RELIABILITY EVALUATION OF KNOWLEDGE E-TESTING BASED ON STATISTICAL METHODS

B. Pozdneev[1], F. Busina[1], P. Ovchinnikov[1], A. Ivannikov[2], E. Obuhova[1], M. Sutyagin[3], A. Levchenko[1], D. Popov[1]

[1] Moscow State University of Technology «STANKIN» (RUSSIAN FEDERATION)
[2] Federal State-Funded Institute of Science Institute for Design Problems in Microelectronics of Russian Academy of Sciences (RUSSIAN FEDERATION)
[3] Gazprom Corporate Institute (RUSSIAN FEDERATION)

Abstract

Further ICT development and accessibility causes heightened interest in the concept of educational process efficiency considered to be a guarantee of successful course within deadlines with less time losses and without decreasing retention level.

The key problem of the successful ICT application is a necessity to largely formalize teacher-learner interaction. It is especially important for getting proper answers to the eternal questions: how is a teacher aware of the trainees' knowledge level and what is their readiness to continue their course training?

So, our main objectives are creating a system of electronic tests and implementing them into teaching practice guided by the standard principles and recommendations, using huge reserves for increasing training effectiveness found in the development and application of special kind of tests – training ones, making possible arranging vocational training in small groups, while creating individual learning paths in heterogeneous groups, as well as training in groups with a non-uniform schedule.

Proposed training tests specificity is explicit and implicit prompts in their tasks, wording and distractors. In addition, these tests can be well applied not only for assessing the current level of knowledge, but also for their rapid recovery and systematization, which allows to noticeably shorten classroom instruction preparatory stage.

It is very important to mention that usual practice for creating and delivering different assessment tools and services, including e-tests, are focusing on high-stake ones, to be evaluated by formal procedures to confirm that they are reliable, valid, recognizable and realistic. As a rule, very extensive focus groups and long-term statistical test results are used in these procedures.

Unfortunately, such practice is useless in small groups – math statistics laws used in big groups have limited application in small ones because of non-representative samples. Therefore, some simplified methods for evaluating e-tests quality are needed, i.e. with appropriate approaches to forming focus groups and special statistical methods applicable to small samples.

The article presents functional models for developing and applying training tests in the learning process as well as the results of the proposed methods approbation for quality assurance and evaluating reliability of knowledge e-testing for undergraduates and graduate students in small groups (5-10 people). Evolution and adaptation of e-tests to the tasks of a separate educational trajectory are carried out by evaluating e-tests results statistically using mathematical methods adapted for small samples.

The practice of applying these statistical methods allows to judge whether the training tests carried out in the proposed facilitated format remain reliable. They also remain valid when used to set minimum scores, i.e. when the results of these tests serve as a formal boundary between the "unsatisfactory" and "satisfactory" assessments (grades). In addition, these statistical methods can be further simplified and successfully applied in intelligent learning systems with feedback, using the Learning Analytics concept.

Keywords: education, information and communication technology (ICT), statistical methods, e-learning, e-tests, small groups.

1 INTRODUCTION

In actual national methodical materials [1], [2] the educational process is defined as an organized, purposeful, systematic, evolving process of interrelated teacher's and student's activity aimed at mastering the system of knowledge and skills, it's causes and conditions.

The educational process background includes "teaching" – teacher's guidance activity aimed at students' course program learning and "studies" – the system of intellectual and practical students' actions that guarantee learning the suggested course.

Issues of the educational process standardization (especially - digital tools and IT application) are very popular in international research and educational literature [3]–[5]. Specified terminology corresponds to the international standards [6] in the domain of IT for learning, education and training. But but it should be noticed that this standard divides education process into "learning" that is knowledge, skills and orientation acquisition and "training" that is skills and knowledge training based on procedurally defined educational actions, aimed at the particular application.

Furthermore, there are four main roles in international standard, instead of two:

- learner – a person learning the course;
- teacher – a person teaching the course;
- trainer – a person supporting, implementing and facilitating the training process;
- tutor – a person or IT-system helping the learner.

The international standard allows using computer systems as self-sufficient tutors. For other roles, human participation is necessary, regardless of the different IT potential usage for educational purposes.

The model shown in Fig.1 can represent the education management process. Concerning the subject of our research, the relationship between managing teaching (A2) and managing assessment (A4) processes is of special interest, allowing a formal evaluation of a number of latent (not directly observable) values relating to the quality of the knowledge obtained by a learner.

