

Knowledge assessment in higher education institutions based on the adaptive Item Response Theory model

Serhii Liubarskyi^{1,*†}, Eduard Bovda^{1,†} and Yuriy Samokhvalov^{1,2,†}

¹ Military Institute of Telecommunications and Informatization named after Heroes of Kruty, Knyaziv Ostrozkyh Street 45/1, Kyiv, 01011, Ukraine

² Taras Shevchenko National University, Volodymyrska Street 64/13, Kyiv, 01601, Ukraine

Abstract

An approach to assessing the level of assimilation of educational material by education seekers is proposed. This approach involves using adaptive Item Response Theory (IRT) models to evaluate knowledge based on the one-parameter Rasch model, design a combined testing procedure, and verify the balance of test tasks. This enables the determination of the objective level of preparedness of the education seeker according to the testing procedure and the assessment of the quality of the preparation for test tasks.

Keywords

knowledge assessment, adaptive testing, Rasch parametric model, test task complexity

1. Introduction

Modern higher education institutions in developed European countries are actively working to improve the education system [1], which can integrate into the European educational and scientific space [2]. One of the driving mechanisms for such improvement is ensuring fair competition in the educational services market with an appropriate quality of higher education.

Training high-quality specialists should involve not only modern methods and means of acquiring new knowledge, but also an objective assessment of their level of assimilation [3]. The effectiveness of the evaluation depends on compliance with the didactic principles of systematicity and objectivity, as emphasized in the work [4, 5].

Currently, a relatively wide range of research assesses the knowledge of education seekers. Their analysis has shown that the results of knowledge control of existing systems are not sufficiently informative and objective. An important parallel direction of improvement is the control of data errors to enhance the reliability of processing results, a methodology explored in detail in related works [6]. Testing procedures must use test tasks that are uniformly selected in terms of complexity to ensure the maximum informativeness and objectivity of the control results. This can be done using adaptive testing.

Adaptive testing (AT) is a type of test assessment that evaluates knowledge, where the sequence of test task presentation (including their complexity) and the number of functions depend on the student's answers to previous test tasks [7, 8]. The use of an adaptive approach enables you to consider the individual capabilities of students, assess their cognitive skills, increase the accuracy of determining their knowledge level, utilize both rating and interval scales for evaluating test results, and enhance student motivation to study [9].

ITS-2024: Information Technologies and Security, December 19, 2024, Kyiv, Ukraine

* Corresponding author.

† These authors contributed equally.

✉ sergii.liubarskyi@viti.edu.ua (S. Liubarskyi); edepig8305@ukr.net (E. Bovda); yu1953@ukr.net (Y. Samokhvalov)

ORCID 0000-0001-8068-1106 (S. Liubarskyi); 0000-0002-8267-2120 (E. Bovda); 0000-0001-5123-1288 (Y. Samokhvalov)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

However, the adaptive approach has the following limitations: the presence of linear processes in testing procedures, the absence of feedback; the lack of ranking of test participants who gave the correct answer to the same number of test tasks, which affects the objectivity of the assessment [10]; the imbalance of test tasks in terms of content and complexity, which affects their quality [11].

Therefore, to ensure proper informativeness and objectivity of the control results, it is proposed to assess the level of assimilation of educational material by education seekers based on adaptive Item Response Theory (IRT) knowledge assessment models, with the possibility of ranking test participants who gave the correct answer to the same number of test tasks, which are balanced in content and complexity.

In the approach to assessing the level of assimilation of educational material by education seekers and ensuring proper objectivity, it is proposed to use adaptive IRT models of knowledge assessment with existing procedures: organizing test assessment of knowledge; determining the conditions for the need to change the level of complexity of the test task; determining the balance of test tasks in the test (test quality).

2. Adaptive testing model

The adaptive testing model involves: a testing organization procedure (determining the initial level of knowledge of the education seeker, determining the set(s) of test tasks and their level of complexity, determining the estimated weight of the correct and incorrect answer to the test task, establishing the form of obtaining the final results for testing); a procedure for conducting testing (establishing the level of complexity of the task from which testing will begin, determining the method for assessing the correctness of the answer to the current test task, determining the rules (conditions) regarding the order of changing test tasks according to the procedure and its completion to obtain the final result); a method for determining the balance of test tasks in the test (test quality).

2.1. Organization and conduct of testing

The organization and implementation of knowledge testing should be examined through the lens of the following classification groups of adaptive testing methods:

1. Pyramidal Testing is an adaptive form of testing where the complexity of the test task changes according to the correctness of the answer to the previous one. At the beginning of the testing procedure, the starting level of complexity of the test task is set. The rule of dividing the complexity scale of test tasks in half at each stage is applied.

The total number of questions in pyramid testing can be adjusted according to the needs of teachers. Still, the formula $4n$ is typically used, where n represents the number of test tasks for each student.

The drawbacks of this testing procedure include the significant resources required for its development and the necessity of independently assessing item complexity, which can potentially compromise the objectivity of the results obtained.

2. FlexiLevel is a type of adaptive testing organization and implementation aimed at comparing each test participant's qualification level with the degree of complexity of the test tasks given to them. The testing procedure begins with a level of complexity that is arbitrary. After that, depending on the test participant's answer, he is given a task of a higher or lower level of complexity. There is a gradual approach to the real level of preparedness of the test participant.

The primary objective of this procedure is to utilize a fixed-branching algorithm to adjust the difficulty level of test items.

