

## ANALYSIS OF TASKS AND TESTS IN MATHEMATICS FOR EIGHTH GRADE IN NINE-YEARS PRIMARY SCHOOL

Mejdin Saliji, Aneta Velkoska, Doncho Dimovski

**Abstract.** The objective of teaching mathematics is active and creative proficiency of knowledge that will be durable and applicable in practice. Therefore, it is necessary to have competent teachers. A very important segment of the teachers' education is their ability for preparing and evaluating the tests of knowledge which satisfy all measuring characteristics and make the test a measuring instrument of the students' knowledge. In this paper we examine to what extent the teachers at mathematics in primary schools could prepare achievement tests with all measuring characteristics. Moreover, in this paper by performing thorough test and items analysis we prepare a final achievement test in mathematics in the eighth grade from nine-year primary schools.

### 1. INTRODUCTION

The educational goal is not only acquiring facts, but understanding, applying and analysis of the factual knowledge in order to synthesize a new patterns or structures and to evaluate the value of a certain material (poem, novel, speech, research report or a project) as the highest educational goals on cognitive hierarchy scale [1]. So, in addition to the appropriate educational contents, the relevant modern teaching tools, the adequate school supplies and the motivation among the parents and the students, it is necessary to have competent teachers, as the principal means of any national school system.

During the training of the teachers and educators it is very important they to understand all measuring characteristics of the achievement tests and test analysis in order to be able by themselves to construct and evaluate achievement test.

---

2010 *Mathematics Subject Classification.* 97C70

*Key words and phrases.* Achievement test, educational measurements, evaluation, reliability coefficient, validity coefficient, discrimination index

The subject of this paper is to determine to what extent it is possible to prepare achievement tests in mathematics in the eighth grade from nine-year primary schools and to determine their measuring characteristics.

The goal of our research is by comparative analysis of the results from the testing the students with the prepared achievement test and their knowledge achievement at Mathematics obtained by traditional methods of evaluation to show that it is very important math teachers to practice achievement tests as measuring instrument if they want to achieve more objective information about the students' performance.

During this research a preliminary annual achievement math test was carried out in eighth graders in primary school "Planjane" and one class of eighth graders in primary school "Meto Bjaraktari" in Prizren. After test analysis of this trial achievement test we constructed main annual achievement math test which we used to evaluate 102 eighth graders in primary schools "Planjane", "Meto Bajraktari", "Brightness" and "25 May" in Prizren.

The rest of the paper is organized as follows: Section 2 describes the achievement tests as standardized educational gauge. In Section 3 we discuss the measurement characteristics of the achievement test. The preliminary and central experiments from our research are presented in Section 4 and 5 respectively. In Section 6 by inferential statistics we analyse the relationship between the grades assigned by the main annual achievement test scores and their knowledge achievement at Mathematics obtained by traditional methods of evaluation. Section 7 concludes the paper.

## **2. ACHIEVEMENT TESTS**

Assessment, measurement and evaluation are part of the processes of science and issues related to each topic often overlap. Assessment refers to the collection of data to describe or better understand an issue, measurement is the process of quantifying assessment data, and evaluation refers to the comparison of data to a standard for the purpose of judging worth or quality [2].

Educational measurement and evaluation refer to the use of educational assessments and the analysis of data such as scores obtained from educational assessments to infer the abilities and proficiency of students. Assessment is a set of all activities that teachers use to help students learn and to gauge student progress. But assessment of understanding is very difficult and complex. Skills can be practiced; understandings cannot. We can assess a person's knowledge in a variety of ways, but there is always a leap, an inference that we make about what a person does in relation to what it signifies about what he knows.

In other words, educational assessment is a process in which the teachers not only measure the students' knowledge and proficiency, but also they are gauges. So it is not unusual that each teacher has his own standards and grading scale. In order to standardize the assessment each teacher should be able to design achievement tests.

Achievement test is an important tool in school evaluation and has great significance in measuring instructional progress and progress of the students in the subject area.