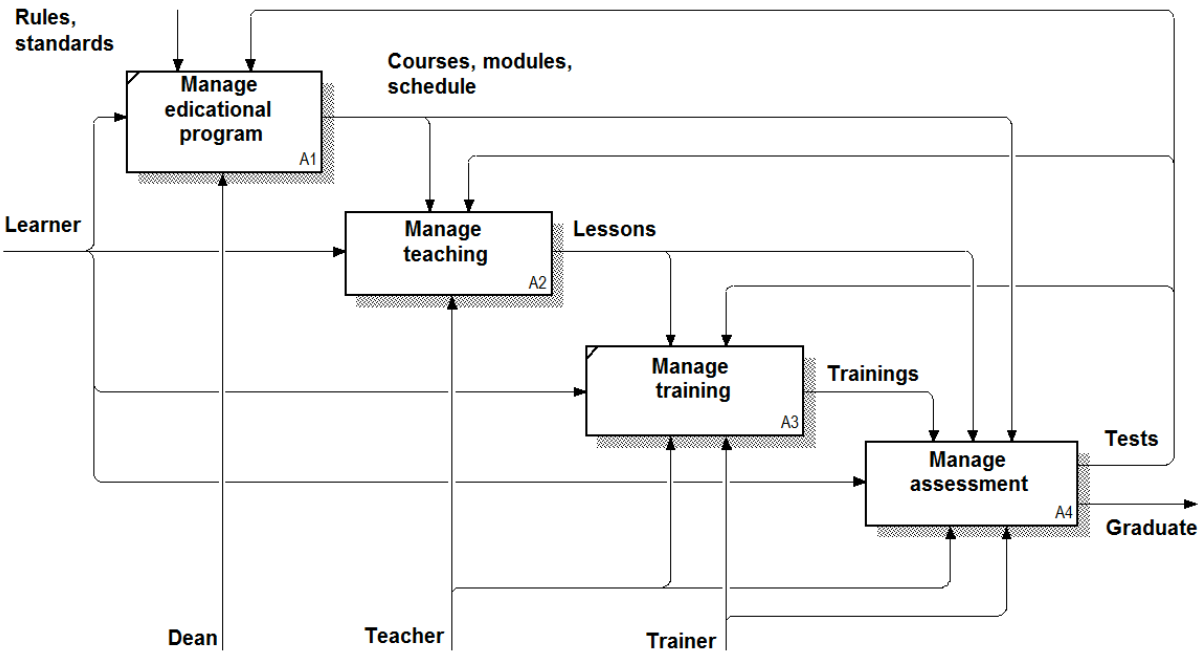

Figure 1. Manage education process

Note, that the given model includes only the most common processes relating to quite different aspects of managing an educational program. At the same time, processes related to training management have much greater opportunities for direct observation and are not considered in the further research.

In addition, the model has not yet considered the role of the tutor and the relevant IT systems. Suppose that the problems of allowing IT systems to make any decisions are in many aspects similar to the transfer of professor's authority to his assistants, and mostly lie in the credibility of the obtained results.

In particular, this credibility problem refers to the results of the knowledge test, which, when applied to e-testing, requires formalizing the basic quality indicators for each e-test. Each test, designed to check some knowledge, must be reliable, allowing a stable definition and evaluation of various levels of the knowledge. In addition, the test must be valid, providing responsible decision-making information about the test subject.

The use of such tests will improve the efficiency of assessment of competencies [7] acquired by learners [8], which will bring a new level of educational services quality [9]. It is worth noting that the acquired competence should be shown in the e-Portfolio of the learner, which will allow the future employer to evaluate the quality of student's training and to invite experts with the necessary knowledge and skills [10]. The use of these e-learning technologies is an indispensable element for the transition to SMART learning [11], [12].

2 EDUCATIONAL PROCESS EFFICIENCY CONCEPT

National methodical materials [1], [2] define some classical parts of the educational process:

- student's preparation for studies;
- studying the new material;
- perception of the new material;
- summarizing;
- consolidation;
- using knowledge in practice;
- knowledge evaluation and the results analysis.

The fundamental pedagogical principle of systematicity and consistency supposes teaching and learning to be organized according to some system and demands logical structure for both the content and the process. In terms of the specified principle, the concept of educational process efficiency can be defined as a ratio of knowledge learning level and depth to the full time costs of all educational process participants:

$$E = E_x \times D \ / \ \sum t_i$$

where E_x – learning level, D – learning depth, t_i – time costs of the i-th educational process participant.

Educational process economical efficiency can be defined using time and labour costs of the teachers [5] who take any role described above:

$$E = E_x \times D \ / \ \sum t_i \times c_i$$

where E_x – learning level, D – learning depth, t_i and c_i – time costs and salary rate of the i-th teacher (role category).

The most common for teachers and learners is the management of teaching by the model shown in Fig.2. This model is also frequently used in arranging e-learning, significantly reducing all the costs associated with the need for personal contact of learners and teachers.

