Testing technologies in education. The problem of test quality

Bakhrushin V.E. and Gorban A.N.

Classic Private University, 70b Zhukovsky St., 69002 Zaporizhzhia, Ukraine
Vladimir.Bakhrushin@zhu.edu.ua, gorban@education.zp.ua

Received: 21.03.2011

Abstract. The article deals with the problems related to quality of the tests applied in education, in particular with regard to the External Independent Evaluation recently introduced in Ukraine.

Keywords: education, testing technology

UDC: 378

1. Introduction

Recently, testing technologies have become more prevalent in education. Beginning from 2008, Ukraine has introduced the External Independent Evaluation (EIE) of university entrants. Many Ukrainian universities have been implementing a practice of testing as one of the main tools for both intermediate and final evaluation of learning results. For instance, such tools have been extensively implemented at the Classic Private University (Zaporizhzhia, Ukraine) since 2003. Today, only the testing portal of the Classic Private University contains the tests for final checkup of knowledge on over 900 academic disciplines. A large amount of tests for the intermediate and final controls is also available on the website of the Classic Private University assigned for supporting educational programs.

At first, there has been an impression that the testing technologies introduced in Ukraine would help solve a lot of problems, or at least those related to education quality, evaluation objectivity, corruption, etc. But soon enough, it has become clear that, like any other tool, the testing technologies have a limited area of application. In spite of some advantages over other means of assessing, they also have essential drawbacks.

The aim of this paper is to analyse some problems related to application of the testing technologies in Ukrainian education, especially the problems of design and quality of educational tests.

2. Prerequisites for application of testing technologies

The history of educational testing is several thousand years old [1, 2]. Testing technologies for knowledge assessment have been used in ancient China, Babylon and Greece. A famous English scientist Francis Halton is considered to be the founder of the modern testing [3], while the works by Alfred Binet and Theodore Simon, who have developed tests for selecting students with developmental disabilities for the specialised schools of the Ministry of Education of France, have marked the beginning of the educational testing [4, 5].

The fundamentals of pedagogical theory of testing have been worked out in detail in the monographs by V. S. Avanesov [6], L. Ya. Aschepkova [7], V. S. Kim [2] and M. B. Tchelyshkova [8], as well as in numerous publications in the journals such as “Herald of Testing and Monitoring in Education”, “Information Technologies in Education”, “Information Technologies and Learning Tools” and “Pedagogical Measurements”.

There are different definitions of the educational tests. In particular, according to the definition given by V. S. Avanesov, "Pedagogical test is a system of parallel tasks of increasing complexity and specific form, which enables high-quality, efficient evaluating of the level and structure of students' knowledge" [9]. By V. S. Kim's definition [2], "Pedagogical test is a system of tasks of different complexity that allows high-quality, efficient measure of the level and structure of students' knowledge".

It follows from these definitions that the pedagogical test is a certain model of knowledge supplemented with tools for checking the relevance of specific student's knowledge to this model. However, any model is a target, approximate, and incomplete reflection of the original [10]. Among the other factors, the reflection inaccuracy in the case of testing is caused by statistical nature of methods for assessing the test results. Hence we are to make the following conclusions:

1. Different tests are required in order to achieve different goals.
2. A good test must be developed according to certain rules that ensure its sufficient quality and, in particular, provide for verification of this quality.
3. Even a good test yields the results suitable for a certain 'average' student only, whereas the testing results concerned with particular students may be essentially wrong.

In order to check the quality of tests, such characteristics as reliability and validity are usually used [2, 6–8]. The test reliability is defined as the correlation of results obtained after several attempts or after passing its equivalent (parallel) forms, as well as the correlation of results obtained for different parts of a given test (in the latter case, the complexity of tasks should be uniform). Various kinds of correlation coefficients could be used for this purpose [2]. The reliability is considered to be sufficient if the relevant quantitative index is not less than 0.8. In many cases, a so-called Cronbach *α*, which does not assume a necessity for repeating the test, is taken as a reliability index [11]. It may be determined with the formula

$$\alpha = \frac{N}{N-1}\frac{\sigma_X^2 - \sum_{i=1}^{N}\sigma_{Y_i}^2}{\sigma_X^2}, \tag{1}$$

where N denotes the number of testing tasks, σ_X^2 the variance of the final test score, and $\sigma_{Y_i}^2$ the variance of scores obtained by the tested for the ith task.

