

DESENVOLVIMENTO DE UM MÉTODO PARA AVALIAR OS RESULTADOS DA APRENDIZAGEM ATRAVÉS DO GERENCIAMENTO AUTOMÁTICO DE ENSAIOS

DEVELOPMENT OF A METHOD FOR ASSESSING LEARNING OUTCOMES THROUGH AUTOMATED TESTING MANAGEMENT

РАЗРАБОТКА МЕТОДА ОЦЕНКИ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПОСРЕДСТВОМ УПРАВЛЕНИЯ АВТОМАТИЗИРОВАННОГО ТЕСТИРОВАНИЯ

KAZANBAYEVA, Albina S.^{1*}; IKLASSOVA, Kainizhamal E.²; KULIKOV, Vladimir P.³;

^{1,2,3} M. Kozybayev North Kazakhstan State University, Department of Information and Communication Technologies, 86 Pushkin Str., zip code 150000, Petropavlovsk – Republic of Kazakhstan

* Correspondence author

e-mail: akazanbaeva83@mail.ru

Received 03 June 2019; received in revised form 20 October 2019; accepted 09 November 2019

RESUMO

A avaliação do conhecimento é uma parte importante do processo pedagógico, pois ajuda a melhorar a qualidade do conhecimento profissional, contribui para a eficácia e eficiência da educação. Este artigo discute a aplicação de várias abordagens para a construção de um sistema especialista difuso de diagnóstico para testar o conhecimento. O objetivo do estudo é otimizar de maneira abrangente o processamento de informações de testes de massa, a fim de obter dados de relatórios, considerar testes claros e difusos, traçar uma analogia entre testes claros e difusos, oferecer e justificar vários esquemas para resumir os resultados dos testes, oferecer nosso próprio protótipo do sistema de teste especialista automatizado que abrange todas as etapas do processo de teste. Em geral, o teste como um problema algorítmico pode ser representado como a tarefa de construir um sistema especialista para avaliar o conhecimento dos sujeitos. Esses sistemas são usados para resolver muitos problemas relacionados ao gerenciamento eficaz, pesquisa científica e afins, onde é necessário obter o resultado com uma quantidade limitada de informações confiáveis. Está estabelecido que a teoria clássica dos testes tem uma série de suposições controversas, e os resultados de sua aplicação têm sérias desvantagens práticas. Portanto, o princípio básico da abordagem no trabalho é a separação da criação de um banco de tarefas de teste e o processo de compilação de testes (banco de material de teste) e o teste.

Palavras-chave: *sistema de treinamento, teste, teste de treinamento, resultados acadêmicos, previsão de desempenho dos estudantes.*

ABSTRACT

Testing knowledge is an important part of the pedagogical process, as it helps to improve the quality of professional knowledge, contributes to the effectiveness and efficiency of training. This article discusses the use of several approaches to building a fuzzy diagnostic expert system for knowledge testing. The aim of the research is a comprehensive optimisation of the process of mass test information processing in order to obtain reporting data, consideration of clear and fuzzy tests, drawing the analogy between clear and fuzzy tests, proposal and justification of different schemes of summing up the testing and proposal of one's own prototype of automated expert testing system which covers all stages of the testing process. Generally speaking, testing, as an algorithmic problem, can be represented as a problem of building an expert system to evaluate knowledge of testees. Such systems are used in solving many problems related to effective management, scientific research, and the like, where it is necessary to get the result with a limited amount of reliable information. It is established that the classical theory of tests has a number of controversial assumptions, and the results of its application have serious practical drawbacks. Therefore, the main principle of the approach in the work is the separation of the creation of a bank of test tasks and the process of compiling tests (bank of test material) and testing.

Keywords: *studying system, testing, training test, educational achievements, student's performance prediction.*

АННОТАЦИЯ

Проверка знаний является важной частью педагогического процесса, поскольку она способствует повышению качества профессиональных знаний, способствует результативности и эффективности обучения. В данной статье рассмотрено применение нескольких подходов к построению диагностирующей нечеткой экспертной системы для тестирования знаний. Целью исследования является комплексная оптимизация процесса обработки массовой тестовой информации с целью получения отчетных данных, рассмотрение четких и нечетких тестов, проведение аналогии между четкими и нечеткими тестами, предложение и обоснование различных схем подведения итогов тестирования, предложение своего прототипа автоматизированной экспертной системы тестирования, охватывающей все этапы процесса тестирования. Вообще говоря, тестирование, как алгоритмическая задача может быть представлено как задача конструирования экспертной системы для оценки знаний испытуемых. Подобные системы применяются при решении многих задач, связанных с эффективным управлением, научным поиском и подобными, где требуется получить результат при ограниченном количестве достоверной информации. Установлено, что классическая теория тестов имеет ряд спорных предположений, а результаты ее применения – серьезные практические недостатки. Поэтому в работе основным принципом подхода является разделение создания банка тестовых заданий и процесса составления тестов (банка тестового материала) и тестирования.

Ключевые слова: система обучения, тестирование, обучающий тест, учебные достижения, прогнозирование успеваемости студентов.

1. INTRODUCTION

The system of testing the results of studying has been transformed in the direction of quality control of education. One should note that the introduction of innovative educational technology in continuing education is becoming an increasingly effective tool for professional development.

Tests can be used at any stage of studying. Some of them are intended to evaluate the readiness of students to master a new academic course; others are intended to identify specific gaps in knowledge. Also, tests are used to plan the necessary targeted corrective work, predict the further process of studying, and determine the competence of specialists of the organisations (Maslennikov *et al.*, 2017; Aleksandrova *et al.*, 2018). Basic testing includes conducting input, current and interim controls, interim attestation for every discipline, including external evaluation of educational achievements (EEEE), and state final attestation.

The widest possible introduction of information technology in the education system and the solution of the problem of education quality cause the need for research and development of models adequate to the processes of educational activities of higher educational establishment (Naykhanova *et al.*, 2002).

Improving the quality of studying is the main direction of the education system's development. One of the main tasks of the

management of education quality in a higher educational establishment is the task of control and evaluation of the quality of studying. In the conditions of the modern information society, automated testing can become the main tool for quality control (Naykhanova and Danilova, 2002; Akhmetshin *et al.*, 2019; Bocheliuk *et al.*, 2019).

In this article, the authors have studied the problems of the development of evaluation of the results of studying by testing. In the course of the research, the analysis of knowledge control systems of students showed that the development of these systems is mainly on an empirical basis. It has neither proper scientific and methodological justification, nor theoretical models of specialist's readiness for professional activity. In connection with the objective needs of university practice, the implementation of the programme to improve the quality of specialists' education, there arises a need to develop a science-based system of control of students' knowledge.