To determine the complexity of test tasks, initial statistical data is collected based on fixed-form testing procedures. The next step involves determining the test item's complexity parameter by calculation according to the formula:

$$Difficulty = 1 - \frac{n_p}{n_r}, \quad (1)$$

where: n_p – the count of tested participants who successfully completed this test item;
 n_r – the count of tested participants who answered this test item.

The advantage of such testing is the simplicity of the algorithm. Such tests demonstrate significantly less variability in test items presented to each test taker, and an identical number of items is administered to several test takers who follow the same sequence of correct or incorrect responses.

The system may produce incorrect results if the question base is small or does not cover all difficulty levels. FlexiLevel testing often focuses on assessing individual aspects of knowledge rather than their integrated application. This can result in students who are good at the material but have gaps in individual topics receiving a lower score.

3. Stradaptive (stratified adaptive) – stratified adaptive testing. This procedure is based on Alfred Binet’s strategy of using peak tests. In such tests, there is low variability in task complexity, and tasks are distributed over a narrow range of complexity. The significance of peak tests in the stratified adaptive testing procedure is to provide a more accurate measurement of the student's abilities.

A stratified adaptive test selects stratified or organized tasks into scaled, progressively more difficult tests by difficulty level.

Typically, peak tests are located at the beginning of the testing procedure and are used to assess the starting abilities of the education seeker. Their number is variable and depends on the purpose of the test. After the initial knowledge of the education seeker is determined, the assessment procedure switches to adaptive mode, and test tasks are provided according to the selected adaptive algorithm.

The advantages of this procedure in organizing and conducting testing include factors such as the accuracy of assessment, time efficiency, task balance, flexibility in test design, and participant psychological comfort. The disadvantages include the complexity of development and implementation. Stratified testing operates within specific strata, which may limit flexibility in adaptation. If a participant makes mistakes in the initial stages of testing, they may be moved to a lower tier, where the tasks will become too easy for them to complete. This may lead to an underestimation of their real knowledge.

Therefore, each of the above procedures has certain peculiarities in its application, which determine its advantages and limitations. This necessitates the use of a combined approach in the organization and conduct of testing in the proposed model.

According to the procedure for organizing testing, the initial level of preparedness of the applicant for education is average. It is proposed that three sets of homogeneously selected test tasks be formed, differing in complexity. The estimated weight of the correct answer to the test task is 1, and the incorrect one is 0. The method of obtaining final results involves calculating the ratio of correctly answered questions to the total number of test tasks, taking into account their complexity.

According to the testing procedure, the level of complexity of the task from which the testing will begin is set to medium, which is typical of the pyramidal approach. When answering the test task, the student launches an adaptive branching algorithm by processing test tasks of different levels of complexity. If the student answers the test task correctly, the complexity level of the next test increases; if incorrectly, it decreases (flexible procedure).

2.2. Evaluating the correct answer to a test task

The processing of the condition for the need to complicate (ease) the complexity of the test task, depending on the current answer of the education seeker, will be entrusted to the appropriate procedure that assesses the probability of the education seeker's correct answer to the current test task.

Currently, Rasch and Birnbaum's logistic models [12] are widely used to determine such a probability [13].

The one-parameter Rasch model (1PL) utilizes two values: θ – the knowledge level of the test taker, and δ – the task's complexity level, represented as $\theta \in (-\infty; \infty)$, $\delta \in (-\infty; \infty)$.

The one-parameter Rasch model (1PL) involves the introduction of two main parameters: θ , representing the test participant's knowledge level, and δ , representing the task complexity level [14]. This is the probability that a test participant with a level of preparedness θ will correctly complete a test task of complexity δ . The likelihood of success depends, in fact, only on one parameter – the difference $\theta - \delta$, and is calculated by the following formula:

$$P_j(\theta) = \frac{e^{(\theta-\delta_j)}}{1+e^{(\theta-\delta_j)}}, \quad (2)$$

where: $\theta_i = \ln\left(\frac{p_i+1}{q_i+1}\right)$, p_i – the proportion of correct answers to test tasks, q_i – the proportion of incorrect answers to test tasks;

$$\delta_i = 1 + \ln\left(\frac{q_i+1}{p_i+1}\right);$$

$i = 1, 2, \dots, m$ – number of testing participants;

$j = 1, 2, \dots, n$ – number of test tasks.

At the same time:

$$\begin{aligned} \lim_{(\theta-\delta) \rightarrow +\infty} P &= 1, & \lim_{(\theta-\delta) \rightarrow -\infty} P &= 0 \\ P &= 0.5 \text{ if } \theta = \delta. \end{aligned} \quad (3)$$

The one-parameter Rasch model is most effective for analyzing dichotomous tasks, where the answer can only be binary: "correct" or "incorrect". This model considers only one parameter – task difficulty (δ), which makes it ideal for tasks where additional characteristics such as guessing level or discrimination need not be considered. The model assumes that all items have the same level of discrimination (i.e., all items are equally good at distinguishing between strong and weak test takers). The one-parameter Rasch model is well-suited for tests with a limited number of items, as it requires fewer parameters to be estimated. This model is best suited for open-ended test items where there is no way to guess the correct answer.

In the two-parameter Rasch model (2PL), the adaptive test consists of n test items. In the model, the response variables U_j , $j = 1, 2, \dots, n$, can take the value one if the answer is correct, and zero if it is incorrect, that is:

$$U_j = \begin{cases} 0, & \text{if the answer is incorrect} \\ 1, & \text{if the answer is correct} \end{cases}$$

This model looks like this:

$$P_j(U_j = 1|\theta) = \frac{e^{a_j(\theta-\delta_j)}}{1+e^{a_j(\theta-\delta_j)}}. \quad (4)$$

where: a_j – differentiating ability (discriminative power) of the j -th test task (discrimination coefficient).

