Performance criteria of learning math tests

Abstract—**In present days modern computer tools and information technologies implementation in educational process are widespread. The main purpose of computer-based testing shifted from the assessment instrument of learning outcomes to the purpose of teaching. Performance criteria of learning mathematical tests are not yet studied properly. The goal of our work is to find and to study the performance criteria of learning math tests. First, we attempted to formulate the performance criteria of learning math tests. We assume these criteria to be informativity, discriminative criterion, validity, and reliability. Thus, we created 21 learning tests based on the course of mathematical analysis. We tested approximately 48 first year students of the physical department of the National Research Tomsk State University (Russia) to check how fully our tests correspond to the purpose of teaching. Then we performed the study of the statistical data, based on the criteria test results. Thus, we found out our tests informative since we used the scale intervals and since the results of testing were easily processed statistically. Due to the fact that the data obtained are distributed uniformly or normally, we concluded that our tests satisfy the discriminative criterion. Our tests are also valid, because they measure exactly what they are intended to measure. Nevertheless, we describe the tests at hand are not always have a high reliability. It can be improved by increasing the test value, but it leads to the loose of student's interest and attention to the task. In the following we are planning to do the reliability improvement without increase in the test tasks value, and to perform the study of normative data availability and criteria established by the experts.**

Keywords— performance criteria; learning math tests

I. INTRODUCTION

The development of modern teaching techniques caused the widespread introduction of tests in the learning process. As a rule, they are regarded as an assessment instrument of learning outcomes. In our work testing techniques are used for teaching rather than assessment, which helps to introduce students to the new methods, knowledge, ideas. They also encourage them and intensify the process of cognition. In our tests, there are different types of test questions: multiple choice, matching, classification, ordering, and short-answer questions. The basic criteria for the selection of the item for the learning test are the following: novelty, originality, and feasible difficulty at the same time. Therefore, the designed tests are not quite homogeneous, do not cover all the studied material, and cannot be used for the assessment purposes.

A. Study Background

For many decades, the tests are used extensively in the world's the pedagogical practice [1-4]. At present in addition to controlling function of the tests, the field of learning tests is developing rapidly [5]. A learning test with respect to its mathematics applications is known to be a pedagogical tool, which allows diagnosing the correct student's evaluation of the way to solve the educational problem until the right solution is found. However, despite these positive aspects, the use of learning tests in teaching practice is hampered by the lack of elaboration of a number of fundamental issues. One of these issues is the quality estimation of learning tests used in the educational process. An insufficient development of scientific and methodological approaches to the learning tests quality parameters selection issue appears due to varying interpretations of the "quality of a learning test" concept.

Before make the selection of the performance criteria for the learning math tests, we perform an analysis of the selection process of the corresponding criteria for psychological tests, as this area of knowledge is well developed. For example in [4] as the performance criteria of psychological tests the use of the scale intervals, reliability, validity, discriminativity, the availability of normative data and the criteria established by the experts are considered. In the pedagogical diagnostics [6] the quality of the measurement is of high importance. Therefore, objectivity, reliability, and validity criteria help to make an assessment of the test performance. Quality assessment in the field of controlling test is given in [7].

We suggest the performance criteria of the learning math test to be informativity, validity, reliability, and discriminativity. Learning math tests should be informative, i.e. test scores can be related to a measurement scale, and therefore, statistical analysis can be easily carried out. Our Learning tests should be also valid, because they should measure exactly what they are intended to measure. Reliable test yields predictable results. On the basis of discriminative criterion one can form a student's groups in accordance with the level of their knowledge. We do not consider the objectivity, because learning math tests should not depend on the mood of the teacher, nor the methods and means of control. The availability of normative data and criteria established by the experts will be examined in the following.

A significant number of test reliability criteria are well-known [8 - 10]. For example, Pearson correlation coefficient between the two parallel tests on the same set of students can be used as the reliability criterion. The coefficient of correlation between the test results and the results of expert's assessments can be used as well. Reliability coefficient of Guttman [11], correlation coefficient of Spearman-Brown [4, 10, 11], and its modification are well used as the reliability criteria in the most cases in practice. Thus, so-called formula KR-20 [10] for calculating the reliability of the test has a widespread application. The formula takes its name from names of its founders F. Kuder and M. Richardson (number 20 is the number of formula in the publication).

