HiLL PUBLISHING

Journal of Applied Mathematics and Computation, 2024, 8(3), 238-255
https://www.hillpublisher.com/journals/JAMC/
ISSN Online: 2576-0653
ISSN Print: 2576-0645

Knowledge Assessment Through Tests—Semantic Approach

Nikolay P. Takuchev

Trakia University, Stara Zagora 6000, Bulgaria.

How to cite this paper: Nikolay P. Takuchev. (2024) Knowledge Assessment Through Tests—Semantic Approach. *Journal of Applied Mathematics and Computation*, 8(3), 238-255.
DOI: 10.26855/jamc.2024.09.006

Received: August 22, 2024
Accepted: September 19, 2024
Published: October 16, 2024

***Corresponding author:** Nikolay P. Takuchev, Trakia University, Stara Zagora 6000, Bulgaria.

Abstract

Knowledge assessment through tests is an objective and effective technology, widely used in modern education. Tests (multiple-choice tests), consisting of dichotomous items—a question with one correct and one or more incorrect answers (distractors) are widely used in modern education to assess students' knowledge. Test developers face the problem of converting the number of correct answers into a numerical grade representing the knowledge of the assessed. A numerical grade is a point in the numerical interval—a scale of grades. Usually, the number of correct answers is converted into a grade based on the evaluators' inner sense of fair evaluation. In the present work, a model for knowledge evaluation through dichotomous tests is proposed, based on the so-called Semantic branch of Information Theory. A critical opinion is given for the Classical Test Theory and the modern Item Response Theory as tools for knowledge assessment. Some concepts in these theories leave the feeling that it could be desired more concerning assessing knowledge through these theories. A new, entirely different approach to knowledge assessment by tests is proposed in the paper. In the proposed information model for knowledge assessment, the process of knowledge assessment is considered as an information process with information transfer. The information is generated by a source (the assessed), which has a goal—to get as close as possible to the error-free solution of the test. The information in the form of an information signal (the answers to the test that the assessed gives) is directed to the recipient—the assessor. The assessor evaluates the value (importance) of this information signal, which is a measure of the knowledge of the assessed. The value of the information signal is measured by the progress of the assessed toward reaching the goal. Formulas were obtained, linking the value of the information signal with a numerical grade of knowledge of the assessed. In particular, evaluation formulas are derived for tests of the most used types—with items with 3, 4, and 5 answers (the example with the scale of grades used in Bulgaria). However, detailed assessment requires answering a large number of items (items bank, included in the test at the stage of development), which increases the time for the examination. The examination time could be shrunken with an adequate algorithm that reduces an item's number included in the exam according to the answers of the assessed, without deteriorating the quality of the examination and assessment. An adaptive algorithm of knowledge assessment is proposed, based on analytical expressions, which can be integrated into computer tests to shorten the examination process by reducing the number of items asked, depending on the previous answers of the assessed. The adaptive algorithm reduces the number of items that the assessed answers, compared to the number of items in the bank. The grade that the assessed receives for his/her knowledge of the examined topic differs from the "exact" grade (that he/she would receive after solving a test with all items in the bank) with a value not exceeding a given tolerance.

The grade is calculated from: 1. the number of items in the items bank, 2. the number of items the assessed has answered, which are a part of all items in the items bank, and 3. the relative number of correct answers.

Keywords

Education; Knowledge assessment; Semantic approach; Information importance (value); Multiple-choice tests; Adaptive algorithm

1. Introduction

Tests have been used to select candidates for Chinese administration since ancient times. Their main advantage is their applicability as a technology for a rapid assessment of knowledge. Since the beginning of the 20th century, psychological tests have been developed, initially in the USA. The so-called Classical Theory of Tests (CTT, discussed for example in [1]) began to develop in the 40s.

In the last few decades, there has been a growing worldwide use of a certain type of test—dichotomous, composed of items (tasks) with identical structures. Each item consists of a question and two types of logically mutually exclusive answers—one correct and several incorrect answers. The incorrect answers are misleading (distractors) and serve to reduce the probability of guessing the correct answer.

After the 80s, was developed the so-called Item Response Theory (IRT) (see for example in [1-7]). IRT developers claim that this theory could apply to the analysis and evaluation of all types of tests, in particular, those for knowledge assessment. IRT offers several types of dependencies (models) on the probability of a correct answer to a given item on the intellectual qualities of the evaluated, in particular, her/his knowledge of the subject under examination. The models differ in the number of parameters included in them, subject to determination after testing of a group answering the test. As a result of the testing, the values of the parameters in the models are determined and it is assessed which of the models is the most suitable for assessing the verifiable qualities of the group. At the author's discretion, in the particular case of knowledge assessment, there is still much to be desired from the IRT. For example, the difficulty of the test item is an individual feeling of the evaluated, and in the models of IRT the difficulty is a parameter in the models of the test item, independent of the knowledge of the evaluated person, i.e. according to IRT models, the item is equally difficult for both the knowledgeable and the ignorant to answer correctly. Unconvincing is the solution in the IRT of the problem of the individual's tendency to accidentally guess the correct answer, included in some models as a parameter of the test item, and not as the individual's tendency to use or not guess when solving a test. In this regard, Ivailo Partchev, author of [3], commented in chapter 7.7. "Guessing and the 3PL model": "Items never guess—people do". Additional measures with a dubious effect have to be taken to counter the guessing [8].

In the last 40 years, a theory of knowledge assessment has gained popularity, based on the concept of a "knowledge space" to a given topic—a set of questions and connections between them that the learner could study, and on the idea that the examinee knows some of them ("knowledge state") [9-12]. The aim of the test (open-ended questions) is to assess this "knowledge state". The authors found the so-called "anthropometrical approach" influenced by the physical measurements unsuitable for assessing human qualities, in particular, they did not find it appropriate to compare a numerical assessment of the knowledge of the examinees. The result of this type of knowledge assessment is two short lists of problems: 'What the student can do' and 'What the student is ready to learn' concerning the specific "knowledge space". The exam is computerized. The result of the exam is obtained through an adaptive procedure, changing the course of the exam depending on the student's response to the question (correct, false, or not known).

Adequate assessment/self-assessment of knowledge and adaptive testing find application not only in the education system. More and more applications with knowledge assessment and adaptive testing with artificial intelligence are used in various highly specialized areas of professional life. In [13], the use of a multi-expert knowledge-aggregated adaptive assessment scheme using knowledge-based AI approaches to facilitate the learning of clinical core medical knowledge in otolaryngology is described.

2. Objective

The above-discussed theories CTT and IRT use the idea to evaluate the knowledge of the individual relatively—they evaluate individual knowledge in relation to the knowledge of a group of individuals (as a sample or the entire population).

The present paper describes a new informational approach to knowledge assessment through a test, completely different

from the concepts set in CTT and IRT, but remaining within the framework of the "anthropometrical approach". As mentioned above, anthropometrical assessment of knowledge is considered a special case of measurement, analogous to physical measurements, in which a numerical value is assigned to the knowledge of the examinee.

The process of knowledge assessment is regarded as a special case of an information process related to the generation, transmission, and perception of information. In the frame of the described informational approach, the knowledge of the individual is assessed absolutely—as the feature of the individual and it is not necessary to access the knowledge of a group of individuals to be assessed this feature of the individual.

3. Material and methods

3.1 Information system

A system with an information process in it is hereinafter referred to as an information system. In the particular case of knowledge assessment, the information system consists of:

1) Source of the information signal. In this case, it is the examinee (the evaluated), who has the *goal*, when sending an information signal.
2) An information signal directed to the recipient.
3) Recipient of the information signal. In this case, it is the examiner (the evaluator).

The examinee is a generator of the information signal to be perceived by the evaluator. An information signal in the case of the assessment process means the overall presentation of the knowledge that the examinee provides to the evaluator—depending on the signal carrier, this may be a written work, an oral presentation, or computer test answers.

By submitting the information signal, the evaluated has a specific goal—he/she strives to be as close as possible to achieving it—to answer correctly all the items in the test. Progress towards the goal is measured according to the criteria developed by the evaluator. They can be:

1) Applied directly by a person-evaluator, based on a general assessment of the information signal submitted by the evaluated – the so-called holistic assessment based on the professional experience of the evaluator. The assessment criteria remain unclear in the case.
2) Pre-clearly formulated criteria, with which the evaluated is already familiar at the stage of preparation for the exam. In particular, the criteria may be embedded in a computer test.

The evaluator receives the information signal and assesses the value (importance) of the received information, measuring the progress of the examinee toward achieving his goal. The value of the information signal is maximal when the goal is reached. The partial progress in moving towards the goal corresponds to the partial value of the information signal. The evaluator measures the knowledge of the evaluated through the value of the information signal.

3.2 Semantic branch of information theory

The main branch of Information Theory is related to solving the problems of machine transmission and coding of information. Outside the scope of researchers working in the main branch remain the human aspects of the information, such as its conscious generation, transmission, perception, understanding, and subjective evaluation. They are problems of the so-called "semantic" (meaningful) branch of the Information Theory. "The evaluation of human knowledge" can be classified among the problems of the semantic branch of Information Theory.