Discrimination helps to select tasks that best distinguish between different levels of knowledge or skills. It allows you to identify tasks that effectively distinguish between test takers with high and low levels of expertise. The model is suitable for organizing and presenting test tasks where the test taker must choose one correct answer from several proposed ones.

If the test contains open-ended questions or questions that do not have clearly defined right/wrong answers, the 2PL model may be less suitable. For tasks where it is essential to consider additional factors (for example, individual characteristics of test participants), more complex models may be required (for example, a three-parameter model).

By Birnbaum's logistics model (3PL), the probability of the accurate response $P_j(U_j = 1|\theta)$ on j questions with knowledge level θ is calculated by the formula:

$$P_j(U_j = 1|\theta) = c_j + (1 - c_j) \frac{e^{1.7a_j(\theta-b_j)}}{1 + e^{1.7a_j(\theta-b_j)}}. \quad (5)$$

where: c_j – the probability of guessing the correct answer when performing the j -th test task.

This model takes into account three key parameters for each test item: difficulty (helps to select items that match the participant's current level), discrimination ability (allows you to choose items that best distinguish between different levels of knowledge), and guess probability (taken into account when scoring, especially when the items are multiple choice).

The 3PL model involves calibrating tasks before testing to ensure assessment accuracy, even with a fixed test format. The model is most suitable for tests where participants select one or more correct answers from a set of provided options.

A limitation of the 3PL model is that it is difficult to calibrate (for accurate application of the model, a large amount of data is required to estimate task parameters), is sensitive to task quality (if tasks have low discriminatory power or inadequate complexity, the model may produce inaccurate results), and has limitations for test tasks with partial scores (scoring on score scales).

In the proposed adaptive testing model, the single-parameter Rasch model will be used to evaluate the correct answer to a test task and determine the condition for changing the test task's complexity level.

This is because it is most suitable for tests with dichotomous tasks, equal task discrimination, a small number of functions, and a low level of guessing.

Probability of a correct answer $P_j(\theta)$ obtained during testing. If $P_j(\theta) \geq 0.5$, then the task's difficulty level must be increased; otherwise, it must be decreased.

2.3. Determining the quality of test tasks

To assess the quality of selected test tasks, which should reflect the structural hierarchy of the training model in the academic discipline, it is necessary to address the problem of test composition, specifically to evaluate the internal consistency (balance) of tasks within the test.

In this case, the expert needs to assess the effectiveness of the scheme proposed by the developer and the method used to arrange tasks in the test. The concept of balance involves the proportional filling of the test with functions of varying levels of complexity.

Thus, the analysis of the test composition reveals the degree of harmonious presentation of key elements of the academic discipline's content, the adequacy of their reflection in the test, and the appropriate level of differential difficulty of the test tasks.

Currently, the following approaches and methods are most common in the field of assessing the quality of test tasks:

1. Methods based on the criteria of mathematical and statistical methods of analyzing test task statistical characteristics (properties). They enable the detection of hidden defects in test tasks. Models of dispersion, factor, cluster, discriminant analysis, and time series regression analysis are used.
2. Methods that use expert assessment of the consistency of tasks in the test. The procedure for specialist assessment of test quality involves the following stages: semantic capacity evaluation, prediction of test task success, and comprehensive examination of test quality and task performance, which includes approbation testing. The goal is to establish, verify, and evaluate its measuring capabilities by approving representative samples.
3. Methods based on analyzing the degree of balance of test items. The calculated correlation coefficient (balance) is used as a numerical indicator of the validity of the test.

Among the methods of the third group, the Chelyshkova method of checking the balance of test weight coefficients attracts special attention [15]. This method has proven very effective in identifying extreme test tasks, that is, those that are too difficult or too easy. For this, the balance of their weight coefficients is checked. This check is carried out to analyze the complexity of test tasks and inform the expert about the feasibility of replacing extreme questions in the test, which negatively affect the objectivity of assessing the student's knowledge [16].

Thus, preference is given to a combined method of organizing adaptive testing [17]. The single-parameter Rasch model will be used to determine the condition for changing the complexity of the

test task. Experimentally selected coefficients provide the optimal value of the indicator of differential ability for test tasks, the balance of which is achieved using the method of M. Chelyshkova.

Thus, preference is given to a combined method of organizing adaptive testing. The single-parameter Rasch model will be used to determine the condition for changing the complexity of the test task. Experimentally selected coefficients provide the optimal value of the indicator of differential ability for test tasks, the balance of which is achieved using the method of M. Chelyshkova.

3. Knowledge assessment based on the proposed adaptive model

Organizing and conducting a knowledge assessment test involves the following steps:

1. Initialization of the testing procedure. The task is randomly selected from a set of functions corresponding to the specified difficulty level (“easy”, “medium”, or “difficult”). The first task is always of average difficulty. Depending on the answer to it, the next task will be proposed by the adaptive algorithm as easier or more difficult than the previous one:

$$t_i \in \vec{X} \cup \vec{Y} \cup \vec{Z}, \quad i = \overline{1, N}. \quad (6)$$

where: t_i – test task of the appropriate level of complexity;

$\vec{X}, \vec{Y}, \vec{Z}$ – set of tasks of difficulty level “easy”, “medium”, “difficult”;

N – number of test tasks.