Achievement means one's learning attainments, accomplishments, proficiencies, etc. It is directly related to the student's growth and development in educational situations. Tests should give an accurate picture of students' knowledge and skills in the subject area or domain being tested. Accurate achievement data are very important for planning curriculum and instruction and for program evaluation. Test scores that overestimate or underestimate students' actual knowledge and skills cannot serve these important purposes [3].

Test preparation activities which promote quality, long-term learning are appropriate, even essential. Good test - taking skills and appropriate content learning can reduce the likelihood that extraneous factors will influence students' test scores.

### **3. MEASUREMENT CHARACTERISTICS OF ACHIEVEMENT TESTS**

This Section throws light upon the four important measurement characteristics of a good test. The four characteristics are: 1. Reliability 2. Validity 3. Objectivity and 4. Sensitivity.

#### **First characteristic – Reliability**

In measurement reliability is the consistency with which a test yields the same result in measuring whatever it does measure. A test score is called reliable when we have reason for believing the score to be stable and trustworthy. Estimates of the reliability of tests provide essential information for judging their technical quality and motivating efforts to improve them.

Theoretically, reliability is defined as the ratio of the true score and observed score variance. But reliability is a statistical concept. The most revealing statistical index of educational tests' quality is the *Reliability Coefficient*.

In order to determine reliability coefficient we have to obtain two sets of measurements in identical condition and then compare the two sets. But it is only a theoretical condition, because it is impossible on our part to get two measurements on exactly two identical conditions. So that several methods have been developed to determine the relative reliability. The methods are similar in

that all of them involve correlating two sets of data, obtained either from the same evaluation instrument or from equivalent forms of the same procedure.

In our research we used *Split-half method* of determining reliability coefficient. In this method a test is administered to a group of students in usual manner. Then the test is divided into two equivalent values and correlation for these half-tests is found.

The common procedure of splitting the test is to take all odd numbered items i.e. 1, 3, 5, etc. in one half and all even-numbered items i.e. 2, 4, 6, 8 etc. in the other half. Then scores of both the halves are correlated by using the Spearman - Brown formula:

$$r_2 = \frac{2 \cdot r_1}{1 + r_1}, \quad (1)$$

where  $r_1$  is the Pearson correlation coefficient for the half-test.

The reliability coefficient  $r_2$  value equals to 0,90 and up indicates excellent reliability, from 0,80 to 0,89 indicates good reliability, from 0,70 to 0,79 indicates adequate reliability, below 0,70 test may have limited applicability.

### **Second characteristic – Validity**

Validity is the most important characteristic of an evaluation programme, for unless a test is valid it serves no useful function. Psychologists, educators, guidance, counselors use test results for a variety of purposes. Obviously, no purpose can be fulfilled, even partially, if the tests do not have a sufficiently high degree of validity. Validity means truth-fullness of a test. It means to what extent the test measures that, what the test maker intends to measure.

It is not also necessary that a test which is reliable may also be valid. For example suppose a clock is set forward ten minutes. If the clock is a good time piece, the time it tells us will be reliable, because it gives a constant result. But it will not be valid as judged by 'Standard time'. This indicates "the concept that reliability is a necessary but not a sufficient condition for validity."

There are three methods of conducting validity: Criterion-Related Validity, Content Validity and Construct Validity.

The Criterion-related Validity investigates the correspondence between the scores obtained from the newly-developed test and the scores obtained from some independent outside criteria. The criterion-related validity of a test is measured by the *validity coefficient*. In order to determine validity coefficient the newly-developed test has to be administered along with the criterion measure to the same group. The validity coefficient is a correlation coefficient for the newly-developed test scores and the scores from another criterion measure and indicates the extent to which the test scores correlate with a

relevant outside criterion. The larger the validity coefficient, the more confidence you can have in predictions made from the test scores.

In this paper we use Criterion-related Validity to find the validity coefficient of the preliminary achievement test in mathematics and mid-term achievements of a sample of eighth grade students.