4. Results

4.1 Definition of the value of information signal applicable to the problem of knowledge assessment

The approach for assessing knowledge discussed below, as well as all the methods used for assessing knowledge, is indirect (as far as direct reading of thoughts is not possible)—the knowledge of the evaluated on a given topic is judged by his/her answers—his/her information signal. The information value (importance) of this signal is assessed through the progress of the evaluated towards reaching his/her goal—the correct answer to all items in the test. Progress is measured according to the criteria set by the evaluator. The value of the information signal is a characteristic of the knowledge of the evaluated on the topic. The value of the information signal ("knowledge") in the proposed information approach is obtained through "ignorance"—by assessing the progress towards the goal of the examinee without any knowledge of the examined topic.

For the assessment of the value of the information signal described below a probabilistic approach is used.

The proposed analytical type for the value of the information signal is based on two computable probabilities for random progress toward achieving the goal. For tests consisting of items with one correct answer, an indicator of progress towards reaching the goal is the number of correct answers that the evaluated has selected.

In principle, the accidental achievement of the mentioned goal is not impossible. If the number of items in the test is large, and the answers in each item are a finite number, usually 2, 3, 4, or 5, the probability of accidentally guessing the correct answers to all items is very small but greater than zero. In this case, “guessing” means that the examinee randomly chooses the answers to the items in the test, without having any knowledge of them that would affect his/her choice. For example, it happens if the examinee does not understand the language, in which the test is written.

This probability of accidental guessing of all items in the test decreases rapidly with an increase in the items included in the test and the number of answers to each of the items. For example, if an ignorant person solves a test with 20 items, each of which has 4 answers and only one of the answers is correct, the probability of accidental guessing of all items in the test is of the order of 10^{-14} (Figure 1). For comparison, the probability of a gamer guessing six specific numbers from 49 (as in the games of Fortune) is $7.2.10^{-8}$ (he becomes a millionaire). This is comparable with the situation if the examinee accidentally guessed 16 correct answers from the above-mentioned test.

The greater the progress of the evaluated towards achieving the goal, i.e. the more questions he/she answered correctly, the less likely it was to be accidental, and the more likely it was to be the result of available knowledge. Hereinafter, "ignorance" (the probability of accidental progress towards the goal) is used as a measure of the value (importance) of the information signal, which in turn is a characteristic of "available knowledge". A distinction must be made between the concepts of "available knowledge"—a quality of the individual that cannot be measured directly, and "value of the information signal"—a measurable quantity, a characteristic of available knowledge. The greater the value of the information signal, the lower the probability of accidental achievement of certain progress towards the goal, and in particular the value of the information signal is maximal at maximum progress, i.e. when it is sufficient to achieve the ultimate goal—correct answers to all questions in the test.

When solving a test by chance, the ignorant encounters correct answers too. The average number of correct answers that would receive a large group of ignorants, solving simultaneously and independently a test (or one ignorant without memory repeats the test many times), indicates the most likely (probable) progress towards the goal in the absence of knowledge. Accordingly, the information value of a signal leading to the most probable progress is zero. The ignorant could guess by chance each number of correct answers (random progress towards the goal). Each number of randomly guessed correct answers corresponds to a certain probability of guessing it by chance. This probability reaches its maximum at the most probable random progress, after which it rapidly decreases monotonically to its minimum at the maximum progress (when all randomly selected answers are correct, See the example in Figure 1). I.e., the curve of the dependency between the probability for a random number of correct answers on the number of correct answers has a maximum for correct answers number greater than zero. The most probable random progress depends on the number of items in the test and the number of answers per item.

If the examinee deliberately tries to avoid the correct answers, then he has set an "anti-goal", reaching which also requires knowledge.

In a test with n items, each number of correct answers (progress towards the goal) corresponds to the probability of achieving it by chance, according to the scheme in Table 1. As the number of correct answers increases from 0 to n, the corresponding probability initially increases, reaches a maximum P_{max}, and then monotonically decreases to $P(n)$. Its change ΔP (when changing the correct answers by one) is a negative function of the number of correct answers. The relative change $\Delta P/P$ of the probability is also a negative function of the number of correct answers, but it changes to a lesser extent ($0 \div -46.78$ in the example of Table 2 and Figure 1).

Table 1. The number of correct answers in the test (progress towards the goal) and the probability of achieving it by chance

Number of correct answers (progress towards the goal)	0	1	…	m	…	k	…	n
Probability	$P(0)$	$P(1)$	…	P_{max}	…	$P(k)$	…	$P(n)$

Below, the relative change in the probability $\Delta P/P$ is taken as a measure of the change in the value ΔV of the information signal. The two dependencies have opposite signs, i.e. the decrease in the probability of random progress towards the goal corresponds to a proportional increase in the value of the information signal. The analytical expression of this definition of change of the information signal value is:

$$\Delta V = -\frac{\Delta P}{P}, \tag{1}$$

The result of an experiment described below confirms the correctness of the choice of the dependence definition (1).

After integration,

$$V = -\ln P + const. \tag{2}$$

The value of the constant can be determined by the condition that the most probable random progress towards the goal corresponds to the zero value of the information signal, i.e. $V = 0$ at $P = P_{\max}$. Therefore, from (2):

$$0 = -\ln P_{\max} + const\ , \tag{3}$$

and from (2) and (3) for the value of the information signal in the case of knowledge assessment follows:

$$V = -\ln P + \ln P_{\max} = \ln \frac{P_{\max}}{P}. \tag{4}$$

The formula applies to a number of correct answers equal to or greater than the number m of the most probable random progress towards the goal, i.e. for each number of correct answers k from the interval $m \le k \le n$.

The value of the information signal is an additive characteristic—the total value for independent tests is the sum of the values for each of them separately.

In the proposed definition of the value of information signal in knowledge assessment, probabilities are computable values applicable to knowledge assessment through tests. In the model of the value of the information signal, there is a clear criterion for zero value of the information signal, i.e. when the assessed has no knowledge of the topic of the exam. The zero value corresponds to a calculable value—the maximum probability of accidental progress towards the goal. This turns the set of estimates of information signal values into a scale of relations—the most informative type of measurement scale (with a natural zero, the relations between the values are allowed [14]). For comparison, the Celsius temperature scale is a scale of intervals (without natural zero)—a less informative scale, in this scale relations between the temperatures are not allowed (It is incorrect to say that 2°C is two times greater temperature than 1°C). Kelvin temperature scale is a scale of relations (with natural zero, and it is correct to say that 2K is two times greater temperature than 1K).

The probabilities in the proposed formulae (4) for the value of information are computable quantities and the value of the information signal has a natural zero.

Guessing in solving a test for knowledge assessment is a problem in the analysis of tests through the CTT and the IRT. In the proposed information approach to the assessment of knowledge through a test, guessing is integrated and taken into account in the assessment process—the examinee is free to guess—no special measures and sanctions are required against the guessing by the examinee.

4.2 Assessment of a test through the semantic information approach

The most technological type of test is a multiple-choice test, consisting of items with the same number of answers, only one of which is the correct one. If the ignorant guesses solving the multiple choice test, then the Bernoulli formula is applicable for calculating the probability $P_n(k)$ to randomly guess k correct answers from a total of n items in the test [15]:

$$P_n(k) = \binom{n}{k} p^k q^{n-k} \quad , p + q = 1 \tag{5}$$

where p denotes the probability of accidentally guessing the correct answer to a particular test item. If all the items in the test are of the same type, this probability is the same for all items. With q is denoted the probability of accidental choice of an incorrect answer from the answers in the item. The sum of the probabilities for the random choice of an answer among the answers of the item is $p + q = 1$.

The most probable number of correct answers, randomly chosen by the ignorant, is the closest integer to np [15]. To obtain the maximum probability $P_{\max}$ corresponding to the most probable number in random progress to the goal, k in (5) is replaced by np. For $P_{\max}$, we get:

$$P_{\max} = P_n(np) = \binom{n}{np} p^{np} q^{nq}. \tag{6}$$

For the value of the information signal after substitution of (5) and (6) in (4) is obtained the expression:

$$V_n(k) = \ln \frac{\binom{n}{np} p^{np} q^{nq}}{\binom{n}{k} p^k q^{n-k}} \tag{7}$$

This exact expression (7) of the value of the information signal is harder to use in calculations. An easier-to-apply formula for $P_n(k)$ is obtained by the de Moivre-Laplace formula, which is the more accurate approximation of (5) the more items are included in the test. For $P_n(k)$, expressed by the de Moivre-Laplace formula [15], we obtain:

$$P_n(k) \cong \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}} . \tag{8}$$

Since the most probable progress k in the random movement to the goal is equal to np, the exponent in the above formula when calculating the maximum probability is 1, and the maximum probability is equal to the coefficient in front of the exponent in (8).

After substituting (8) and the expression for the maximum probability in (7), the value of the information signal is obtained:

$$V_n(k) \cong \frac{(k-np)^2}{2npq} , \; k \geq np \tag{9}$$

The maximum value of the information signal is reached when $k = n$, i.e.:

$$V_n(n) \cong \frac{(n-np)^2}{2npq} = \frac{n^2q^2}{2npq} = \frac{nq}{2p} \tag{10}$$

The relative value is the ratio between the value of the information signal, corresponding to randomly chosen k correct answers, to the maximum information value. Relative value is a characteristic of the relative progress of the evaluated toward the goal, i.e. the available knowledge of the examinee as regards the knowledge needed to achieve the goal. For the relative value we get:

$$\frac{V_n(k)}{V_n(n)} = \left(\frac{\frac{k}{n} - p}{q} \right)^2 , k \geq np . \tag{11}$$

The available knowledge corresponding to the achieved progress towards the goal can be evaluated according to formula (11) in values between 0 (at $k = np$) and 1 (at $k = n$). This scale is a scale of ratios, as explained above.

