MODESTUM
OPEN ACCESS
EURASIA Journal of Mathematics, Science and Technology Education
ISSN: 1305-8223 (online) 1305-8215 (print)
2017 13(11):7421-7429 DOI: 10.12973/ejmste/79796

Construction of a Mathematical Model for Calibrating Test Task Parameters and the Knowledge Level Scale of University Students by Means of Testing

Duisebek Nurgabyl [1*], Gulmira Kalzhanova [1], Nurzhan Ualiyev [1], Gulsim Abdoldinova [2]

[1] Zhetysu State University named after I. Zhansugurov, KAZAKHSTAN
[2] Kazakh University of Technology and Business, KAZAKHSTAN

Received 13 May 2017 ▪ Revised 13 August 2017 ▪ Accepted 29 September 2017

ABSTRACT

The relevance of this study is determined by the algorithm developed for test task selection. The purpose of this article is to develop this test task selection algorithm having single and multiple-choice answers with various blocks of reactions. The main approach to the study of the problem is the construction of a mathematical model to calibrate test task parameters and the university students' scale of knowledge level. The study has proven insufficient use of the classical theory of testing in an objective assessment of students' knowledge. The method for calibrating test task parameters is developed. Scales of testees' readiness level are defined. The adequate size of test reliability coefficient is found. The rule for test task selection is formulated. A formula for the complexity degree of a test task is found. An algorithm for allocation of test task types with unambiguous and multiple-choice answers with various blocks of reactions is offered.

Keywords: test tasks, knowledge assessment, higher education, single choice answer, multiple choices

INTRODUCTION

In the modern education system, the test method is an increasingly recognized method of assessing the quality of education, which is widely used in many countries. Testing as a new method of knowledge assessment is contrasted to the traditional methods presented by oral and written forms of knowledge testing (Voutilainen et al., 2017). In recent years, testing is widely used in various institutions, government agencies and by individuals to determine professional suitability and training (Senior et al., 2017; Coe, 2003; Botti et al., 2017).

However, for an objective assessment of students' readiness, the testing model used nowadays involving only the classical theory of tests is not sufficient, as the level of the acquired knowledge and the complexity level of the test task have different metrics (Wilmot et al., 2011, Kibzun & Inozemtsev, 2014). In this regard, there is a need for a testing algorithm that would objectively define an assessment of readiness level of the testees, confirmed by various tests having a predetermined level of task difficulty (Edens & Shields, 2015). In addition, it is necessary to apply different-level test tasks with unambiguous and multiple-choice answers to determine the level of professional competences.

All these factors manifest the relevance of studying the issues of test task design and selection. The relevance of this research is determined by the algorithm developed for test selection, which along with an adaptive testing algorithm would make it possible to assess the level of students' knowledge objectively, as well as quickly and effectively eliminate gaps in the acquisition of learning material by a particular student.

The article aims at developing a test type selection algorithm with single and multiple-choice answers and various blocks of reactions.

✉ duisebek.nurgabyl@yandex.ru **(*Correspondence)** ✉ kalzhanova.gulmira@yandex.ru
✉ ualiyev2018@yandex.ru ✉ abdoldinova@rambler.ru

Contribution of this paper to the literature

- Use of a model for designing a level-rated test-base of tasks applying the rating system of knowledge assessment can provide objective evaluation of the level of students' knowledge and skills.
- Introduction of the test task selection algorithm, along with the algorithm of adaptive testing, will quickly and effectively eliminate gaps in mastering the educational material for a particular student.
- The proposed algorithm for constructing a level-rated test-base of tasks using the rating system of knowledge assessment is a new approach in the theory of tests.

LITERATURE REVIEW

In the first three decades of the last century, the knowledge assessment tests were created in Germany, but these tests, owing to research, have become more widespread and more intensively developed in a decentralized, pragmatically oriented and rapidly growing education system in the United States (Avanessov, 2014). In the US, testing is conducted by several companies, the largest of which is the Educational Testing Service (ETS), which has existed since 1947. ETS annually holds more than 12 million tests in 180 countries of the world, and has more than 60 patents for various testing devices and technologies (Howard et al., 2017).