B. Study objectives

The objectives of the present work are: to formulate the performance criteria of learning math tests; to evaluate the relevance of achievement test efficiency criteria for the evaluation of learning test results. The research subjects are

978-1-4799-8706-1/15/$31.00 ©2015 IEEE

test results in mathematical analysis (21 tests, designed on the "Airen" platform [12]). Each of selected criteria:

- Informativity;
- Reliability;
- Validity;
- Discriminative criterion.

will be studied separately.

II. Performance Criteria Study

A. Informativity

Any research starts with the fact that the scientist records the objects' property (or properties) with the help of figures (quantitative characteristics). Thus, one should distinguish between the objects of research (in this case, tests consisting of test items; and test scores), their properties (test efficiency, which is the subject of the study) and characteristics presented in a numeric scale. Therefore, the evaluation of test efficiency should be started with defining the scale of measurement as a tool of further statistical analysis of test results. Let A be a set of objects, and $\{P_i, i = 1,\dots, m\}$ – relations on this set.

Set A together with the system of relations given on it, is called a system with the relations and is expressed in

$$U = <A,\ \{P_i, I = 1\dots m\}> \tag{1}$$

Let k is a measurement scale. It is homomorphism f from the empirical system with the relation $U = < A,\ \{P_i\}>$ to the numerical system k with the relation $V = < R^k,\ \{S_i\}>$ [13]. Thus, the scale is a triple (U, V, f), where f is homomorphism from U to V.

There are many types of measurement scales. For example, nominal scales, in which numbers are used as the names of the objects of study [13]. The nominal scale shows whether the two objects are equivalent or not. For example, test results of the students having scored the same number of points are equivalent, while the examinees themselves are different. The nominal scales is merely used to classify the examinees, for example, they may be allocated to the «being tested» category. The ordinal scale presents data in order (objects are ranked). For example, the examinees are classified according to the average score. The disadvantage of such a scale is that it does not take into account value differences between the ranks. There is the interval scale in which there are exact differences between the values at all the points of the scale, a zero value is arbitrary, and the unit of measure is given. The values obtained at an interval scale are invariant under the group of affine transformations. The ratio scale has a fixed zero value and the unit of measure are chosen by the researcher.

In present study the interval scale is used ($k = 1$), since a large number of statistical methods can be applied to the experimental data treated along this scale, which is crucially important.

B. Reliability

A reliable test is the test that yields similar test results for the same examinee in a hypothetical retesting that is test results do not depend on any random factors. It is called test-retest reliability [3]. The reliability index was calculated by the split-half reliability method using the Spearman-Brown formula

$$r_t' = 2r_t / (1+r_t), \tag{2}$$

where r_t' is a corrected reliability coefficient, and r_t – reliability coefficient (Pearson's correlation coefficient), found by the split half method. It is considered that the lowest satisfactory value for the test-retest reliability coefficient is 0.7 [4, 18]. Having analyzed 21 tests on the course of mathematical analysis, it was found out that most of them have satisfactory reliability (Table 1). We consider two main reasons for unsatisfactory reliability coefficient: a small number of questions in the test (7 – 8 items) and unclear test instructions (item originality). The number of questions in the test cannot be changed, since an increase in the number of test items leads to the increased duration of the test, so students lose concentration, necessary for the successful task performance. Unclear instructions, which sometimes students complained of, are an integral part of learning tests in our understanding. To comprehend the task containing a new mathematical terminology means to learn something. Therefore, a significant improvement of reliability for learning tests is hardly possible.

C. Validity

The test is called valid if it measures what it is supposed to measure [14]. There are several types of validity: obvious (external), i.e. when the examinees get the impression that the test measures exactly what it was designed for. Criterion validity is assessed by comparing students' test scores to the results of similar tests. Predictive validity is correlation between the test scores and the results obtained on another criterion at some point in the future [15, 16]. For example, it is correlation between student's test performance in the first and in the second terms. In the achievement tests content validation is carried out.

