Організація проходження та обчислення норм тестів

Ірина Зянчурина[1],
Владислав Некрасов[2]

Кафедра Електротехніки та мехатроніки, Національний аерокосмічний університет ім. М.Є. Жуковського «ХАІ», УКРАЇНА, Харків, вул. Чкалова 17,
[1]E-mail: irinazyanchurina@mail.ru
[2]E-mail: wladislaw.nekrasow@gmail.com

Освіта є найважливішою сферою соціального життя. У забезпеченні якісної освіти зацікавлений кожен суб'єкт освітнього процесу (педагог, учні, адміністрація та ін.). Підведення підсумків результатів навчання є важливою складовою частиною навчального процесу. Досягнення високої якості освіти можливе тільки за наявності об'єктивних методів діагностики. У практиці будь-якого викладача є конфліктні випадки невдоволення учня (студента) екзаменоційною оцінкою, в той же час подібні конфлікти практично виключені при тестуванні. Крім об'єктивності, тести мають високий ступінь достовірності, також можлива перевірка тестів на надійність та валідність. У зв'язку з цим, процес тестування навчальних досягнень все ширше застосовується в освітній діяльності.

У даній роботі розглянуті питання організації проходження та обчислення норм тестів (нормування індивідуальних результатів, нормативно-орієнтований підхід, нормальний розподіл індивідуальних балів і т.і.).

Розроблена інформаційна система дозволяє розрахувати індивідуальний бал учня, провести статистичну обробку отриманих результатів і шкалювання результатів вимірювань, розрахувати характеристики завдань і показники якості тестів. Система складається з декількох програм: модуль підбору тестів (які будуть перевірятися на норми), модуль реалізації процесу тестування і модуль обробки результатів.

При використанні п'ятибальної шкали викладач виставляє оцінки з розкидом плюс, мінус 1 бал, тобто з точністю 20%. З цього випливає, що за одні знання, випробуваний може бути оцінений різними екзаменаторами по-різному. Більше того, один і той же екзаменатор в різні моменти часу, наприклад з інтервалом в один семестр, найчастіше, по-різному оцінює одну відповідь.

Розроблена програма дозволяє здійснювати розробку тестів, автоматизувати процедуру тестування, а також реалізувати обробку та інтерпретацію результатів.

The organization of passage and calculation of test norms

Irina Zyanchurina[1],
Vladyslav Nekrasov[2]

Department of Electrical Engineering and Mechatronics, National Airspace University named by M.Ye. Jukovsky "Khai", UKRAINE, Kharkiv, Chkalova street 17,
[1]E-mail: irinazyanchurina@mail.ru
[2]E-mail: wladislaw.nekrasow@gmail.com

Because to study of subject matters at the modern higher school new effective forms and work methods are widely used, to testing special significance is attached. Owing to what, there is a necessity of creation of the system, allowing to automate control of knowledge of students, including creation of a set of test tasks, carrying out of testing of students and the analysis of results.

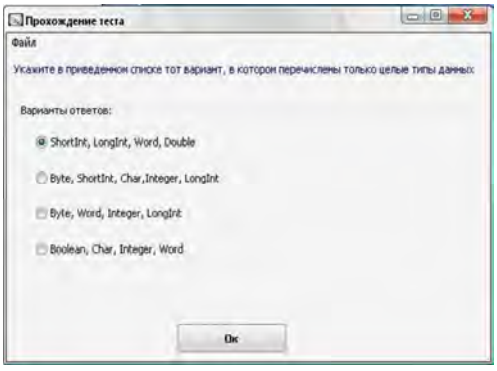
In the given project the question of the automated passage and calculation of norms of tests is considered. The major criterion which is necessary for correct interpretation of results of norms of tests, is reliability. In the given project the information system is developed, allowing to exclude defects in test tasks that considerably raises accuracy of pedagogical measurement, and also stability of results of testing to influence of extraneous factors.

For this reason, as control and measuring action we choose testing.

Keywords – quality of tests, scaling results, a measurement error, distractor, discriminative, reliability factor.