However, this model has a whole range of disadvantages the main of which is the potential loss of some pedagogical techniques used by experienced teachers directly in the course of a separate lesson (A22).

Another problem is the direct dependence of the lessons schedule on the reliability of knowledge testing means, the main ones being various pedagogical tests [13]–[15], each consisting of a specific set of test questions (tasks).

In general, the reliability of an individual test is a certain integral indicator of the test questions (tasks) included in it. For every item it is necessary to determine detailed indicators: efficiency - the ability to

recognize latent (unobservable) parameters of the learner's knowledge level within a time-bound framework [13] and accuracy - the ability to obtain the necessary stochastic characteristics of these parameters with a given resolution ratio.

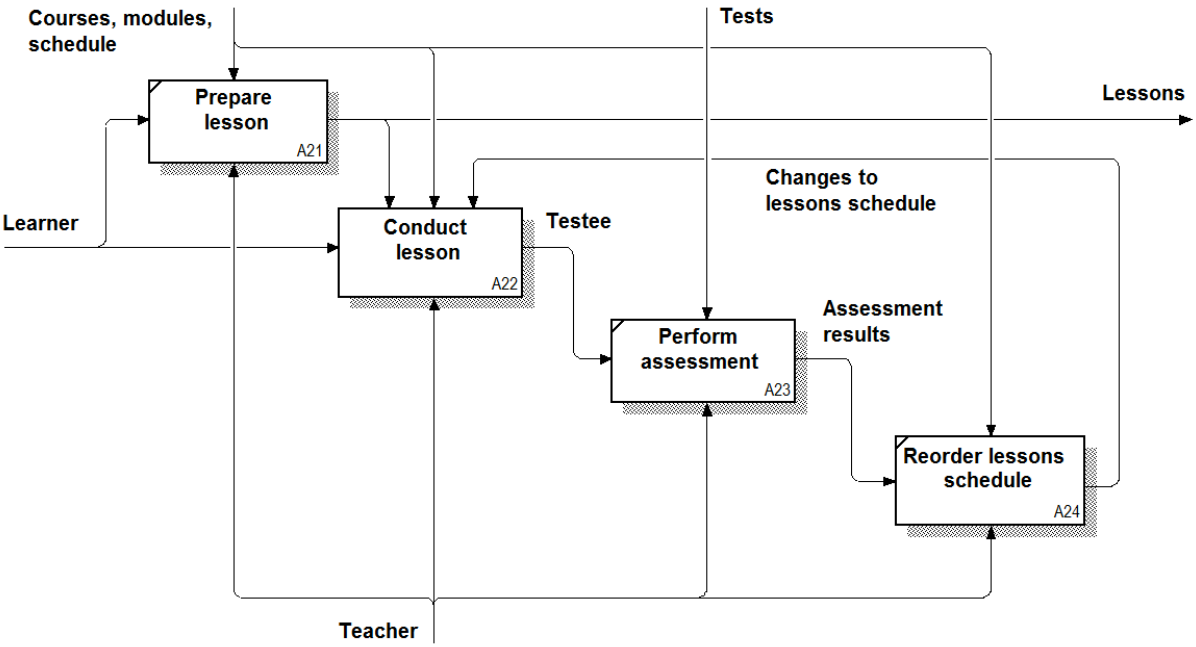

Figure 2. Manage teaching process

Taking into account all above mentioned, can anybody guarantee that the total costs of detailed development and approving the needed e-tests complex will not exceed the savings obtained from reducing personal interaction? In addition, is it possible to ensure that using the given teaching model for e-learning management keeps the time spent by the learners, especially related to perception, summarizing and consolidation of the material within acceptable bounds?

3 VARIOUS TEST TYPES AND THEIR ROLE IN THE EDUCATIONAL PROCESS

The majority of actual pedagogical literature focuses exactly on the control tests elaboration methodology (including the e-tests). It is not surprising because the control testing is the most valuable part of assessing graduates' readiness level in different educational organizations.

Similar tests are used for independent professional certification, by personnel departments and recruiting agencies for candidates profiling, by companies for inner certification. Another examples of high-stake tests are the driver's license tests or certification tests performed by world-wide companies such as IBM, Microsoft, Cisco, etc. Some of these tests are already provided in e-form and this fact imposes additional requirements for e-testing services, among which the requirement of legally significant test subject identification should be noticed.

It is generally mentioned that high-stake tests should be reliable - admitting stable determination and assessment of different readiness levels, and valid - granting valuable data about the subject they are designed to check. Additionally they should be recognizable - testees are provided with a complete set of pass instructions and realistic - the completion time and efforts are wisely limited. The combination of the mentioned characteristics makes the control tests objective, i.e. the result grades are relieved from the examiner's subjective judgment.