To date, there is a technical problem, namely, to improve the process of automated evaluation of the results of the testing of students. The solution which will allow to improve the efficiency of knowledge control, ensure the objectivity of evaluation of passing a test regarding both the levels of assimilation of test tasks and the test as a whole.

To solve the technical problem, a method of evaluating students' knowledge based on the use of fuzzy logic and the theory of question-answer relations is proposed.

2. LITERATURE REVIEW

The research paper (Julliand *et al.*, 2018) presents a test generation method based on the calculation of the achievable and related under-approximation of the abstraction of the event system. Problems of formal description of the processes of studying system and attestation by testing were considered in the research (Bride *et al.*, 2016; Godefroid, 2007; Bué *et al.*, 2011; Godefroid *et al.*, 2005; Tillmann and De Halleux, 2008). In some research papers (Makarov and Sevastyanova, 2020) a method of visualisation and interpretation of the results of stored knowledge monitoring is proposed. The possibility of its use for the analysis of problems of studying at the individual and group levels is justified.

According to the studied material on the use of various methods when developing, as well as conducting the computer testing, there are presented results that will improve the ability to identify not difficult passing the test and perform an individual test points (Makarov and Sevastyanova, 2020; Wise *et al.*, 2019; Wise and Gao, 2017).

There are several approaches to the evaluation of the test task recommended by leading specialists in this field. These developments are based on both international testing experience as well as Russian and Kazakh (special developments were formed, which are distinctive from international recommendations for the preparation of tests, aims, and methodology for evaluating the results).

V.S. Avanesov (Doctor of Education, a leading specialist in the field of the theory of pedagogical measurements) suggests in every task, to give one score for the correct answer and for the incorrect one to give zero (Avanesov, 2003). At the same time, he notes the existence of more "advanced" evaluation schemes, for example, in testing centers in Western countries. The task is scaled according to its contribution to the general variation of test points of testees, that is, the evaluation in every case will be different and depend on a certain set of questions. At the end of the test, before putting a mark, the author proposes to conduct the correction of points for guessing (this is relevant for questions with one correct option). For questions with several correct options, he considers it impractical to conduct a correction for guessing.

Some unsolved problems related to the evaluation of the student's level of competence by

the method of testing are considered in the research paper (Bershadskaya *et al.*, 2019). In modern research (Muchuchuti *et al.*, 2019), different classification algorithms are compared to predict and improve students' performance. The results obtained help predict the total performance of students in time and develop activities that can improve their final mark.

During the simulation, research it was proposed as an adaptive, effort-oriented test (Wise and Kingsbury, 2016), which allowed to weaken the wrong orientation on the subject caused by unmotivated passing the test. Some researchers have developed fuzzy rules for changing the level of difficulty in adaptive testing (Lendyuk *et al.*, 2015; Maravić *et al.*, 2016; Mantzaris *et al.*, 2019). To be exact, these are the authors whose research (Özyurt *et al.*, 2012) is dedicated to the process of designing and developing the module of computerised adaptive testing integrated into UZWEBMAT. UZWEBMAT is an expert system that supports an adaptive and intelligent individual e-learning environment. It was dedicated to teaching a probabilistic subject.

The classical theory of tests has a number of controversial assumptions, and the results of its application are serious practical shortcomings. In particular, it was noted that when evaluating the student's knowledge with the help of tests of different levels of complexity, it is possible to get different ideas about the achievements of students. Statistics calculated within the framework of the classical theory of tests allow obtaining the relative position of every testee in the normative sample. However, they cannot be used to objectively evaluate the values of the parameters characterising the level of knowledge of the testees and the complexity of the test tasks. The question "What is the objective evaluation of the level of the student's preparedness on the subject?" the classical theory of tests is still open. There was an attempt to answer this question in the framework of another methodological approach to the creation of pedagogical tests. Also, to the interpretation of test results, within the framework of so-called Item Response Theory (IRT), but as the practice has shown, this theory also has a number of serious shortcomings.

The shortcoming of the existing theories is the fact that one way or another, all these theories suggest the obligatory presence of "aerobic testing" and a large number of "tests" for every task. In real life, not every higher education establishment can afford it. And this is especially unacceptable for testing students in subjects the content of which varies from year to year (for

example, computer science). Given that in the special subject at the intersessional control a clearly insufficient (to determine the quality of the test according to the proposed theories) number of students are tested, and the tests evaluate the knowledge of students not quite objectively.

Of course, the test set "randomly" obtained from the unified to the binary representation of the test pool of tasks will be by far more cumbersome. However, this is partially compensated by dropping "atypical" mistakes and parameterisation of the test task formatting; that is, counting tasks can be generated from formulas based on a random number sensor. Drawings (figures) are also recovered from parameterised random parameters of functioning algorithms. Also, in the same way, it is possible to set variable texts. Obviously, the costs of such a prepared test base will multiply increase, including the time spent by experts and programmers. Apparently, this problem is one of those that prevents such reading of the computer version of testing (Kulikov and Shatilova, 2005).

Reduction of the time spent, especially of the specialists gathered in groups, can be solved by switching to adaptive testing, which, as noted above, is unrealisable on a paper basis, but is actively used by technology leaders (for example, Microsoft).

One of the ways to overcome these difficulties was the development of automated testing on WebTest shell already available at the higher education establishment. All this allows suggesting that it is advisable to conduct the research dedicated to the method of evaluating the results of automated testing.

3. MATERIALS AND METHODS

3.1. Quantitative knowledge criteria

The evaluation system where knowledge is characterised by general terms such as "good", "strong," and "in-depth" is often unsatisfactory. Also, the four-point system, if it is based on subjective ideas about knowledge and intuitive quantitative criteria, slightly changes the situation.

This system satisfactorily performs its functions in the evaluation of knowledge by a person (professor). However, when using technical means of studying and control for the purposes of studying, such a system of marks is fundamentally inapplicable since technical devices do not have intuition. The machine can work only with formalised and quantified information.

Usually, knowledge is determined (with some degree of accuracy) by setting the student elementary control questions and comparing the answers of the student with the answers, which are considered to be deliberately correct (sample). The concept of "elementary control question" in this case is relative. It is elementary only because within the framework of the considered domain, it is not divided into simpler ones. However, its content can be quite complex. The concept of "student's response" is considered in a broad sense. In general sense, it is a response to a request. It can be both deterministic and probabilistic and can be expressed not only in words but also in actions, such as performing certain operations and the like. We introduce the concepts of "elementary knowledge" and "general knowledge".

Elementary knowledge is understood as the knowledge which refers to one elementary question. It is determined by comparing the response information with the sample information. General knowledge we consider as the knowledge that applies to all issues raised within the framework of this topic or course. At the same time, the issue of how many elementary questions need to be asked in order to reliably evaluate the level of general knowledge, as well as the issue of how complex the questions should be, is within the scope of specific methodologies and are not considered here.