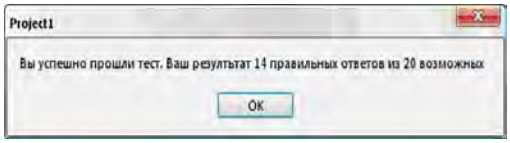
I. Introduction

Statistical processing of results of testing allows on one hand, to objectively estimate the results of trainees, with another – determine the quality of the test, test tasks and to estimate its reliability. A lot of attention in the classical theory of tests is given to reliability problem. Despite occurrence, more modern theories, the classical theory continues to keep the positions

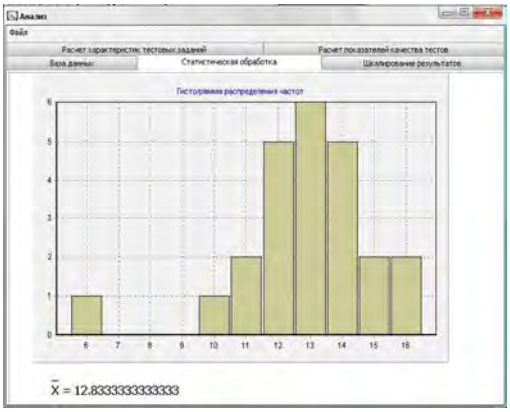


Pic. 1 – The screen form of passage of testing

Difficulty of the task in the classical theory of tests is defined through a parity of quantity of trained, successfully consulted with the given task, to total of trainees.



Pic. 2 – Demonstration of results of testing



Pic. 3 –Carrying out of the analysis and conclusion of total results of testing

The difficulty indicator is very important for definition of the characteristic of the test task and helps to range the tasks entering into the test on degree of complexity.

II. Calculation of an individual point of the pupil

For calculation of an individual point of each trainee the answer to each question is compared to the right answer which has been written down in a database, and the sum of the right answers is displayed.

Scaling by means of Z-Scale (scales of deviations).

This method is based on calculation of a deviation of a "crude" point (X_i) from average value of individual points ($\bar{X}$)on group of the tested. Value Z_i — scaling result of each examinee find under the formula:

$$Z_i = \frac{X_i - \bar{X}}{S_x}, \qquad (1)$$

where X_i — a crude point of i-th examinee;

$\bar{X}$ — average value of individual points N of examinees of group (/=1,2,..., N)

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (2)$$

S_x — A standard deviation on set of the crude points, counted up under the formula:

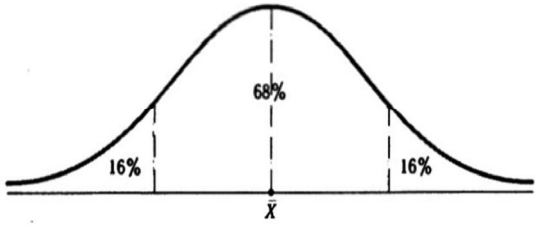
$$S_x = \sqrt{S_x^2}, S_x^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1}. \qquad (3)$$

By means of this formula we calculate values Z_i, we make the table of conformity of values of a "crude" point X_i differences $X_i - \bar{X}$ and values Z_i. Positive values Z_i speak about good results, negative — about bad. Standard deviation on set of values equally 1. With its help it is possible to result the points of pupils received under various tests, in one kind convenient for comparison by rationing of individual results.

Statistical processing of the received results of testing. Construction of the histogram of distribution of frequencies of the right answers. Range of frequencies.

On the basis of the table of frequencies it is possible to construct range of frequencies. On an axis of abscisses test points, and on an axis of ordinates - corresponding frequencies are postponed.

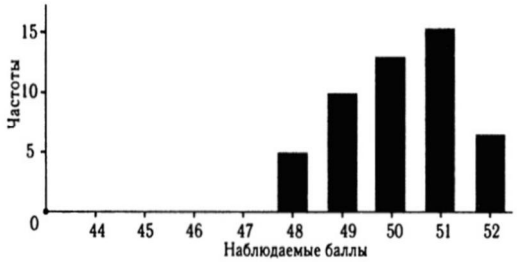
Usually consider that the good is standard-focused test provides normal distribution of individual points of representative sample of pupils when average value of points is in the center, and other values concentrate round an average on normal to the law, i.e. approximately 70 % of values in the center, and the others come to naught to distribution edges, as in drawing.