A new educational paradigm and initiative solutions for modernizing education to achieve high quality of education in the Republic of Kazakhstan stipulate the introduction of an e-Learning system in the State Program for the Development of Education in the Republic of Kazakhstan for 2011-2020. The emergence of e-Learning was caused by the desire to improve the effectiveness of pedagogical measurements, which was associated usually with a decrease in the number of test tasks, time and cost of testing, and also with an increase in the accuracy of students' assessments (De Meo et al., 2017).

However, an e-Learning environment is limited by the inability of online instructors to immediately monitor whether students remain focused during online autonomous learning (Chen, Wang & Yu, 2017).

The training mode is one of the basic operation modes of electronic education system; it provides preparation for exams and tests, and enables to disseminate teaching training on exam materials. It ensures the ability to display correct answers and comments, both before choosing a response and after it, and to perform self-training and knowledge verification functions. There is a possibility of choosing sections for studying, available to all students and to everyone individually (Hung Jui-long, 2012; Guznenkov & Seregin, 2016; Vasiliev, 2007; Vlasin & Chirila, 2014). Borzykh & Gorbunov (2009) analyze a large number of research papers, give a useful idea of what e-Learning is, and also emphasize the importance of computer testing. In computer testing, the usual classical tests are used, where the complexity levels of a test task and the preparedness level of students are not taken into consideration (de Villiers & Becker, 2017; Nurjanah et al., 2017; Martos-Garcia et al., 2017).

Under usual testing, a score or so-called primary cumulative score is the main preparedness criterion of testees. A characteristic feature of usual testing is its simplicity and operational information. Indeed, the more test questions are answered correctly, the higher the score is. However, the primary score is only a comparative assessment. It essentially depends on guessing and random selection of test tasks, and may be different for other types of testing (Haist et al., 2017). The possibility of guessing is the main reason for teachers' negative attitude to the closed form of test tasks. To eliminate this 'white' spot, a correction formula for guessing was proposed for the test scores (Horton & Horton, 2005), the essence of which is that the number of scores for answers that can be guessed is subtracted from the total scoring obtained by each student in accordance with the provisions of the probability theory. This formula can be used in a test with closed-type tasks only in the case of the same number of ready-made answers and has the form:

$$X_i^1 = R_i - W_i/(k-1)$$

where X_i^1 is the testee's score corrected for guessing in the test; R_i and W_i are the number correct and wrong answers, respectively, obtained by the testee in the test; k is the number of ready answers in the test tasks.

S. N. Dvoryatkina (2013) considered the intellectual technology of teaching mathematics on the basis of ICT, through which corrected self-learning of students was realized. J. Park (2010) introduced a newly developed computerized system of constructive multiple choice. The system unites short answers (SA) and multiple-choice formats (MC), asking the testees to answer the same question twice, first in the SA format, and then in the MC format.

The authors proposed an interactive three-leveled algorithm for adaptive computer-based testing, which offers individual trajectory of receiving test tasks depending on the results of already completed tasks by a student (Nurgabyl, 2012; Nurgabyl & Ramazanov, 2013). The individualization of the procedure for selecting test tasks forms the basis of adaptive testing and provides generation of effective tests by optimizing the difficulty of assignments with regard to the readiness level of trainees (Linden & Gess, 2010; Ozturk & Dogan, 2015; van Rijn & Ali, 2017; Cheng et al., 2017). As a result of optimal transition from one level to the next level, the algorithm allows the student to get the corresponding objective test score from the score-rating system of knowledge assessment. For ensuring transition from one level to the next, it is necessary to create test tasks at different levels.

With this background, it follows that e-Learning, adaptive computer testing and objective assessment of students' knowledge directly depend on the quality of test task construction.

While examining the quality of test tasks and the test in general, it is necessary to evaluate each component of the test task structure not only separately, but also in the system of relations with other test tasks. At the same time, it is required to consider that each test task has a number of structural elements, each of which is characterized by a number of internal and external properties. Each element of the test task possesses its own topology, and the properties of the test task and the test in general can be defined by the features of a great number of testees, and the indicators of the test task and test quality. In this regard, the test tasks intended for knowledge control and measurement usually undergo strict objective expert review (Permyakov & Maksimov, 2006). Moreover, practice of global testing demonstrates that the high quality of tasks cannot be guaranteed even by the most experienced compilers (Yaman, 2011).

In this connection, questions arise concerning the selection of poor-quality test tasks and the calibration of test task parameters.