**Third characteristic – Objectivity:**

Objectivity is an important characteristic of a good test. It affects both validity and reliability of test scores. In [5] Good defines objectivity in testing as “the extent to which the instrument is free from personal error (personal bias), that is subjectivity on the part of the scorer”. Objectivity of scoring means same person or different persons scoring the test at any time arrives at the same result without chance error.

Statistically the objectivity of the test is measured by calculating the Pearson correlation coefficient of the students’ test scores of two different assessments administrated by the same test. If the Pearson correlation coefficient of these two assessments is greater than the critical value of 0,80 we conclude that the test is objective.

**Fourth characteristic – Sensitivity:**

According to Popham in [6], instructional sensitivity is “the degree to which students’ performances on a test accurately reflects the quality of instruction specifically provided to promote students’ mastery of what is being assessed”. In Pham’s dissertation [7], instructional sensitivity is defined as “responsiveness to the varying pedagogical practices of teachers and it allows for standardized testing to be used as an accountability tool”. In [8] Haladyna and Roid defined instructional sensitivity as “the tendency for an item to vary in difficulty as a function of instruction”. Niemi et al. [9], instructionally sensitive assessments consider as “assessments that can measure the effects of previous teaching, and they can also be used as outcome measures to evaluate instruction, as well as to identify students who need additional instruction”.

All the above definitions emphasize a fact that instructional quality is an important part of the school environment and that instructional sensitivity is an important index of an effective or well designed achievement test, which serves as a tool of accountability.

One of the important factors of the instructional sensitivity is an index of item discrimination, i.e., how well the item serves to discriminate between students with higher and lower levels of knowledge. The discrimination index can be calculated with the point-biserial correlation between the score on the item and total test scores by the formula:

$$r_{pbi} = \frac{M_p - M_q}{SD} \sqrt{pq}, \quad (2)$$

- $M_p$  - the arithmetic average of the test scores obtained by the students who solved the task correctly
- $M_q$  - the arithmetic average of the test scores obtained by the students who solved the task incorrectly
- $SD$  - the standard deviation of the total test scores
- $p = \frac{\text{correctly solved tasks' frequencies}}{n}$  - the item easiness index
- $q = 1 - p$  - the item difficulty index.

The point-biserial correlation reflects the degree of relationship between scores on the item 0=incorrect, 1=correct and total test scores. Thus the point-biserial will be positive if better students answered the item correctly more frequently than poorer students did, and negative if the opposite occurred. A negative point-biserial is denoted by a minus sign in front of the value.

The value of a positive point-biserial discrimination index can range between 0 and 1; the closer the value is to 1, the better the discrimination. (The value of a negative point-biserial discrimination index can range between -1 and 0, but positive values are desirable). As a general rule, point-biserial values of 0,20 and above are considered to be desirable.

Items with negative discrimination values should be reviewed. A negative discrimination value, like a low difficulty value, may occur as a result of several possible causes: a miskeyed item, an item that is ambiguous, or an item that is misleading. Further insight into the cause of negative discrimination can often be gained by examining the distribution of student responses. In small classes negative values close to zero are not necessarily reason for concern; they may be caused by one good student answering the item incorrectly or one poor student answering the item correctly.

In our research we calculate the point-biserial correlation coefficients of all 40 questions from the preliminary achievement test and total test scores. Using item factor analysis we revise the tasks with the low value discrimination index to improve the preliminary achievement test into main achievement test which will have all measurement characteristics of a good educational gauge.

#### 4. PRELIMINARY TESTING

In the first model of the achievement test in mathematics, designed with 40 questions we combined different types of questions: Multiple choice questions,

Matching type questions, Completion type questions and Short answer questions, presented in Table 1. The test should be completed in 45 minutes. The students making the test have no right to use any textbooks or other materials. With the first variant of the binary score test were examined 100 eighth grade students ( $n = 100$ ) from primary schools.