In the model, the response variables take the value one if the answer is correct and zero if it is incorrect, that is:

$$q(t_i) \in \{0; 1\}. \quad (7)$$

2. Calculate the current B_S (8) and the maximum possible B_A (9) number of points for the testing procedure, as well as the current success rate of the education seeker S (10).

$$B_S = \sum_{i=1}^M q(t_i) k_i. \quad (8)$$

$$B_A = N k_{max}. \quad (9)$$

$$S = \frac{B_S}{B_A} 100\%. \quad (10)$$

where: $q(t_i)$ – the answer to the test task of the appropriate level of complexity,

$q(t_i) \in \{0; 1\}$;

k_i – weight of the test task, $k_{min} = 1$, $k_{average} = 2$, $k_{max} = 3$;

N – number of test tasks.

The success rate of passing the test indicates the percentage of points received by the student at this stage, compared to the maximum number of points he could have if he had answered all the proposed test tasks correctly, S (10).

3. Determining the current proportion of true p_i (11) and false q_i (12) answers.

$$p_i = \frac{\sum_{i=1}^N q(t_i)}{N}. \quad (11)$$

$$q_i = 1 - p_i. \quad (12)$$

4. Calculation of the current logit of the level of knowledge of the education seeker θ_i (13) and the logit of the level of difficulty of the task δ_i (13).

$$\theta_i = \ln \left(\frac{p_i+1}{q_i+1} \right). \quad (13)$$

$$\delta_i = 1 + \ln \left(\frac{q_i+1}{p_i+1} \right). \quad (14)$$

5. Deciding on the need to move to another level of task complexity.

A feature of this step is that after each answer to the test task, the probability of the student's correct answer is calculated, $P_j(\theta)$ (15), and, based on its value, determines the transition to the appropriate level of task complexity from the experimentally established range.

A threshold value of the probability of a student's correct answer, $P_j(\theta) = 0.5$, serves as a condition for changing the task's difficulty level during the testing process. The task's difficulty level increases when $P_j(\theta) \geq 0.5$ and decreases otherwise.

The higher the probability of an answer, the higher the level of difficulty of the task, and vice versa. The essence of this operation is to ensure the correspondence between the task's difficulty level and the student's knowledge level, thereby enhancing assessment objectivity. This correspondence is particularly crucial, as the discriminant ability of the item refers to its capacity to distinguish learners based on their level of preparedness. The higher the discriminant ability, the better the division of learners into those who are prepared and those who are not.

$$P_j(\theta) = \frac{e^{(\theta - \delta_j)}}{1 + e^{(\theta - \delta_j)}} \quad (15)$$

6. Organization of verification of the conditions for completing testing.

Below is a schematic representation of the testing process, taking into account the change in the weight of the test task, and for the case where the final level of success is determined by the condition of completing the total number of test items (Fig. 1).

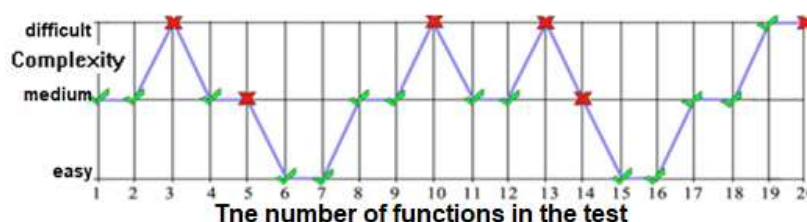


Figure 1: The process of test administration with a gradual change in task difficulty level, constrained by a fixed total number of questions in the test is completed

4. Checking the balance of test tasks

A balance check of their weight coefficients is performed to find extreme test tasks, i.e., those that are too difficult or too easy. This check is performed to analyze the complexity of the functions and inform the expert about the feasibility of replacing extreme questions in the test that negatively affect the objectivity of the student's knowledge assessment.

The balance of the weight coefficients of test tasks is proposed to be checked using the Chelyshkova balance method.

Checking the balance of test item weights includes the following steps:

1. Organization of the formation of the evaluation matrix.

In any testing process, the results of calculating $\tilde{\theta}_i$ the statistical estimates θ_i and $\tilde{\delta}_j$ the statistical estimates δ_j will differ from the existing exact values. In this sense, the estimates are certain functions of the initial random values of the elements of the test results matrix. This matrix contains numerical definitions of the indicator gradation.

This matrix is a table, with the rows corresponding to the number of education seekers (N) and the columns corresponding to the number of indicator variables (M). In the case of testing educational achievements, the indicator variables are test tasks. A number that characterizes the education seeker's response to this task is at the intersection of the rows and columns. In this work, a binary matrix is used, a matrix of test results for the dichotomous case, where the education seeker's responses are characterized by two values: 0 and 1. Zero corresponds to an incorrect answer, and units to a correct one.

2. Determining the proportion of correct p_i and incorrect q_i answers.

The proportion of correct answers p_j for the i -th student is defined as the ratio of their number of correct responses to the total number of items answered.

$$p_i = \frac{\sum_{j=1}^M q(t_i)}{M}, q_i = 1 - p_i. \quad (16)$$

Similarly, the proportion of correct answers for the j -th test item p_j is calculated as the ratio of the total number of correct responses by all students to the aggregate number of answers to that item. In both scenarios, the proportion of incorrect answers q_j is obtained by subtracting the proportion of correct answers from unity.

$$p_j = \frac{\sum_{i=1}^M q(t_j)}{N}, q_j = 1 - p_j. \quad (17)$$

3. Calculate the initial level of knowledge of the education seeker θ_i^0 (18) and the initial level of difficulty of tasks β_j^0 (18).

$$\theta_i^0 = \ln\left(\frac{p_i+1}{q_i+1}\right), \beta_j^0 = 1 + \ln\left(\frac{q_j+1}{p_j+1}\right). \quad (18)$$

4. Determining the average level of preparedness of education seekers $\bar{\theta}$ and the average difficulty level of tasks $\bar{\beta}$.

$$\bar{\theta} = \frac{\sum_{i=1}^N \theta_i^0}{N}, \bar{\beta} = \frac{\sum_{j=1}^M \beta_j^0}{M}. \quad (19)$$

5. Calculation of variances S_θ and S_β to reduce the parameters into a single scale of standard estimates, and calculation of the slope coefficients α_θ α_β .