Educational systems around the world use numerical grading scales to assess knowledge, in which grades vary within the traditionally defined range r from the minimum grade ϑ_{min} to the maximum ϑ_{max} ($r = \vartheta_{max} - \vartheta_{min}$).

The traditional scale would also be a relations scale if the same available knowledge is expressed on the one hand as a ratio of grades of the traditional scale and on the other as a relative value of the information signal (11).

The ratio in the grades on the traditional scale would be:

1. with a numerator the difference between the grade ϑ for the achieved progress towards the goal, and the minimum possible grade ϑ_{min}, i.e. $\vartheta - \vartheta_{min}$,
2. with a denominator $r = \vartheta_{max} - \vartheta_{min}$.

After equating this ratio with (11), is obtained the expression:

$$\frac{\vartheta - \vartheta_{min}}{\vartheta_{max} - \vartheta_{min}} = \frac{V_n(k)}{V_n(n)} , \tag{12}$$

from which the grade ϑ follows:

$$\vartheta = r \frac{V_n(k)}{V_n(n)} + \vartheta_{min} = r \left(\frac{\frac{k}{n} - p}{q} \right)^2 + \vartheta_{min} \tag{13}$$

In particular, for the grades scale used in Bulgaria $\vartheta_{min} = 2$, $\vartheta_{max} = 6$, i.e. $r = 4$.

The most commonly used tests consist of items with 3, 4, and 5 answers. For them for Bulgaria, formula (13) has the form:

$$\vartheta_{3answers} = 4\left(\frac{\frac{k}{n}-\frac{1}{3}}{\frac{2}{3}}\right)^2 + 2 = \left(3\frac{k}{n}-1\right)^2 + 2\,,\ \frac{k}{n} \geq \frac{1}{3}.$$

$$\vartheta_{4answers} = 4\left(\frac{\frac{k}{n}-\frac{1}{4}}{\frac{3}{4}}\right)^2 + 2 = \frac{4}{9}\left(4\frac{k}{n}-1\right)^2 + 2\,,\ \frac{k}{n} \geq \frac{1}{4} \tag{14}$$

$$\vartheta_{5answers} = 4\left(\frac{\frac{k}{n}-\frac{1}{5}}{\frac{4}{5}}\right)^2 + 2 = \frac{1}{4}\left(5\frac{k}{n}-1\right)^2 + 2\,,\ \frac{k}{n} \geq \frac{1}{5}.$$

The evaluation by the derived formulas can be judged for reliability with data in Table 2 below.

The probability of erroneous evaluation when using the scale from the type presented in Table 2 is very small. The criterion used in practice in Bulgaria for the successfully solved test is that the test was taken successfully if the grade is at least 3.00. The example in Table 2 shows, that a grade of 3.14 corresponds to 13 correct answers out of 20 possible. As can be seen from Table 2, the probability of accidentally encountering 13 correct answers is of the order of 10^{-5}, i.e. only one out of one hundred thousand ignorant evaluated by the test from the example in Table 2, would pass the exam erroneously.

Table 2. Test parameters of the example of a test with 20 items with 4 equally probable answers, one of which is correct. The grade is from a scale of grades between 2.00 and 6.00, used in Bulgarian education

k - progress towards the goal (number of correct answers), $k \geq 5$	Probability P to randomly choose k correct answers	Difference in probabilities $\Delta P = P(k) - P(k\text{-}1)$	Relative change in probability $\Delta P/P$	Value of the information signal $V = \ln P_{max}/P$	Numerical grade ϑ
5	$2.06.10^{-01}$				2.00
6	$1.80.10^{-01}$	$-2.57.10^{-02}$	-0.14	0.13	2.02
7	$1.21.10^{-01}$	$-5.94.10^{-02}$	-0.49	0.53	2.07
8	$6.20.10^{-02}$	$-5.88.10^{-02}$	-0.95	1.20	2.16
9	$2.44.10^{-02}$	$-3.76.10^{-02}$	-1.54	2.13	2.28
10	$7.35.10^{-03}$	$-1.71.10^{-02}$	-2.32	3.33	2.44
11	$1.70.10^{-03}$	$-5.65.10^{-03}$	-3.33	4.80	2.64
12	$3.00.10^{-04}$	$-1.40.10^{-03}$	-4.66	6.53	2.87
13	$4.05.10^{-05}$	$-2.59.10^{-04}$	-6.39	8.53	3.14
14	$4.20.10^{-06}$	$-3.63.10^{-05}$	-8.65	10.80	3.44
15	$3.34.10^{-07}$	$-3.87.10^{-06}$	-11.60	13.33	3.78
16	$2.03.10^{-08}$	$-3.13.10^{-07}$	-15.44	16.13	4.15
17	$9.45.10^{-10}$	$-1.93.10^{-08}$	-20.47	19.20	4.56
18	$3.37.10^{-11}$	$-9.11.10^{-10}$	-27.03	22.53	5.00
19	$9.21.10^{-13}$	$-3.28.10^{-11}$	-35.60	26.13	5.48
20	$1.93.10^{-14}$	$-9.02.10^{-13}$	-46.78	30.00	6.00

Figures 1 and 2 show graphically the data in Table 2.

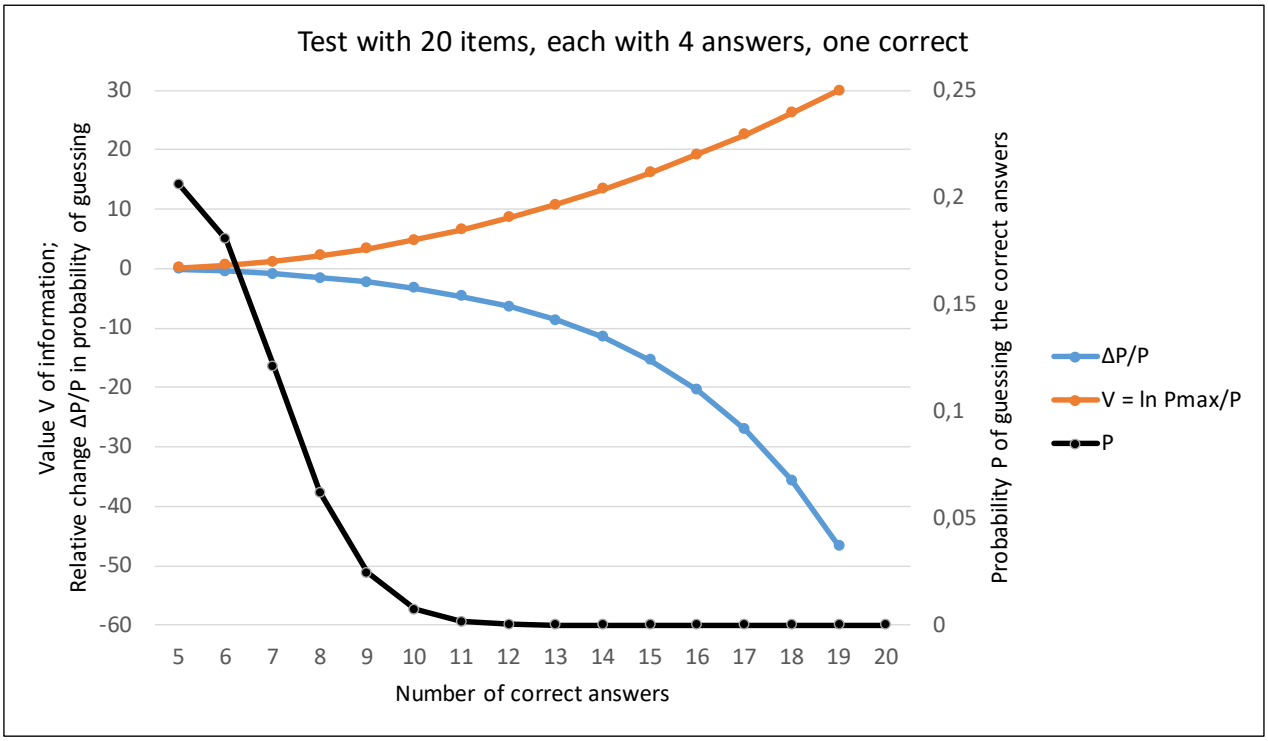

Figure 1. The probability *P* of an ignorant person progressing randomly to the goal – correct answers to all items in a test with 20 items, decreases rapidly with the progress toward the goal. The relative change in probability Δ*P*/*P* is also shown, as well as the change in the value *V* of the information signal (see below).