**Table 1:** Achievement preliminary test

<i>N</i> <i>o.</i>	<i>Problems</i>	T	F
1	$\sqrt{4}$ is: a) $\pm 4$ b) 4                              c) - 4                              d) $\pm 2$	1	0
2	Pythagorean theorem is valid for: a) any triangle                      b) right triangle c) acute triangle                      d) obtuse triangle	1	0
3	Arithmetic mean of 12 and 6 is: a) 2                              b) 3                              c) 6                              d) 9	1	0
4	The coordinate plane is consisted of: a) 1 Quadrant      b) 2 Quadrants      c) 4 Squares      d) 4 Quadrants.	1	0
5	A car uses 30l fuel for 350km. With 50l the car will drive: a) less km's    b) 5 times less km's    c) more km's    d) 5 times more km's	1	0
6	The formula $x + 5$ means: a) a number which is less from $x$ for 5                      b) a number equal to $x$ c) a number which 5 times greater than $x$ d) a number which is greater than $x$ for 5	1	0
7	$a^n$ denotes: a) square of a natural number                      b) exponentiation c) power    d) base	1	0
8	A monomial is a polynomial which has _____.	1	0
9	A regular polygon is a polygon which _____ equal _____ and equal _____.	1	0
10	The characteristic triangle of a regular hexagon is _____.	1	0
11	$c^2 = a^2 + b^2$ denotes: a) Thales' theorem                      b) sum of two squares c) Pythagorean theorem                      d) square of a sum	1	0
12	Square of a binomial subtraction is: a) $(a - b)^2$ b) $a^3 - b^3$ c) $a^2 - b^2$ d) $(a - b)(a + b)$	1	0
13	If the hypotenuse of a right triangle is 10m, and one catheti is 8m, the other catheti is ___ m.	1	0
14	In one class 5 students from 25 students have grade 5. In percentages that is: a) 20%                              b) 10 %                              c) 30 %                              d) 25 %	1	0

15	Right to each formula write the letter in front of the mathematical object, respectively: a) difference of squares $a^3 + b^3$ ____ b) square of binomial sum $a^2 - b^2$ ____ c) square of binomial subtraction $(a + b)^2$ ____ d) sum of cubes $(a - b)^2$ ____	1	0
16	Right to each mathematical object write the letter in front of the formula, respectively: a) $S(n) = (n - 2) \cdot 180^\circ$ Number of diagonals in a polygon ____ b) $D(n) = \frac{n(n-3)}{2}$ Sum of all interior angles in a polygon ____ c) $\alpha = \frac{(n-2)180^\circ}{n}$ Number of diagonals per a vertex in a polygon ____ d) $n - 3$ Interior angle of a regular polygon ____	1	0
17	If the number of diagonals in a polygon per a vertex is, then it has _ sides.	1	0
18	The number $\pi$ is a: a) real number b) rational number c) integer d) natural number.	1	0
19	If arms of an angle are cut by parallel straight lines, then the ratio of the lengths of the line segments obtained on one arm are equal to the corresponding segments obtained on the second arm. This statement is known as _____.	1	0
20	A circle is a set of all _____ in a plane that are at a given distance from a given point, called _____ of the circle.	1	0
21	The interior of a circle is a _____ of a plane, _____ by the circle.	1	0
22	The square of the hypotenuse (the side opposite the right angle) is equal to the sum of the squares of the other two sides. This formulation is known as _____.	1	0
23	Find the sides of a right triangle: a) 5cm, 6cm, 7cm b) 1cm, 2cm, 3cm c) 3cm, 4cm, 5cm, d) 2cm, 4cm, 3cm	1	0
24	The set of all irrational numbers is a subset of the set of all: a) rational numbers b) natural numbers c) real numbers d) integers.	1	0
25	Write the name of the set of numbers denoted by: $\mathbb{Z}$ _____ $\mathbb{I}$ _____ $\mathbb{R}$ _____ $\mathbb{Q}$ _____ $\mathbb{N}$ _____	1	0
26	A tangent line to a circle is a line that _____ the circle at exactly _____ point, never entering the circle's interior.	1	0
27	The chord and the circle intercept in _____ points.	1	0