Due to the influence of various random factors, the estimates of the parameters θ and δ obtained from several samples will, of course, differ. If the sample size is large enough, one can calculate stable values of the parameters θ and δ , which will be the most effective estimates and can be accepted as objective estimates of θ and δ .

Thus, the question arises of finding the mathematical expectations and variances of these random variables. It is necessary that the mathematical expectation of the corresponding estimates coincide with the corresponding exact values, and the variance of the forecast is minimal.

$$S_\theta = \frac{\sum_{i=1}^N (\theta_i^0)^2 - N(\bar{\theta})^2}{N-1}, S_\beta = \frac{\sum_{j=1}^M (\beta_j^0)^2 - M(\bar{\beta})^2}{M-1}. \quad (20)$$

$$\alpha_\theta = \sqrt{\frac{1+S_\beta/2,89}{1+S_\theta S_\beta/8,35}}, \alpha_\beta = \sqrt{\frac{1+S_\theta/2,89}{1+S_\theta S_\beta/8,35}}. \quad (21)$$

6. Calculation of the values of the parameter estimates θ_i and β_j on a single interval scale.

$$\theta_i = \alpha_\theta \theta_i^0 + \bar{\beta}, \beta_j = \alpha_\beta \beta_j^0 + \bar{\theta}. \quad (22)$$

7. Calculation of the balance of the test task complexity indicator $\sum \beta$ (23).

$$\sum \beta = \sum_{j=1}^N \beta_j. \quad (23)$$

The determination of the final grade according to the traditional learning scheme (unsatisfactory, satisfactory, good, very good, excellent) is carried out using conversion scales based on the student's level of success.

To evaluate a user who has passed the test, it is necessary to determine their success and compare this value with those set by the developer when creating the training task. Based on this comparison, a score can be assigned according to the traditional training scheme.

5. Practical use

To evaluate a user who has passed the test, it is necessary to determine their success and compare this value with those set by the developer when creating the training task. Based on this comparison, a score can be assigned according to the traditional training scheme.

The input conditions for executing the algorithm are:

$3 \cdot N$ – total number of tasks, $N=10$;

A set of tasks of difficulty levels “easy”, “medium”, and “difficult”.

$X = \{x_1, x_2, \dots, x_N\} = \{1, 2, \dots, 10\}$;

$Y = \{y_1, y_2, \dots, y_N\} = \{1, 2, \dots, 10\}$;

$Z = \{z_1, z_2, \dots, z_N\} = \{1, 2, \dots, 10\}$;

The set of task complexity weights $K = \{k_1, k_2, k_3\} = \{1; 2; 3\}$;

The set of answer options $q = \{0; 1\}$.

The number of students is 3.

N – number of test questions, $N=20$.

$k_{min}=1, k_{average}=2, k_{max}=3$.

Table 1

Evaluation matrices

	Test task	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Student 1	difficult	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	medium	1	1	0	1	0	0	0	1	1	0	1	1	0	0	0	0	1	1	0	0
	easy	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0
Student 2	difficult	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1
	medium	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0
	easy	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
Student 3	difficult	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	medium	0	0	0	0	0	0	0	1	1	0	1	1	0	1	0	0	0	1	1	0
	easy	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0

You need to find:

1. Success rate, as determined by test results (%), to evaluate the rating indicator of the education seeker on the preference scale.
2. Assessment according to the traditional learning scheme (unsatisfactory, sufficient, satisfactory, good, very good, excellent).
3. An indicator of the balance of the level of difficulty of tasks.

Calculation results:

1. Success according to testing results.

$$B_A^1 = Nk_{max} = 20 \times 3 = 60.$$

$$B_S^1 = \sum_{i=1}^M q(t_i) k_i = 2 \times 1 + 2 \times 1 + 3 \times 0 + 2 \times 1 + 2 \times 0 + 1 \times 1 + 1 \times 1 + 2 \times 1 + 2 \times 1 + 3 \times 0 + 2 \times 1 + 2 \times 1 + 3 \times 0 + 2 \times 0 + 1 \times 1 + 1 \times 1 + 2 \times 1 + 2 \times 1 + 3 \times 1 + 3 \times 0 = 2 + 2 + 0 + 2 + 0 + 1 + 1 + 2 + 2 + 0 + 2 + 2 + 0 + 0 + 1 + 1 + 2 + 2 + 3 + 0 = 25.$$

$$S^1 = \frac{B_S^1}{B_A^1} 100\% = \frac{25}{60} \times 100 = 41.67.$$

$$B_A^2 = Nk_{max} = 20 \times 3 = 60.$$

$$B_S^2 = \sum_{i=1}^M q(t_i) k_i = 2+3+3+3+0+2+2+0+0+1+1+2+2+3+3+3+0+2+2+3=37.$$

$$S^2 = \frac{B_S^2}{B_A^2} 100\% = \frac{37}{60} \times 100 = 61.67.$$

$$B_A^3 = Nk_{max} = 20 \times 3 = 60.$$

$$B_S^3 = \sum_{i=1}^M q(t_i) k_i = 0+1+1+0+0+1+1+2+2+0+2+2+0+2+0+1+1+2+2+0=20.$$

$$S^3 = \frac{B_S^3}{B_A^3} 100\% = \frac{20}{60} \times 100 = 33.33.$$

2. Construction of a series of preferences for education seekers.
Based on the calculations, rating indicators of the success of education seekers were determined.
 S^1, S^2, S^3 Allow us to construct an objective series of precedence:

$$\text{Student 2 (61.67)} < \text{Student 1 (41.67)} < \text{Student 3 (33.33)}$$