METHODS AND DATA

The modern theory of testing – Item Response Theory (IRT), which is part of a more general theory of latent-structural analysis, has been developing for several decades in many countries. Separately, it is necessary to specify the theory of G. Rasch (1980), which is sometimes called the IRT one-parameter (theory). Rasch analysis assesses the difficulty level of a task, regardless of the readiness level of testees, gives the chance of equalizing of the testees' scores obtained in parallel variants of tasks that measure the same property of interest.

A systematic approach to Rasch method has enabled to define the concept of drawing up the pedagogical test in a new way as well. The pedagogical test represents a set of interconnected tasks of increasing difficulty, allowing a qualitative assessment of the structure and measuring the level of knowledge. In this regard, it is necessary to mention the methods of M. Wilson (2005), F. Siddiq, P. Gochyyev, and M. Wilson (2017), M. Pantziara and G. Philippou (2012), C. J. Sangwin and I. Jones (2017), and V. Avanessov (2014).

Although advances have improved our ability to describe the precision of test measurement, it often remains challenging to summarize how well a test is performed in general. Reliability, for example, is provided by an overall summary of measurement precision, but it is sample-specific and might not reflect the potential **usefulness** of the test if the sample suits poorly for the test purposes (Markon, 2013; Huang & Hung, 2010).

To assess the quality of the test and test tasks, statistical processing methods of testing knowledge results are mainly used. In particular, correlation, regression and factor analysis methods are applied to reject an inefficient test. The first two methods allow estimating the so-called statistical pure contribution of each task to the general variation of test scores, while the factor analysis is a good method for checking the homogeneity of the test (Bortz & Döring, 2005; Avanessov, 2009; Prado et al., 2010; Lim & Chapman, 2013). As a result, these methods indicate suitability, or unfitness of the considered test (Boesen & Palm, 2010; Xia, Liang & Wu, 2017).

However, these and other studies of domestic and foreign authors do not consider the development of an algorithm for calibrating test tasks and determining the scale of testees' knowledge level.

A multistep adaptive testing method can be used to increase the objectivity of assessing the results of test tasks (Nurgabyl, 2014), where the next step is made only after evaluating the results of the previous step. The task selection and presentation algorithm is built on the principle of feedback, when after a correct answer of the testee, the next task is chosen to be more difficult; whereas an incorrect answer implies the presentation of an easier subsequent task than that which had been answered incorrectly.

Our task is to develop an algorithm for calibration of test task parameters and the scale of university students' knowledge level, which, along with the adaptive testing algorithm, would enable to objectively determine the students' knowledge level, and would also eliminate quickly and effectively gaps in acquisition of the educational material by this particular student.

The main approach to studying the problem is the proposed method by the authors for calibration of test task parameters and scales of university students' knowledge level. With the help of this method, the complexity degree of test tasks is determined, the test tasks are selected according to their complexity degree, the level-rated classes of test assignments are divided into single and multiple-choice answers with various blocks of reactions, and the preparedness level of testees is determined. This method allows for effective use of the algorithm for multi-step adaptive testing. The proposed procedure makes it possible to objectively determine the level of students' knowledge, and quickly and effectively eliminate the gaps in mastery of the teaching material of a particular student. The reliability value of all test tasks calculated by the Spearman-Brown formula confirms the promising ability of the proposed algorithm.

RESULTS

Design of Tests

In many countries, and Kazakhstan in particular, the universities mainly use formalized test tasks, which are not objective for determination of students' academic performance.

To determine the level of professional competency, it is necessary to use a variety of test tasks with single and multiple-choice answers. For example, theoretical knowledge can be measured by test tasks with multiple choice answers, whereas practical skills should be assessed with single-choice answer tests (Napankangas, Harila & Lahti, 2012; Nurgabyl, 2014). To design test tasks, it is necessary to determine the training goals primarily, and consequently, the corresponding types of testing.

Studies have shown that three types of testing should be differentiated:

1) Test that determines a student's thematic module knowledge during the final control, i.e., generally at the end of a term (systemic method).
2) Test that determines a student's knowledge concerning the main parts of a thematic module during mid-term control, and as a result, offers information about whether the training objectives have been achieved to a sufficient extent (criteria-oriented method).
3) Test that determines a student's knowledge concerning a training element, diagnosing probable difficulties with the training element (diagnostic method).