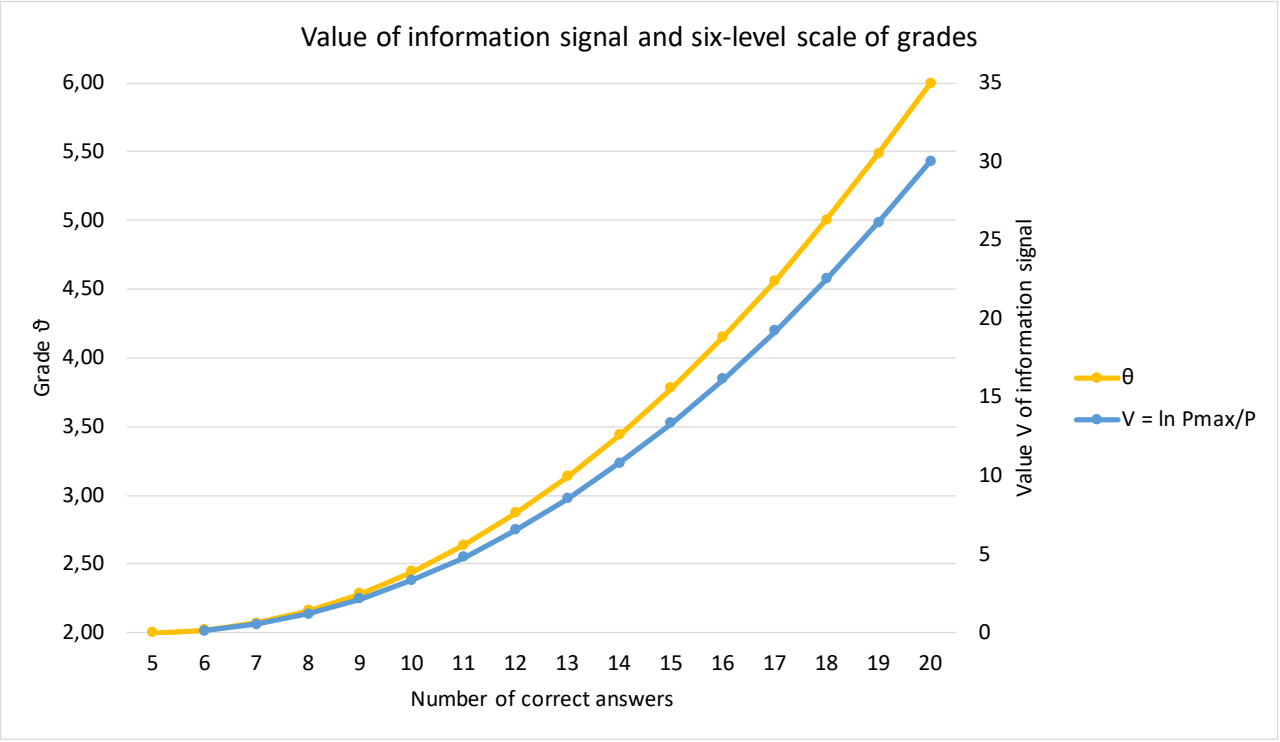

Figure 2. Dependences of the value of the information signal and the numerical grade from the scale of grades between 2.00 and 6.00 on the progress towards the goal according to the data in Table 2.

4.3 An experiment with an examination of a group of students simultaneously with computer tests and human experts

An experiment was conducted, the purpose of which was to check to what extent the assessment of knowledge by a computer test coincides with the human expert assessment of knowledge if the assessment algorithm in the computer test is based on a formula (14).

In the experiment, a group of 82 students answered in writing 20 open-ended (without answers) physics questions, immediately after which they solved a computer test with the same questions, but of a closed type—with 4 answers.

The written responses were assessed independently by 10 physics teachers on a grade scale between 2.00 and 6.00. The mean grades of the written papers were compared with the computer grades. The teachers' average grades varied between 4.43 and 5.08, and the average of all the teachers' grades—the "exact" grade—was 4.87. The average grade on the computer test was 4.76. Only two of the teachers had a better match of the average grade to the "exact" grade. I.e. the grade from the computer test is unbiased from the “exact"—the computer test has no systematic error—it does not increase or decrease the average grade from the "exact”.

In Figure 3, for each student in the experiment, the relationship between the computer grade and the corresponding average teacher grade of the written work is shown graphically.

By differentiating the linear model of the obtained dependence between the grades from the computer test and the expert evaluations of the written works (shown in Figure 3), it turns out that the change in the computer grade is almost equal to the change in the teacher's grade—the coefficient (0.962) is very close to the “ideal” (1.000). I.e., computer evaluations obtained by formula (14) are adequate with the human evaluation of knowledge. Therefore, the choice of (1) as the definition of the value of the information signal leads to an objective evaluation of knowledge, adequate to the human one.

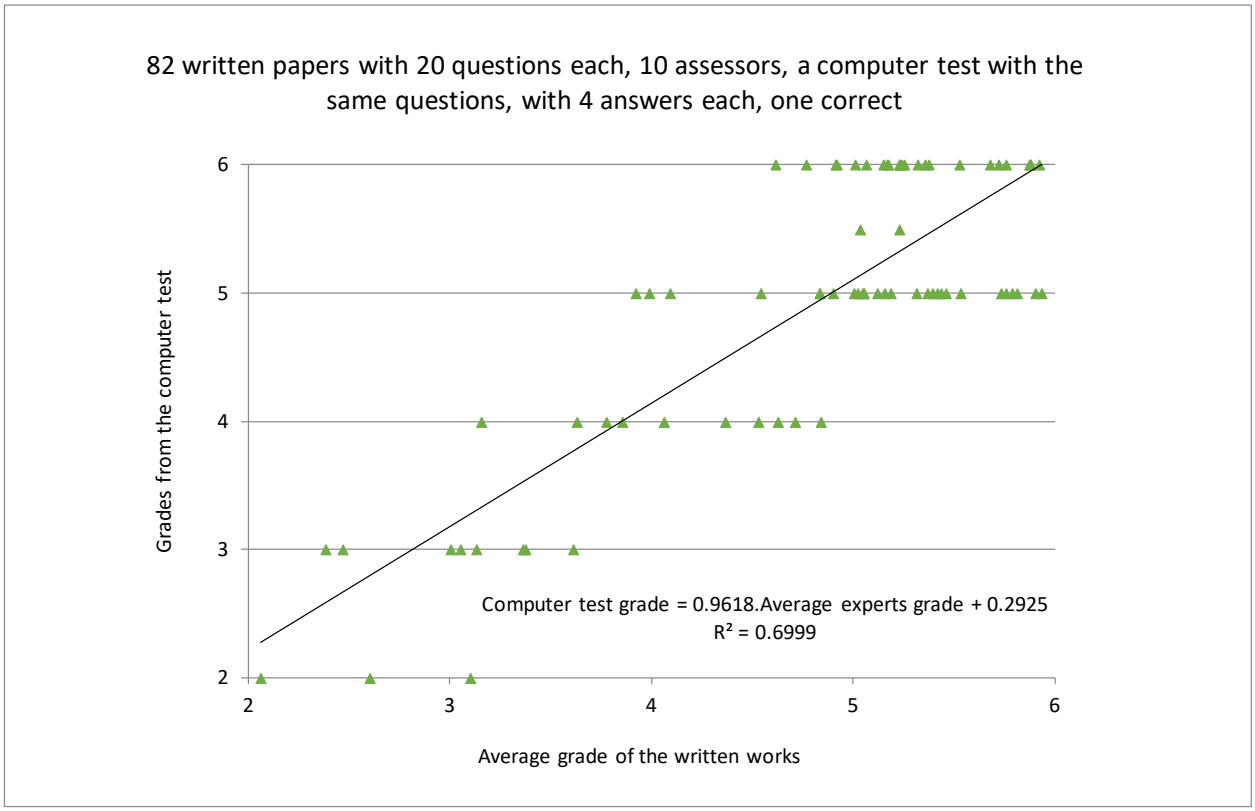

Figure 3. Correlation between computer test grades and their corresponding mean teacher grades of written works.

4.4 Minimum number of items in a dichotomous test for a scale with non-degenerate scores

Traditionally, the interval r of grades in the scale is divided into equal sub-intervals (scores) of grades. Scales with different scores, in some cases dozens, are used in evaluation practice around the world. Scores also have a verbal expression, for example, a “six-score” scale used in the universities in Bulgaria includes 5 scores: "poor", a grade scale interval between 2.00 and 2.99, "satisfactory", a grade scale interval between 3.00 and 3.49, "good", a grade scale interval between 3.50 and

4.49, "very good", a grade scale interval between 4.50 and 5.49, and "excellent", a grade scale interval between 5.50 and 6.00.

The grade from the computer test (14) can be compared with a point that falls in one of the scores on the scale. The distribution of point grades on the scale depends on the number of items in the test and the number of answers in each item. The density of point grades is uneven. As the grade increases, the density of the point grades decreases according to the quadratic law, i.e. the highest grades are most distant on the scale.

In traditional assessment practice, the final result of the knowledge assessment is presented through the score in which the point grade falls.

If the items in the test are few, due to the uneven distribution of the grades in the scale interval, the scale can have more scores than the possible point grades. I.e. the scale has "empty" scores that do not correspond to a point grade. The term "degenerate" is used for such a scale below. The use of degenerate scales is not logically justified.

Below is a criterion for the minimum number of items in a test so that the scale is nondegenerate.