Pic. 4 – Normal curve of distribution of individual points

Histograms for distribution of individual points on too easy or too difficult selection of tests can look as follows:

Pic. 5 – The histogram of distribution of points under the difficult test



Pic 6. – The histogram of distribution of points under the easy test

Scaling results of test measurements

For an objective estimation of level of readiness interrogated in comparison with other participants who pass test, the technology scaling results is often applied. In process scaling primary results are translated in the numerical system generated by certain rules in which the relation between properties of object is expressed in the corresponding numbers, to each primary point the certain test point is put in conformity. Following methods scaling are usually applied.

Scaling with the percentile ranks. The percentile - is the derivative indicator specifying in a share, those who has correctly carried out the test task, from total tested in group.

Calculation of characteristics of test tasks.

By results of trial testings are defined characteristics of test tasks - difficulty and discriminative.

Difficulty of test tasks.

Difficulty of the task is calculated under the formula (a measure of ease of the test task):

$$p_j = \frac{R_j}{N} \quad (4)$$

where Pj - share of the right answers on j-th task;

Rj - The quantity of the examinees who have executed j-th the task is true;

N - Number of examinees in group, j - task number.

Also difficulty of the test task can be calculated in percentage Pj:

$$P_j = \frac{R_j}{N} \cdot 100\% \quad (5)$$

From the formula it is visible that the above the difficulty indicator, the task is easier, and accordingly, it is less indicator of difficulty of the task, the task is more difficult. For example, if P = 30 % it means that only 30 % of examinees have coped with this task and if P = 70 %, then 70 % have coped with the task, and turn out that the first task is more difficult, than the second.

Within the limits of the is standard-focused approach tasks of average difficulty p=q=0,5 which provide the maximum dispersion of the test are considered as the most successful ($s = p \cdot q$).This product reaches the maximum value (0,5 x 0,5 = 0,25) at p = 0,5.

The quality analysis distractors in tasks of the closed form.

One of the major requirements who is shown to tasks of the closed form, is a plausibility distractors (equivalent probability of a choice distractor at the wrong answer). The distractor analysis assumes calculation of shares of the examinees who have chosen each distractor. In an ideal variant everyone distractor should get out in an equal share from all wrong answers. Table 1 shows the ideal distribution of shares:

TABLE 1

DISTRIBUTION OF DISTRACTERS

№ tasks	1 answer	2 answer *	3 answer	4 answer
J	0,1	0,7	0,1	0,1

In table 1 it is shown that 70 % of trainees have correctly carried out the task (have chosen 2nd answer). The others of 30 %, which distances wrong answers, have in regular intervals chosen 1, 3, 4 answers, i.e. in the task have been given equiprobable distractors. But such ideal picture of distribution of a choice of wrong answers in real practice meets seldom.

Discriminative the test task.

Discriminative (the differentiating ability distinguishing ability) of the task is an ability of the task to differentiate examinees on level of achievements, on strong and weak.

One of ways of calculation the discriminative - calculation with application of a method of extreme groups where for calculation indicators of the weakest and strongest examinees undertake. It is 27 (30) % of the worst and 27 (30) % of the best by results of performance of the test task more often.

In table 2 results of calculation of the discriminative index are presented.

TABLE 2

RESULTS OF CALCULATION OF THE DISCRIMINATIVE INDEX

№ tasks	Pj for the weak	Pj for the strong	Index r_{dis}
Question1	100	100	0
Question2	100	100	0
Question3	81,8	100	0,182
Question4	72,7	100	0,273
Question5	100	100	0
Question6	100	100	0
Question7	100	92,9	-0,071
Question8	90,9	100	0,091
Question9	90,9	100	0,091
Question10	9,1	57,1	0,48

The discriminative index is defined as a difference of shares of the right answers of strong and weak groups:

$$(r_{dis})_i = p_i^1 - p_i^0, \quad (6)$$

where r_{dis} - the discriminative index, p^1 - a share of the right answers in a strong subgroup (27 % from all quantity), p^0 - a share of the right answers in weak group (27 %).