Critics of the objective tests often point out that such tests provoke the students into "cramming" or getting superficial subject knowledge in other way. Besides, the fact of using objective test results for assessment of the educational organization itself can provoke teachers to excess attention to the factual knowledge instead of the deep subject learning.

In addition, there are another important problems restricting the objective tests (the e-tests especially) role and usage in the educational process, which are connected with extensive falsification opportunities – either the testee substitution or using unauthorized sources of background information.

Now let us compare the approach described above with the real practice cases in conducting a classroom lesson process, the model of which is presented in Fig.3. The first and very important activity in this process is to establish and then keep contact with the audience (A221), somehow receiving the necessary feedback. Note that most of the material is explained to the learners in a declarative form, and the interrogative form is used as an additional pedagogical technique.

The simplest, typical and quite effective method of receiving feedback is a dialogue with the audience. Any teacher can ask questions to the whole audience or someone personally, give necessary explanations depending on the answers received, change the sequence and minuteness presenting individual materials. Among other things, the teacher always has the possibility to manage the whole process intuitively, providing learners with the opportunity to ask their own questions, require clarifying materials, etc.

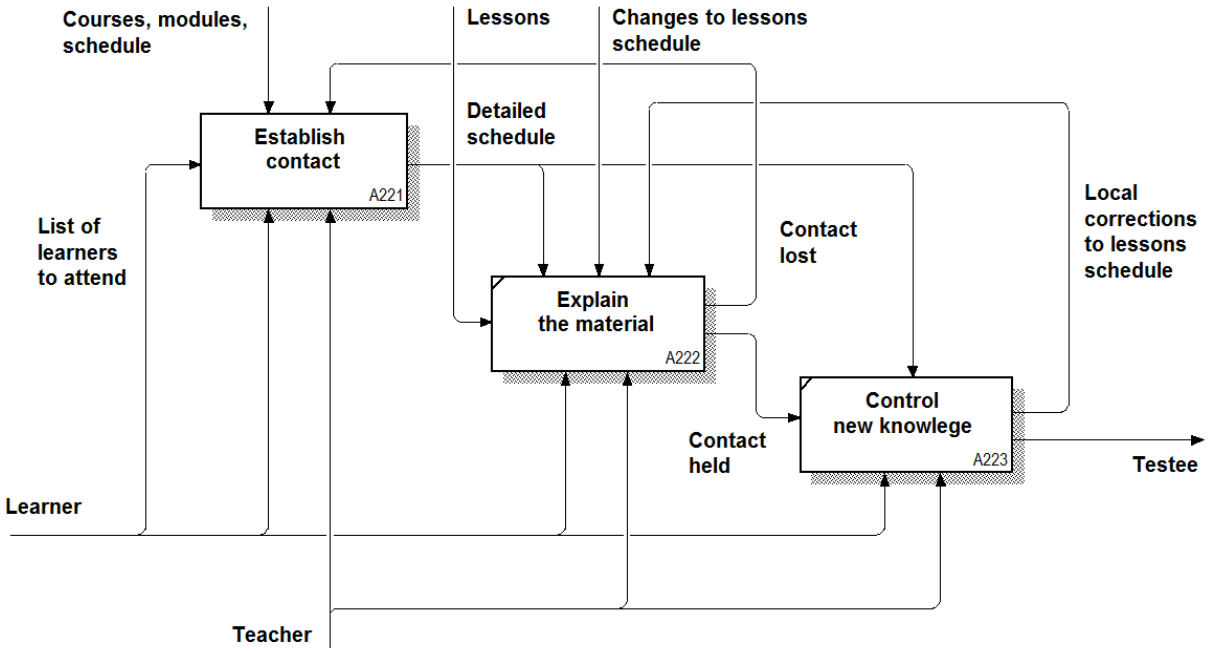

Figure 3. Conduct lesson process

Does the teacher worry about problems of falsification and cheating, which is, beyond doubt, very significant for controlling tests? It is unlikely, as all the questions and answers sound quite publicly, and any learner's desire to use additional material may give a reason to rejoice.

Finally, imagine that the described process of organizing a dialogue in the form of a question-answer is somehow formalized and brought to the form of an e-test. For control tests, many authoritative sources [13] recommend are paying great attention to the quality of distractors (wrong answer options for choice questions) and informational isolation of individual test questions. The distractors are considered to be good, if they do “deceive” about a half of the group testees. The test questions are considered to be well isolated if there is no opportunity to get some additional promptings for any question from any others.