28	A polygon is a plane figure that is _____ by a finite chain of _____ segments closing in a loop to form a closed chain or circuit.	1	0
29	If the length of a side in a regular octagon is 5cm then its perimeter is a) 13cm, b) 25cm, c) 40cm d) 60cm	1	0
30	If the central angle of a regular polygon is 45°, then it is called a) square b) regular pentagon c) regular octagon d) regular nonagon.	1	0
31	The number of diagonals per vertex in a pentagon is ___ diagonals.	1	0
32	The largest chord in a circle is the _____ of the circle.	1	0
33	If the central angle of a circle is 80°, then the inscribed angle over the same intercepted arc is: _____.	1	0
34	An angle inscribed in a semi-circle is a _____.	1	0
35	Find the irrational number: a) $\sqrt{1}$ b) $\sqrt{2}$ c) 3 d) $\sqrt{4}$	1	0
36	The number which we can write as a fraction is called _____ number.	1	0
37	Write the notation of the given set of numbers, respectively: a) The set of natural numbers _____ b) The set of integers _____ c) The set of rational numbers _____ d) The set of irrational numbers _____	1	0
38	The area of circle with a radius $r$ is: a) $2r\pi$ b) $4r\pi$ c) $r^2\pi$ d) $2r^2\pi$	1	0
39	The perimeter of a circle with a radius $r$ is _____.	1	0
40	If the diameter of a circle is 5cm, then its radius is ____ cm.	1	0

Descriptive statistics of the achievement test scores are reported in Table 2.

**Table 2:** Achievement preliminary test scores

<i>Measures of Central Tendency</i>	<i>Measures of Variability</i>
<b>Mode</b> - most frequent score: <b>three modes</b> - 21, 22, 23	<b>Range</b> - difference between the highest and lowest scores: <b>37 – 4 = 33</b>
<b>Median</b> - middle data: <b>20</b>	<b>Standard Deviation</b> - variability of a distribution of test scores: <b>SD = 7,21</b>
<b>Mean</b> - arithmetic average: $\bar{X} = 19,7$	

If the scores are Normally distributed the following statements are true:

$$\bar{X} \pm 1 \cdot SD = \text{approximately 68\% of the scores} \quad (3)$$

$$\bar{X} \pm 2 \cdot SD = \text{approximately 95\% of the scores} \quad (4)$$

$$\bar{X} \pm 3 \cdot SD = \text{approximately 99\% of the scores} \quad (5)$$

Since the mean and the standard deviation of the preliminary achievement test scores are 19,7 and 7,21, respectively, we found that only 66% are included

in 1 *SD*. Since the statement (3) is not true we conclude that the distribution of the preliminary achievement test scores is not Normal.

For determining reliability coefficient of the preliminary achievement test we used Split-half method. First we split the test by taking all odd numbered items in one half and all even-numbered items in the other half. Then we found the Pearson correlation coefficient  $r_1 = 0,81$  for the half-tests. Plugging this coefficient in Spearman-Brown formula (1) the reliability coefficient value of the preliminary achievement test is 0,90 which indicates excellent reliability.

In order to determine validity coefficient of the preliminary achievement test we obtained the Pearson correlation coefficient for the grades assigned by the preliminary 2014/15 annual achievement test scores and the half yearly grades from the 2014/15 school year. Since the validity coefficient value of the preliminary achievement test is 0,83 we conclude that the test is reliable.

The discrimination indexes of the preliminary achievement test tasks are reported in Table 3.