The difference between the point grades $\Delta\vartheta$ depends on the difference Δk of the number of correct answers in the test, and can be calculated by differentiating from (13):

$$\Delta\vartheta = 2r\left(\frac{\tau - p}{q}\right)\frac{\Delta\tau}{q} = 2r\left(\frac{\tau - p}{q}\right)\frac{\Delta k}{qn},\ \tau \geq p\,, \tag{15}$$

where the notations are the same as in formula (13).

If the difference between the correct answers Δk is fixed at its minimum value 1 ($\Delta k_{\min} = 1$), it follows from (15) that the difference between neighbor point grades $\Delta\vartheta$ is proportional to the ratio τ between the number of correct answers and the number of items in the test. The difference in the point grades reaches a maximum $\Delta\vartheta_{\max}$ at $\tau_{\max} = 1$. From (15), the maximum difference of the neighbor point grades is obtained $\Delta\vartheta_{\max} = 2r/(qn)$. If the number n of items in the test is small, $\Delta\vartheta_{\max}$ is large.

If the number of scores in the scale is N and the scale is uniform—with equal-sized scores, the interval of grades for one score is $\Delta B = r/N$. A situation can arise in which $\Delta\vartheta_{\max} > 2.\Delta B$, i.e. to have a blank score between two point grades (the scale is degenerate for this test). Such a situation is the result of the traditional choice of the number of scores in the scale in a given education system and the assumption that the intervals of grades corresponding to all scores are the same in size. Point grades are not related to this degeneration.

Therefore, for the scale to be non-degenerate for a specific test with n items in it, it is necessary to meet the criterion:

$$\Delta\vartheta_{\max} < 2.\Delta B \tag{16}$$

or

$$\frac{2r}{qn} < \frac{2.r}{N} \rightarrow \frac{N}{q} < n_{\min}\,. \tag{17}$$

For knowledge assessment in the universities in Bulgaria, the “six-score” scale, mentioned above (in fact five-score scale, N = 5), is used. For example, for the scale to be non-degenerated for a test with 3, 4, and 5 answers, the minimum number $n_{\min}$ of items in the test, calculated from (17) is given in Table 3:

Table 3. The minimum number $n_{\min}$ of items in the test

Answers per item	q	N/q	$n_{\min}$
3	2/3	7.33	8
4	3/4	6.67	7
5	4/5	6.20	7

4.5 Requirements for test items

The above conditions, under which formula (5) is valid, impose additional requirements on the test items, which must be observed in the preparation of the tests for the obtained formulas for assessment of knowledge by test to be applicable. Many of these requirements are met in the most widely used tests. The more the test complies with these requirements, the more applicable the resulting assessment formula (14) is to that test.

1) The answers must be constructed in such a way that they appear to be equally meaningful so that the examinee cannot guess which is the correct answer by side signs as a difference in length or shape between the correct and incorrect answers.

In particular, if the correct answer contains a list of concepts, distractors can be constructed as lists containing incorrect concepts cyclically mixing with the correct concepts. The symmetry in the answers equalizes the probabilities for the examinee to guess the correct answer. For example, if an item with 4 answers has the question: "Which colors are in the USA flag?" with a correct answer: "White, blue, red", the distractors can be constructed cyclically: "Green, white, blue", "Red, green, white" and "Blue, red, green".

2) When composing tests, certain words or phrases are more often used to compose distractors and should be avoided, for example, words like "only", and expressions like "All answers are correct". The examinee knows this and can use it to eliminate some of the answers, increasing the likelihood of guessing the correct answer.

3) There should be no items in the test with partially correct distractors. For example, to the question: "Specify the vegetables:" the test compiler has indicated for a correct answer "cabbage, lettuce, tomato" and a distractor "cabbage, lettuce, apple", i.e. the distractor is a partially correct answer (vegetables are also indicated) and the item is not dichotomous. In this case, the question should be modified as follows: "Indicate the answer containing only vegetables:". The suggestive word "only", often used to construct distractors, moved from the distractor to the question, improves the logical structure of the item, making it dichotomous.

4) Assessors often tend to give more weight to a certain item than to others multiplying the correct answer with a coefficient (more than 1). The information content of all items in the test (with the same number of answers) is the same, for example for items with four answers it is 2 bits, so it is equally easy to guess the correct answer of the "more valuable" item as well as of the "less valuable". In case of an accidental correct answer to an item with a high coefficient of weight, the evaluated person would be unfairly overestimated, and for the accidental incorrect answer to such an item, the evaluated person would be unfairly punished. I.e. it is inadmissible to assign a higher value to individual items in a test of the discussed type than to others, by assigning weight coefficients to them. The problem could be solved at the stage of test development. The pre-evaluated topic is broken down by the test developer (evaluator) into separate concepts. A rank (weighting factor) is assigned to each of them, showing its importance for the evaluated topic. Then for each concept, a certain number of items proportional to its rank are developed (See substantive validation of the test [1]).

4.6 An adaptive algorithm

The detailed assessment of knowledge through a test requires testing with a large number of items in the test, the answer to which would take a considerable time for assessment. The term "items bank" is used below for all items included in the test at the stage of its development. If a computer version of the test is used, the exam can be shortened in time by reducing the number of items assigned during the exam in comparison with their number in the bank of items. Reducing the time simply by reducing the number of assigned items hides risks of unacceptable reduction of the detail of the test and the accuracy of assessment of knowledge on the topic. The detail and accuracy of the assessment will not be compromised if:

1. The items in the bank meet the requirement for substantive validation of the test [1] concerning the tested topic, i.e. all items test the knowledge on this topic. For such a test, dropping some of the items does not significantly reduce the detail of the assessment.

2. For the examination, the items are randomly drawn from the items bank and submitted sequentially one after the other, with only one of them visible on the screen. The next item appears after the answer to the previous one. The examinee cannot return to previous items, even just to see them.

3. The assessment of the shortened version of the test remains within the permissible deviation from the exact assessment—the one that the examinee would have received if he had solved the test with all the items from the bank.

To reduce the time of the exam in compliance with the above conditions, an adaptive assessment algorithm is needed, which, depending on the frequency of the correct answers indicated by the examinee, changes the number of items set during the exam. The aim is to reduce the time for the exam in as many cases as possible—for the exam to end before all the items of the bank have been exhausted, without negatively affecting the accuracy of the assessment of the knowledge of the examinee.

The paper proposes an adaptive algorithm of the type described above, based on the semantic information model for knowledge assessment through a test containing dichotomous items, with an equal number of answers. The test should be designed so that the answers to each item seem equally likely to the examinee who does not know the correct answers. With $\tau = k/n$ denoted the ratio between the number of correct answers k and the number n of items in the test, formula (13) takes the form:

$$\vartheta = r\left(\frac{\tau - p}{q}\right)^2 + \vartheta_{\min}, \quad \tau \geq \mathrm{p}, \tag{18}$$

in which the remaining notations are the same as in formula (13).

From (1), the opposite task can be solved too: to determine the number of correct answers through which a specific grade is achieved when solving the test:

$$\tau = \frac{k}{n} = q\sqrt{\frac{\vartheta - \vartheta_{\min}}{r}} + p \;\rightarrow\; k = n\left(q\sqrt{\frac{\vartheta - \vartheta_{\min}}{r}} + p\right),\; k \geq np\;. \tag{19}$$

The numerical grade corresponds to a point on the axis on which the grade scale can be plotted, so the term "a point grade" is also used below. The final result of the exam often is the score (sub-interval on the grade scale) in which the point grade falls.

Table 4 presents the example of all possible values of point grades for a test with $n = 20$ items with 4 answers each—from $n = 8$ to $n = 20$ and $k \geq 5$ ($\tau \geq 1/4$ according to the constraint in (18)). The traditional criterion for successfully passing a test applied in Bulgaria is the grade to be at least 3.00. Table 4 shows that the test is not solved successfully if the examinee, solving a test with all 20 items in the bank, has chosen no more than 12 correct answers (for which the grade is 2.87). I.e. if no correct answer is selected after the examinee has answered 8 items, the adaptive algorithm can terminate the test at the earliest. This number of items is the difference between the number of items in the bank (20 in the example) and the maximum number of correct answers (12 in the example) for which the criterion for taking the exam is not met in the case of a test with all items in the bank.

Let a test contain a bank with N items. In the process of adaptive testing, the examinee responds to a certain number $n \leq N$ of randomly withdrawn items from all items in the bank (non-repeat sample). To some of the items (or to all), the examinee gives correct answers.

Let k_N indicate the number of correct answers that the examinee would give depending on his/her knowledge of the assessed topic in solving all N items in the bank. The ratio $\tau_N = k_N/N$, substituted in (18), would give the exact grade for the knowledge of the examinee. If in the course of the test, the examinee answers $n < N$ randomly drawn items (sample) from all items in the bank and the number of given correct answers is denoted by k, then the ratio is $\tau = k/n$. The ratio τ is a sample estimate of the exact ratio τ_N. In the test process, k, n, and τ are known after each item response but the exact τ_N ratio remains unknown because k_N is unknown.