Will the same approach work in the dialogue we are considering? There are also great doubts: for a more precise explanation of the individual positions of the materials studied, many teachers use the idea of contrasts and direct opposition of right and wrong answers. Equally often, methods to involve students in a chain of logical reasoning and thought-provoking questions are used.

A formal description of the differences and features listed above can be regarded as the common definition of a training test:

- clear and short instructions before each test question (task);
- declarative (narrative) form of the question (tasks);
- implicit and explicit prompts in wording and sequence of tasks and answers;
- clearly noticeable difference between wrong and right answers.

4 TRAINING TESTS AND TASKS COMPOSITION FEATURES

The terms “test” and “test task” are defined in the most popular national methodical materials [13], [14], [16] dedicated to testing problematics. Each task has a certain form and is designed to check some specific knowledge or skills. Each test represents a set (system, group) of test tasks.

The following types by their relation to specific subject (i.e. discipline) are defined, among them:

- homogeneous – measures a single discipline knowledge;
- heterogeneous – includes several homogeneous tests of various disciplines;
- integrative – consists of tasks requiring several disciplines knowledge.

There are following typical compositions of tasks in a test:

- chain tasks – the correct answer for one task depends on the previous ones;
- thematic tasks – a set of tasks in any form for controlling single theme knowledge;
- situational tasks – tasks used for checking knowledge and skills of acting in practical, extreme and other situations.

Most of the authors define six typical forms among different test task types:

- single-choice question (from 2 to 5 answers);
- single-choice of the most correct answer;
- multiple-choice question (from 5 to 14 answers);
- open-ended question;
- compliance question;
- correct-sequence question.

Depending on the task form a variety of assessment techniques can be applied to its result, including weight coefficients usage.

All these possibilities can be successfully applied for compiling training tests. The proposed model of their use in establishing contact activity is shown in Fig.4.

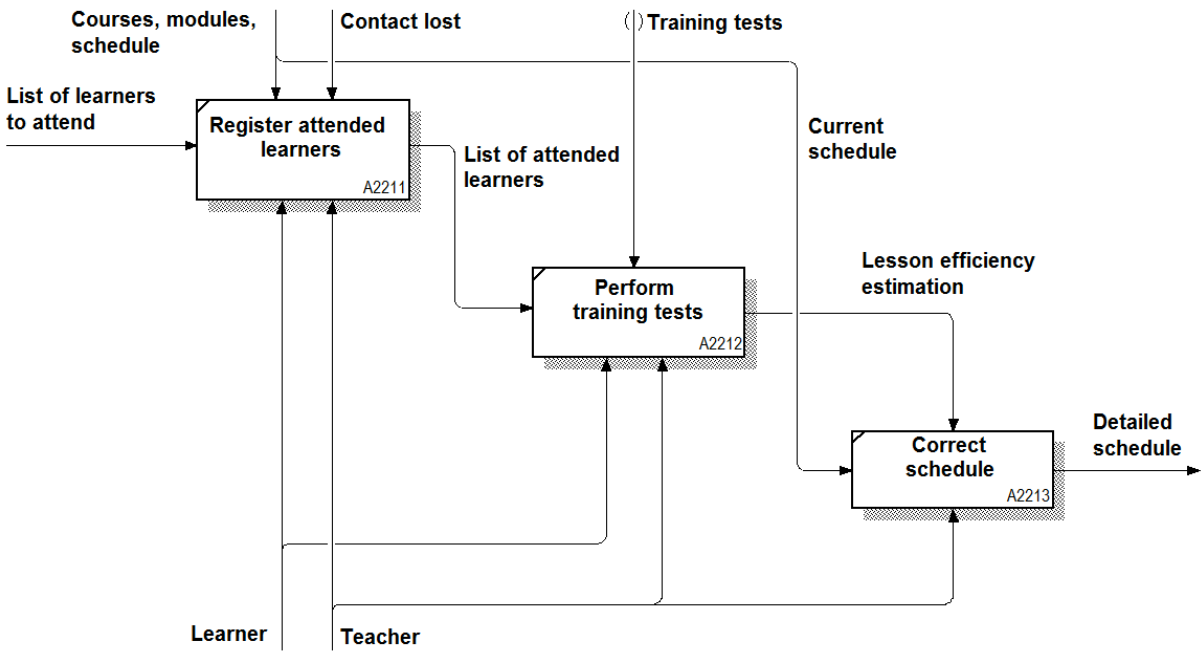

Figure 4. Establish contact activity (proposed)

5 ASSESSMENT OF EDUCATIONAL PROCESS EFFICIENCY

As said above, the entire educational process efficiency depends on its components efficiency. Let's consider a very simplified model of a single lesson and appraise its efficiency, assuming the lesson consists of 3 parts:

- Introduction, when teacher and students establish contact.
- Main part, when the material is explained.
- Control part, when the lesson material knowledge is checked.

