Improvement of the student evaluation system based on the ICT use

Olena Pshenychna[1,*], *Roman* Klopov[1], *Oleksandr* Gura[1], and *Tetiana* Gura[2]

[1]Zaporizhzhya National University, Zaporizhzhya, 69600, Ukraine
[2]Zaporizhzhia Regional Institute of Postgraduate Education, Zaporizhzhya, 69035, Ukraine

Abstract. Today, considerable attention is paid to the higher education quality issues. The problem is solved by using tests that should provide a reliable student evaluation. The article presents the technology for improving test tasks. It includes functional procedures that specify the test and test task improvement sequence. It is found that it is better to use specialized computer applications for their implementation, that is why this technology involves the use of the author program "Statistical Analysis of Test Results". This program calculates the indicators – the item difficulty, discrimination, reliability and validity – according to empirical student testing data. The indicators help identify unsatisfactory quality test tasks and improve the student assessment means, as the program derives the recommendations. The steps set out by the testing result processing technology with the help of a statistical package increase the improvement process efficiency. The correlation and factor analyses help identify the tasks that put the highest load into the test score. These procedures influence on making a decision on the test task review need. The technology involves repeated checking procedures. The presented technology has been tested at Zaporizhzhya National University and Zaporizhzhya Regional Institute of Postgraduate Teacher Education. ANOVA has helped prove its effectiveness.

1 Introduction

The sustainable development is associated with solving problems that humanity will face in the near future. That is why ambitious goals such as providing universal and high-quality education; creating conditions that enable children to get free, equal and high-quality secondary education; ensuring equal access for women and men to high-quality education, including the university one; facilitating the students' acquisition of knowledge and skills necessary to promote the sustainable development are set before education [1]. According to the World Education Monitoring Report, learning helps solve global environmental problems, promotes economic growth, helps overcome gender and social inequalities and is considered to be a conflict and violence prevention means.

From the standpoint of higher education, the important issues are access, accessibility and quality. Access to higher education reflects a number of indicators one of which is the university entering preparedness level. In Ukraine, the level is determined by the external independent evaluation (testing) results, so the use of high-quality tests is very important. In addition to funding, accessibility is associated with higher education enrollment of differently-abled young people, including the disabled ones. The problem is solved by introducing information and communication technologies into the educational process that will enable to create comprehensive and effective learning conditions for everyone. And the use of computer-based testing expands the disabled students' education availability.

Achieving sustainable development goals is ensured by high education quality. Nowadays the future specialist training quality problem is considered by many researchers. The standard introduction in educational institutions, the use of learning practice research data, the inclusion of alternative qualifications in training, the communication with educational centers, the external evaluation implementation are considered to be the ways to improve the future teacher training quality [2]. To improve the future English Philology Masters' education quality, the innovation introducing new educational technologies and new learning methods into the educational process is important [3]. To improve the future programmer training quality, it is proposed to modernize the content and methods of programming learning in accordance with international standards; to develop variable modules taking into account the modern labor market standards and needs; to carry out the constant future software engineer training quality monitoring at all levels; to monitor the labor market in order to determine the employers' requirements and to adjust the training content in accordance with the latter [4].

The education quality is determined on the basis of the university ranking or according to the student

[*] Corresponding author: esp961081@gmail.com

performance results. As acknowledged by the authors of the report, university ranking is not a reliable way to determine the education quality and relates to the marketing tools [1]. A more reliable criterion is the student assessment that should be transparent and understandable. The testing as a way of knowledge level assessment is an essential and integral part of the operative intermediate, stage and final learning result assessment. This once again confirms that the relevance of research in this direction is a tool to reduce the student knowledge level assessment cost.

The evaluation is carried out mostly through testing with computer programs dominantly applied for its implementation. They are required to implement testing, obtain initial information, accumulate and store students' performance data. Such programs are computer knowledge testing systems (Brainbench, INDIGO, Hot Potatoes, MyTest, OpenTEST2, TCExam, etc.) and learning management systems (Blackboard, Inkling, MOODLE, Sakai, WebTutor, etc.). Most of them provide "technological testing cycle", that is preparation of the test task bank; test development; testing; testing result report making [5].

Nowadays the use of learning result testing computer programs is considered quite actively in terms of ongoing monitoring, final assessment and qualification examinations.

In the paper titled *Introducing Computer-Based Testing in High-Stakes Exams in Higher Education: Results of a Field Experiment,* the authors presented a comparative analysis of the use of the paper-based and computer-based tests in high-stakes exams [6]. The authors drew attention to the significance of the random test forming and the importance of using computer-based tests at the intermediate learning stage. The results of their study demonstrate that most students are ready to pass high-stakes exams based on the use of computer-based tests. Their positive attitude is explained by the possibility of getting a mark after passing the test.

The possibilities of using computer-based tests in the technical drawing assessment of students are discussed in *Development of Computer-Based Tests Mode of Assessment for Technical Drafting Students* by L. Aquino. The computer-based test development was carried out in four stages: Planning Stage; Development Stage; Validation and Acceptability Stage; Final Revision Stage [7]. The computer-based tests were analyzed and evaluated by five experts in the field of technical design according to the following parameters: Utility, Accuracy, Content and Navigation. At the same time, a computer-based test was evaluated according to the criteria of preferences in use, item difficulty level, readiness for computer-based testing and fraud prevention. As a result of the research, the author concludes that computer-based tests are appropriate and acceptable for technical drawing learning result assessment.