Table 4. Point grades ϑ, corresponding to each of the admissible combinations between the number k of correct answers and the number n of the drawn items in a test with a bank of $N = 20$ items. With smaller symbols are the grades that do not meet the criterion to pass successfully the exam. The grades of at least 3.00 (the exam is passed with success) are shown in bigger and bold characters

ϑ	k	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n																	
8		**3,00**	**3,78**	**4,78**	**6,00**												
9		2,66	**3,23**	**3,98**	**4,90**	**6,00**											
10		2,44	2,87	**3,44**	**4,15**	**5,00**	**6,00**										
11		2,30	2,62	**3,06**	**3,62**	**4,30**	**5,09**	**6,00**									
12		2,20	2,44	2,79	**3,23**	**3,78**	**4,42**	**5,16**	**6,00**								
13		2.13	2.32	2.59	2.95	**3.39**	**3.92**	**4.53**	**5.22**	**6.00**							
14		2.08	2.23	2.44	2.73	**3.10**	**3.53**	**4.04**	**4.62**	**5.27**	**6.00**						
15		2.05	2.16	2.33	2.57	2.87	**3.23**	**3.66**	**4.15**	**4.70**	**5.32**	**6.00**					
16		2.03	2.11	2.25	2.44	2.69	**3.00**	**3.36**	**3.78**	**4.25**	**4.78**	**5.36**	**6.00**				
17		2.01	2.08	2.19	2.35	2.56	2.81	**3.12**	**3.48**	**3.88**	**4.34**	**4.84**	**5.40**	**6.00**			
18		2.01	2.05	2.14	2.27	2.44	2.66	2.93	**3.23**	**3.59**	**3.98**	**4.42**	**4.90**	**5.43**	**6.00**		
19		2.00	2.03	2.10	2.21	2.36	2.54	2.77	**3.04**	**3.34**	**3.69**	**4.07**	**4.49**	**4.96**	**5.46**	**6.00**	
20		2.00	2.02	2.07	2.16	2.28	2.44	2.64	2.87	**3.14**	**3.44**	**3.78**	**4.15**	**4.56**	**5.00**	**5.48**	**6.00**

If the difference between the ratios τ_n (known) and τ_N (unknown) is negligible (statistically insignificant) at a number $n < N$ of the randomly withdrawn items, the test grade calculated after substituting τ_n in (17) will not differ significantly from its exact value, calculated with τ_N (if its value was known). If at a given time during the test, a non-repeated sample of n items was withdrawn randomly and the difference between τ_n of τ_N became less than one limit value—the maximum deviation $\Delta\tau_{max}$, then this difference is negligible and the grade may be calculated from (18) with τ_n instead of τ_N without significant loss of accuracy and the adaptive algorithm could finish the exam. The maximum deviation $\Delta\tau_{max}$ is calculated by the formula [15]:

$$\Delta\tau_{max} = t_{\alpha,n-1}\sqrt{\frac{\tau(\tau-1)}{n}\cdot\frac{N-n}{N-1}}, \tag{20}$$

where $t_{\alpha,n-1}$ denotes the critical points of Student's t-criterion for significance level α and degree of freedom n-1.

The values of $t_{\alpha,n-1}$ for a significance level of 0.05 and a different number of n items in the test are given in Table 5 [16].

Table 5. The values of $t_{\alpha,n-1}$ for a significance level of 0.05 and a different number of n items in the test [16]

n	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$t_{0.05,\,n-1}$	2.45	2.37	2.31	2.26	2.23	2.20	2.18	2.16	2.14	2.13	2.12	2.11	2.10	2.09
n	21	22	23	24	25	26	27	28	29	30	40	60	120	∞
$t_{0.05,\,n-1}$	2.09	2.08	2.07	2.07	2.06	2.06	2.06	2.05	2.05	2.05	2.02	2.00	1.98	1.96

After $n < N$ test items were answered, the maximum deviation $\Delta\tau_{nmax}$ of τ_n from its corresponding exact value τ_N can be calculated from formula (20) and Table 5.

Figure 4 shows the dependence of τ on the number of correct answers k for tests with a different number n of items representing a non-repeated sample from a bank of test items with volume $N = 20$. Each such test corresponds to a segment of a line marked in the figure with τ_n. The conclusions below are made using this example. It can be seen that as the number of correct answers increases, the straight sections corresponding to the tests with a different number of items move away from each other, i.e. the difference between τ_n and τ_N also increases (in the example τ_N coincides with τ_{20}).

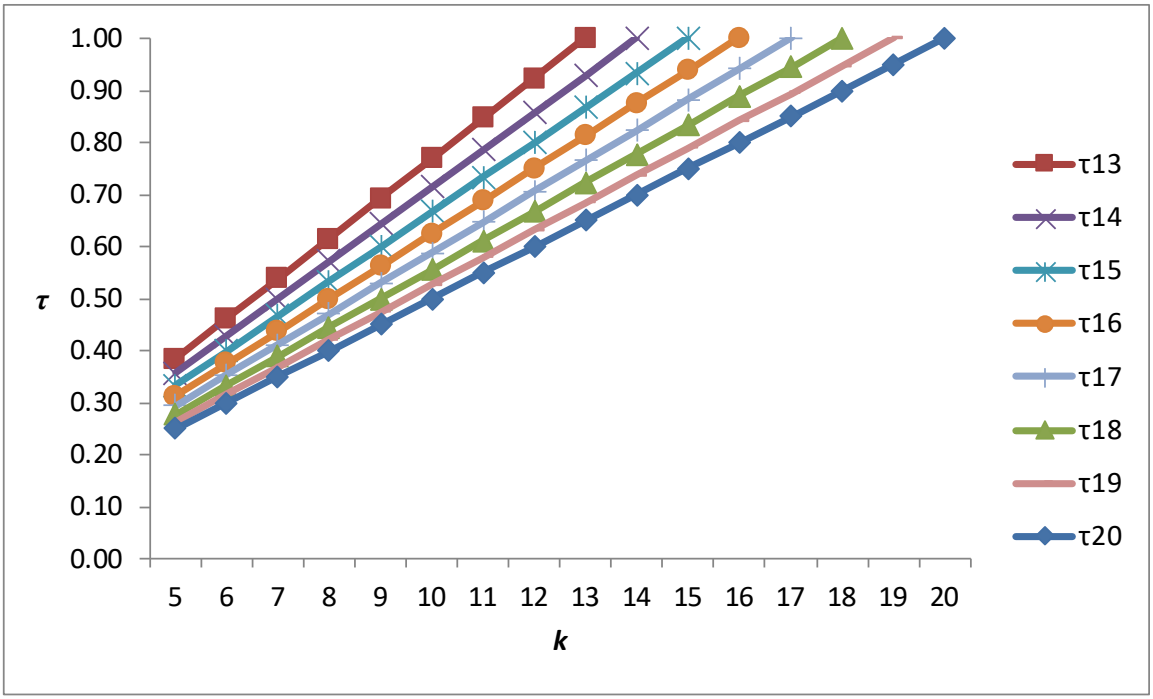

Figure 4. Dependence of the ratio $\tau = k/n$ on the number of correct answers k in tests in which the number n of items is a non-repeated sample withdrawn from a bank with $N = 20$ items. With τ_n is denoted the corresponding number of items – from $n = 13$ to $n = 20$.

Figure 5 with an arrow shows graphically the maximum deviation $\Delta\tau_{max}$ of $\tau_{13,11}$ (for a test with 13 items and 11 correct answers) in the direction of its exact value τ_N, which is a point on the segment τ_{20} (corresponds to a test with the maximum number of items, $N = 20$).

If for a given k the absolute value of the distance between τ_n and τ_N is less than $\Delta\tau_{max}$, i.e. the condition is fulfilled:

$$\left|\tau_{nk} - \tau_N\right| \le \Delta\tau_{\max}, \tag{21}$$

then the ratio τ_{nk} is a statistically unbiased estimate of the exact ratio τ_N and the difference between them is statistically insignificant. I.e. if this condition is met, the test grade can be calculated with the value of τ_{nk} substituted in formula (18). The grade thus obtained would differ statistically insignificantly from the exact grade, although it was obtained through a sample of the bank of items. As far as τ_N is unknown, the fulfillment of condition (21) cannot be verified in this form.

This condition can be transformed into the condition:

$$\left|\tau_{nk} - \tau_{Nk}\right| \le \Delta\tau_{nk\ \max}, \tag{22}$$

where τ_{nk} and τ_{Nk} denote the values of the ratios τ and τ_N for k correct answers. In condition (22) the difference τ_{nk} - τ_{Nk} contains computable quantities, i.e. is known. But the maximum deviation $\Delta\tau_{nkmax}$ for which the ratios τ_{nk} and τ_{Nk} differ statistically insignificantly is unknown. Geometric considerations were used to determine it.