Every lesson part can be described with a set of potentially observable complex events, statistically evaluable during full period of the lesson part, such as attending of students and teacher during all the period, focusing on learning during all the period etc.

A more detailed examination of real data can reveal statistical or causal dependency between such events, but to simplify the model, we will consider all the events to be independent. After all, it is quite possible that the absent student is better acquainted with the terminology or is stronger than the others motivated to learning.

As shown in special studies [5], [15], such approach can be used not only to carry out a real experiment that requires the collection and analysis of a large number of statistical measurements, but also when modeling various "what if" scenarios.

In the conducted studies, a simplified model "what if" was considered, in which it was assumed that the successful passage of the training test by individual students increases the probability of successful development of the introductory part from 90% to 100%. The effectiveness of an individual lesson was calculated as the total number of students who successfully mastered all the necessary material, with the same probabilities of the remaining events amounting to 90%.

Table 1 shows the results of calculations of another value - the total number of repeated classes (rounding up) to overcome a certain threshold, which determines the maximum permissible number of students who have not mastered the material.

Table 1. *Estimation for a set of lessons efficiency*

Threshold	Lessons with no training tests	Lessons with using training tests
0,5	4,3 (5)	2,1 (3)
0,3	7,4 (8)	3,6 (4)
0,1	14,2 (15)	6,9 (7)

Note that even with the optimistic probability values chosen, the calculated effectiveness of employment grows at least twice. Note also that this estimate directly refers only to the possible cost savings of the teacher's time, and to calculate the overall effectiveness, this estimate needs to be multiplied by the total number of learners.

6 TRAINING TESTS VALIDATION PROBLEMS

The model of the managing assessment process is presented in Fig. 5, and the development of training tests is carried out using this model.

Nowadays test validation (A43, A44 activities in Fig.5) practices need a focus group (A42 activity in Fig.5) big enough for tests to be checked for efficiency (i.e. to determine how well the test can separate people who have necessary knowledge from those who do not). Usually the gradation is based on that trial testing too.

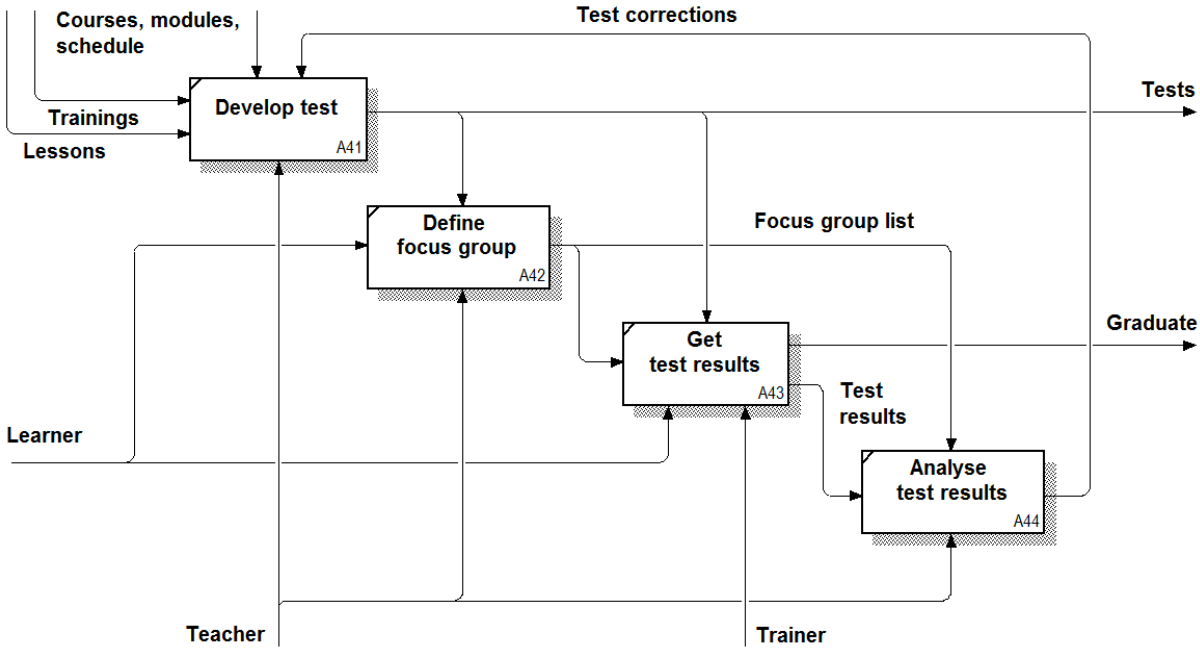

Figure 5. Manage assessment process

In small groups, such practices are useless – math statistics laws used in big groups have limited application in small groups because of non-representative sample. Hence, some special grades distribution and test efficiency assessment methods are needed.