The validity of our learning tests as a whole is not so evident, since they do not cover all the studied material, and often focus on details. The performance on the designed tests requires knowledge of the basic concepts and theoretical facts, application of the techniques used at the lessons. However, they can't cover the content fully. In addition, many math tasks require performing a great sequence of operations and, therefore, can't be selected for a learning test. So, validation of our tests can be carried out only in qualitative terms, by professional expertise. Moreover need to evaluate the individual test items instead of whole set of tests. We think our tests are valid, as they are recognized by our colleagues and become part of the educational process [17].

D. Discriminative criterion

Discriminative criterion means discriminative ability of the test (ability to separate examinees with high scores on the test from those with the low scores) [9]. Achieving a satisfactory item discrimination index is one of the objectives of the test designer. Item discrimination index can be introduced by Guildford's correlation coefficient and by Ferguson's coefficient delta δ [4]. The latter one is used in our study. Discrimination, measured by Ferguson's delta reaches the maximal value $\delta = 1$ at uniform distribution. Ferguson's delta is calculated by the formula

$$\delta = \frac{(n+1)\left(N^2 - \sum_{i=1}^{N} w_i^2\right)}{nN^2}, \quad (3)$$

where N – number of examinees, n – the number of test items, w_i, $i = 1\ldots N$ – number of results from the i-th interval. If $\delta = 0$ all the examinees got the same number of points, which means the test doesn't have discriminative power. In learning tests, $\delta = 0$ shows that, in fact, the test doesn't train the students at all, as all of them did the questions of the test. It demonstrates that the content of the test presents the material which has already been learned. Table 1 shows the statistical characteristics of our tests. To study the distribution of test scores Pearson correlation coefficient was used [18]. This table shows that our tests are discriminative.

TABLE I. STATISTICAL CHARACTERISTICS OF THE LEARNING TESTS

Test	**Number of test items**	**Number of examinees**	**Sample mean**	**Sample standard deviation**	**Distribution of test scores**	**Reliability r_t'**	**Ferguson's delta δ**
1. «Sets»	7	67	46.19	22.71	normal	0.690	0.987
2. «Numerical functions»	7	56	46.25	26.8	random	0.727	1
3. «Real numbers»	7	51	52.94	28.63	random	0.717	1
4. «Limit of a sequence – 1»	7	52	41.92	29.06	random	0.752	1
5. «Limit of a sequence – 2»	8	43	58.37	24.72	normal	0.581	0.926
6. «Limit of a sequence – 3»	7	48	45	25.2	normal	0.619	0.909
7. «Limit of function – 1»	7	49	51.84	29.45	random	0.734	1
8. «Limit of function – 2»	8	39	48.59	24.18	normal	0.678	0.926
9. «Continuous function»	8	39	54.19	27.62	random	0.814	1
10. «Derivative – 1»	8	31	72.22	23	Not determined	0.837	0.812
11. «Derivative – 2»	8	31	56.76	25.86	random	0.588	1
12. «Derivative –3»	7	36	61.53	22.11	normal	0.576	0.884
13. «Complex numbers»	8	45	62.11	25.92	random	0.793	1
14. «Indefinite integral – 1»	4	52	54.81	26.46	random	0.720	1
15. «Indefinite integral – 2»	7	43	52.09	27.83	random	0.745	1
16. «Indefinite integral – 3»	4	41	63.17	26.93	random	0.691	1
17. «Indefinite integral – 4»	7	44	53.75	27.24	random	0.752	1
18. «Definite integral – 1»	7	38	54.08	25.41	normal	0.397	0.844
19. «Definite integral – 2»	7	33	49.09	27.95	random	0.610	1
20. «Definite integral – 3»	7	29	59.83	26.57	random	0.696	1
21. «Definite integral – 4»	7	30	50.33	26.17	normal	0.759	0.907

III. DISCUSSIONS

Statistical processing data of the learning math tests results are summarized in Table 1. Statistics-based analysis was allowed by the interval-scale method. The first column of the table above contains the name of the test, and the second column contains the number of test tasks. It should be noted that the number of examinees (the third column) is changing from test to test. First, this value decreases is due to the decrease in the student's interest to the learning test educational form. Then the number of examinees increasing due to some motivation provided by the teacher.