If the training objective is to form systemic knowledge of a discipline or a thematic module, the test-oriented systemic method should be applied. If the training objective is to obtain knowledge of various concepts, statements, algorithms, problem solving methods with a common theoretical basis, then the criteria-oriented test should be applied. If the training objective is to acquire knowledge of fundamental concepts, statements or methods for solving complex problems, then testing aimed at the diagnosis should be applied. When designing test tasks, authors must first find out what training elements (concepts, statements, methods) are included in each test task and if they coincide with the educational goals, and so on. Secondly, the complexity level of the task depends on the number of correct and incorrect answers in a question and the logic of answer selection.

For this reason, the experts of these disciplines first subdivide the task test-base into different complexity levels: easy, medium, hard. It is recommended to use questions with single choice answers for an easy level. It is recommended to use multiple choice questions with weight added on the basis of logic "OR", multiple choice questions with weight added according to logic "AND" and single choice questions for a medium level. Multiple choice questions with weight added on the basis of logic "OR" and logic "AND" should be used for a high complexity level (Nurgabyl, 2014).

The logic "AND" is a rule according to which students are assigned a maximum number of scores for the answer on condition that they have chosen all correct answers, and have not chosen the incorrect answers. The logic "OR" is a rule according to which the students are assigned scores on condition that they have chosen at least one correct answer and scores are deducted if an incorrect answer has been chosen.

The Boolean model is used to assess the validity of the answers to the questions of easy level. At the same time, the answer validity is expressed by a two-valued logic, and can be set to "true" or "false": 1 – if the answer is correct; 0 – if the answer is incorrect. The verity of answers to single choice questions of medium level can take the following values: 2 – if the answer is correct; 0, -1 – if the answer is incorrect.

According to the 'OR' logic, the verity of the answer for multiple choice questions is determined by the weight of the question. In this case, the weights of correct answers for the 2nd complexity level can be equal to: 1; 2, the weights of incorrect answers can be equal to: -1; 0. Finally, the weights of correct answers for the third complexity level are equal to: 3; 2; 1, the weight of incorrect answers can be equal to: -2; -1; 0.

According to the "AND" logic, the verity of the answer for multiple choice questions is determined by the correct answer weight. At the same time, the weights of correct answers to single choice questions of medium level are equal to 2, the weights of incorrect answers can be equal to -1; 0, and the weights of correct answers for the third complexity level are equal to 3; the weights of incorrect answers can be equal to -2; -1; 0.

Hence, while the author is designing the test tasks, s/he must first determine the complexity degree of each task. However, the studies have shown that test compilers, regardless of the actual difficulty of tasks, tend to underestimate their complexity. On average, teachers determine the complexity degree of a test task correctly only by 20% (Avanessov, 2009). For this reason, let us define the complexity degree of a test task by means of the algorithm developed by the authors.

Parameters of Test Task Calibration

First, we will define the quantitative data corresponding to the complexity levels of the tasks. The complexity order of the test tasks is determined by the percentage of testees who obtained the correct result. Let r be the number of testees who completed the test task, r_v- the number of testees who completed the same task correctly. Then the complexity order of the task can be determined using the following formula:

$$P = \frac{r_v}{r} 100\%.$$

For example, if 73% of the students do the test tasks correctly, then the quantitative complexity order of task will be 73.

It should be noticed that the greater the complexity order of P in quantitative terms, the easier the test task. Therefore, quantitative characteristic of test task level, i.e., the complexity degree of the task, can be defined by the following formula:

$$B = 100 - P = 100 - \frac{r_v}{r} 100$$

Thus, the test tasks will be classified by complexity levels.

Onwards, the “Selective thresholds” algorithm is used to further clarify the complexity degree of question and, thus, to determine the complexity level of the task. The coefficients of the selective thresholds were introduced by the authors in their previous work (Nurgabyl, 2014).