We consider that the developed system (model) allows formulating such a (sufficiently large) number of questions that, in all cases will provide sufficiently reliable identification of the level of general knowledge on the topic, course, or their sections. It means that in the framework of model the following condition is always met where q is a number of questions which can be posed by means of the model (model capabilities); q' is a number of questions necessary for reliable identification of the level of general knowledge (psychological and pedagogical requirements).

As a quantitative criterion of knowledge, one can use the degree of compliance of T_i information contained in student's answers to i question and information that in relation to this question is considered to be correct (sample) (Equation 1). Where X_i is the code response information to i question; X'_i is the code of sample information corresponding to i question. Until now, this approach/method had been widely applied. It was proposed a method in which when determining the level of general knowledge, not only the degree of compliance of the information contained in the student's answer with sample

information but also the number of mistakes made is taken into account. The degree of noncompliance can be used as a mistake criterion of F_i of the same types of information (Equation 2). We consider T_i value determined by Equation 1 to be the evaluation of knowledge of i question and F_i value determined by Equation 2 to be the mistake that characterises lack of knowledge of i question.

The value of T_i evaluation related to the degree of compliance of directly proportional dependence: the higher the degree of compliance, the higher the evaluation. The value of F_i mistake is inversely proportional to the degree of compliance: the smaller the degree of compliance, the higher the mistake.

This representation of quantitative criteria characterising elementary knowledge is in good agreement with the intuitive representation of the correlation of evaluations and mistakes (the greater the mistake, the lower the evaluation and vice versa). S_i is the general evaluation for i question that can be determined by Equations 3 and 4.

In the first case, it can be obtained on the basis of analysis the evaluation of knowledge of i question (traditional "clear" testing) and in the second one on the basis of analysis proposed by us, that is, taking into account the mistakes made (as the difference of the evaluations of knowledge and mistakes of i question).

3.2. Point and integral evaluation of knowledge level

In accordance with the provisions of the mathematical theory of evaluation of the quality of test tasks, the level of knowledge θ can be considered as a general part of correct answers to all conceivable test tasks that display the subject area. By the data of testing, it is possible to find

pinpoint $\tilde{\theta}_k$ and/or interval $\left[\tilde{\theta}_k^1, \tilde{\theta}_k^2 \right]$ statistical

evaluation θ . As it is known from the mathematical statistics course, the point evaluation of the general share is the sampling rate (Wise and Gao, 2017), which is (Equation 5). Where p is a part of correct answers of tested/student? Determined, according to (5). This evaluation is unbiased, consistent, and effective, but it is only an approximate value θ . However, the accuracy of the approximation can be considered sufficient for practical conclusions only in case when testee/student was offered a large number of test tasks ($k \rightarrow \infty$). For short

tests, the issue of accuracy of the estimate can be solved as follows. By setting δ probability, we

determine $\tilde{\theta}_k^1$ and $\tilde{\theta}_k^2$ in the way that the following correlation should perform (Equation 6). In other words, we find θ an interval estimation with δ reliability. Traditionally, the boundaries of

$\tilde{\theta}_k^1$ and $\tilde{\theta}_k^2$ confidence intervals for general share are determined like solutions of Equations 7, 8. If $k > 20$, one can use approximate formulas.

3.3. A priori determination of the complexity of the test task (TT)

A priori determination of the complexity of the test task (TT). In classical testology, there are only statistical methods for determining the complexity of the TT. However, in practice, especially for entrance testing, the attempt to establish the level of pre-experimental complexity is particularly relevant because not for every task from the bank of the TT (due to the principle of non-disclosure of the TT before entrance exams) there is a possibility to determine the complexity statistically. However, on the other hand, it is impossible to talk about balanced versions of the entrance TT without taking into account the quality of the TT as a complexity.

Of course, it is difficult to achieve a complete coincidence of the values of a priori and statistical complexity because the quality of the determination of complexity depends not only on how good and thought-out is the system of determining the complexity of the TT proposed in this research paper, but also on the level of professionalism of the experts who work with it. However, one way or another, the system will self-improve in the process of operating, namely:

1. Correct a priori evaluations of complexity with available statistical evaluations;
2. Evaluate how correct the expert's opinion is and provide relevant information (warn) about it.

For the attempt of a priori determination of the complexity, the first step is to establish the factors ($x_1, \dots, x_n$) affecting the complexity and establish their weight coefficients. A priori complexity of the TT will be considered to be the value determined by Equation 9, where $T(x_i)$ is the quantitative importance of i complexity factor to the TT. Together with experienced professors-experts in the field of entrance test tasks, we have developed a system of factors (reflected in Table

1.) that affect the task complexity for a priori determination of the complexity of the test task in any discipline.

According to the system of a priori determination of the TT complexity, the complexity of 20 entrance TT in mathematics was measured and compared with the statistical complexity measured for about 500 university entrants. The results are shown in Figure 1. The measurement error of this experiment was plus or minus 7%, which can be considered acceptable.

3.4. Correction method for guessing

Modern testology, which leaves open the issue of the identification of the testees who simply guessed when answering the test. The approaches to determining the quality of the task in which the proportion of very easy and very complicated questions does not change the status of guessed answers.

Some testologists, in particular, V. S. Avanesov (Doctor of Education, a leading expert in the field of the theory of pedagogical measurements), tried to introduce at the end of the test, before putting marks, the correction of points for guessing. But many testologists fairly disagree. Thus, for example, in the methodical recommendations of the Department of Pedagogy and Psychology of Leningrad Sanitary Hygiene Medical Institute based on the developments of I. M. Sechenov First Moscow State Medical University, the authors note that this formula equally penalises those who "can guess" and those who "can not".

In this research paper, we propose a fundamentally new approach that does not imply mass fining the testees by correcting the points for guessing.

There is a category of questions that testees answered mostly without thinking, namely, by guessing, but the questions did not fall into the category of low-response. Without identifying such questions before determining the level of reliability and validity of the test, we have a high probability that the reliability and validity of the test will be determined incorrectly.

The proposed method allows us to identify a group of testees who with a high probability guessed, and a group of "bad" questions that most students answered at random.

The model of a group of students who answer the test without thinking that is, randomly choose answers was created. If a standard test task consists of a question and 5 answers, a

random student gains at random the result of 20% plus or minus 4% of the maximum of one hundred percent. At the same time, the number of people who simultaneously chose correct answers to one question will be at average 20% for each question with a deviation of plus or minus 6%. (If the number of questions ~100, then the number of testees ~300).

Thus, when evaluating the quality of tests, it is logical to exclude from consideration the questions answered by 20 plus or minus 6% of the testees from their total number, and it is logical to assume that students who gained 20 points plus or minus 4% of the total number of points are likely to have guessed.