Recently, universities have been using learning management systems to enhance real learning opportunities. The use such programs enables students to study in a convenient place at a convenient time for them that is the basis for transforming the existing higher education system into Education for You [8]. The use of such programs is an adaptation of young specialists to passing qualification examinations that are already the base for the enterprise personnel selection. The learning management system opportunities are expanding by creating mobile applications that meet the challenges of the fourth industrial revolution. The use of mobile learning management systems will allow universities to refuse from traditional learning approaches, to implement innovations, and to form effective human capital [9].

Regardless of the means chosen for testing, all of them should implement an adequate learning result assessment and ensure the effective functioning of the educational process monitoring system. In this regard, the evaluation tool quality analysis and improvement is more relevant today than ever, regardless of the tools used in its implementation, whether by using a specialized program, with the help of a statistical package, or by formula calculation in a word processor environment.

These calculations are based on the Classical Testing Theory (CTT) and Item Response Theory (IRT) provisions. In general, the IRT results are considered to be more reliable than the CTT ones [10]. However, studies showing a link between the parameters obtained through these two theories have recently been conducted.

The paper titled *Validation of a developed university placement test using classical test theory and Rasch measurement approach* [11] presents a sequential economy test analysis that was conducted by using item difficulty, discrimination, and reliability indicators. Testing data was analyzed by using Classical Testing Theory and Item Response Theory. To calculate the CTT and IRT indicators, the authors used such specialized software as ITEMAN 4.3 and WINSTEPS 3.72.3. The data obtained proved a correlation between the results processed with the two models. It is important that the paper tested the task suitability to measure the desired result.

In the source [12], the authors considered the use of the CTT and IRT models in evaluating open test tasks. In order to obtain reliable results, open-ended test tasks were evaluated by experts and by using a developed scale. The estimates obtained were compared by using two models, and the open test task item difficulty indicators were calculated. The results demonstrated a high level of correspondence between them. The methods of mathematical statistics (factor analysis, correlation analysis, Che-square criterion) that proved the correspondence of the constructed model to the real data were used in the paper.

The paper titled *Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination* [13] provides the mathematics examination result analysis by using the CTT and IRT methods. The indicators obtained by using the two theories were compared by the factor analysis methods (principal component analysis) and correlation analysis (Fisher Correction, Olkin and Pratt Correction, Point-Biserial). Factor analysis proved the unidimensionality of all the tasks included in the examination. Correlation indicators indicated the absence of discrepancies between the item difficulty and discrimination indicators calculated by the two author-selected methods. The authors have also found that the item difficulty and discrimination indicators obtained are independent of sample size: n=100 and n=1000.

This review proves that statistical calculations (descriptive statistics, correlation analysis, statistical hypothesis testing, factor analysis, variance analysis, etc.) necessary to draw conclusions are used to carry out the test and test task analysis. However, the calculations turn into a big problem for teachers unschooled in mathematical statistics, and it is better to use a specialized program for this. Of course, nowadays there are specialized programs designed for test analysis [11, 14, 15]: Iteman, Winsteps, Test_Results, Computer-based system of quality analysis of test items etc. Some of these programs are local solutions that are not available to the general public: Test_Results, Computer-based system of quality analysis of test items.

Their functionality analysis has shown that they only output test quality indicators (in numerical or graphical form), and it is more logical to provide recommendations to assessment means developers. The availability of such programs cannot be a panacea to address the problem of improving the assessment means quality for students.

Hypothesis of our study. Based on a scientific publication and pedagogical experience theoretical analysis, we assume that the use of special technology to improve test tasks will allow: to gradually create adequate and reliable tests for evaluating student learning result assessment; to constantly check their validity; to implement the procedure efficiently and simply. To this end, we have developed specialized software.

2 Methods and instruments

The study hypothesis checking was carried out by using a set of methods. To determine the indicators necessary to improve the learning result assessment means quality, the methods of scientific and methodological literature data theoretical analysis and generalization were used. The analysis of the publications allowed to determine the test quality indicators. Their calculation is based on the test theory and statistical methods.

In the process of developing the test improvement technology, series of computational procedures were carried out that made it possible to select the most effective test theory and statistical methods. They are the test and test task item difficulty determination; task discriminative ability test; test reliability and validity evaluation; correlation analysis; factor analysis, ANOVA. Computational procedures used empirical student test data (the control paper, training test, test and examination results) derived from the LMS Moodle.

In the process of an experimental work, the pedagogical experiment method that took place in vivo was used. 20 lecturers agreed to take part in it. In the process of an experimental work, the testing results of 2283 students were processed. The results were generalized that led to the test improvement technology development.

In addition to the LMS Moodle (version 3.7), the specialized author computer program “Statistical Analysis of Test Results” and the SPSS statistical package (version 20) were also used in the research.

3 Results

3.1 The test task improvement technology

As a number of studies indicate, learning management systems are quite popular nowadays [16, 17]. And the MOODLE (Modular Object-Oriented Dynamic Learning Environment) LMS is considered to be the most effective and widespread [16]. The orientation to the MOODLE LMS environment is also due to the fact that this system is widely used for the learning process didactic support in universities. The control event results are exported to a spreadsheet document (.xlsx or .ods file) that contains:

- general information about the student;
- test duration (the test start and end time and the time spent);
- test score as a whole;
- answer results for each task (task types are Multiple choice, Matching, Calculated, Short answer, Numerical, Embedded answers, Drag and drop, etc.).

We developed a technology of the assessment means improvement, which is based on the educational measurement theory. There are a number of scientifically sound criteria for the quality of the test as a whole and for the individual test tasks from which we have chosen the item difficulty, discrimination, reliability and validity [10].