3. Construction of a series of test task priorities by complexity.
To determine the balance of test tasks, we will determine the proportion of correct answers to the j -th test task for the i -th student.

For example,

$$p_i^1 = \frac{\sum_{i=1}^M q(t_i)}{M} = \frac{2}{3} = 0.7, i = 1, \dots, 3$$

$$q_i^1 = 1 - p_i = 1 - 0.7 = 0.3.$$

$$p_i^2 = \frac{\sum_{i=1}^M q(t_i)}{M} = \frac{3}{3} = 1.$$

$$q_i^2 = 1 - p_i = 1 - 1 = 0.$$

...

$$p_i^{20} = \frac{\sum_{i=1}^M q(t_i)}{M} = 0.4.$$

$$q_i^{20} = 1 - p_i = 1 - 0.4 = 0.6.$$

Table 2

The number of test tasks in terms of difficulty

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
p_i^N	0.7	1	0.7	0.4	0	1	1	0.7	0.7	0.4	1	1	0.4	0.7	0.7	1	0.7	1	1	0.4
q_i^N	0.3	0	0.3	0.6	1	0	0	0.3	0.3	0.6	0	0	0.6	0.3	0.3	0	0.3	0	0	0.6

We built a series of test tasks in order of difficulty:

$$(2,6,7,11,12,16,18,19) < (1,3,8,9,14,15,17) < (4,10,13,20) < (5)$$

We notice that the questions are not well-balanced.

Calculate the initial level of knowledge of the education seeker θ_i^0 and the initial level of difficulty of tasks β_i^0 .

For example,

$$\theta_i^1 = \ln \left(\frac{p_i + 1}{q_i + 1} \right) = \ln \frac{1.7}{1.3} = 0.27.$$

$$\beta_j^1 = 1 + \ln \left(\frac{q_j + 1}{p_j + 1} \right) = 1 + \ln \frac{1.3}{1.7} = 0.73.$$

Table 3

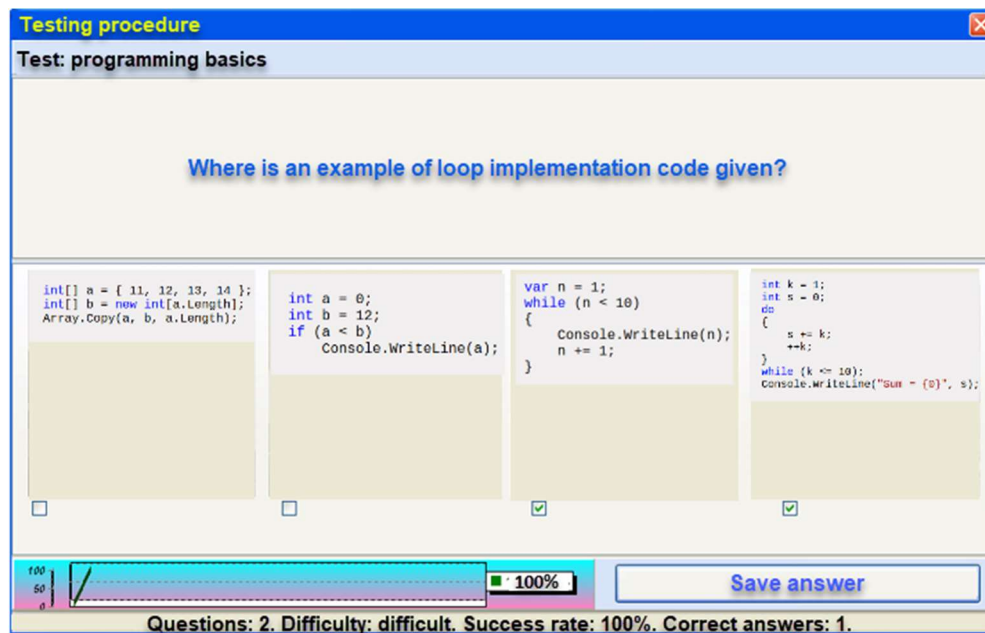
The number of test tasks in terms of difficulty

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
p_i^N	0.7	1	0.7	0.4	0	1	1	0.7	0.7	0.4	1	1	0.4	0.7	0.7	1	0.7	1	1	0.4
q_i^N	0.3	0	0.3	0.6	1	0	0	0.3	0.3	0.6	0	0	0.6	0.3	0.3	0	0.3	0	0	0.6
θ_i^o	0.27	0.69	0.27			0.7	0.7	0.3	0.3		0.7	0.7		0.3	0.3	0.7	0.3	0.7	0.7	
β_j^o	0.73	0.31	0.73			0.3	0.3	0.7	0.7		0.3	0.3		0.7	0.7	0.3	0.7	0.3	0.3	

For questions 4, 5, 10, 13, and 20, a reformulation is required to ensure that the proportion of accurate responses exceeds 0.5.