Further, on the model of "guessing" students the average number of pairs of overlapping questions is calculated, i.e., for each pair of questions the number of identical answers is calculated (for example, if a student answers or does not answer these 2 questions at the same time). With 99% of reliability, this number is 67 % of the total number of students plus or minus 1%.

From real suspicious questions revealed according to the scheme stated in cl.1 (20%+-6%), they distinguish those which the number of pair overlaps is 67% +-1%. The answers to the other questions were not random (for example, cheating or not guessing, but an attempt to think).

The set of remaining real suspicious questions (cl. 3) is checked for the similarity of the distribution according to the Fisher criterion with the set of modelled random questions. In case of a positive result with the reliability adopted when calculating the Fisher statistics, it is possible to decide that the majority of students answered them at random, namely, by guessing. It is logical to exclude these questions from consideration and provide relevant information to the professor.

By excluding the questions which were guessed, it is possible to calculate the test results for one of the proposed methodology. However, in this case, the lowest threshold for students should be the result not lower than $20\% + 2 \cdot 4\% = 28\%$ from maximum (at the final interval, the spread of the summed up factor formed by a large number of independent causes is evaluated as sigma 2). To the interval from 12% to 28% will fall, the students who guessed, below 12% are those who were consciously mistaken. As can be seen in Table 2 and Figure 2, the results are flat-lining with the help of this methodology.

A more detailed picture can be seen in (Figure. 3) where, in fact, the scale of evaluation

"lengthened" by half and became of more "normal" character.

4. RESULTS AND DISCUSSION:

4.1 Bank of Test Tasks

In this research paper, the main principle of the approach is to separate the creation of the bank of test tasks and the process of compiling the tests (bank of test material) and testing. The Bank of Test Tasks should function as an individual complete system, and it can be further used by different testing systems. The main purpose of the Bank of Test Tasks is as entering test results in database storage, pairing this base with the test material base, processing test results, and obtaining the accounting data with the minimum cost of all kinds of resources.

The preparatory stage of organising the testing includes the creation (throughout the academic year) of the Bank of Test Tasks and the Bank of Test Material in the form of an electronic database (DB). Test tasks are classified by topics and complexity levels, which are at the primary level determined by experts (highly qualified professors).

Every test task must comply with a specific rule base for the selection of test tasks, the requirements for which are developed in Chapter 2.

The Expert Data Collection Bank contains the information on expert recommendations (highly qualified professors) concerning the quality of the test material, division into topics, complexity levels, etc. Test material "verified" by experts is placed to the Bank of Test Tasks, where tasks are classified according to the specified criteria.

The first requirement imposed on any test is the completeness of the coverage of all sections of the subject. To meet this requirement, it is advisable to divide all the questions into several topics and group the questions by topics. In this case, the elements of the division can be both unrelated sections of the subject and a group of issues that satisfy the same properties. For example, if we want to include a question in all tickets of the future test, it is enough to consider this question as an individual topic because every examination paper will include at least one question from every topic. In more detail, the scheme of creating the Bank of Test Tasks reflected in (Figure. 4) and is the following:

The automated test management system block described above can be prepared throughout the academic year.

The object-oriented design technology was

used to develop a model of the automated Bank of Test Tasks. The main object of the whole system is a task. Every task is characterised by a category (clear-fuzzy), quality (satisfactory, good and excellent), subject, topic, module it belongs to, complexity, the time required for the answer, the author, and the expert who verified and evaluated it. The elements of studying, module, and subject (for easier qualitative and quantitative evaluation of the Bank of Test Tasks) are also characterised by average time, complexity and number of the TT.

The content of the task and answers can contain text, graphic, and formula information as well as figures. The content of the task can be created directly in the system or imported from an external system.

Every task can correspond to an arbitrary number of answers. At the same time, the task can have one or more correct answers. The number of correct answers is determined by the task category.

4.2 Test generator

The test generation from test tasks can be carried out automatically, and the user can set the required scope and complexity of the test as well as impose other restrictions. Before the beginning of test generation, the system suggests you specify the test structure and the number of required options.

Textstructure:

1. Subject name.
2. Total number of the TT in the test.(N)
3. Number of modules used in the test.(M)
 - 3.1. Module name
 - 3.2. Number of topics used in the test.(L)
 - 3.2.1. Topic name.
 - 3.2.2. Number of TT(K)
 - 3.2.3. Average time for one TT (T)
 - 3.2.4. Average complexity of TT (S)

Rule base for test generation of the topic with the specified K, T.

1. Choose K from the BTT and give this massive the name TEST
2. Find Nt (that is Nt is = new test) – average time of TT in TEST
3. D =0
4. If topic No. >1, then T=T+A
5. If Nt >T, then move on to 13 or to 6
6. If Nt =T, then move on to 16 or to 7
7. Delete from TEST massive the task with minimum time (t minimal test), TEST= new TEST

8. Copy to new TEST from the BTT the TT which is the nearest by time value (t of the BTT) to t minimal test, to f the $BTT > t$ minimal test, move on to 11

9. Delete from TEST massive the task with maximum time (t maximum test), $TEST = \text{new TEST}$

10. Copy to new TEST from the BTT the TT(the nearest by time value (t of the BTT) to t minimal test, to f the $BTT < t$ minimal test, move on to 11.

11. Find T_{new} in new TEST

12. $D_{\text{new}} = T - T_{\text{new}}$

13. If $D_{\text{new}} * D \leq 0$ then move on to 14 otherwise to 5

14. If topic No. = K then if $D > 0$, $D = D_{\text{new}}$, $TEST = \text{new TEST}$ otherwise 15

15. If $|D_{\text{new}}| - |D| \leq 0$, to 16 otherwise accurate to within $D = D_{\text{new}}$, $TEST = \text{new TEST}$

16. $A = -D$

17. Accurate to within DTEST is compiled

To generate a test with a set complexity, a similar rule base is used, with changes in cls. 8 and 10 (the replacement condition of the test task with the same time values is added, if there are none, then the next task in order is checked).

Unit for evaluation of testing results:

Rule base of evaluating the knowledge level of testees. Unit for identifying "guessed" TT:

1. If the complexity of the question is from 0.14 to 0.26, *then* include the question in the category of suspicious and move on to 2;

2. If according to the Fisher criterion, the distribution of a "suspicious" question is similar to a set of modelled random questions, so to consider a question to be guessed means to consider it ordinary.

Unit for evaluation of clear testing:

1. If it is necessary to evaluate the test in "non-traditional" ways, *then* move on to the Unit for evaluating fuzzy testing *otherwise* evaluate in the traditional way and move on to 2;

2. If it is necessary to evaluate the test, excluding "guessed questions", *then* move on to the unit for identifying "guessed" TT and move on to 3 *otherwise* evaluate in the traditional way *and* move on to 4;

3. If the question is guessed, *then* exclude the question from further consideration and move on to 4;

4. Evaluate the test in the traditional way on a scale of 1 or 0.

5. Question complexity Equation 10 (S_{iq} is the evaluation of answer of q student to i question, w is a number of students in the group).