The item difficulty is associated with both the individual task and the test as a whole. For example, according to the item difficulty, the tasks are divided into the most difficult, the most successful, quite simple and very simple ones. The simplest and quite simple tasks should be at the beginning and in the end of the test, and the most difficult ones should be at the center of the test. The total test item difficulty is divided into 4 levels: very high test item difficulty, the test is not balanced; the test is balanced according to the item difficulty; the test item difficulty is sufficient; the test item difficulty is bad.

The index of discrimination means the task ability to differentiate students from the better to the worse ones [18]. High discrimination is considered to be an important indicator of a successful test task. The index value is in the range of [–1; 1] and the qualitative values may be as follows: the task is functioning quite satisfactorily; a small task correction is required; the task should be reviewed; the task should be deleted.

The reliability is considered as the test result stability degree during repeated measurements [10]. That is, the test is reliable if it provides high measurement accuracy and the results are resistant to external factors.

The test must be valid. It is a characteristic that reflects its ability to get the results corresponding to the testing purpose [10].

In addition to the mentioned test quality criteria, you should also consider the time indicator: the correlation between performance and testing time. The time interval when the students made the least mistakes is determined in accordance with the testing data.

According to the pedagogical test development algorithm, the following stages are gradually carried out: the test task bank development; the testing for the task approbation purpose (the item difficulty and discrimination checking); the test forming and the second testing session conducting (the test item difficulty, reliability and validity checking); the standardization procedure implementation (the preparation of several parallel test variants, the testing time calculation) [19].

Also, after testing, important indicators that provide additional information about the test tasks are: point-biserial coefficient for each task, nominative correlation coefficients, factor and analysis of variance results [10].

The authors have developed a phased test task improvement technology: 1) the test task bank forming (LMS Moodle); 2) the probation testing using the bank tasks (LMS Moodle); 3) the discriminativity and item difficulty level determination after the probation testing ("Statistical Analysis of Test Results"); 4) based on the "Statistical Analysis of Test Results" recommendations, some test tasks are deleted from the bank, the rest are improved or remain unchanged; 5) the testing is carried out (LMS Moodle); 6) the test task item difficulty level is determined ("Statistical Analysis of Test Results"); 7) based on the "Statistical Analysis of Test Results" recommendations, the tasks are redistributed in the test; 8) the testing is carried out (LMS Moodle); 9) the test reliability and validity are checked ("Statistical Analysis of Test Results"); 10) the optimal testing time is determined ("Statistical Analysis of Test Results"); 11) based on the "Statistical Analysis of Test Results" recommendations, adjustments are made to the test, if necessary; 12) the calculation of correlation coefficients like the point-biserial and nominative one (SPSS); 13) based on the SPSS calculation results, the tasks that should be deleted are determined; 14) the factor analysis implementation (SPSS); 15) based on the SPSS calculation results, the tasks that are the most significant to get an objective assessment are determined, adjustments are made, if necessary; 16) the testing is carried out (LMS Moodle), and empirical data are accumulated; 17) the ANOVA implementation to compare the student test results over several years (SPSS); 18) based on the SPSS calculation results, the final decision is made on the test effectiveness.

The technology is represented in the model (Fig. 1).