As a result of practical implementation, performance rating indicators were obtained, allowing for a sufficiently objective ranking of students on a preference scale. Test tasks requiring reformulation were identified based on task complexity balance indicators.

Using the integrated development environment Microsoft Visual Studio [18], the high-level object-oriented programming language Microsoft Visual C# [19]. The developed class library, NET, and the .NET Framework platform [20] were used to create a Windows Application that incorporates the logic for testing according to the proposed approach. Figure 2 illustrates the screen forms that demonstrate the operation of the test information system in test passing mode, as well as the process of obtaining the test result.

**Figure 2a:** Screen form demonstrating the testing procedure

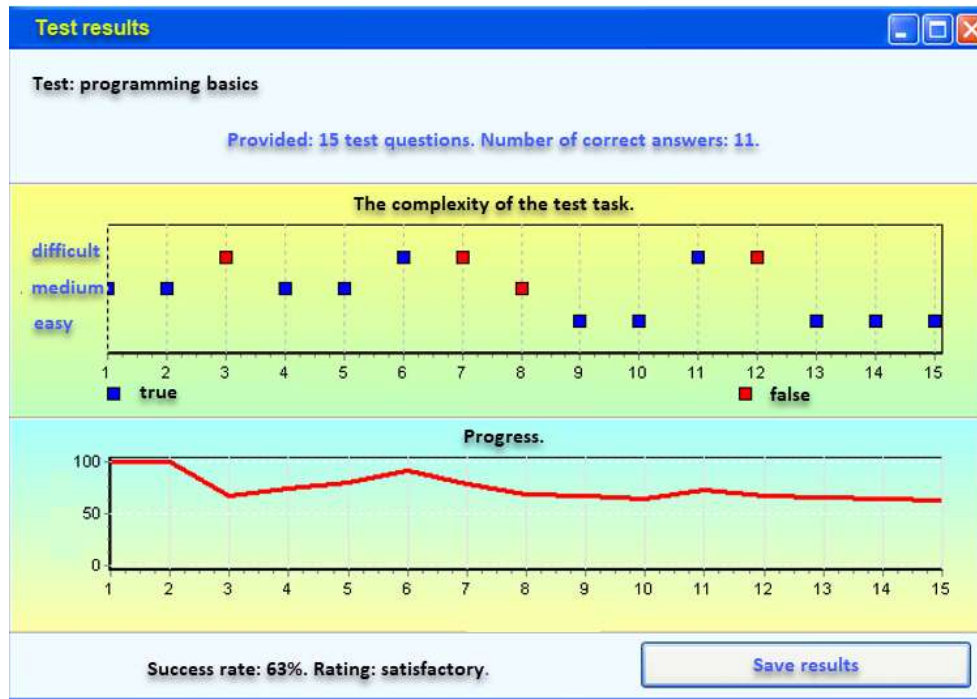


Figure 2B: Screen form for obtaining test results

Examples of modern adaptive testing systems include :

- Khan Academy: uses adaptive exercises and tests that are adjusted based on student performance [21].
- Duolingo: employs adaptive algorithms for language learning, covering grammar, vocabulary, and listening tests [22].
- ALEKS (Assessment and Learning in Knowledge Spaces) is based on the Knowledge Space Theory, analyzes the student's knowledge structure, and offers an individual learning path [23].
- Carnegie Learning Adaptive Learning Platform: platforms for creating adaptive learning modules with integrated testing [24].
- Pearson MyLab: commercial solutions for higher education featuring adaptive tests and learning tracks [25].

6. Conclusions

An approach to assessing the level of educational material assimilation by education seekers is proposed.

According to the procedure for organizing testing, the applicant's starting level of preparedness for education is average. It is proposed that three sets of uniformly selected test tasks be formed, differing in complexity. The estimated weight of the correct answer to the test task is 1; the incorrect one is 0. The method of obtaining final results involves calculating the ratio of correctly answered questions to the total number of test tasks, taking into account their complexity.

According to the testing procedure, the test task's complexity level, with which the testing will begin, is set to medium. When answering the test task, the student launches an adaptive branching algorithm by processing test tasks of different levels of complexity. If the student answers the test task correctly, the complexity level of the following test increases; if incorrectly, it decreases (flexible procedure).

In the proposed adaptive testing model, the single-parameter Rasch model will be used in the procedure for evaluating the correct answer to a test task and determining the condition for changing the test task's complexity level.

This is because it is most suitable for tests with dichotomous tasks, equal task discrimination, a small number of tasks, and a low level of guessing.

Probability of a correct answer $P_j(\theta)$ obtained during testing. If $P_j(\theta) \geq 0.5$, then the task's difficulty level must be increased or decreased, otherwise.

Experimentally selected coefficients provide the differentiating ability (discriminative power) of test tasks, the balance of which is achieved by the method of M. Chelyshkova. This enables the determination of the student's objective level of preparedness and the assessment of the quality of preparation for test tasks.