Unit for evaluation of fuzzy testing:

1. If it is a guessed question, *then* exclude the question from further consideration and move on to;

2. If a priori information on the degree of preparedness of testees is known (an expert can single out the category of "excellent students" *then* move on to the unit for methodology with excellent students, *otherwise* move on to the unit for universal methodology.

Unit for universal methodology:

1. Correct answer weight t_k (Equation 11) ($a_1 \dots a_m$ is the number of correct options of answers).

2. Incorrect answer weight f_j if $\forall b_j \neq 0$ *then* (Equation 12), If $b_j = 0$ *then* $f_j = 0$. If $b_k = 0$, $k \neq j$ *then* (Equation 13) ($b_1 \dots b_n$ is the number of incorrect options of answers).

3. Question weight (Equation 14) (S_{iq} is the evaluation of answer of q student to i question, w is a number of students in the group).

Unit for methodology with excellent students

1. Data entry ("excellent students" according to experts);

2. If a student who was entered in the category of "excellent students", scored less than 50% of the maximum, then he should be excluded from the category of "excellent students" and the relevant message to the expert should be provided;

3. Linear programming problem is solved (Equation 15): $x_j \in [-1; 0]$, if x_j is an incorrect answer, $x_j \in [0; 1]$, if x_j is a correct answer, ($a_{i1}, \dots, a_{i5}$) is variety of options of answers of i "excellent student" to k question;

4. If the sample is the only *then* move on to 7 *otherwise* move on to 5;

5. Move on to the unit for universal methodology;

6. Select from the variety of N samples obtained by the methodology with excellent students $X_n = (x_{1n}, \dots, x_{5n})$ "the closest" to the sample obtained by the universal methodology $X_y = (x_{1y}, \dots, x_{5y})$, the one which $\sum_{i=1}^5 (x_{1i} - x_{yi})^2$ was the

least;

7. Question weight (Equation 16).

4.3. Analyser

The analyser is intended to provide feedback in the system. Internal feedback is the information that comes from a student to the testing program in response to his actions when performing tests. It is intended for self-correction by the testing program of its database and rule base. The information of external feedback in the system under consideration comes to the teacher and is used by him to correct the informative content of the tests or the course of lectures as well as the student's activity. More detailed scheme of the analyser of test tasks is reflected in (Figure 5) and is the following:

Analyser's rule base

If the testing is clear, then perform unit 1.

If the testing is clear, then perform unit 2.

Unit 1:

If the guessed questions are identified, then provide the relevant information to the expert.

If there is a priori information on the testees, then provide the information on the validity degree to the expert (the question weight by the methodology with "excellent students") and recommend questions with the validity coefficient of less than 0.3, exclude from the Bank of Test Tasks or make a correction.

If a priori complexity in the Bank of Test Tasks is not equal to a posteriori within the accuracy of 0.01, then adjust the value in the Bank of Test Tasks (and all parametric forms of the TT).

If a posteriori complexity in the Bank of Test Tasks is not equal to a posteriori complexity for the current stage of testing within the accuracy of 0.01, then adjust the value in the Bank of Test Tasks (and all parametric forms of the TT), taking into account the accumulated statistics.

If a priori evaluation of the time required to solve the task in the Bank of Test Task is not equal to a posteriori one within the accuracy of 1 second, then adjust the value in the Bank of Test Task (and all parametric forms of the TT).

If a posteriori evaluation of the time required to solve the task in the Bank of Test Tasks is not equal to a posteriori one for the current stage of testing within the accuracy of 1 second, then adjust the value in the Bank of Test

Tasks (and all parametric forms of the TT).

If the reliability coefficient of the test satisfies the expert, then memorise the test structure.

The test structure includes number of tasks (the best option is parametric) of certain complexity level and time characteristics for every module and subject.

Unit 2:

If the guessed questions are identified, then provide the relevant information to the expert.

If there is a priori information on the testees, then provide the information on the validity degree to the expert (the question weight by the methodology with "excellent students") otherwise provide the information on complexity to the expert (by the universal methodology) and recommend questions with the validity coefficient of less than 0.3, exclude from the Bank of Test Tasks or make correction.

If a posteriori complexity in the Bank of Test Tasks is not equal to a posteriori complexity for the current stage of testing within the accuracy of 0.01, then adjust the value in the Bank of Test Tasks (and all parametric forms of the TT), taking into account the accumulated statistics.

If a posteriori evaluation of the time required to solve the task in the Bank of Test Tasks is not equal to a posteriori one for the current stage of testing within the accuracy of 1 second, then adjust the value in the Bank of Test Tasks (and all parametric forms of the TT).

If the reliability coefficient of the test satisfies the expert, then memorise the test structure.

The test structure includes a number of tasks (the best option is parametric) of certain complexity level and time characteristics for every module and subject. Knowledge control is an integral part of the pedagogical process, especially if we want to consider this process as technological, i.e., one that ensures under certain conditions a guaranteed goal achievement. Although there are individual software tools focused on computerisation of control, however, these tools are not in all cases adequate for solving the relevant problems of entrance and intersessional testing the solutions of which the ASTM presented in this paper is responsible for.

The ASTM can be divided into the following parts: automatic and automated generation of test tasks by the description of the knowledge system, test generation from test tasks, conducting a computerised test survey, and processing of test results.

For effective use of pedagogical testing in practice, it is necessary to have the opportunity to generate a sufficient number of variants of test tasks with minimal time.

Test generation from test tasks can be carried out automatically, and the user can set the required volume and complexity of the test, as well as impose other restrictions.

A computerised survey can be carried out on the basis of tests generated immediately before testing that really allows providing every testee with its own variant (in case of "computer test").

Processing the results of pedagogical testing is an important part of knowledge control and is considered as an element of the iterative process of test development, i.e., the use of test results to modify the test in order to improve its a posteriori evaluations.

The automated test system of management (ATSM) was used to generate tests for entrance in various academic disciplines and intersessional examinations in North Kazakhstan State University and demonstrated its high efficiency.

In more detail, a testing process is as follows. From the Bank of Test Material, the electronic system selects the tasks which satisfy a certain rule base and evaluates the suitability level and the complexity level. The tasks, which do not satisfy the rules base, are not entered into the Bank of Test Tasks. Next, the generator compiles a test with the necessary set of subjects, the number of questions on every topic in the subject, and the set complexity level of questions and time characteristics. After processing the test results, the analyser provides recommendations on how to optimise the test.

The unit for knowledge base and data of model consists of the Bank of Test Materials and Bank of Test Tasks. The Bank of Test Tasks can be replenished with new tasks at any time. The interaction of the units is reflected in Figure 6 with the help of lines.