Let us assume that n groups take part in testing. Let r_m^k $(k = 1, \dots, n)$be the number of testees in group G^k $(k = 1, \dots, n)$ who performed the $m-th$ task, $\bar{r}_m^k$ - the number of testees who correctly did the same $m-th$ task. The complexity degree of the $m-th$ task is determined by the formula:

$$B_m^k = 100 - \frac{\bar{r}_m^k}{r_m^k} 100.$$

Now we form a matrix of test task complexity degrees:

$$B = \left\| \begin{matrix} B_1^1 & B_2^1 & B_3^1 \dots & B_p^1 \\ B_1^2 & B_2^2 & B_3^2 \dots & B_p^2 \\ \dots & \dots & \dots & \dots \\ B_1^n & B_2^n & B_3^n \dots & B_p^n \end{matrix} \right\|,$$

where p is the number of test tasks. Column numbers of matrix B correspond to the numbers of the proposed tasks, line numbers correspond to the number of testees’ groups.

With single-choice answers, if all the elements of any column of matrix B, for example j -th column, satisfy the score $82 < B_j^k \leq 100$ $(k = 1, \dots, n)$, j -th task will be excluded from the tests. In addition, those single choice tasks are excluded for which the testees failed to score the lower threshold, i.e. if all elements of any column, for example, the i -th column, satisfy the score $B_i^k < 48$ $(k = 1, \dots, n)$, the i^{th} task will be also excluded from the tests. For the remaining single choice tasks which got in the range of test applicability, i.e. if they satisfy the inequality $48 \leq B_j^k \leq 82$ $(k = 1, \dots, n)$, selectivity correction factor α is introduced for a random result, where: $-3 \leq \alpha \leq +3$. If, for the tasks under consideration, there is inequality $48 \leq B_j^k \leq 74 (k = 1, \dots, n)$, then using the positive selectivity factor, these tasks will be referred to “LU” (See **Table 1**) type of test tasks with threshold interval [50, 74].

If they satisfy the inequality $74 < B_j^k \leq 82 (k = 1, \dots, n)$, then using the negative selectivity factor, the task will be referred to as “SU” type (See **Table 1**) of test tasks with threshold interval [75, 79]. In case of multiple choice answers with an “OR” logic, those tasks which are in the range of inapplicability interval are immediately excluded from the test base, i.e. if all elements of any column of matrix B, for example, the j^{th} column, satisfy the score $B_j^k < 77$ $(k = 1, \dots, n)$ and the elements of the l^{th} column satisfy the score $B_j^k > 98$ $(k = 1, \dots, n)$, then the j^{th} and l^{th} tasks will be excluded from the test-base. For the remaining multiple-choice tasks with weight addition according to the “OR” logic which got in the range of test applicability, i.e. if they satisfy the inequality $77 \leq B_j^k \leq 98$ $(k = 1, \dots, n)$, selectivity correction factor α is introduced for a random result, where: $-4 \leq \alpha \leq +4$. If they satisfy the inequality $77 \leq B_j^k \leq 86 (k = 1, \dots, n)$, then using selectivity factors, these tasks will be included into “SOR” (See **Table 1**) type of test tasks with threshold interval [80, 84].

If they satisfy the inequality $86 < B_j^k \leq 98$ $(k = 1, \dots, n)$, then using selectivity factor, these tasks will be included into “VOR” (See **Table 1**) type of test tasks with threshold interval [90, 94]. Similarly, in case of multiple choice answers with “AND” logic, those tasks which got in the range of inapplicability interval are immediately

Table 1. Test task calibration parameters and scale of testees' knowledge level

Complexity levels	**Easy complexity level**	**Medium complexity level**			**High complexity level**	
Choice of test tasks	Single choice answer	Single choice answer	OR	AND	OR	AND
Types of test tasks	LU	SU	SOR	SAND	VOR	VAND
Scores	50-74	75-79	80-84	85-89	90-94	95-100

excluded from the test-base, i.e. if all elements of any column of matrix B, for example, the j^{th} column, satisfy the score $B_j^k < 83$ $(k = 1, \ldots, n)$ and the elements of the l^{th} column satisfy $B_j^k > 98$ $(k = 1, \ldots, n)$, then the j^{th} and l^{th} tasks will be excluded from the test-base. For the remaining multiple-choice tasks with weight addition according to "AND" logic which got in the range of test applicability, i.e. if they satisfy the inequality $83 \leq B_j^k \leq 98$ $(k = 1, \ldots, n)$, selectivity correction factor α is introduced for a random result, where: $-3 \leq \alpha \leq +3$. If they satisfy the inequality $83 \leq B_j^k \leq 92 (k = 1, \ldots, n)$, then using selectivity factor, this task will be referred to "SAND" (See **Table 1**) type of test tasks with threshold interval [85, 89]. If they satisfy the inequality $92 < B_j^k \leq 98 (k = 1, \ldots, n)$, then using selectivity coefficient, this task will be referred to as "VAND" (see **Table 1**) type of test tasks with threshold interval [95, 100].