**Table 3:** Point-biserial correlation coefficients of the preliminary achievement test tasks

<i>No.</i>	$M_p$	$M_q$	$r_{pbi}$	<i>No.</i>	$M_p$	$M_q$	$r_{pbi}$
1.	24,58	18,56	<b>0,33</b>	21.	21,82	17,67	<b>0,29</b>
2.	20,59	14,21	<b>0,31</b>	22.	20,72	15,86	<b>0,27</b>
3.	22,43	16,50	<b>0,41</b>	23.	25,33	17,92	<b>0,44</b>
4.	21,19	18,58	<b>0,18</b>	24.	24,28	17,13	<b>0,47</b>
5.	20,05	18,52	<b>0,09</b>	25.	23,63	17,08	<b>0,44</b>
6.	22,53	14,21	<b>0,54</b>	26.	21,50	17,21	<b>0,29</b>
7.	22,91	15,93	<b>0,48</b>	27.	24,09	16,25	<b>0,54</b>
8.	22,57	16,33	<b>0,43</b>	28.	21,09	19,53	<b>0,07</b>
9.	21,00	17,39	<b>0,24</b>	29.	22,42	13,03	<b>0,59</b>
10.	25,56	17,75	<b>0,47</b>	30.	23,44	18,88	<b>0,24</b>
11.	22,16	13,97	<b>0,52</b>	31.	26,55	17,77	<b>0,50</b>
12.	20,59	14,67	<b>0,29</b>	32.	23,79	15,09	<b>0,60</b>
13.	24,37	16,84	<b>0,50</b>	33.	21,52	19,22	<b>0,13</b>
14.	23,64	16,84	<b>0,46</b>	34.	25,00	19,04	<b>0,26</b>
15.	22,27	12,74	<b>0,58</b>	35.	23,50	16,71	<b>0,46</b>
16.	22,63	15,49	<b>0,48</b>	36.	23,29	16,25	<b>0,49</b>
17.	24,41	17,49	<b>0,45</b>	37.	22,78	16,49	<b>0,43</b>
18.	21,03	19,07	<b>0,13</b>	38.	20,98	17,42	<b>0,24</b>
19.	21,27	14,14	<b>0,41</b>	39.	22,29	13,67	<b>0,55</b>
20.	26,58	17,53	<b>0,53</b>	40.	21,83	11,20	<b>0,59</b>

According to the discrimination indexes values presented in Table 3 we found out that we should revise the tasks number 4, 5, 9, 18, 28 and 33 since their point-biserial coefficients are below 0,20.

## 5. MAIN TESTING

The second stage of our research was the central experiment of main achievement testing. After thorough analysis of the preliminary achievements test scores and considering the time limitation for administrating the test we improved the preliminary achievement test by eliminating the tasks with the low value discrimination indexes and the tasks number 21 and 22 since their values discrimination indexes are near by 0,20. With the main annual achievement binary score test were examined 102 eighth graders ( $n = 102$ ) from primary schools.

Descriptive statistics of the achievement test scores are reported in Table 4.

**Table 4:** Annual achievement test scores

<i>Measures of Central Tendency</i>	<i>Measures of Variability</i>
<b>Mode:</b> 20	<b>Range:</b> $32 - 3 = 29$
<b>Median:</b> 18,5	<b>Standard Deviation:</b> $SD = 6,13$
<b>Mean:</b> $\bar{X} = 17,4$	

Since the mean and the standard deviation of the preliminary achievement test scores are 17,4 and 6,13, respectively, we found that only 65,7% are included in 1 *SD*. Since the statement (3) is not true we conclude that the distribution of the annual achievement test scores is not Normal.

We assigned numerical 1 – 5 grades to students by the following Table 5.

**Table 5:** Grading scale assigned by annual test scores

<i>Grade</i>	<i>Test scores - t</i>	<i>Criterion</i>
<b>1</b>	0 to 6	$t \leq \bar{X} - 1,8 \cdot SD$
<b>2</b>	7 to 13	$\bar{X} - 1,8 \cdot SD < t \leq \bar{X} - 0,6 \cdot SD$
<b>3</b>	14 to 21	$\bar{X} - 0,6 \cdot SD < t \leq \bar{X} + 0,6 \cdot SD$
<b>4</b>	22 to 28	$\bar{X} + 0,6 \cdot SD < t \leq \bar{X} + 1,8 \cdot SD$
<b>5</b>	29 - 40	$t > \bar{X} + 1,8 \cdot SD$

Since we designed the main annual achievement test by excluding some tasks from the preliminary achievement test which is reliable and valid we can conclude that this improved test is not only reliable and valid, but also it is sensitive with desirable values item discrimination indexes.