The expert survey unit consists of the rule base (RB) for selection of test tasks to the bank; rule bases (RB) to evaluate the complexity of tests and answers and unit for data collection from experts (where can be included data about a priori evaluation of the individual qualities of tasks,

information on the testes and required test parameters).

Expert survey unit (ESU) is intended to generate and configure the unit for evaluation of test results as well as the decision-making unit (DMU) of the operating part of the ESU.

The test generator is directly engaged in the formation of the test (taking into account the information coming from the EEU about the required set of subjects, the number of questions on each topic in the subject, and general level of test complexity).

Decision-making unit includes the analyser of test results which gives guidelines for optimizing a test (for example, concerning the determination of the TT qualities, adjusts the attributes of tasks in the Bank of Test Tasks, memorises the structure of "successfully formed" tests according to experts) and unit of test results evaluation which depending on the purpose of testing (intersessional or entrance) makes a decision on the TT qualities, assignment of optimal weights to the TT, points distribution or the exclusion of TT from consideration.

The application of the test management system allows to conduct examinations at the higher technological and technical level and increases the evaluation ability of the test.

5. CONCLUSIONS:

Due to the widespread introduction of the test system of knowledge control, the development of control systems based on testing and development of methods for compiling the tests have become particularly relevant.

Optimised test data processing is understood as entering test results in database storage, pairing this base with the test material base, processing test results, and obtaining the accounting data with the minimum cost of all kinds of resources.

The aim of the research was to identify the main evaluation criteria of the test material and test task, the development of generalized formalized approaches to the evaluation of different types of questions of clear and fuzzy tests, statistical analysis of the results.

According to the results of the study, the automated test management system was created, which allows obtaining a priori information on the TT, create tests, conduct testing, and automatically obtain its results. The program includes the schemes of finding the result

described in the research paper.

6. REFERENCES:

1. Akhmetshin, E. M., Pavlyuk, A. V., Ling, V. V., Mikhailova, M. V., Shichiyakh, R. A., Kozachek, A. V. *Journal of Entrepreneurship Education*, **2019**, 22, 1-12.
2. Aleksandrova, O., Batchenko, L., Dielini, M., Lavryk, U. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*, **2018**, 4, 157-165.
3. Bershadskaya, M.D., Serova, A.V., Chepurenko, A.Yu., Zima, E.A. *VyssheeObrazovanie v Rossii*, **2019**, 28(2), 38-50.
4. Bocheliuk, V., Panov, M., Nechyporenko, V., Pozdniakova-Kyrbiatieva, E. *Astra Salvensis*, **2019**, 7(13), 275-288
5. Bride, H., Julliand, J., Masson, P.-A. *Science of Computer Programming*, **2016**, 132(2), 190-208.
6. Bué, P.-C., Julliand, J., Masson, P.-A. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6706 LNCS, **2011**, 51-68.
7. Godefroid, P. Compositional dynamic test generation. *The ACM Special Interest Group on Programming Languages*, **2007**, 42 (1), 47-54.
8. Godefroid, P., Klarlund, N. & Sen, K. *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*, Phoenix, USA, 2005, 213-223.
9. Julliand, J., Kouchnarenko, O., Masson, P.A., Voiron, G. *Programming and Computer Software*, **2018**, 44, 1-14.
10. Kulikov, V.P., Shatilova, E.M. *Bulletin of EKSTU named after D. Serikbayev*, **2005**, 1, 2.
11. Lendyuk, T., Sachenko, S., Rippa, S., Sapojnyk, G. *Proceedings of the 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS 2015)*, **2015**, 2(7341443), 945-948.
12. Makarov, S.I., Sevastyanova, S.A. *Advances in Intelligent Systems and Computing*, **2020**, 908, 502-509.
13. Mantzaris, A.V., Walker, T.G., Taylor, C.E., Ehling, D. *Journal of Big Data*, **2009**, 6(1), Article No.24.
14. Maravić, S., Čisar, P., Pinter, R. *Computers and Education*, **2016**, 92-93, 142-160.
15. Maslennikov, V., Grishina, O., Lyandau, Y., Kalinina, I. *Proceedings of the 30th International Business Information Management Association Conference, IBIMA 2017 – Vision 2020: Sustainable Economic development, Innovation Management, and Global Growth*, Madrid, Spain, **2017**, 2017, 4376-4384
16. Muchuchuti, S., Narasimhan, L., Sidume, F. *Lecture Notes in Networks and Systems*, **2019**, 69, 742-755.
17. Naykhanova, L.V., Danilova, S.D. *Theoretical and applied issues of modern information technologies: Materials of the Third All-Russian Scientific and Technical Conference*. Ulan-Ude Russian Federation, **2002**, 170-173.
18. Naykhanova, L.V., Danilova, S.D., Tumashov, S.V., Angaeva, E.Yu. *Collection of Scientific and Methodological Articles*, **2002**, 8, 114-120.
19. Özyurt, H., Özyurt, O., Baki, A., Güven, B. *Expert Systems with Applications*, **2012**, 39(10), 9837-9847.
20. Tillmann, N., De Halleux, J. *Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, **2008**, 4966 LNCS, 134-153.
21. Wise, S.L., Gao, L. *Applied Measurement in Education*, **2017**, 30(4), 343-354.
22. Wise, S.L., Kingsbury, G.G. *Journal of Educational Measurement*, **2016**, 53(1), 86-105.
23. Wise, S.L., Kuhfeld, M.R., Soland, J. *Applied Measurement in Education*, **2019**, 32(2), 183-192.
24. Avanesov, V.S. *Pedagogical Diagnostics*, **2003**, 1, 3-7.

$$T_i = f(X_i, X'_i), \quad (\text{Eq. 1})$$

$$F_i = f_1(X_i, X'_i), \quad (\text{Eq. 2})$$

$$S_i = T_i \quad (\text{Eq. 3})$$

$$S_i = T_i - F_i \quad (\text{Eq. 4})$$

$$\tilde{\theta}_k = p, \quad (\text{Eq. 5})$$

$$p \left(\tilde{\theta}_k^1 \leq \theta \leq \tilde{\theta}_k^2 \right) = \delta, \quad (\text{Eq. 6})$$

$$\sum_{m=kp}^k C_k^m \left(\tilde{\theta}_k^1 \right)^m \left(1 - \tilde{\theta}_k^1 \right)^{k-m} = \frac{1-\delta}{2}, \quad (\text{Eq. 7})$$

$$\sum_{m=0}^{kp} C_k^m \left(\tilde{\theta}_k^2 \right)^m \left(1 - \tilde{\theta}_k^2 \right)^{k-m} = \frac{1-\delta}{2}. \quad (\text{Eq. 8})$$