The sample mean (test scores values divided by the number of examinees) is quite simple characteristic of the test. Thus, if the average score is close to 100, the test is worthless. It teaches nothing. And if the average score is significantly less than 50, the test is difficult for this group of students. The following steps should be taken for improvement: either to adapt the test for these students, or to discuss all the unclear issues and retest. The fifth column ("Sample standard deviation") demarks the deviation of the score from the sample mean.

The sixth column contains the type of the distribution of test scores. It is uniform (62% of test scores) and normal (38% of test scores). The fact that the distribution of test scores is normal indicates that the test is discriminative. Indeed, as can be seen from table, the corresponding Ferguson delta (the last column) is equal to one. That is one more prove of the test's discriminativity.

The reliability of the tests at hand (the seventh column) is calculated using Spearman-Brown formula. To do this, first we found the correlation coefficient between the two parts of the test, and then we calculated the corrected coefficient of reliability. Most of the tests (about 52%) are reliable (value of the reliability coefficient is more then 0.7). About 43% of tests have acceptable reliability (value of the reliability coefficient is between 0.5 and 0.7). And about 5% of tests are unreliable (value of the reliability coefficient is less than 0.5). Reliability can be improved by increasing the test value, but it leads to the loose of student's interest and attention to the task. In the following we are planning to do the reliability improvement without increase in the test tasks value, and to perform the study of normative data availability and criteria established by the experts.

IV. Conclusion

Having studied test results, we found out that most of the tests are reliable, their validation should be carried out qualitatively and they prove to obtain satisfactory discriminative power. Learning through the test performing, according to [17] is the process where at the beginning a student knows and is able to do less than at the end. Therefore, the improvement of test reliability should not be considered a key factor in evaluating the efficiency of a learning test.

References

[1] A. Anastasi, S. Urbina, Psychological testing, (7th ed.), Upper Saddle River, NJ: Prentice Hall, 1997.

[2] R.M. Kaplan, D.P. Saccuzzo, Psychological Testing: Principles, Applications, and Issues, (8th ed.), Belmont, CA: Wadsworth, Cengage Learning, 2010.

[3] K.R. Murphy, C.O. Davidshofer, Psychological testing Principles and Applications, Upper Saddle River, N.J.: Pearson/Prentice Hall, 2005.

[4] P. Kline, Handbook of Test Construction, London: Methuen, 1986.

[5] D. Upton, P. Upton, Cognitive Psychology, London: Learning Matters., 2011.

[6] M.J. Kolen, R.L. Brenan, Test equating methods and practices, NY: Springer-Verlag, 2004.

[7] G.V Glass, “Standards and criteria”, Journal of Educational Measurement, vol. 15, pp. 237-261, December 1978.

[8] V.G. Glass, J.C. Stanley, Statistical Methods in Education and Psychology, N.J.: Prentice-Hall, 1970.

[9] C. Levis, “Classical test theory”, in Handbook of Statistics, vol. 26, C. R. Rao and S. Sinharay, Eds. Amsterdam: Psychometrics, 2007, pp. 29-43.

[10] G.F. Kuder, M.W. Richardson, “The theory of the estimation of test reliability”, Psychometrika, vol. 2, pp. 151-160, September 1937.

[11] L. Guttman, “A basis for analyzing test-retest reliability”, Psychometrika, vol. 10, pp. 255-282, April 1945.

[12] http://irenproject.ru/

[13] M.J. Allen, W.M. Yen, Introduction to Measurement Theory, Long Grove, IL: Waveland Press, 2002.

[14] www.socialresearchmethods.net/kb/measval.php

[15] R. Clark, “The parent-child early relational assessment: a factorial validity study”, Educational and Psychological Measurement, vol. 59, pp. 821-846, October 1999.

[16] C.H. Yu, “Test–retest reliability”, in Encyclopedia of Social Measurement, vol III, K. Kempf-Leonard, Eds. Amsterdam: P-A. Elsevier, 2005, pp. 777–784.

[17] E.G. Lazareva, I.G. Ustinova, A.G. Podstrigich, “The use of test programs in learning higher mathematics”, Tomsk State Pedagogical University Bulletin, vol. 7, pp. 217-222, July 2012.

[18] R.J. Larsen, M.L. Marx, An Introduction to Mathematical Statistics and its Applications, Boston: Prentice Hall, 2012.