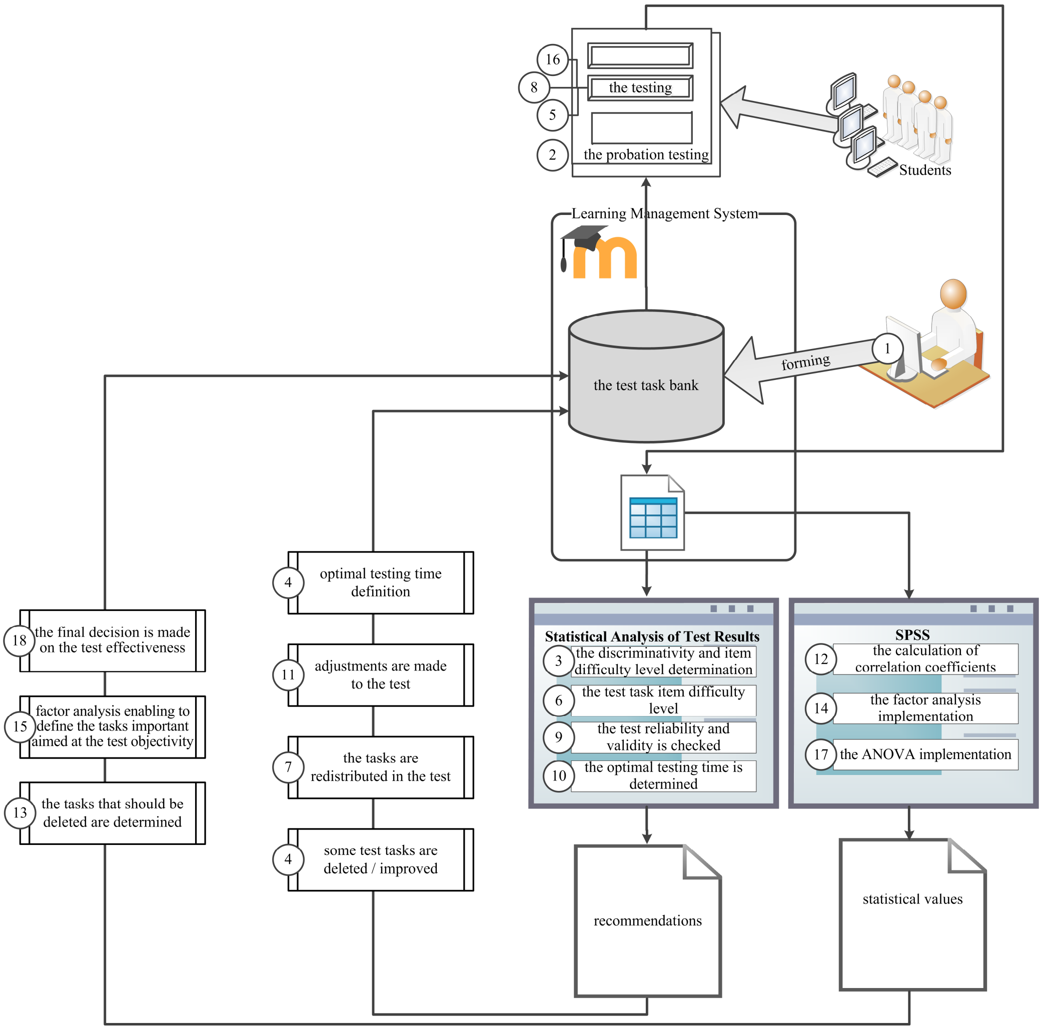

16
8
5
2
the testing
the probation testing
Students
Learning Management System
1
forming
the test task bank
Statistical Analysis of Test Results
3 the discriminativity and item difficulty level determination
6 the test task item difficulty level
9 the test reliability and validity is checked
10 the optimal testing time is determined
SPSS
12 the calculation of correlation coefficients
14 the factor analysis implementation
17 the ANOVA implementation
recommendations
statistical values
4 optimal testing time definition
11 adjustments are made to the test
7 the tasks are redistributed in the test
4 some test tasks are deleted / improved
18 the final decision is made on the test effectiveness
15 factor analysis enabling to define the tasks important aimed at the test objectivity
13 the tasks that should be deleted are determined

Fig. 1. The test task improvement technology.

3.2 Specialized computer program “Statistical Analysis of Test Results”

The “Statistical Analysis of Test Results” is a base for the introduced technology assessment means improvement, so let’s take a look at this specialized computer program The C# programming language in Microsoft Visual Studio 2017 and the Windows Presentation Foundation technology have been selected for program implementation. When choosing the development means, we were guided by the following considerations: a convenient form designer and powerful means for working with arrays; the universal interface provides an integrated design and application component implementation.

The work with the Statistical analysis of students’ test results software starts from the main window that is organized on the basis of pressing the buttons opening the corresponding system modules (Fig. 2).

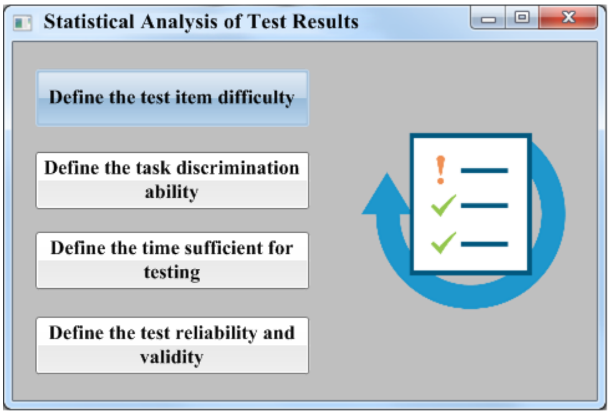
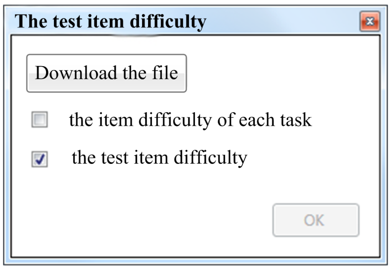

Statistical Analysis of Test Results
Define the test item difficulty
Define the task discrimination ability
Define the time sufficient for testing
Define the test reliability and validity

Fig. 2. The “Statistical Analysis of Test Results” software main window.

The clicking of the [Define the test item difficulty] button opens The Test item difficulty dialogue window (Fig. 3). The system is focused on the testing results in the LMS MOODLE, so it provides downloading files with these results (the [Download the file] button). You can get:
- the item difficulty of each task;
- the test item difficulty;
- the item difficulty of each task and the test item difficulty.

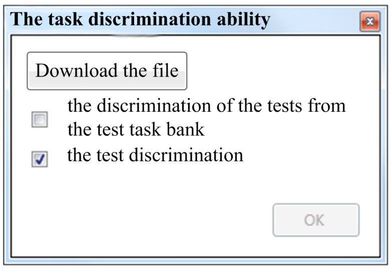


□ the item difficulty of each task
☑ the test item difficulty

Fig. 3. The test item difficulty dialogue window

The results from the downloaded file are transferred to a dichotomous matrix, the initial test indicators are calculated: the ratios of correct and incorrect answers; if there is the item difficulty of each task checkbox, the variance is calculated; if there is the test item difficulty checkbox, the average task item difficulty level is calculated. After that, a window showing the test task and whole test item difficulty checking results is displayed. The numerical value of the item difficulty indicator and its level are derived for the test. The item difficulty level is determined for each test task.

The work of the task discrimination ability module (Fig. 4) helps determine the task discrimination ability of one test or recommendations on test tasks from the test task bank. That is why, the user can choose only one of the checkboxes after downloading the result file: the discrimination of the tests from the test task bank or the test discrimination.