Figure 5 represents the line segment τ_{20} corresponding to a test with the full number $N = 20$ items. For k correct answers $\tau_{Nk} = k/N$. The change $\Delta\tau_{Nk}$ with the change Δk determines the slope of this line relative to the abscissa. At $\Delta k = 1$,

$$\Delta\tau_{Nk} / \Delta k = \tau_{Nk+1} - \tau_{Nk} = (k+1)/N - k/N = 1/N. \tag{23}$$

The same slope is equal to tgα, where α is the angle between the abscissa and the segment τ_N (in example τ_{20}). It follows from (23) that this angle is:

$$\alpha = \arctan(1/N). \tag{24}$$

The same angle is concluded between $\Delta\tau_{max}$ and $\Delta\tau_{nkmax}$ (Figure 5), i.e.

$$\Delta\tau_{nk\max} = \Delta\tau_{\max} / \cos\alpha. \tag{25}$$

From (24) and from [17] it follows:

$$\cos(\arctan(1/N)) = 1/(1+1/N^2)^{1/2}, \tag{26}$$

and

$$\Delta\tau_{nk\max} = \Delta\tau_{\max}.(1+1/N^2)^{1/2}. \tag{27}$$

For example, at $N = 20$, the expression $(1+1/N^2)^{1/2} = 1.001249$, decreases with increasing N and can be assumed to be 1 without affecting noticeable the accuracy of the calculation. I.e. condition (22) can be converted to

$$\left|\tau_{nk} - \tau_{Nk}\right| \le \Delta\tau_{nk\ \max} \cong \Delta\tau_{\max}, \tag{28}$$

or

$$\left|\tau_{nk} - \Delta\tau_{nk\max}\right| \le \tau_{Nk}, \tag{29}$$

In condition (28) all quantities are computable. If it is fulfilled, the difference $\tau_{nk} - \tau_{Nk}$ is statistically insignificant.

Figure 6 shows the dependence of three of the differences τ - $\Delta\tau_{max}$ on the number of correct answers: τ_{13} - $\Delta\tau_{13}$ (for a sample test of 13 of all 20 items), τ_{16} - $\Delta\tau_{16}$ (sample test of 16 of all 20 items), and τ_{19} - $\Delta\tau_{19}$ (sample test of 19 of all 20 items). The notations are the same as in Figure 5. The straight line τ_{20} also is shown in Figure 6. The differences $\tau_{nk} - \Delta\tau_{nkmax}$ are curved lines that approach the straight line τ_{20} at its various points for which condition (29) is satisfied. In them, the difference between the sample τ_n and the exact τ_N as well as between sample grade and exact grade is insignificant. For example, the figure shows that if after answering 13 items the examinee answered correctly only 7 of them, the difference between ratio $\tau_{13,7}$ and $\tau_{20,7}$ is negligible. The adaptive algorithm could terminate the test after the 13th answer with a grade of 2.59 (Table 4). If the grade is in a corresponding score “poor” which is between grades 2.00 and 2.99 in mentioned above “six-score” scale, the exam will not be passed. Similarly, if after answering 16 items the examinee gave 10 correct answers, the ratio $\tau_{16,10}$ is negligibly different from $\tau_{20,10}$ and the test can be terminated on the 16th item with a grade of 3.00 (score "satisfactory", the exam will be taken successfully). For a test with 19 items and 16 correct answers, the ratio $\tau_{19,16}$ is negligibly different from $\tau_{20,16}$ the adaptive algorithm should terminate the test with a grade of 4.49 (score "good").

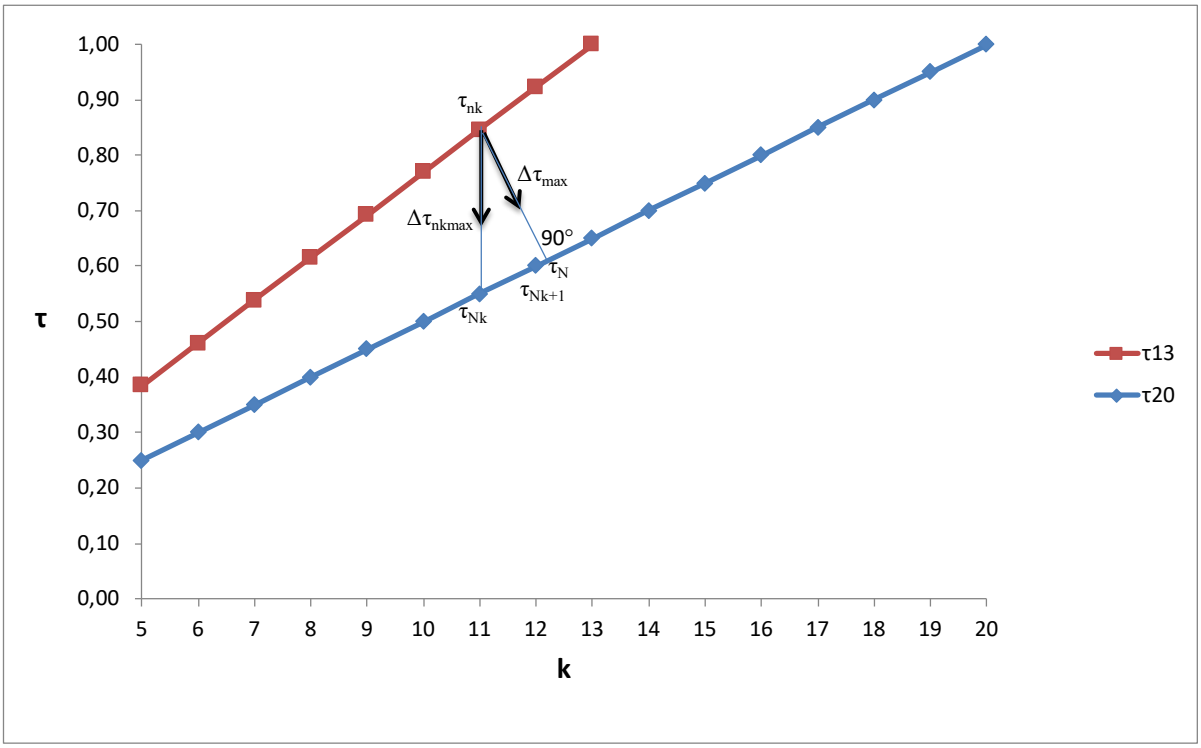

Figure 5. Graphical representation of the deviation of the ratio $\tau_{13,11}$ at the point $k = 11$ from the exact ratio τ_N and from the ratio τ_{Nk}.

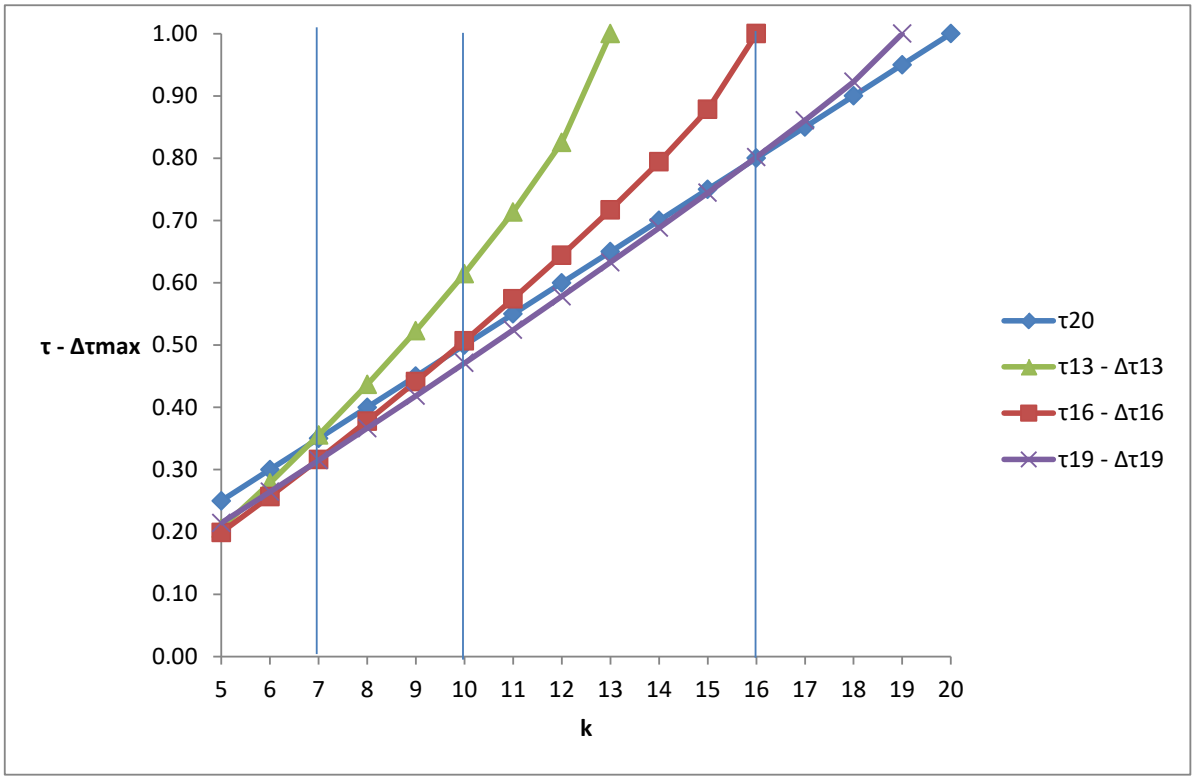

Figure 6. Dependence of three of the differences $\tau_n - \Delta\tau_{nmax}$ on the number of correct answers: $\tau_{13} - \Delta\tau_{13}$ (for a test with a sample of 13 of all 20 items), $\tau_{16} - \Delta\tau_{16}$ (test with 16 of 20 items), and $\tau_{19} - \Delta\tau_{19}$ (sample test 19 of 20 items). The same figure shows the straight line τ_{20}.