In this stage we tested the objectivity of the annual achievement test by calculating the Pearson correlation coefficient of the students' annual achievement test scores from our assessment and our colleague assessment administrated by the same test in mathematics.

Since the Pearson correlation coefficient of these two assessments is 0,86 which is above the critical value of 0,80 we conclude that the main annual achievement test is objective.

## 6. INFERENCE STATISTICS

The goal of the inferential statistics in our research: is there a difference between the traditional assessment and the assessment with the achievement tests at mathematics in eighth graders primary school. Therefore, we tested the following null hypothesis:

$X^1$  – We assume that there is no difference between the grades assigned with the main 2014/15 annual achievement test in mathematics and the grades assigned by the traditional methods of assesement in 2013/14 school year.

$X^2$  – We assume that there is no difference between the grades assigned with the main 2014/15 annual achievement test in mathematics and the grades assigned by the traditional methods of assessment in 2012/13 school year.

$X^3$  – We assume that there is no difference between the grades assigned by the traditional methods of assessment in 2012/13 and 2013/14 school years.

Since student's grade is cathegorical variable, we use  $\chi^2$ - test for homogeneity, [10]. The level of significance which is used to determine the critical value is  $\alpha=0.01$ . All chi-square tests of independence are right-tailed tests, so the critical value is  $\chi_{k;\alpha}^2$  with  $k = (r - 1)(c - 1)$  degrees of freedom, where  $r$  is the number or rows and  $c$  is the number of columns. .

The observed and expected frequencies of the grades assigned with the main 2014/15 annual achievement test in mathematics and the grades assigned by the traditional methods of assessment in 2013/14 school year are reported in Table

6. Notice that the two requirements of the  $\chi^2$ - test for homogeneity are valid:

A.1 All expected frequencies are greater than or equal to 1,

A.2 No more than 20% of the expected frequencies are less than 5.

**Table 6:** The observed and expected frequencies from the 2014/15 annual test and the traditional methods of 2013/14 assessment in mathematics

<i>Assessment \ Grade</i>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>Total</b>
<b>Annual test 2014/15</b>	3 (3,24)	27 (15,.38)	44 (35,21)	24 (30,76)	4 (17,40)	102

<b>2013/14</b>	5 (4,76)	11 (22,62)	43 (51,79)	52 (45,24)	39 (25,60)	150
<b>Total</b>	8	38	87	76	43	252

Since the statistical value  $\chi^2 = 38,30$  is greater than the critical value of the test  $\chi^2_{4;0,01} = 13,28$  we concluded that *there is a significant difference* between the grades assigned with the main 2014/15 annual achievement test in mathematics and the grades assigned by the traditional methods of assesment in 2013/14 school year with Type I error of 1 %.

The observed and expected frequencies of the grades assigned with the main 2014/15 annual achievement test in mathematics and the grades assigned by the traditional methods of assessment in 2012/13 school year are reported in Table 7. The two requirements of the  $\chi^2$ - test for homogeneity A.1 and A.2 are valid.

**Table 7:** The observed and expected frequencies from the 2014/15 annual test and the traditional methods of 2012/13 assessment in mathematics

<i>Assessment \ Grade</i>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>Total</b>
<b>Annual test 2014/15</b>	3 (4,54)	27 (17,76)	44 (28,91)	24 (34,69)	4 (16,11)	102
<b>2012/13</b>	8 (6,46)	16 (25,24)	26 (41,09)	60 (49,31)	35 (22,89)	145
<b>Total</b>	11	43	70	84	39	247

Since the statistical value  $\chi^2 = 43,62$  is greater than the critical value of the test  $\chi^2_{4;0,01} = 13,28$  we concluded that *there is a significant difference* between the grades assigned with the main 2014/15 annual achievement test in mathematics and the grades assigned by the traditional methods of assesment in 2012/13 school year with Type I error of 1 %.