The validity of tests is defined as a correspondence of their results to those of the other independent evaluations of students' knowledge (or some other properties under checking).

In addition, it is necessary to know the quality of separate tasks when developing the tests. The following indicators of their quality are most important:

1. The level of complexity p_i, which is determined by

$$p_i = \frac{\sum_{j=1}^{m} Y_{ij}k_{ij}}{kY_{i\,\max}}, \tag{2}$$

where k means the number of persons tested, m the number of answers on the ith task, Y_{ij} the score for the jth answer on the ith task, k_{ij} the number of persons under test who selected the j-th answer on the ith task, and $Y_{i\,\max}$ the maximum score for the ith task (in the most common case when the scores for the answers can take only the values 0 and 1, it is equal to arithmetical average of all the scores on the given task).

2. The correlation coefficient R_i, which shows how much the successful accomplishment of a given task and the total student's score for the whole test correlate with each other. It is determined as [12]

$$R_i = \frac{\dfrac{\sum\limits_{l=1}^{k} Y_{il} x_l}{k} - \overline{Y_i}\,\overline{x}}{S_x S_{Y_i}} \frac{k}{k-1}\,; \tag{3}$$

where $Y_{i\ell}$ is the score of the ℓth person under test for the ith task, x_ℓ the total score of the ℓth person tested, $\overline{Y_i}$ the average score for the ith task, $\overline{x}$ the average score for the whole test, and S_x, S_{Y_i} the standard deviations of the corresponding indices.

3. The discrimination coefficient (index) I_i, which indicates to what extent the results of a given task can be used for differentiating between the groups of students which have successfully (or unsuccessfully) fulfilled the test as a whole (for its calculation 1/3 of the best and 1/3 of the worst final score results are usually taken).

After having determined these characteristics, one needs to remove too complicated ($p_i < 0.2$) or too easy ($p_i > 0.9$) tasks, together with those characterised by too low correlation coefficients [2, 6–9] (different authors suggest choosing the boundary value in the region of 0.15–0.40). In our opinion, the boundary 0.3 is more reasonable, because it better corresponds to the boundary value for the correlation coefficient generally applied in statistics [13]. Lower boundary values often used are caused by the complexity of high-quality tests preparation rather than a significance of the corresponding tasks. Overly complex tasks may be not removed completely, though in this case the score for the correct answer should take account of complexity of the task.

Unfortunately, we have to mention that neither the educational tests on the national level (the EIE ones) nor those practised on the level of separate educational institutions are being inspected with respect to the above criteria of the test quality, or at least the corresponding results of their verification are not used for further improving the tests. Usually this is explained by a necessity for preserving confidentiality (for the EIE tests), or simply by lack of time or other resources. However, a natural question then arises: is the practical utilisation of such testing results advisable at all?

3. Quality analysis for the EIE tests

In this section we review some results of the EIE-2009 and EIE-2010 campaigns. These tests can be analysed and deserve a careful analysis from many points of view, in particular issuing from the availability of official statistics, large volumes of samples that increase the accuracy of statistical indicators. Moreover, the availability of specialised nation-wide testing centre should eventually ensure compliance with the science-based procedures for designing the tests. The source data were taken from the official reports [14, 15] published on the website of the Ukrainian Center for Educational Quality Evaluation (http://www.testportal.gov.ua).

For obvious reasons the tests validity was not determined. To do this, it would have been necessary to test preliminarily large representative groups of entrants, with simultaneous independent evaluation of their knowledge level by specialists. In terms of preserving the tests secrecy, such an approach is unacceptable. The other way is to analyse the correlation between the test results and the secondary-school marks. This would require considerable costs due to necessity of introducing secondary-school marks into an appropriate database. In addition, the results of this analysis would be notably distorted since the secondary-school ratings are strongly affected by relative ranking of pupils at separate schools. Nonetheless, we are to note that the available information [16, 17] about weak or moderate (on the level of 0.28–0.50) correlation between the EIE results and the first-year-student marks is an indirect evidence of imperfection of the corresponding tests.