$$AT(x_1, \dots, x_n) = \sum_{i=1}^n T(x_i) \quad (\text{Eq. 9})$$

$$d_i = \frac{w - \sum_{q=1}^w s_{iq}}{w} \quad (\text{Eq. 10})$$

$$t_k = \frac{a_k}{\sum_{k=1}^m a_k} \quad (\text{Eq. 11})$$

$$f_j = - \frac{\prod_{j=1}^n b_j}{b_j \sum_{j=1}^n \frac{\prod_{j=1}^n b_j}{b_j}} \quad (\text{Eq. 12})$$

$$f_j = - \frac{\prod_{j=1, j \neq k}^m b_j}{b_j \sum_{j=1, j \neq k}^m \frac{\prod_{j=1, j \neq k}^m b_j}{b_j}} \quad (\text{Eq. 13})$$

$$r_i = 1 + \frac{\sum_{q=1}^w s_{iq}}{w} \quad (\text{Eq. 14})$$

$$\sum_{i=1}^l r_i = \sum_{i=1}^l \sum_{j=1}^5 a_{ij} \cdot x_j \rightarrow \max \quad (\text{Eq. 15})$$

$$r_k = \frac{\sum_{i=1}^l r_i^k}{l} \quad (\text{Eq. 16})$$

Table 2. Test results with the exclusion of suspicious questions

	Test with all questions		Test without "suspicious" questions	
0-19%	9	31%	4	15%
20-39%	17	59%	15	56%
40-59%	2	7%	5	19%
60-79%	1	3%	1	4%
80-100%	0	0	0	0
	29		25	

(Table 2 was presented before Table 1 for editorial reasons, sorry for any inconvenience. Editors)

Table 1. Factors affecting the task complexity for a priori determination of the TT

Complexity criterion	Criterion qualitative characteristics	Final rank
time required for the perception of task	big	0.0666666666666667
	average	0.065
	small	0.03
time required for solving	big	0.0666666666666667
	average	0.055
	small	0.01
complexity (by avanesov)	level 1	0.1111111111111111
	level 2	0.09
	level 3	0.06
	level 4	0.03
complexity (by bespalko)	level 5	0.0888888888888889
	level 0	0.0888888888888889
	level 1	0.0444444444444444
	level 2	0.06
	level 3	0.055
	level 4	0.1111111111111111
content	homogeneous (1 topic)	0.07
	homogeneous (several topics)	0.08
meaning	obvious	0.065
	non-obvious, without subtext	0.06
	with subtext	0.085
	a test for intelligence	0.08
by number of elements of actin	1	0.04
	2 or 3	0.08
	>3	0.07
if any additional constructions are required	strictly yes	0.051
	depending on university entrant's level	0.055
	strictly no	0.0666666666666667
time required for solving a task by expert	>60 seconds	0.1111111111111111
	40-60	0.1
	20-40	0.03
	5-20	0.06
	<5	0.03
if it can seem obvious at first sight or "confuse" topic complexity	yes	0.06
	no	0.0222222222222222
	high	0.0666666666666667
	medium	0.03
	low	0.04
by links of data structure	worldview minimum	0.0666666666666667
	basic knowledge	0.0222222222222222
	curriculum knowledge beyond basic level	0.08
	extra knowledge besides curriculum	0.0888888888888889
content of tt	calculate	0.0666666666666667
	simplify (use of formulas, laws)	0.04
	text task	0.06
if can be solved orally (without using paper and notes)	strictly yes	0.0666666666666667
	depending on university entrant's level	0.04
	strictly no	0.06