The observed and expected frequencies of the grades assigned by the traditional methods of assessment in 2012/13 and 2013/14 school years are reported in Table 8. The two requirements of the  $\chi^2$ - test for homogeneity A.1 and A.2 are valid.

**Table 8:** The observed and expected frequencies from the traditional methods of 2012/13 and 2013/14 assessments in mathematics

<i>Assessment \ Grade</i>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>Total</b>
<b>2012/13</b>	8 (6,39)	16 (13,27)	26 (33,92)	60 (55,05)	35 (36,37)	145
<b>2013/14</b>	5 (6,61)	11 (13,73)	43 (35,08)	52 (56,95)	39 (37,63)	150
<b>Total</b>	13	27	69	112	74	295

Since the statistical value  $\chi^2 = 6,51$  is less than the critical value of the test  $\chi^2_{4;0,01} = 13,28$  we concluded that *there is not a significant difference* between the grades assigned by the traditional methods of assessment in 2012/13 and 2013/14 school years with 99% level of confidence.

## 7. CONCLUSIONS

In the education assessment profession, there is a promising movement toward more research and development on standardized assessment systems that are instructionally sensitive and useful for classroom teaching. But, the distinctions among different types of tests and their purposes are often unclear to policymakers, educators and other test users, leading to test misuses. In this paper we offer recommendations for the design of instructionally sensitive achievement tests in mathematics for eighth graders from primary school that can serve as improved gauge of educational evaluation needs.

Our comparative analysis showed that there is a significant difference between the grades assigned with the annual achievement test and the grades assigned with the traditional methods of assessment. Additionally, today teachers are believed to be more effective when their students have higher performance on high-stakes achievement assessments. So, it is more than ever important the teachers to create by them self the achievement test which they will practice as an important tool of the assessment.

## COMPETING INTERESTS

Authors have declared that no competing interests exist.

## References

- [1] B. S. Bloom, *Taxonomy of educational objectives: The classification of educational goals: Handbook I, Cognitive domain*, Longmans, Toronto, Green, New York, 1956
- [2] W. Huitt, J. Hummel, D. Kaeck, *Assessment, measurement, evaluation, and research*, Educational Psychology Interactive, Valdosta, GA: Valdosta State University, 2001, retrieved from: <http://www.edpsycinteractive.org/topics/intro/sciknow.html>
- [3] *Achievement test- Definition, objectives, functions, characteristics*, Nursing Journal - Online Archive of Medical and Nursing Articles,

December 2014, retrieved from:

<http://www.nsgmed.com/education/achievement-test-definition-objectives-functions-characteristics/>

- [4] R. L. Linn, N. E. Gronlund, *Measurement and assessment in teaching*, Seventh edition, Macmillan, New York, 1995
- [5] C.V. Good, *Dictionary of education*, : McGraw-Hill, New York, 1973
- [6] W. J. Popham, *Assessment for Educational Leaders*, Pearson / Allyn and Bacon, 2006
- [7] V. H. Pham, *Computer Modelling of the Instructionally Insensitive Nature of the Texas Assessment of Knowledge and Skills (TAKS) Exam*, Dissertation, The University of Texas at Austin, 2009
- [8] T. Haladyna, G. Roid, *The role of instructional sensitivity in the empirical review of criterion-referenced test items*, Journal of Educational Measurement, 18, (1981) 39-53
- [9] D. Niemi, J. Wang, D. H. Steinberg, E. L. Baker, H. Wang, *Instructional sensitivity of a complex language arts performance assessment*. Educational Assessment, 12(3&4), (2007) 215-237.
- [10] M. E. Worley, *Multiple Comparison Procedures for Chi-square Test for Homogeneity*, New Mexico State University, 1996.

Osnovna i niza srednja skola Planjane, Prizren

*E-mail address:* mejdins@googlemail.com

Faculty of Communication Networks and Security, University of Information Science and Technology, Ohrid

*E-mail address:* aneta.velkoska@uist.edu.mk

Institute of Mathematics, Faculty of Natural Sciences and Mathematics, University of St. Cyril and Methodius, Skopje