The reliability of the EIE tests has been estimated using the Cronbach coefficient α given by Eq. (1). For the EIE-2010 tests, it varies within 0.80–0.96. The values less than 0.9 have been obtained for the following tests: Geography – 0.80, Biology – 0.81, History of Ukraine (the first session) – 0.83, History of Ukraine (the second session) – 0.84, and Physics – 0.86. According to general requirements of the testing theory, this index must not be less than 0.7–0.9 [6–8] (notice that different authors give somewhat different boundary values). At the same time, we emphasise that, according to the results of modelling for the answer distribution basing on the Cronbach coefficient performed under certain conditions, it can be rather high (0.95–0.98), even for high enough levels of randomness of the answers. This is why its high values cannot be considered as sufficient for concluding that the tests under analysis are good enough. The latter should require taking the other indicators of reliability into consideration. Unfortunately, the details concerned with the calculations of the EIE results are missing in the EIE reports [14, 15].

The fractions of very difficult and very easy tasks in the EIE-2010 tests vary from 7 to 23%. Namely, the lowest one (7%) is peculiar for the tests in the History of Ukraine (the first session), and the largest fraction (23%) for the tests in Physics, whereas the tasks with the open answers in Physics and Mathematics may be regarded as the most difficult of all.

It should be stressed that the high fraction of tests “optimal” in their complexity (at least from the viewpoint of reports by the author of Refs. [14, 15]) is largely caused by the high fraction of tasks with a choice of answers. As a result, even if the data indicates a random choice of answers on a certain task (the choice frequencies are approximately the same for all answers), the task appears to be optimal from a formal point of view only. Such an error occurs because the authors of the tests and the above reports have not adjusted the final results and the corresponding scores so as to correct them on the probability of simple guessing, as required by the theory of testing.

As an example, let us consider the fractions of different answer variants for the 45th task of the tests in History of Ukraine: “Liquidation of the Economic Councils in the second half of 1960s has led to”: (1) “Liberation of economic rights of Republics” – 30.33%, (2) “Strengthening of planning centralisation” – 21.78%, (3) “Liquidation of directive planning” – 20.22%, and (4) “Formation of free economic zones” – 27.45%. Hence, here the fraction of choosing the answers is close to 25%. This example shows that the fraction of entrants who really knew the correct answer to this question was close to zero. Of course, a somewhat different situation often occurs when the fraction of entrants that have chosen the correct answer is somewhat larger. For instance, let us examine the fractions of different variants for the 39th task of the EIE-2010 tests in History of Ukraine (the second session). The corresponding question is “What measures directed at sovietisation of the Western Ukraine in 1939–1941 were positively accepted by the Ukrainian population?” The possible answers are: “(1) Political repressions”, “(2) Collectivisation of peasant farms”, “(3) Nationalisation of trade and industry”, “(4) Creating a social welfare system”, “(5) Prohibition of political parties and public associations”, and “(6) Liquidation of Polish and Romanian state apparatus”. The following variants of answers have been proposed to assessing: A – (2), (4), and (5); B – (1), (3), and (6); C – (1), (2), and (5), and D – (3), (4), and (6). Finally, the resulting distribution of the variants of answers is as follows: A – 17.65, B – 11.45, C – 14.12, and D – 56.63% (notice that the fourth answer, D, is correct). However, one can see that the percentage of those who chose their answer by guessing was about 15%. As a consequence, the number of those who really knew the correct answer to this task was essentially lower than 56.63% (namely, approximately 40–45%).

As already mentioned above, another important test quality index is given by the correlation coefficient. The analysis of the reports [14, 15] testifies that, among the EIE-2009 and EIE-2010 tasks, the fraction of those revealing low (or even negative) correlation coefficients (see Eq. (3)) is very high.

The data displayed in Table 1 gives strong reasons to doubt the adequacy of the EIE-2009 and EIE-2010 results. Indeed, it is hard to understand the subject being characterised by the final tests score, since the latter does not correlate with the results of majority of its individual tasks, at least for some of the tests (e.g., History of Ukraine, Biology, and Geography-2010).

The character of the final score distribution can also be considered as an indirect indicator of the tests quality. The analyses of the reports [14, 15] on the EIE results and the results presented in [18, 19] show that the distributions of the primary scores for various disciplines are of very different types and can deviate notably from the normal distribution. In some cases, there occurs separation into several subgroups. Fig. 1 shows some examples of histograms illustrating distributions of the EIE-2010 primary test scores, which are taken from the report [15].

Character of the distribution depends not only on the quality of the test itself but also on the objective fluctuations of pupils' knowledge level. However, the manner of processing of the testing results has to take the latter into account. In addition, the methods for evaluation of test quality are in many cases based on certain assumptions about the type of the distribution (usually assuming normal or, at least, homogeneous distributions) and so should be corrected in cases when these assumptions are not valid.

