SAM Validity Study

SAM (Student Achievements' Monitoring) is a testing toolkit to assess subject competencies of school students. Theoretical framework of the toolkit relies upon the teaching/learning process concept based on L.S. Vygotsky's ideas. On the design side, SAM tests are noted for the embedded diagnostic mechanism providing information on the quality of content assimilation.

The SAM model can be implemented for different school subjects. At the moment, mathematics and Russian language tests have been developed for primary school graduates.

SAM is designed for the organization, improvement and follow-up of the learning process in schools, and is intended for teachers, methodologists and education management authorities.

Table of contents

1. Description of the SAM validity study

Validity is the extent to which a test fulfils its purpose. In the current validity conception, different forms of evidence on the validity of tests should not be considered to represent distinct types of validity, but validity should be considered a "unitary concept" (American Educational Research Association et al., 1999). From this point of view, it is important to collect evidence of validity that supports the intended interpretation and proposed use of the test's scores.

The Dutch rating system was chosen as a basis for conducting the SAM validity study (Evers, A., 2001). This rating system follows the traditional three-category classification with respect to the purpose of the validity research. These categories are: construct validity, criterion validity and content validity.

Content validity evidence is the necessary part of the developmental process of a test. The expertise of the SAM tests content was conducted by external experts specializing in particular subjects. At least two independent experts have evaluated the content of the SAM tests and have given their conclusion. The results of this expertise are not provided here, but can be provided on request.

Construct-related evidence should support the claim that the test measures the intended trait or ability. This concerns answers to questions such as "What does the test measure?" and "Does the test measure the intended concept or does it partly or mainly measure something else?" There are six types of research in support of construct validity distinguished in the modified Dutch Rating system (Evers, A. et al., 2010). These types are: research on the dimensionality of the item scores, the psychometric quality of the test (including reliability analysis), invariance of the factor structure and possible bias (e.g., evidence for fair test use), convergent and discriminant validity (i.e., supporting construct validity), differences between relevant groups (e.g., clinical and normal groups), and other research (e.g., research on criterion validity that is also relevant for construct validity). Additionally, one important way to support construct validity is to test hypotheses that follow from the theoretical foundation of the test construct.

Thus, in modern approach to validity study only one type of validity (besides content validity) can be considered and it is the construct validity. Other aspects of validity including criterion validity can be considered as aspects of construct validity.

In according with it the SAM validity study scheme has been developed and it is indicated in Figure 1. There are two parts in this scheme. First of all we should support the claim that SAM tests represent the high quality measurement tool – unidimensional, reliable and fair. Secondly, we should support the claim that SAM tests measure the intended concept. The results of research will be presented shortly below. The more detailed research description and its results are presented in the SAM Technical report.

SAM validity study was conducted during 2011-2013 SAM pilot testing in different regions of the Russian Federation. The total number of examinees was almost 6000 fourth-grade students (graduates from primary school). The design of each research and sample will be described in detail for each part of validity study.

As mentioned in (Evers, A., 2001), construct validity is a matter of the accumulation of research evidence. Construct validation research is never completed. At the end of this paper ongoing research will be presented.