If difficulty is set in percentage

$$(r_{dis})_i = \frac{P_i^1 - P_i^0}{100\%}. \quad (7)$$

Value of the discriminative index settles down in an interval [-1; 1].

III. Calculation of indicators of quality of tests

Statistical processing of results of testing allows on the one hand, objectively to define results of examinees, with another – to estimate quality of the test, test tasks, in particular to estimate its reliability.

A lot of attention in the classical theory of tests is given reliability problem. This theory hasn't lost the urgency and now. Despite occurrence, more modern theories, the classical theory continues to keep the positions.

The classical theory of tests is based on following five substantive provisions:

1. Empirically received result of measurement (X) represents the sum of true result of measurement (T) and measurement errors (E):

$$X = T + E . \quad (8)$$

Sizes T and E are usually unknown.

2. If we consider an observable test estimation as a casual variable X it is possible to express true result of measurement as a population mean

$$T = M(X) . \quad (9)$$

3. Correlation between a true estimation and its erroneous component for general totality of examinees is equal to zero

$$r(T, E) = 0 . \quad (10)$$

4. When examinees carry out two separate tests and estimations of each examinee under two tests (or on two testings by means of the same form of the test) are assumed casually chosen from two independent distributions of possible observable estimations, correlation between erroneous components of estimations on these two testings is equal to zero:

$$r(E1, E2) = 0 . \quad (11)$$

5. Erroneous components of one test don't correlate with true components of any other test:

$$r(E1, T2) = 0 . \quad (12)$$

Besides, the basis of the classical theory of tests is made by two definitions – parallel and equivalent tests.

Measurement error - the statistical size reflecting degree of a deviation of an observable point from a true point of the examinee. The dispersion of observable test points will be equal to the sum of dispersions of true and erroneous components:

$$S_x^2 = S_T^2 + S_E^2 . \quad (13)$$

Accordingly, the more close an indicator of a dispersion of observable points to a dispersion of points true, the above correlation between set of observable points (X)and set of true points (T),i.e. the test is more reliable. Therefore reliability of the test (factor of reliability of the test - r_r) It is defined through the relation of a dispersion of a true point to a dispersion of an observable test point:

$$r_r = \frac{S_t^2}{S_x^2} = 1 - \frac{S_e^2}{S_x^2} . \quad (14)$$

The standard error of measurement is as a root square of a dispersion erroneous components:

$$S_e = \sqrt{S_e^2} . \quad (15)$$

Conclusion

Quality achievement of training is possible only in the presence of objective methods of diagnostics. Unfortunately, the traditional form of estimation of level of knowledge in the form of poll, the examination spent by the person, is rather subjective.

Clearly that so inexact «the measuring device» what the person is, essentially reduces efficiency of diagnostics of educational process. For this reason, as control and measuring action testing gets out.

The developed program allows to carry out working out of tests, to automate testing procedure, and also to realize processing and interpretation of results.

Literature

[1] Майоров А.Н. Теория и практика создания тестов для системы образования. – М.: «Интеллект-центр», 2001. -296 с.

[2] Ким В.С. Коррекция тестовых баллов на угадывание // Педагогические измерения, 2006, №4. – С.47-55.

[3] Орлов А.И. Теория измерений и педагогическая диагностика // Педагогическая информатика, 2004, №1. – С.22-31.

[4] Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие. – М.: Логос, 2002. -432с.

[5] Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических 2. тестов. -М.: Прометей, -169 с.

[6] Ингенкамп К. Педагогическая диагностика. -М.: Педагогика, 1991. -240 с.

[7] Аванесов В.С. Основы научной организации педагогического контроля в высшей школе. -М., 1989. -167 с.