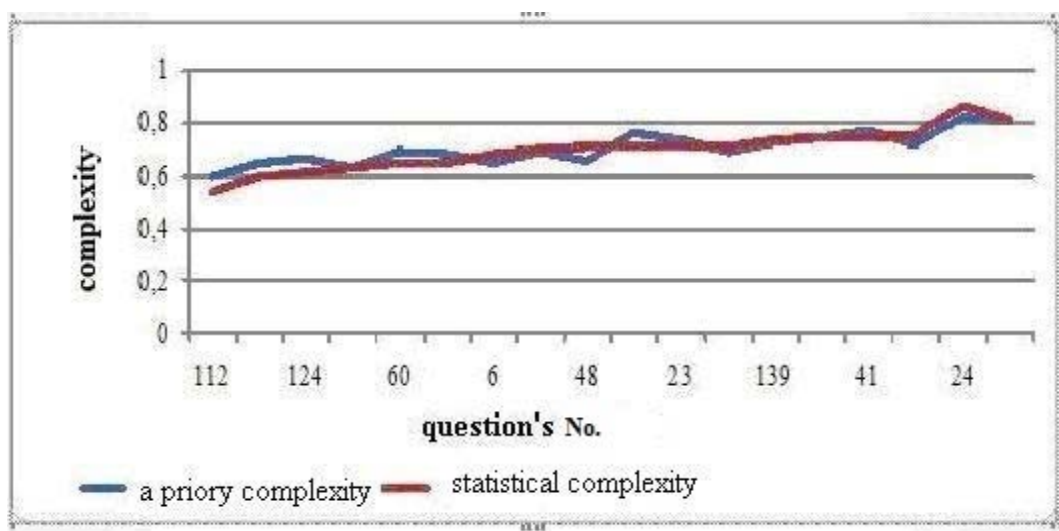


Figure 1. *A priori determination of the difficulty of the test task*

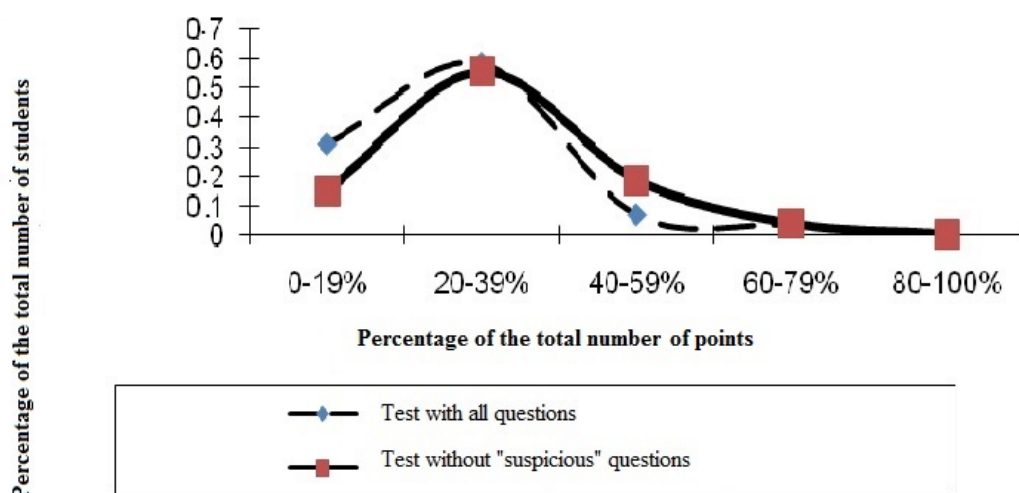


Figure 2. *Testing results diagram*

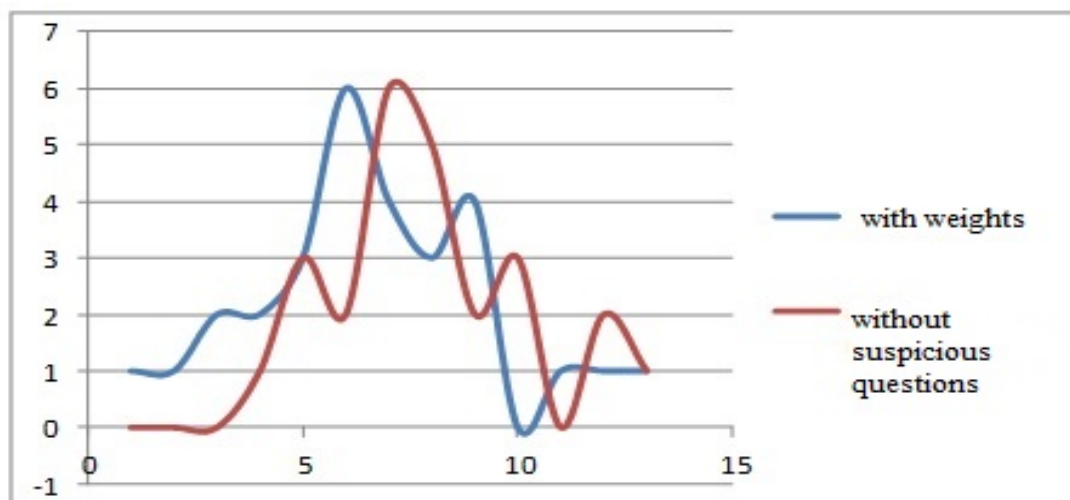


Figure 3. *Evaluation scale of test tasks*

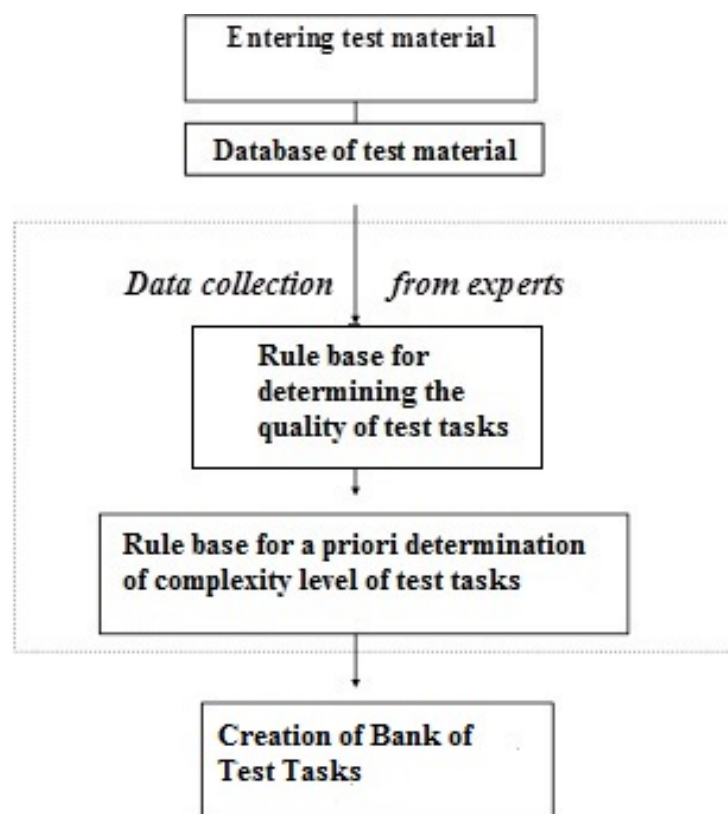


Figure 4. Scheme of creation of Bank of Test Tasks

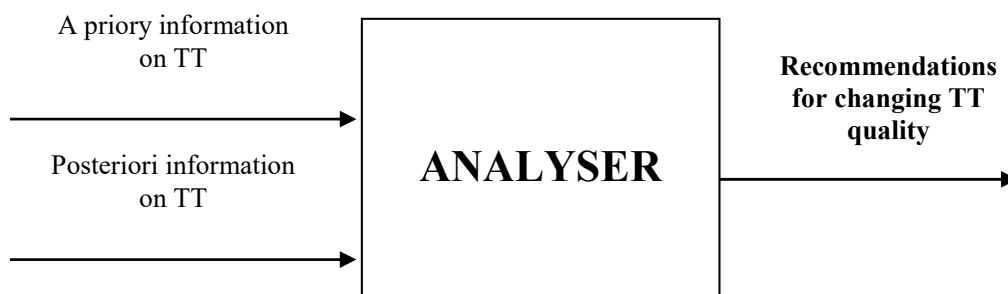


Figure 5. Scheme of test tasks analyser

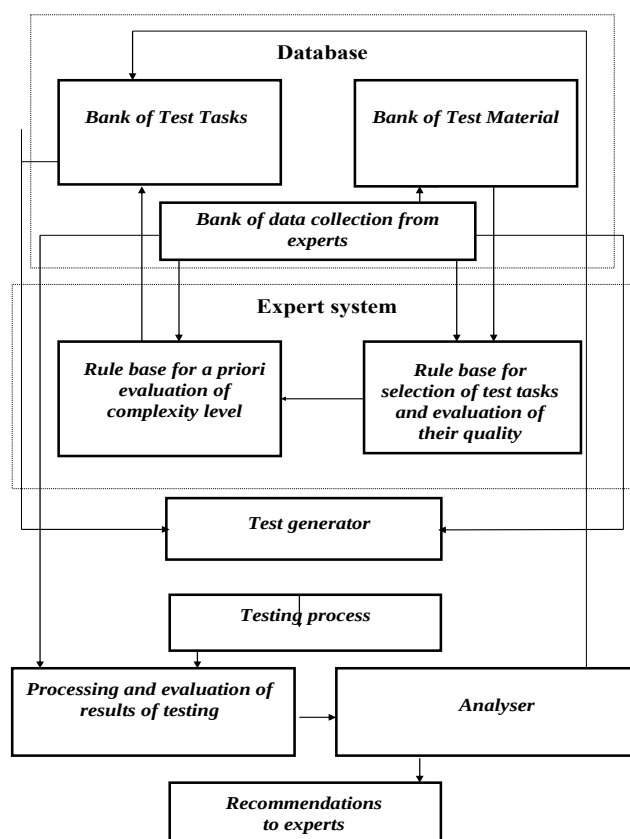


Figure 6. Structure of an automated system of test management