Compared to existing counterparts, the described approach enables the objective ranking of learners based on test evaluation results, offering high clarity and moderate computational complexity.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

- [1] G. Nagy, Performance Measurement and Quality Assurance in Higher Education: Application of DEA, AHP, and Bayesian Models, *Trends High. Educ.* 4(3) (2025) 54. doi:10.3390/higheredu4030054.
- [2] M. Drozdova, M. Dado, M. Malcik, E. Mehlova, Open education at universities, quo vadis, in: *Proceedings of the 2013 IEEE 11th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, IEEE, 2013, pp. 299–304. doi:10.1109/ICETA.2013.6674407.
- [3] O. Vitchenko, Problems of Realizing the Development Potential of Assessing Learners in Higher Military Educational Institutions in Different Training Systems, in: S. M. Salcutsan (Ed.), *Materials of the Scientific-Practical Seminar*, Kyiv, Oct. 28, 2020, CP "Komprint", 2020, p. 116.
- [4] B. Díaz-Lauzurica, Active Learning Methodologies for Increasing the Interest and Engagement in Computer Science Subjects in Vocational Education and Training, *Education Sciences* 15(8) (2025) 1017. doi:10.3390/educsci15081017.
- [5] S. Onyshchenko et al., Method for detection of the modified DDoS cyber attacks on a web resource of an Information and Telecommunication Network based on the use of intelligent systems, in: *Proceedings of the 6th International Workshop on Modern Data Science Technologies Workshop, MoDaST'2024*, CEUR, Lviv, Ukraine, 2024, pp. 219–235. URL: <https://ceur-ws.org/Vol-3723/paper12.pdf>.
- [6] V. Krasnobayev, A. Kuznetsov, A. Yanko, T. Kuznetsova, The data errors control in the modular number system based on the nullification procedure, in: *Proceedings of the 3rd International Workshop on Computer Modeling and Intelligent Systems, CMIS-2020*, CEUR, Zaporizhzhia, Ukraine, (2020), pp. 580–593. doi:10.32782/cmisis/2608-45.
- [7] K. Rajamanickam and V. Kathiravan, Adaptive assessment system for generating sequential test sheets using item response theory, in: *Proceedings of the International Conference on Pattern Recognition, Informatics and Mobile Engineering (ICPRIME 2013)*, Salem, India, 2013, pp. 120–124. doi:10.1109/ICPRIME.2013.6496458.
- [8] C. Nie, H. Leung and K. -Y. Cai, Adaptive Combinatorial Testing, 2013 13th International Conference on Quality Software, Naging, China, 2013, pp. 284–287, doi: 10.1109/QSIC.2013.22.
- [9] L. Na and G. Chenhua, Scientific Study on the Learning Motivation of Higher Vocational Students in Shanghai Based on SPSS 11.5, 2021 2nd International Conference on Education,

- Knowledge and Information Management (ICEKIM), Xiamen, China, 2021, pp. 75-80, doi: 10.1109/ICEKIM52309.2021.00025.
- [10] S. O. Naumenko, Foreign Experience in Using Computerized Adaptive Testing as a Subject-Oriented Tool for Evaluating Educational Outcomes of Learners, in: Impact of Innovation on Science: Fundamental and Applied Aspects, Verona, Italy, Jun. 26, 2020, vol. 2, pp. 41–45.
- [11] L. T. Hung and N. T. Ha, Experimental Research and Application of Computerized Adaptive Tests to assess Learners' Competencies, 2021 3rd International Conference on Computer Science and Technologies in Education (CSTE), Beijing, China, 2021, pp. 69-74, doi: 10.1109/CSTE53634.2021.00021.
- [12] F. B. Baker, The Basics of Item Response Theory, 2nd ed., ERIC Clearinghouse on Assessment and Evaluation, (2001). URL: <https://files.eric.ed.gov/fulltext/ED458219.pdf>.
- [13] S. Yuan, S. Zhao and H. Bai, The principle of Rasch model and compare with the other models, 2011 International Conference on Computer Science and Service System (CSSS), Nanjing, 2011, pp. 3731-3733, doi: 10.1109/CSSS.2011.5974856.
- [14] D. Triono, R. Sarno and K. R. Sungkono, Item Analysis for Examination Test in the Postgraduate Student's Selection with Classical Test Theory and Rasch Measurement Model, 2020 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2020, pp. 523-529, doi: 10.1109/iSemantic50169.2020.9234204.
- [15] S. Maslovskiy and A. Sachenko, Adaptive test system of student knowledge based on neural networks, 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Warsaw, Poland, 2015, pp. 940-944, doi: 10.1109/IDAACS.2015.7341442.
- [16] N. Albalushi and W. S. Awad, Generating Questions Bank for Adaptive Assessment Using Machine Learning Techniques: Review, 2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD), Manama, Bahrain, 2025, pp. 1-5, doi: 10.1109/ITIKD63574.2025.11004696.
- [17] Y. Choi, C. McClenen, Development of Adaptive Formative Assessment System Using Computerized Adaptive Testing and Dynamic Bayesian Networks, Applied Sciences 10(22) (2020) 8196. URL: <https://www.mdpi.com/2076-3417/10/22/8196/pdf> . doi: 10.3390/app10228196.
- [18] Microsoft Visual Studio software development tool. URL: <https://visualstudio.microsoft.com/> .
- [19] High-level programming language Microsoft Visual C#.NET. URL: <https://learn.microsoft.com/dotnet/csharp/language-reference/language-specification/> .
- [20] Microsoft .NET platform. URL: <https://dotnet.microsoft.com/> .
- [21] Khan Academy Blog. URL: <https://blog.khanacademy.org/> .
- [22] Duolingo English Test Technical Manual. URL: <https://englishtest.duolingo.com/technical-manual> .
- [23] ALEKS (Assessment and Learning in Knowledge Spaces) platform. URL: https://www.aleks.com/about_aleks .
- [24] Carnegie Learning Adaptive Learning Platform. URL: <https://www.carnegielearning.com/adaptive-learning-platform/> .
- [25] Pearson MyLab platform. URL: <https://www.pearson.com/en-us/subject-catalog/p/mylab>.