The examples analyzed above help to clarify the logical sequence underlying an adaptive algorithm that reduces the number of items assigned in the test. It should consist of the following 7 successive steps of calculations:

1. Calculate the number of correct answers for the upper limit of the score "poor" for a test with all N items in the bank. In the case of the “six-score” scale discussed above, the upper limit for the score "poor" is 2.99 (3.00 means passed exam) and after replacing with this value the grade in (19), 12.46 correct answers can be calculated, which is the maximum number of the correct answers for the score "poor". It is rounded to the nearest whole number 12.

2. Calculate the “terminal” difference between the number of items in the bank and the maximum number of correct answers for a score of "poor". In the example, this difference is 8.

3. The examinee solves a sequence of randomly drawn items with a number equal to the “terminal” difference. In the example, these are 8 items.

4. The current number of items n and the number of correct answers k are used to calculate τ_{nk}.

5. After reaching the number of items equal to the “terminal” difference, after each next answer it is checked whether the number of incorrect answers is not equal to the “terminal” difference. If the number of incorrect answers has reached the “terminal” difference, the test is terminated with a grade in the score "poor", as this is the grade that the examinee would receive, even if his test includes all items from the bank.

6. $\Delta\tau_{nkmax}$ is calculated from formula (20), and τ_{nk} - $\Delta\tau_{nkmax}$ is calculated too.

7. The fulfillment of criterion (29) is checked.

a. If the criterion is not met, the next item is randomly submitted and proceeds to step 4.

b. If the criterion is met, according to formula (18) the point grade ϑ is calculated, corresponding to $\tau_{nk} = k/n$, as well as the score in which it falls. The test is completed and is terminated after n answered items from all N items in the bank.

In the example with a test with a bank of 20 dichotomous items with 4 answers each, according to the above formulas it can be calculated that the test is terminated:

1. with a score "poor" due to the number of incorrect answers reaching the terminal difference:
1.a. after the 8th item with 0 correct answers,
1.b. after the 9th item with 1 correct answer,
1.c. after the 10th item with 2 correct answers,
1.d. after the 11th item with 3 correct answers,
1.e. after the 12th item with 4 correct answers,
1.f. after the 13th item with fewer than 7 correct answers,
1.g. after the 14th item with less than 8 correct answers,
1.h. after the 15th item with less than 9 correct answers,
1.i. after the 16th item with less than 10 correct answers,
1.j. after the 17th item with less than 11 correct answers,
1.k. after the 18th item with less than 12 correct answers,
1.l. after the 19th item with less than 12 correct answers,
1.m. after the 20th item with less than 13 correct answers,
2. with a score "poor" due to a fulfilled criterion, but for a grade below 3.00:
2.a. after the 10th item with 5 correct answers,
2.b. after the 11th item with 5 correct answers,
2.c. after the 12th item with 5 and 6 correct answers,
2.d. after the 13th item with 5 and 6 correct answers,
2.e. after the 14th item with 7 correct answers,
2.f. after the 15th item with 8 correct answers,
2.g. after the 16th item with 9 correct answers,
2.h. after the 17th item with 10 correct answers,
3. with a score of "satisfactory" as a result of a fulfilled criterion:
3.a. after the 17th item with 11 correct answers,
3.b. after the 18th item with 12 and 13 correct answers,
3.c. after the 19th item with 12 and 13 correct answers,
3.d. after the 20th item with 13 and 14 correct answers,
4. with a score of "good" as a result of a fulfilled criterion:
4.a. after the 19th item with 14, 15, and 16 correct answers,
4.b. after the 20th item with 15 and 16 correct answers,
5. with a score of "very good" as a result of a fulfilled criterion:
5.a. after the 20th item with 17, 18, and 19 correct answers,
6. with a score of "excellent":
6.a. after the 20th item with 20 correct answers.

5. Conclusions

In the present article, a new mathematical model based on the Semantic branch of Information Theory and Probability Theory is offered to the reader interested in the objective assessment of knowledge. The model defines the parameter "value (importance)" of the information signal, as a measure of the knowledge of the evaluated. The information values form the most informative type of scale of relations, allowing absolute knowledge assessment, without the need to compare this knowledge with an external standard such as the knowledge of other subjects. The model offers formulas convenient for inclusion in the test algorithm, and suitable for assessment with computer tests.

Unlike the IRT, which was created with the ambition to be applicable in all areas of life in which tests are applicable, the proposed model is suitable only for the information systems with a goal, as the systems of assessing knowledge are. The model is an alternative to IRT in several aspects: (1) different paradigm; (2) solves the problem of guessing, for which there is no convincing solution in IRT; (3) uses the most informative type of scale of relations for information value with absolute zero, while in IRT and CTT assume that grades form less informative scale—the interval scale, with no absolute zero. IRT offers several models with a different number of parameters, the values of which are calculated through an optimization procedure from the data of a group solving the test. I.e. the evaluation with the obtained IRT model is relative—it depends on the specifics of the group, while the evaluation with the proposed model is absolute, and depends only on knowledge of the evaluated.

An adaptive algorithm is proposed, adaptively reducing the number of set items in the test whenever possible, depending on the alternation of correct and incorrect answers of the examinee in the testing process. The algorithm saves time in the testing process without changing the grade obtained from the one the examinee would receive if he/she answered all the items in the test bank. The analysis shows that the adaptive algorithm saves time mainly for testing those without knowledge who receive a score of "poor". Exam practice shows that this type of examinee is the most hesitant and their exam is time-consuming. Therefore, a quick preliminary computer test with an adaptive algorithm in its software as the first part of the examination process would weed out the unprepared and would make this process shorter in time without loss of accuracy in the assessment.

References

[1] L. Croker and J. Algina. *Introduction to Classical and Modern Test Theory*, (Cengage Learning, 1st edition, 2006) pp. 1-527.

[2] F. B. Backer. *The Basics of Item Response Theory* (ERIC, Clearinghouse on Assessment and Evaluation, 2001), pp. 1-187. https://files.eric.ed.gov/fulltext/ED458219.pdf (28/05/2023).

[3] I. Partchev. *A visual guide to item response theory*, (Friedrich-Schiller-Universitat Jena, 2004), pp. 1-61, https://studylib.net/doc/18658865/a-visual-guide-to-item-response-theory---friedrich (28/06/2023).

[4] Zanon, C., Hutz, C.S., Yoo, H., et al. *An application of item response theory to psychological test development*. Psicol. Refl. Crít. 29, 18 (2016). https://doi.org/10.1186/s41155-016-0040-x2016 pp. 2-10.

[5] A. Xinming and Yiu-Fai Yung. *Item Response Theory: what it is and how you can use the IRT procedure to apply it*, SAS Institute Inc. Paper SAS364-2014, pp. 1-14. https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf.

[6] C. Ebesutani, J. Regan, A. Smith S. Reise, C. Higa-McMillan, and B. Chorpita. Application of Item Response Theory for More Efficient Assessment. *J Psychopathol Behav Assess.*, 34, 191-203 (2012). https://doi.org/10.1007/s10862-011-9273-2.

[7] C. De Mars. *Item Response Theory. Understanding Statistics Measurement*. Oxford University Press, 2010, pp. 1-138.

[8] Suvadeep Mukherjee, Björn Rohles, Verena Distler, Gabriele Lenzini, Vincent Koenig. The effects of privacy-non-invasive interventions on cheating prevention and user experience in unproctored online assessments: An empirical study. *Computers & Education*, Volume 207, December 2023, 104925. https://doi.org/10.1016/j.compedu.2023.104925.

[9] Jean-Paul Doignon, Jean-Claude Falmagne. Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, Volume 23, Issue 2, August 1985, Pages 175-196. https://doi.org/10.1016/S0020-7373(85)80031.

[10] Jean-Claude Falmagne, Eric Cosyn, Jean-Paul Doignon, Nicolas Thiéry. The Assessment of Knowledge in Theory and in Practice. Conference Paper in Lecture Notes in Computer Science January 2006, DOI: 10.1109/KIMAS.2003.1245109.

[11] Yin-Feng Zhou, Hai-Long Yang, Jin-Jin Li, Yi-Dong Lin. Automata for knowledge assessment based on the structure of observed learning outcome taxonomy. *Information Sciences*, Volume 659, February 2024, 120058. https://doi.org/10.1016/j.ins.2023.120058.

[12] Pasquale Anselmi, Egidio Robusto, Luca Stefanutti, Debora de Chiusole. An Upgrading Procedure for Adaptive Assessment of Knowledge. *Psychometrika*, 2016 Jun. 81(2):461-82. doi: 10.1007/s11336-016-9498-9.

[13] Jun-Ming Su, Su-Yi Hsu, Te-Yung Fang, Pa-Chun Wang. Developing and validating a knowledge-based AI assessment system for learning clinical core medical knowledge in otolaryngology. *Computers in Biology and Medicine*, Volume 178, August 2024, 108765. https://doi.org/10.1016/j.compbiomed.2024.108765to.

[14] Stevens, S. S. On the theory of scales of measurement. *Science*, 1946, 103, 677-680.

[15] A. I. Karasev. *Theory of probabilities and mathematical statistics* (Moscow, Statistics, 1979), pp. 62-78, (in Russian).

[16] G. F. Lakin. *Biometrics* (Moscow, Higher Education, 1990), p. 323. (In Russian).

[17] Inverse trigonometric functions.
https://en.wikipedia.org/wiki/Inverse_trigonometric_functions#References.