Thus, calibration parameters for the test tasks and levels of testees' knowledge have been determined as shown in **Table 1**.

The proposed algorithm for selecting single and multiple-choice test tasks with different blocks of reactions and a scale of students' knowledge level allows for correlation of the results of various tests with each other.

Thus, the method for calibration of test task parameters has been developed, scales of testees' preparedness have been determined, and the level-rated classes for single and multiple-choice test tasks with different blocks of reactions have been established.

DISCUSSION

The proposed algorithm for selection of single and multiple-choice test tasks with various blocks of reactions, along with adaptive testing algorithms allow determining a student's knowledge level, and also quickly and effectively eliminating particular student's gaps in educational material acquisition.

The obtained results give an opportunity for further research on the theory of test task selection, and will be useful in the preparation of test tasks for determining the professionalism and the preparedness of employees, students, and others.

It should be noticed that great laboriousness and the difficulty of drawing up a bank of test tasks pose a certain difficulty and limitation in conducting the research.

CONCLUSION

The verification of the proposed algorithm was carried out for two years in Zhetysu State University named after I. Zhansugurov (Kazakhstan). The sample size was 310 students. Such sampling, despite its small size, has a rather high representativeness, and can show statistically accurate results. The verification of the proposed algorithm serves to perform several tasks at the same time:

- Selection of low-quality tasks;
- Establishing the test task level;
- Determination of the scale of students' preparedness level.

The value of the test task reliability factor, found by the algorithm described above, was $K = 0{,}87$. Therefore, the proposed test has a good reliability rating.

Using the proposed algorithm, the complexity degree of a test task can be determined, the test task selection by the complexity degree of the task is implemented, the level-rated classes of single and multiple choice test tasks with different blocks of reaction are allocated, and the scale of testees' preparedness level is determined.

Using the algorithm described, it is possible to determine the decision-making function from many independent variables characterizing the knowledge and skills of the students, and the informational utility of the test tasks, which is the object of future research.

REFERENCES

Avanessov, V. S. (2009). The language of pedagogical measurements. *Pedagogical Measurements*, *2*, 29-60. http://testolog.narod.ru/Theory65.html [in Russian]

Avanessov, V. S. (2014). New educational technology in university. *Bulletin of the Russian University of Peoples' Friendship, a series of education issues: languages and specialty*, *4*, 138-144. [in Russian]

Boesen, J., Lithner, J., & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning students use. *Educational studies in mathematics*, *75*(1), 89-105. doi:10.1007/s10649-010-9242-9

Bortz, J., & Döring, N. (2005). Forschungsmethoden und Evaluation. Heidelberg: Springer-Verlag. doi:10.1007/978-3-662-07299-8

Borzykh, A. A., & Gorbunov, A. S. (2009). Virtual worlds, information environments and ambitions e-Learning. *Educational Technology & Society, 12*(2), 423-437. https://cyberleninka.ru/article/v/virtualnye-miry-informatsionnye-sredy-i-ambitsii-e-learning [in Russian]

Botti, A., Grimaldi, M., Tommasetti, A., Troisi, O., & Vesci, M. (2017). Modeling and Measuring the Consumer Activities Associated with Value Correction: An Exploratory Test in the Context of Education. *Service Science*, *9*(1), 63-73. doi:10.1287/serv.2016.0156

Chen, C., Wang, J., & Yu, C. (2017). Assessing the attention levels of students by using a novel attention aware system based on brainwave signals. *British J. of Educational Technology, 48*(2), 348-369. doi:10.1111/bjet.12359.