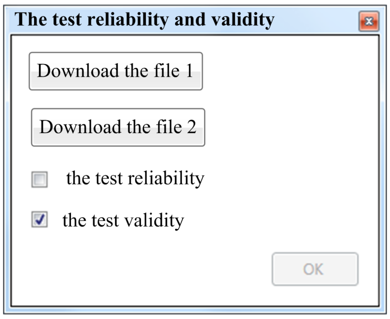
□ the discrimination of the tests from the test task bank
☑ the test discrimination

Fig. 4. The task discrimination ability dialogue window

It is envisaged that the results are displayed in groups after the discrimination checking of all the bank tasks: 1) at first, the tasks functioning satisfactorily are listed; 2) then, a list of those ones requiring a small correction is displayed; 3) next, there is a list of test tasks that should be reviewed; 4) at the end, there are the tasks that should be deleted. To do this, the test task bank statistics file is downloaded and the discrimination of the tests from the test task bank checkbox is put, the tasks are grouped according to discrimination indicators.

The test task discrimination checking derives recommendations for each test task. This is necessary when the test is generated by bypassing the test task bank. To do this, the index of discrimination is calculated for each task and a recommendation for each test task is derived according to the numerical value.

After clicking the [Define the test reliability and validity] button, the test reliability and validity dialogue window opens (Fig. 5).



□ the test reliability
☑ the test validity

Fig. 5. The test reliability and validity dialogue window

After downloading two files, the user starts the process of calculating the main test indicators – reliability and validity. After pressing the [OK] button:

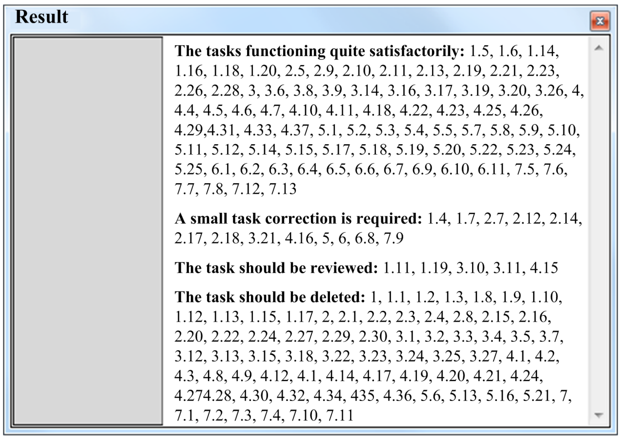
the program checks the test reliability checkbox and calculates the reliability indicator and derives a qualitative reliability characteristic; the program checks the test validity checkbox, calculates the validity indicator and derives a qualitative validity characteristic. These indicators can be obtained individually or together with two downloaded files.

A feature of the “Statistical Analysis of Test Results system is that it derives not numerical values but qualitative characteristics of the test and its tasks. This is convenient because the teacher does not need to analyze numerical values, define the item difficulty, discrimination, reliability and validity level and make decisions about the test and its tasks.

3.3 Test improvement technology implementation

The assessment means improvement technology used at Zaporizhzhya National University in the course of current, final and rectorial control, and also tested at Zaporizhzhya Regional Institute of Postgraduate Teacher Education during the special course and training “The basics of testology and student computer-based testing”.

After the development of test tasks, the approbation testing is carried out. According to its results, the task item difficulty and discrimination are checked using the “Statistical Analysis of Test Results” software. The data help identify the tasks that need to be improved or deleted (Fig. 6).



The tasks functioning quite satisfactorily: 1.5, 1.6, 1.14, 1.16, 1.18, 1.20, 2.5, 2.9, 2.10, 2.11, 2.13, 2.19, 2.21, 2.23, 2.26, 2.28, 3, 3.6, 3.8, 3.9, 3.14, 3.16, 3.17, 3.19, 3.20, 3.26, 4, 4.4, 4.5, 4.6, 4.7, 4.10, 4.11, 4.18, 4.22, 4.23, 4.25, 4.26, 4.29,4.31, 4.33, 4.37, 5.1, 5.2, 5.3, 5.4, 5.5, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12, 5.14, 5.15, 5.17, 5.18, 5.19, 5.20, 5.22, 5.23, 5.24, 5.25, 6.1, 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 6.9, 6.10, 6.11, 7.5, 7.6, 7.7, 7.8, 7.12, 7.13

A small task correction is required: 1.4, 1.7, 2.7, 2.12, 2.14, 2.17, 2.18, 3.21, 4.16, 5, 6, 6.8, 7.9

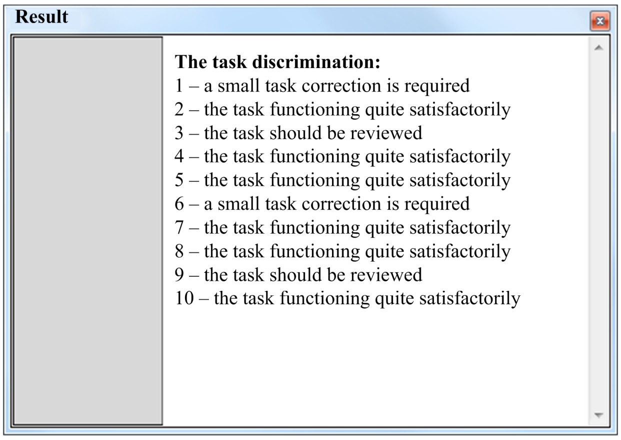
The task should be reviewed: 1.11, 1.19, 3.10, 3.11, 4.15

The task should be deleted: 1, 1.1, 1.2, 1.3, 1.8, 1.9, 1.10, 1.12, 1.13, 1.15, 1.17, 2, 2.1, 2.2, 2.3, 2.4, 2.8, 2.15, 2.16, 2.20, 2.22, 2.24, 2.27, 2.29, 2.30, 3.1, 3.2, 3.3, 3.4, 3.5, 3.7, 3.12, 3.13, 3.15, 3.18, 3.22, 3.23, 3.24, 3.25, 3.27, 4.1, 4.2, 4.3, 4.8, 4.9, 4.12, 4.1, 4.14, 4.17, 4.19, 4.20, 4.21, 4.24, 4.274.28, 4.30, 4.32, 4.34, 435, 4.36, 5.6, 5.13, 5.16, 5.21, 7, 7.1, 7.2, 7.3, 7.4, 7.10, 7.11

Fig. 6. The (bank) test task discrimination checking data

It should be noted that the process is sufficiently long and lasts all the time during which teachers use the testing. In addition, the knowledge and skill level of students of different study years is still different, so the system provides test task discrimination checking (Fig. 7).