Table 1. Relative fractions of tasks with the correlation coefficient $R_i \le 0.3$.

Discipline	EIE-2009	EIE-2010
Ukrainian Language and Literature	0.45	0.28
Mathematics	0.03	0.11
History of Ukraine	0.59	0.77
Physics	0.43	0.24
Chemistry	0.25	0.27
Biology	0.63	0.67
Geography	0.47	0.78
English	0.38	0.37
Spanish	0.22	0.12
German	0.16	0.17
French	0.38	0.28

At the end of this section we would like to emphasise that the problems mentioned above are not typical for the EIE only. The institutions dealing with the Russian Unified State Exam, as well as many of Ukrainian universities which have started implementing technologies for the test assessing of knowledge, face the same problems [18–20]. In our opinion, this represents a result of some "growing pains", which consist in overvaluation of the testing technologies and underestimation of a strict necessity to adhere to the proper requirements when designing the tests.

4. Advantages and drawbacks of testing technologies

The following points are usually considered as advantages of the testing [2, 6–8]:

- higher objectivity, when compared with the other forms of assessing;
- higher fairness;
- more complete coverage of the educational material;
- higher accuracy of estimation;
- higher economical efficiency;
- relatively little time spent on assessing procedures.

Nonetheless, we are to mention that the above advantages are in fact achievable only if the correct technology for designing the tests is strictly followed. In particular, higher objectivity and fairness of testing are only feasible if the tests are valid. Otherwise, one can arrive at replacement of evaluation subjectivity by subjectivity associated with selecting test tasks and assigning scores for their answers. On the other hand, more complete coverage of the educational material may cause increase in the fraction of secondary issues which do not reflect overall educational level of a person under test. The EIE is an example that illustrates too large number of questions, with the correctness of answers that does not correlate with the general test score.

Quite similarly, any real improvement of the economic efficiency and reduction of the timetable in comparison with the other forms of knowledge assessing would take place only if the tests are used for assessing of large groups of students. It would also be desirable that the tests could be used many times, though then one should create a sufficiently large set of variants equivalent in their complexity and use more varying forms of tasks for this purpose.

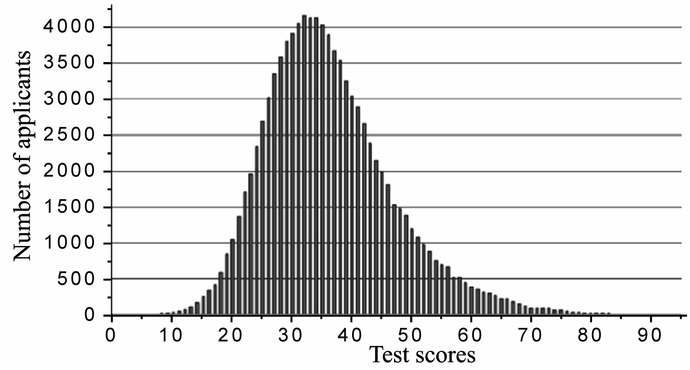
When discussing the EIE, many people often accentuate an increasing quality of entrants selection it allows. In fact, this statement should be wrong, which is clearly illustrated by Table 2. One can see from Table 2 that the minimum passing score (124 on the scale that ranges from 100 to 200 points) does not correspond even to the FX mark ('unsatisfactory, with a possibility of re-examination') on the ECTS scale [21], for which the lowest boundary amounts to 35 per cent of the maximum score. In addition, some of the tests allow a high probability for simply guessing such a number of answers which is required for obtaining the minimum passing score. Nevertheless, the EIE-2010 tests are better than those of the EIE-2009 in this respect.

Apart of the evident advantages, the tests as a form of knowledge assessing have also some drawbacks. The following points are usually reckoned [2, 6–8] to be the main deficiencies:

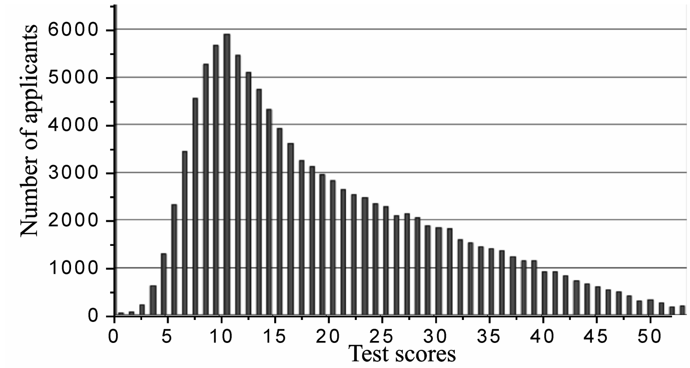
- durability, high complexity and cost of designing;
- common tests do not allow understanding the causes why the answers are unsatisfactory;
- testing does not allow checking the level of knowledge associated with creativity, deep analysis of problems, etc.;
- tests results always include a random component.