Cheng, Y., Diao, Q., & Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior Research methods*, *49*(2), 502-512. doi:10.3758/s13428-016-0712-6

Coe, R. (2003). Web-based nuclear testing & training. *Nuclear Plant Journal, 21*(1), 47-61. https://www.highbeam.com/doc/1P3-318821101.html

De Meo, P., Messina, F., Rosaci, D., & Sarné G. M. L. (2017). Combining trust and skills evaluation to form e-Learning classes in online social networks. *Information Sciences*, *405*, 107-122. doi:10.1016/j.ins.2017.04.002

de Villiers, M. R. (Ruth), & Becker, D. (2017). Investigating learning with an interactive tutorial: A mixed-methods strategy. *Innovations in Education and Teaching International, 54*(3), 247-259. doi:10.1080/14703297.2016.1266959

Dvoryatkina, S. N. (2013). Designing adaptive computerized training system tasks probabilistic-statistical areas of mathematics. *Bulletin of the Russian Peoples Friendship University, Series Informatization of Education, 1*, 97-104. [in Russian]

Edens, K., & Shields, C. A. (2015). Vygotskian approach to promote and formatively assess academic concept learning. *Assessment & Evaluation in Higher Education*, *40*(7), 928-942. doi:10.1080/02602938.2014.957643

Guznenkov V. N., & Seregin V. I. (2016). Computer testing as a form of control knowledge of students on geometric-graphic disciplines. *International Research Journal, Series: Pedagogical Sciences*, *9*(51), 56-58. [in Russian]

Haist, S. A., Butler, A. P., & Paniagua, M. A. (2017). Testing and evaluation: the present and future of the assessment of medical professionals. *Advances in Physiology Education, 41*(1), 149-153. doi:10.1152/advan.00001.2017

Horton, W., & Horton K. (2005). *E-learning: tools and technologies*. Moscow: Kudits-Image. [in Russian]

Howard, S. J., Woodcock, S., Ehrich, J., Bokosmaty, S., & others (2017). What are standardized literacy and numeracy tests testing? Evidence of the domain-general contributions to students' standardized educational test performance. *British J. of Educational Psychology*, *87*(1), 108-122. doi:10.1111/bjep.12138

Huang, H. T. D., & Hung, S. T. A. (2010). Examining the practice of a reading-to-speak test task: anxiety and experience of EFL students. *Asia Pacific Education Review*, *11*(2), 235-242. doi:10.1007/s12564-010-9072-6

Hung, J. (2012). Trends of E-learning Research from 2000 to 2008, Use of text mining and bibliometrics. *British J. of Educational Technology*, *43*(1), 5–16. doi:10.1111/j.1467-8535.2010.01144.x

Kibzun, A. I., & Inozemtsev, A. O. (2014). Using the maximum likelihood method to estimate test complexity levels. *Automation and Remote control*, *75*(4), 607-621. doi:10.1134/S000511791404002X

Lim, S. Y., & Chapman, E. (2013). Development of a short form of the attitudes toward mathematics inventory. *Educational studies in mathematics, 82*(1), 145-164. doi:10.1007/s10649-012-9414-x.

Markon, K. E. (2013). Information Utility: Quantifying the Total Psychometric Information Provided by a Measure. *Psychological Methods*, *18*(1), 15-35. doi:10.1037/a0030638

Martos-Garcia, D., Usabiaga, O., & Valencia-Peris, A. (2017). Students' Perception on Formative and Shared Assessment: Connecting two Universities through the Blogosphere. *Journal of new Approaches in Educational Research*, *6*(1), 64-70. doi:10.7821/naer.2017.1.194

Nurgabyl, D. N. (2014a). About one mathematical model of calibration of parameters of test tasks. *Bulletin of KazNTU named after K. Satpayev, 3*, 482-487. http://vestnik.kazntu.kz/files/newspapers/81/2669/2669.pdf [in Russian]

Nurgabyl, D. N. (2014b). On a mathematical model of multi-step adaptive testing. *Bulletin of the Abai Kazakh National Pedagogical University, a Series of Physical and Mathematical Sciences, 1*(45), 143-149. [in Russian]

Nurgabyl, D. N. (2012). On a computer adaptive testing technology in vocational training. *Proceedings of the international scientific-practical conference "Mathematical, science education and information", 2*, (316-319). Moscow: Institute of Mathematics and Informatics. [in Russian]

Nurgabyl, D. N., & Ramazanov, R. G. (2013). About one model of adaptive computerized testing. *Proceedings of the International Conference on the Transformation of Education, Mathematics*, (pp.13-21). London.