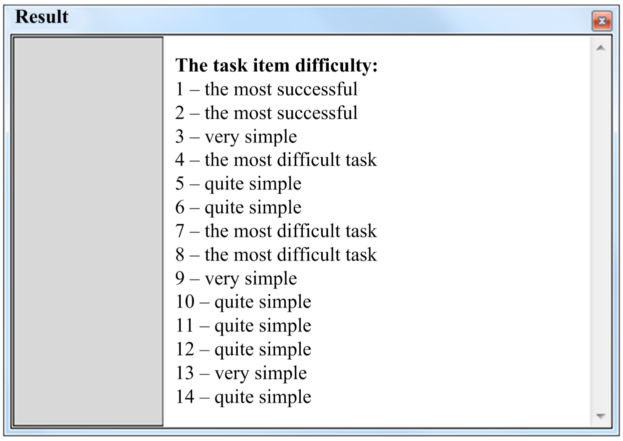
According to the of educational measurement specialists’ recommendations, the test should include 20% of the most difficult tasks, 20% of very simple and quite simple tasks, other tasks should be the most successful [10]. The test task distribution should be as follows: the simplest and quite simple ones should be at the beginning and in the end of the test, and the most difficult ones should be in the center of the test, unless the test mode involves the task randomization.



The task discrimination:
1 – a small task correction is required
2 – the task functioning quite satisfactorily
3 – the task should be reviewed
4 – the task functioning quite satisfactorily
5 – the task functioning quite satisfactorily
6 – a small task correction is required
7 – the task functioning quite satisfactorily
8 – the task functioning quite satisfactorily
9 – the task should be reviewed
10 – the task functioning quite satisfactorily

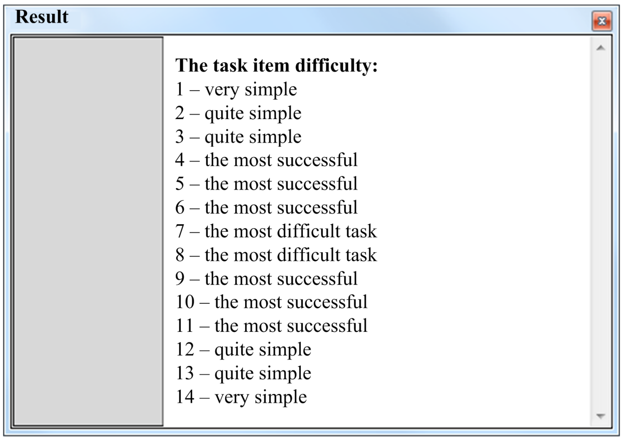
Fig. 7. The test task discrimination checking results

The Fig. 8 presents the test task item difficulty checking results. From these data it is clear that the test tasks are placed not in a balanced way there. After the task improvement and redistribution according to the item difficulty, an optimal distribution was obtained (Fig. 9). According to our observations, the simple ones were mostly the closed tasks (multiple choice and conformity) and the most difficult ones were the built-in answers.



The task item difficulty:
1 – the most successful
2 – the most successful
3 – very simple
4 – the most difficult task
5 – quite simple
6 – quite simple
7 – the most difficult task
8 – the most difficult task
9 – very simple
10 – quite simple
11 – quite simple
12 – quite simple
13 – very simple
14 – quite simple

Fig. 8. The test task item difficulty checking results (version 1)

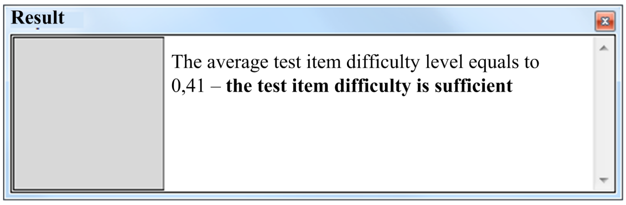


The task item difficulty:
1 – very simple
2 – quite simple
3 – quite simple
4 – the most successful
5 – the most successful
6 – the most successful
7 – the most difficult task
8 – the most difficult task
9 – the most successful
10 – the most successful
11 – the most successful
12 – quite simple
13 – quite simple
14 – very simple

Fig. 9. The test task item difficulty checking results (version 2)

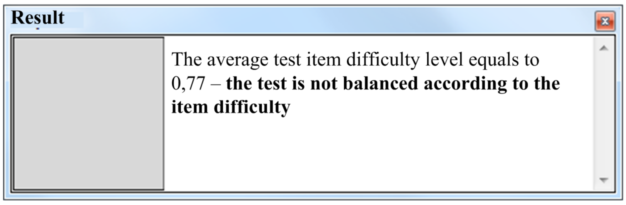
The item difficulty, reliability and validity indicators helping evaluate the test quality are calculated for the tests. The developed tests are repeatedly used in the higher education institution educational process, often the final control (credit or examination) is carried out with the help of them. There is also the practice of using a pilot test, through which students conduct the test preparation self-monitoring.