To solve this problem, the authors used a special method of forming a focus group of participants with a priori estimated level of expected knowledge:

- Group 1 - presumably not familiar with the subject of testing at all.
- Group 2 - presumably familiar with the subject of testing.
- Group 3 – guaranteed to be familiar with the subject of testing.

An important advantage of this method is the possibility of spreading the focus group across a teaching small group by various volunteers, including teachers, junior and senior students from other groups and even just random people. In addition, the results of the same participant in the trial testing can be assigned to different groups, if the evaluation of his level of knowledge is changed upon repeated attempts.

Therefore, when planning the teaching in a small group of 11 people, extra 10 volunteers were involved, allowing a full-fledged check of the developed training test in one of the subjects, consisting of 38 test results distributed as shown on Fig.6:

- Group 1 (8 passes): average of correct answers is 45.3% ± 10.7%.
- Group 2 (12 passes): average of correct answers is 46.2% ± 15.8%.
- Group 3 (18 passes): average of correct answers is 72.7% ± 14.2%.

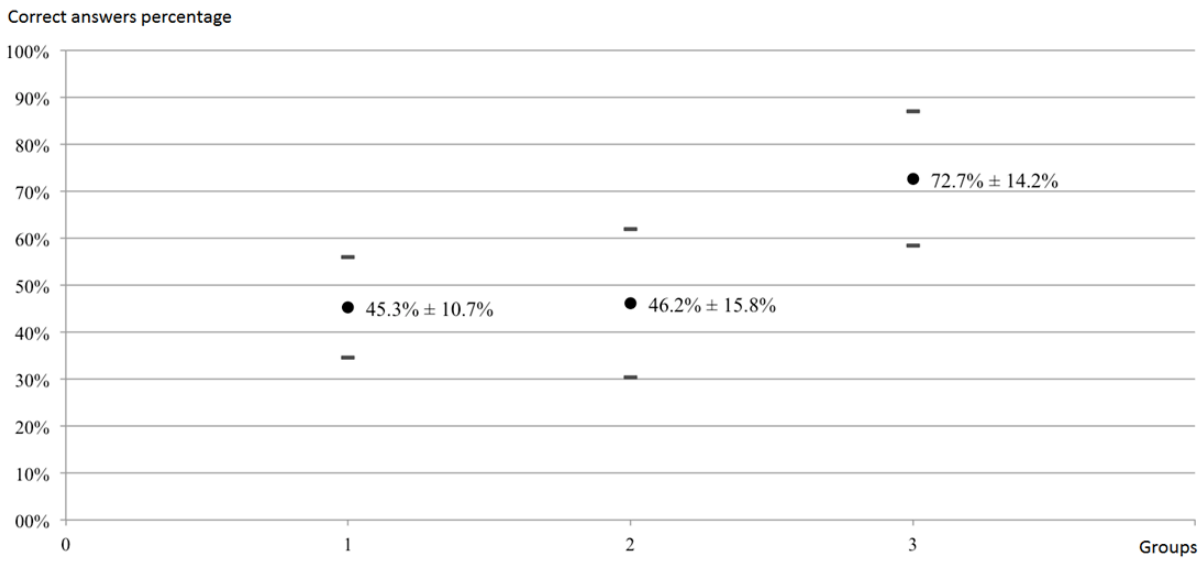

Figure 6. Average of correct answers in different groups

Analyzing these results, one can come to the following conclusions regarding the test reliability, consisting of its effectiveness and accuracy:

- The results of group 1 show that the test is effective (reliable) to identify students who do not have the necessary knowledge.
- The results of group 2 show that the test is effective (reliable) to identify students who have the necessary knowledge.
- The difference between the results of group 1 and group 2 indicates that the test is accurate.

One of the most unexpected results of this experiment was a slight difference in the results of testing in groups 1 and 2, which leads to the conclusion that it is possible further simplify the process of forming a focus group. To confirm the reliability of a test, it is sufficient to expand it artificially only by additional participants from group 1 (presumably not familiar with the subject of testing).

Also from the results of the experiment, it is obvious that setting the "satisfactorily" boundary at the level of 60% makes it possible completely exclude the probability of accidental receiving a positive mark, which confirms the validity of this test.

7 CONCLUSIONS

As it follows from the example above, the training tests having less complicated format, compared with the traditional ones, are still reliable. They are also valid when used for minimal grade, i.e. when the test results are used as the formal bound between “satisfactory” and “unsatisfactory” marks.