It should be noted that some of these disadvantages and problems can be solved by applying of advanced forms of tasks and improving methods for processing the tests results. In particular, the authors of the work [22] suggest adding of the following testing tasks to the traditional ones:

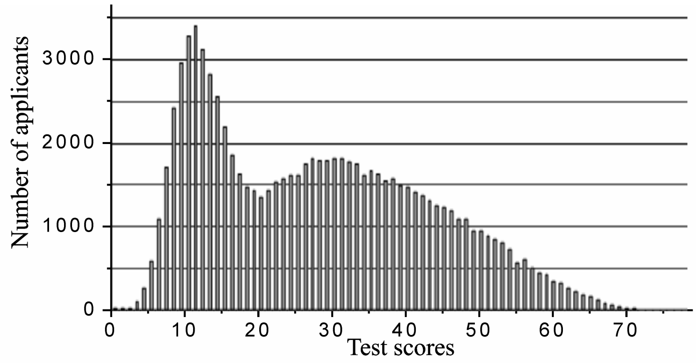
- the tasks with numerical answers where the points are assigned with taking into account the deviation from the correct answer [Their advantage is that the number of similar variants of tasks with identical difficulty can be very large (sometimes infinitely large). Here the ability of student to use adequate algorithms for solving the tasks of certain types is tested rather than his ability to remember correct answers];



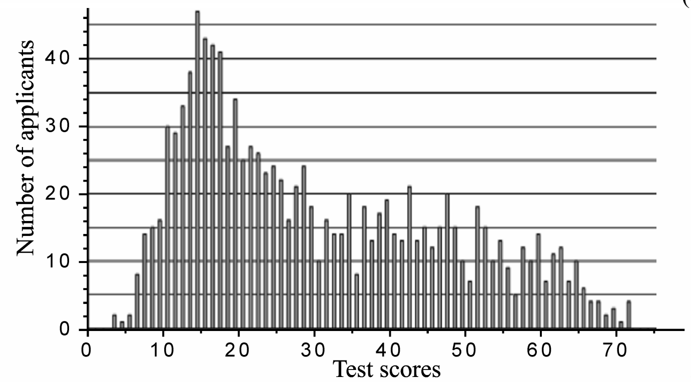
(a) Geography (98690)



(b) Mathematics-1 (110759)



(c) English (86678)



(d) French (1129)

Fig. 1. Examples of histograms for the EIE-2010 primary test score distributions (the numbers of persons tested are shown in brackets).

- the tasks with dual answers, which are being given sequentially, while the time for each answer is limited (a kind of tasks with choice);
- the tasks with verbal answers being evaluated by the coefficient of correlation with the correct answer (a kind of open-form tasks);
- the tasks with choice where all the answers are partly correct though differ by degree of completeness (accordingly, different points are assigned for different answers).

Table 2. Characteristics of the EIE minimum passing score.

No	Discipline	Minimum passing score (% of maximum possible)		Probability of guessing, %	
		2009	2010	2009	2010
1	Ukrainian Language and Literature-1	23	25	< 1	< 1
2	Mathematics-1	7	13	59	22
3	History of Ukraine-1	22	23	9	3
4	Chemistry	19	21	< 1	< 1
5	Physics	12	13	62	44
6	Geography	24	26	< 1	< 1
7	Biology	24	25	2	1
8	English	20	12	< 1	< 1
9	German	18	10	< 1	< 1
10	French	18	14	15	1
11	Spanish	18	13	< 1	< 1

Another direction in the development of test technologies is to improve the methods for processing the tests results. In particular, several specific algorithms are considered in the study [23], which allow determining the final results automatically and are based on the analysis of their empirical distribution functions or on the comparison of answers with certain “etalon” ones, as well as automated methods for assigning scores for separate tasks, taking into account a percentage of tested persons which have given the correct answers.