The results of any test are processed and a level of difficulty is obtained. The teacher can continue the task improvement, add more or less difficult tasks if the test item difficulty is bad, the test item difficulty is sufficient (Fig. 10) or the test is not balanced (Fig. 11).



The average test item difficulty level equals to 0,41 – **the test item difficulty is sufficient**

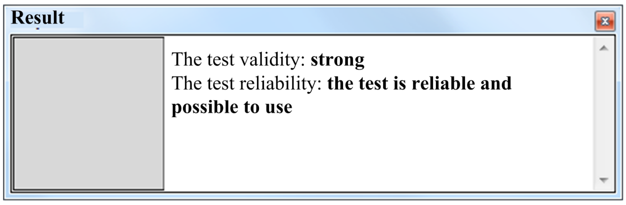
Fig. 10. The test item difficulty checking variants (version 1)



The average test item difficulty level equals to 0,77 – **the test is not balanced according to the item difficulty**

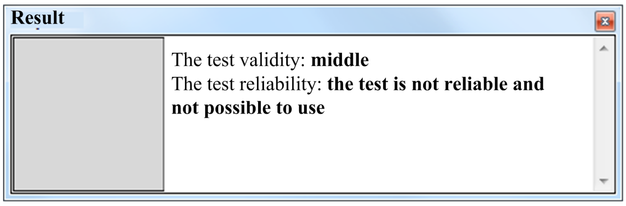
Fig. 11. The test item difficulty checking variants (version 2)

The reliability checking is performed according to two parallel testings (the pilot and control one), and the validity checking is also based on the control and final work results. The Fig. 12 and Fig. 13 show two sufficiently divergent variants of the reliability and validity test checking. An unsatisfactory test validity or reliability is a signal to the task change.



The test validity: **strong**
The test reliability: **the test is reliable and possible to use**

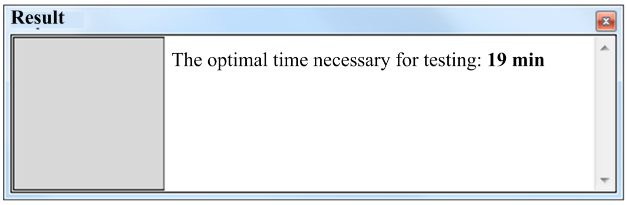
Fig. 12. The test reliability and validity checking variants (version 1)



The test validity: **middle**
The test reliability: **the test is not reliable and not possible to use**

Fig. 13. The test reliability and validity checking variants (version 2)

As noted above, an important problem of testing is the time allotted for it. The disadvantages are both the insufficient amount of time and its excess. In this regard, the developed program defines the optimal time to pass the appropriate test according to the testing results (Fig. 14).



The optimal time necessary for testing: **19 min**

Fig. 14. The definition of the optimal time necessary for testing

The calculation of the point-biserial correlation for each task helped check the task differentiation (Table 1). Since all indicators are greater than 0,2, all the tasks differentiate students well.

Table 1. The calculated point-biserial correlations

Task	Test 1	Test 2
1	0,202	0,213
2	0,353	0,654
3	0,372	0,679
4	0,234	0,619
…	…	…
19	0,514	0,235
20	0,505	0,581

It was found in the process of obtaining complexity indicators that the closed tests (Multiple choice and Matching) are among the simplest ones according to the item difficulty. The point-biserial correlation also proved this.

A part of the lecturers does not go beyond the theoretical closed tasks (Multiple choice and Matching) when developing tests, therefore, the research of two tests from the same discipline was conducted by using a factor analysis (the same students took the test). One test included solely theoretical tasks, and another one open-ended tasks of different types in addition to the former ones. The first test included tasks identical to the second test tasks: the task 1_1 was identical to the task 2_1, the task 1_2 was identical to the task 2_5, the task 1_3 was identical to the task 2_4. The factor analysis results showed the following: the factor 1 (informativeness of 24,7%) included all the open tasks; the identical tasks were in the same factor (2, 3 or 4) in pairs. Therefore, open test tasks put a higher load into the test score.

The variance analysis results also proved that the test improvement factor inclusion contributed to a more adequate assessment. The score of students who were tested by improving means was lower than that the one of the groups that was tested by non-improving tasks. The test improvement factor had a significant impact on the assessment adequacy.

428 tests (37,7%) for different disciplines (higher mathematics, computer science, programming, pedagogy, economics) used to evaluate the students at Zaporizhzhya National University and to train teachers at Zaporizhzhya Regional Institute of Postgraduate Teacher Education were analyzed by using the presented technology. This program was used to check 1375 (32,9%) tests for item difficulty and discrimination that

in turn helped improve them. As a result, it was found that 63,7% of the test tasks functioned quite satisfactorily while others required a small correction according to the discrimination; 66,9% of the tasks had a sufficient test item difficulty level, and 31,1% of the tasks needed to be balanced according to the item difficulty (the item difficulty is very high or not sufficient); 82,3% of the tests were reliable; 74,9% of the tests showed high and medium validity levels.

Table 2. Factor structure of 10 tasks

Component matrix[a]				
task	Factor			
	1	2	3	4
1_1				,807
1_2		,903		
1_3			,833	
2_1				,891
2_2	,610			
2_3	,592			
2_4	,446			
2_5		,912		
2_6			,874	
2_7	,870			
Definition method: Principal component analysis.				
a. Defined components: 4				

4 Conclusions

So, the need to improve the future specialist training quality is based on the effective higher education system. It should not only create conditions, but also have reliable tools for the student knowledge level assessment.