Nurjanah, Dahlan, J. A., & Wibisono, Y. (2017). Design and Development Computer-Based E-Learning Teaching Material for Improving Mathematical Understanding Ability and Spatial Sense of Junior High School Students. Proceedings of the 3rd International Seminar on Mathematics, Science, and Computer Science Education (MSCEIS), Bandung, Indonesia, 2016, *Journal of Physics Conference Series*, *812*, UNSP 012098. doi:10.1088/1742-6596/812/1/012098

Ozturk, N., & Dogan, N. (2015). Investigating Item Exposure Control Methods in Computerized Adaptive Testing. *Educational sciences-theory & practice, 15*(1). 85-89. doi:10.12738/estp.2015.1.2593

Pantziara, M., & Philippou, G. (2012). Levels of students' "conception" of fractions. *Educational studies in mathematics, 79*(1), 61-83. doi:10.1007/s10649-011-9338-x

Park, J. (2010). Constructive multiple-choice testing system. *British J. of Educational Technology, Special Issue, Learning objects in progress*, *41*(6), 1054–1064. doi:10.1111/j.1467-8535.2010.01058.x

Permyakov, O. E., & Maksimov, O. A. (2015). Formalization of expert evaluation of the quality of test materials from the positions of the system approach. *Vestnik pedagogicheskikh innovatsii, 3*(7), 157-178. [in Russian]

Prado, E., Hartini, S., & Rahmawati, A. et al. (2010). Test selection, adaptation, and evaluation: A systematic approach to assess nutritional influences on child development in developing countries. *British J. of Educational Psychology, 80*(1), 31-53. doi:10.1348/000709909X470483

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: The University of Chicago Press.

Sangwin, C. J., & Jones, I. (2017). Asymmetry in student achievement on multiple-choice and constructed-response items in reversible mathematics processes. *Educational studies in mathematics, 94*(2), 205-222. doi:10.1007/s10649-016-9725-4

Senior, C., Fearon, C., & Mclaughlin, H. et al. (2017). How might your staff react to news of an institutional merger? A psychological contract approach. *International Journal of Educational management*, *31*(3), 364-382. doi:10.1108/IJEM-05-2016-0087

Siddiq, F., Gochyyev, P., & Wilson, M. (2017). Learning in Digital Networks. ICT literacy: A novel assessment of students' 21st century skills. *Computers & Education, 109*, 11-37. doi:10.1016/j.compedu.2017.01.014

van der Linden, W. J., & Glas, C. A. W. (Eds.) (2010). *Elements of Adaptive Testing*. Springer. doi:10.1007/978-0-387-85461-8.

van Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British J. of Mathematical* & *Statistical Psychology*, *70*(2), SI, 317-345. doi:10.1111/bmsp.12101

Vasiliev V. N. (2007). University as an open system. *Innovations,* 2, 57-60. https://elibrary.ru/item.asp?id=12898737 [in Russian]

Vlasin, I., & Chirila, C. B. (2015). The model of a competence based e-learning platform for primary and middle school students. *Smart 2014-Social media in academia: Research and Teaching,* 179-184 http://www.academia.edu/9872137/The_model_of_a_competence_based_e-learning_platform_for_primary_and_middle_school_students

Voutilainen, A., Saaranen, T. S., & Ormunen, M. (2017). Conventional vs. e-learning in nursing education: A systematic review and meta-analysis. *Nurse Education Today*, *50*, 97-103. doi:10.1108/IJEM-05-2016-0087

Wilmot, D. B., Schoenfeld, A., Wilson, M., Champney, D., & Zahner, W. (2011). Validating a Learning Progression in Mathematical Functions for College Readiness. *Mathematical Thinking and learning, 13*(4), 259-291. doi:10.1080/10986065.2011.608344

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Xia, Q., Liang, R., & Wu, J. (2017). Transformed contribution ratio test for the number of factors in static approximate factor models. *Computational statistics & Data Analysis*, *112*, 235-241. doi:10.1016/j.csda.2017.03.005

Yaman, S. (2011). Comparison of test use and multiple-evaluation to test effectiveness of PBL in different grouping strategies. *Energy Education Science and Technology, Part B-social and Educational studies*, *3*(1-2), 131-142.

http://www.ejmste.com