008) *ITEMAN and Lertap 5. Curtin University of Technology*. http://www.lertap.curtin.edu.au, accessed 17 May 2014.

Roever, C. (2001) “Web-based language testing.” *Web-based Language Learning and Technology.* Vol. 5, No. 2. pp. 84-94.

Rudner, L. (1990). “Computer Testing: Research Needs Based on Practice.” *Educational Measurement: Issues and Practice*, Vol. 2, pp.19-21.

Rudner, L**.** (2014) *An On-line, Interactive, Computer Adaptive Testing Tutorial*. http://echo.edres.org:8080/scripts/cat/catdemo.htm, accessed 22 May 2014.

Sedlacik, M., Cechova, I. and Doudova, L. (2013) “Be born as successful mathematics or language learner: myths, true or false?” *Journal on Efficiency and Responsibility in Education and Science*, 2013, Vol. 6, No. 3, pp. 155-166.

Salvatori, P. (2001). “Reliability and validity of admissions tools used to select students for the health professions.” *Advances in Health Sciences Education*, Vol. 6, pp.159-175.

Sikolova, M., Slozilova E. and Svoboda, P. (2009) “Developing an Entrance English Test at the University of Defence.” In: *Nové trendy vo vyucovaní anglickeho jazyka*. Ekonomicka univerzita Bratislava: Ekonomicka univerzita.

Suen, H. K. and McClellan, S. (2003) “Test Item Construction Technique and Principles.” In: *Encyclopedia of Vocational and Technological Education (Vol. 1)*. pp.777-798. Taipei: ROC Ministry of Education.

Thompson, N. A. and Weiss, D. A. (2011) “A Framework for the Development of Computerized Adaptive Tests.” *Practical Assessment, Research & Evaluation*, Vol. 16, No. 1, pp. 1-9**.** http://pareonline.net/getvn.asp?v=16&n=1, accessed 19 May 2014.

Tuckman, B. W. (1972) *Conducting Educational Research*. San Diego, CA: Harcourt Brace Jovanovich, Inc.

“Understanding Item Analysis Reports” (n. d.) http://www.washington.edu/oea/services/scanning_scoring/scoring/item_analysis.html

Van der Linden, W. J. and Glas, C. A. W. (2000).*Computerized Adaptive Testing: Theory and Practice.* Boston: Kluwer.

Vanova, T., Prochazka, J. and Denglerova, D. (2012). *Adaptivni test COMPACT.* Masaryk University, Brno.

Weiss D. (2004) “Computerized Adaptive Testing for Effective and Efficient Measurement in Counselling and Education.” *Measurement and Evaluation in Counseling and Development*, Vol. 37, No. 2 (July). pp. 70-84.