The effective future specialist training system functioning depends largely on the perfection and quality of the assessment means, the most common of which are tests. Tests must meet the requirements for the item difficulty, discrimination, reliability and validity indicators. A study of the formulas used to make the calculations showed that a computer program could be an effective solution to the test quality checking problem. The paper presents a specialized computer program “Statistical Analysis of Test Results” that consists of four independent modules and derives the qualitative characteristics of the indicators involving the basis for making a decision on the need for test task improvement, as well as to define the optimal testing time. Fourthly, a special procedure to increase the test quality helping improve the means is needed. For this purpose, special indicators are applied: item difficulty, discrimination, reliability and validity. This procedure is presented in the form of a special technology that includes testing in the LMS MOODLE environment, calculating the main test quality indicators by using a specialized author program and statistical processing of empirical data with the help of the SPSS program environment. The results of the assessment means improvement program and technology approbation in the process of testing the applicants of the Zaporizhzhya National University and the postgraduate education system students proved their effectiveness.

The test improvement technology introduction has let make the tests transparent and objective. Such a test improvement will improve the future specialist training quality. Then it can be expected that in the future they will be able to think critically, generate creative ideas, make original decisions, strive to ensure the global environmental safety, economic prosperity, justice and equality.

It is possible to choose such developed program improvement directions as the test task distractor analysis, optimal the test length determination, the test task calibration for further research. It is also desirable to introduce a special course on educational measurements for students in pedagogical disciplines and practicing teachers.

References

1. *Education for people and planet: creating sustainable futures for all* (UNESCO, 2016). https://unesdoc.unesco.org/ark:/48223/pf0000245752. Accessed 07 Feb 2020
2. D. Vaillant, J. Manso, J. Manso, Teacher education programs: learning from worldwide inspiring experiences. JoSPoE **1**, 94 (2013). https://revistas.uam.es/index.php/jospoe/article/view/5622/6036. Accessed 26 Jan 2020
3. H. Pyatakova, N. Ratushnyak, in *Sustainable Education as a Way of Bringing People Together – Multiple Stories From Europe Editors*, ed. by V. Haluzyak, R. Kucha, A. Vykhrusch (Studia-Monografie, Lodz-Warshawa, 2018)
4. V. Kruglyk, V. Osadchyi, Developing Competency in Programming among Future Software Engineers. Integration of Education **23**, 587–606 (2019). doi:10.15507/1991-9468.097.023.201904.587-606
5. I.E. Bulakh, M.R. Mruga, *Stvoryuyemo yakisnij test* (Creating a qualitative test). (Majster-klas, Kiev, 2009).
6. A.J. Boevė, R.R. Meijer, C.J. Albers, Y. Beetsma, R.J. Bosker, Introducing Computer-Based Testing in High-Stakes Exams in Higher Education: Results of a Field Experiment. PLoS ONE **10**(12). doi:10.1371/journal.pone.0143616
7. L. Rodolfo, Aquino Development of Computer-Based Tests Mode of Assessment for Technical Drafting Students. Journal of Advanced Studies **1**(1), (2018). http://psurj.org/wp-content/uploads/2018/12/JAS-003.pdf. Accessed 20 Jan 2020.
8. G. Suganya, A Study on Challenges before Higher Education in the Emerging Fourth Industrial Revolution. IJETSR **4**, 3 (2017). http://www.ijetsr.com/images/short_pdf/1507295641_1-3-cdac833_ijetsr.pdf. Accessed 17 Jan 2020
9. K. Schwab, *The Fourth Industrial Revolution: what it means, how to respond*, (Foreign Affairs, 2015),

https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution, Accessed 07 Feb 2020

10. L. Crocker, J. Algina, *Introduction to Classical and Modern Test Theory* (Mason, 2008).
11. A.A. Bichi, R. Talib, N.A. Atan, H. Ibrahim, S.M. Yusof, Validation of a developed university placement test using classical test theory and Rasch measurement approach. IJAA **6**(6), 22 (2019). doi:10.21833/ijaas.2019.06.004. Accessed 12 Feb 2020
12. M. Ilhan, N. Guler, A Comparison of Difficulty Indices Calculated for Open-Ended Items According to Classical Test Theory and Many Facet Rasch Model. Eurasian Journal of Educational Research **18**, 99 (2018). doi:10.14689/ejer.2018.75.6
13. O.A. Awopeju, E.R.I. Afolabi, Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. European Scientific Journal **12**(28), 263 (2016). doi:10.19044/esj.2016.v12n28p263
14. B.E. Starichenko, M.G. Gizatullin, E.A. Istomina, Assessment of the level of readiness and quality of test materials using the online form Test_Results. Pedagogical Education in Russia **7**, 104 (2016)
15. O. Dykhovychnyi, A. Dudko, Computer-based Analysis System of Results of Online Testing in Higher Mathematics. Scientific works. Series: "Pedagogy, Psychology and Sociology" **2**, 103, (2013)
16. A.I. Belous, A.I. Kupalov, Comparative Analysis of Modern Distance Learning Systems. Vestnik of Moscow City University. Series "Informatics and Informatization of Education" **3**, 85 (2019)
17. D. Borboa, M. Joseph, D. Spake, A. Yazdanparast, Perceptions and Use of Learning Management System Tools and other Technologies in Higher Education: A Preliminary Analysis. Journal of Learning in Higher Education **10**, 2, 17 (2017)
18. R.L. Ebel, D.A. Frisbie, *Essentials of Educational Measurement* (Englewood Cliffs, New-Delhi, 1991)
19. N.F. Efremova, *Testovyj kontrol v obrazovanii* (Test Control in Education). (Logos, Moscow, 2007).