5. Conclusions

The test technologies have certain advantages over traditional methods for knowledge assessment. However, they also reveal some disadvantages and need special caution when applied in practice. One of the main problems associated with the test technologies is a necessity to follow the proper requirements when designing the tests, including the following points:

- to take into account the influence of aim of the tests upon their structure, content and processing algorithms;
- to examine the quality of test in general and of separate testing tasks, use the control groups and correct the tests after accounting for their quality indicators;
- to select the methods of putting down marks for correct answers, etc.

The other problems are related to applying tests results, which are reliable ‘on average’ only, to each person under test, determining the confidence levels for the results obtained and the probabilities of errors, etc. The EIE tests need a particular attention, since they represent a face of educational testing in Ukraine, while their quality greatly affects the process of student selection at the Ukrainian universities and the fate of each entrant.

References

1. Kadnevsky V M, The history of tests. Moscow: Narodnoye Obrazovaniye (2004).
2. Kim V S, Testing of educational achievements. Ussuriysk: USPI Publishing (2007).
3. Galton F. Inquiries into human faculty and its development. 2nd Ed., London: Dent & Dutton (Everyman) (1907).
4. Binet A and Simon Th, 1904. Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. L'Année Psychologique 11: 191–244.
5. Binet A and Simon Th, 1904. Application des méthodes nouvelles au diagnostic du niveau intellectuel chez des enfants normaux et anormaux d'hospice et d'école primaire. In: L'année Psychologique. 11: 245–336.
6. Avanesov V S, Composition of testing tasks. Moscow: Center for Testing (2002).
7. Aschepkova L Ya, Construction of test tasks and processing of test results. Vladivostok: FESU (2003).
8. Tchelyshkova M B, Theory and practice of designing pedagogical tests: A textbook. Moscow: Logos (2002).
9. Avanesov V S, The form of testing tasks. Moscow: Center for Testing (2005).
10. Bakhrushin V E, Mathematical foundations of system modelling. Zaporozhye: Classic Private University (2009).
11. Cronbach L J, 1951. Coefficient alpha and the internal structure of tests. Psychometrika. 16: 297–335.
12. Krisilov V A, Onischenko T V and Rusinova N V, 2004. Methods for analysis of pedagogical tests based on the tests results. Proc. Odessa Polytechnic Univ. 2: 1–6.
13. Dubina I N, Mathematical foundations of empirical social and economic research. Barnaul: Altai University Publishing (2006).
14. Likarchuk I, The official report on the independent external evaluation of knowledge of Ukrainian secondary educational institutions graduates in 2009. Kyiv: MESU, UCEQA (2009).
15. Likarchuk I, The official report on the independent external evaluation of knowledge of Ukrainian secondary educational institutions graduates in 2010. Kyiv: MESU, UCEQA (2010).
16. Investigation of the quality of university admission in Ukraine based on the EIE. http://www.irf.ua/files/ukr/reliz_%20zno%20validity.doc
17. Entrants choose the EIE because the testing selects the best. Portal on the External Testing. http://www.useti.org.ua/ua/news/548
18. Bakhrushin V E, Zhuravel S V and Ignakhina M A, 2009. Empirical distribution function of school graduates testing results. Control Systems and Machines. 2: 82–84.
19. Bakhrushin V E, Ignakhina M A and Shumada R Ya, Empirical distribution function of testing results. Proc. 3rd Int. Conf. “New Information Technologies in Education for All”, Ed. V Gritsenko. Kyiv: ISTC ITS (2008) p. 79.

20. Petriv V F and Rykaliuk R E, Visualization of quality criteria of test tasks. Proc. XVI Ukrain. Sci. Conf. “Modern Problems of Applied Mathematics and Informatics”. Lviv: Lviv Nat. Univ. (2009) p. 166.
21. National Aviation University: Information Package ECTS. http://www.nau.edu.ua/uk/EduProcess/ECTS
22. Oganesyan A G, Deschinsky Yu L and Biriulev K Yu, 2010. Testing or exam on computer? Educational Technologies & Society. **13**: 1–17.
23. Bakhrushin V E, Zhuravel S V and Ignakhina M A, 2010. Automation of assessing tests results. Control Systems and Machines. **2**: 10–12.

Bakhrushin V.E. and Gorban A.N., 2011. Testing technologies in education. The problem of test quality. Ukr.J.Phys.Opt. **12**, Suppl. 2 Sc. Horiz.: S1-S10.

Анотація. *У статті розглянуто проблему якості тестів в освіті, зокрема при здійсненні зовнішнього незалежного оцінювання, недавно впровадженого в Україні.*