A system of tests developed in the way described above can be converted to the e-test form and used as a distant tutor. In addition, the statistical methods used can be further simplified and successfully applied in intelligent learning systems with feedback using the Learning Analytics concept.

Properly developed system of training tests increases the learning efficiency at least in two times, reduces superfluous teachers’ work and decreases the time learners spend on studying learning material. And with a systematic approach of the training tests positive impact on the entire educational process efficiency may become stronger on account of synergies.

ACKNOWLEDGEMENTS

Thanks to all the member of ISO/IEC JTC1/SC36 for their great work in the field of ICT in Learning, Education and Training, and especially for all kinds of contributions.

REFERENCES

[1] A. Novikov, Osnovaniya pedagogiki: posobie dlya avtorov uchebnikov i prepodavatelej pedagogiki [Grounds of Pedagogy: A guide for authors of textbooks and teachers of pedagogics] (in Russian). *Moscow, Egves,* 2010.

[2] Zagvyazinskiy V.I. Teoriya obucheniya: sovremennaya interpretatsiya [Learning theory: modern interpretation] (in Russian). *Moscow, Academia*, 2008.

[3] C. Stracke, "How Innovations and Competence Development support Quality in Lifelong Learning", 2014. Retrieved from http://opening-up.education/publications.

[4] B. Pozdneev, M. Sutyagin, V. Zuev, P. Ovchinnikov, A. Levchenko, and I. Kupriyanenko, "O perspektivakh razvitiya mezhdunarodnoj standartizatsii informatsionnykh tekhnologij v obuchenii, obrazovanii i podgotovke [On the prospects of the development of the international standardization of information technology in learning, education and training]," (in Russian), *IT-Standard*, vol. 1, no. 3-1 (8), pp. 1–7, 2016.

[5] T. Krupa, P. Ovchinnikov, Y. Severina, O. Shalyapina, and E. Rybalko, "Obespechenie effektivnogo obrazovatelnogo protsessa na osnove standartizatsii sredstv elektronnogo testirovaniya [Provision of Effective Educational Process Based on standardizing E-tests Means]" (in Russian*), Open Education,* no. 1 (108), pp. 41–49, 2015.

[6] ISO/IEC 2382-36:2013 Information technology – Vocabulary – Part 36: Learning, education and training.

[7] P. Ovchinnikov, N. Pustovoytov, and Y. Severina, "O neobkhodimosti garmonizatsii i standartizatsii terminologii pri professionalno-obschestvennoj akkreditatsii obrazovatelnykh programm [On the necessity of harmonization and standardization of terminology in professional public accreditation of educational programs," (in Russian*), IT-Standard*, vol. 1, no. 3-1 (8), pp. 12–15, 2016.

[8] B. Pozdneev, F. Busina, M. Sutyagin, P. Ovchinnikov, D. Popov, and A. Levchenko, "Development of educational programs and ICT skills of personnel based on harmonization of standards requirements," *ICERI2016 Proceedings,* pp. 4750–4759, 2016.

[9] B. Pozdneev, F. Busina, M. Sutyagin, D. Popov, P. Ovchinnikov, A. Levchenko, "Quality assurance of learning processes based on standards", *ICERI2016 Proceedings*, pp. 4920–4927, 2016.

[10] B. Pozdneev, F. Busina, M. Sutyagin, A. Levchenko, and V. Tikhomirova, "Information model of e-Portfolio", *EDULEARN16 Proceedings*, pp. 2118–2125, 2016.

[11] A. Solovov, and A. Menshikova, "Elektronnoe obuchenie: vector razvitiya [E-learning: development vector]," (in Russian), *Higher Education in Russia*, no. 11, pp. 66–75, 2015.

[12] B. Pozdneev, F. Busina, and A. Ivannikov, "Smart University Management Based on Process Approach and IT-Standards," *Smart Innovation, Systems and Technologies*, pp. 73–82, 2016.

[13] N. Efremov, Testovyj kontrol v obrazovanii [Testing control in education] (in Russian). *Moscow, Logos, Universitetskaya kniga*, 2014.

[14] V. Avanesov, Kompozitsiya testovykh zadanij [The composition of the test tasks] (in Russian). *Moscow, Testing center Publ*., 2002.

[15] Y. Neyman, and V. Hlebnikov, Vvedenie v teoriyu modelirovaniya i parametrizatsii pedagogicheskikh testov [Introduction to pedagogical tests modeling and parametrization theory] (in Russian). *Moscow, Testing center Publ*., 2000.

[16] A. Sviridov, Standartizirovannye metody na primere kontrolya i diagnostirovaniya znanij [Standardized methods on the example of monitoring and diagnosing knowledge] (in Russian). *Moscow, Publishing house of Russian State Social University,* 2011.