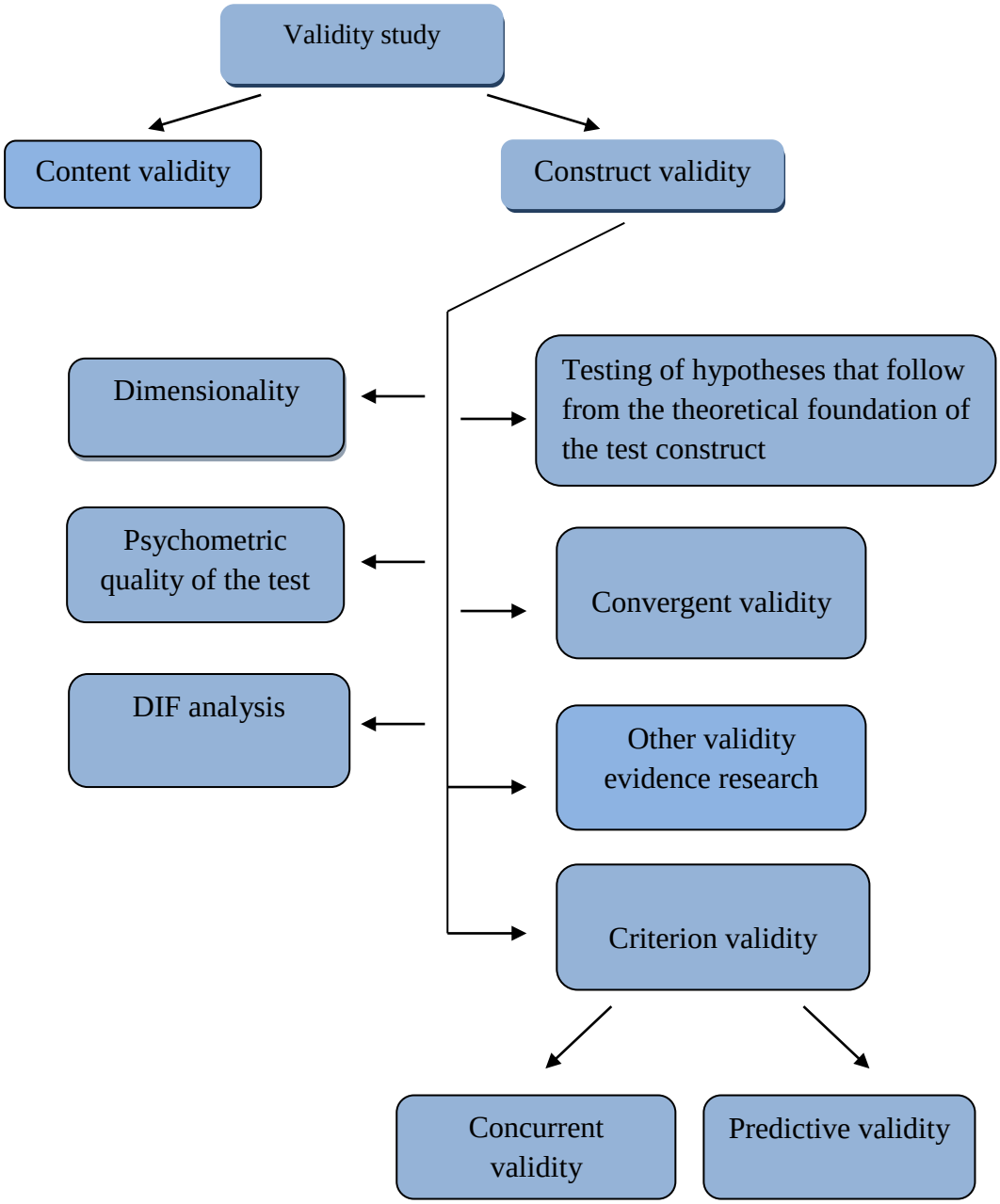

Figure 1. Structure of the SAM validity study

2. Dimensionality of SAM tests

There are known a few methods to detect multidimensionality. In this study, two methods were used: principal component analysis of raw data and principal component analysis of the standardized residuals based on Rasch analysis. We were looking for a dominant factor that could explain the largest part of the variance, and where most of the items were loaded on that factor. SPSS software was used for direct principal component analysis and WINSTEPS software was used to conduct principal component analysis of the standardized residuals of the items.

Note. All studies were run for all test forms of SAM tests. The report primarily features the results for test form 1 in mathematics only. The total number of examinees completed this test form was 3018 fourth-grade students (spring 2012).

2.1. Classical approach

Principal component analysis of raw data extracted 3 factors (based on the eigenvalue more than 1 criterion), the first one explaining 51.3% of total variance (eigenvalue is 23.1), the second one, 5.2% (eigenvalue is 2.3) and the third one, 2.5% (eigenvalue is 1.1) of the variance (Table 1). The screen plot of the eigenvalues is shown in Figure 2.

The results of the principal component analysis of raw data showed that the instrument can be considered as essentially unidimensional.

Table 1. *Principal Component Analysis of Raw Data*

Component	Eigenvalue	% Of variance	Cumulative %
1	23.1	51.3	51.3
2	2.3	5.2	56.5
3	1.1	2.5	59.0

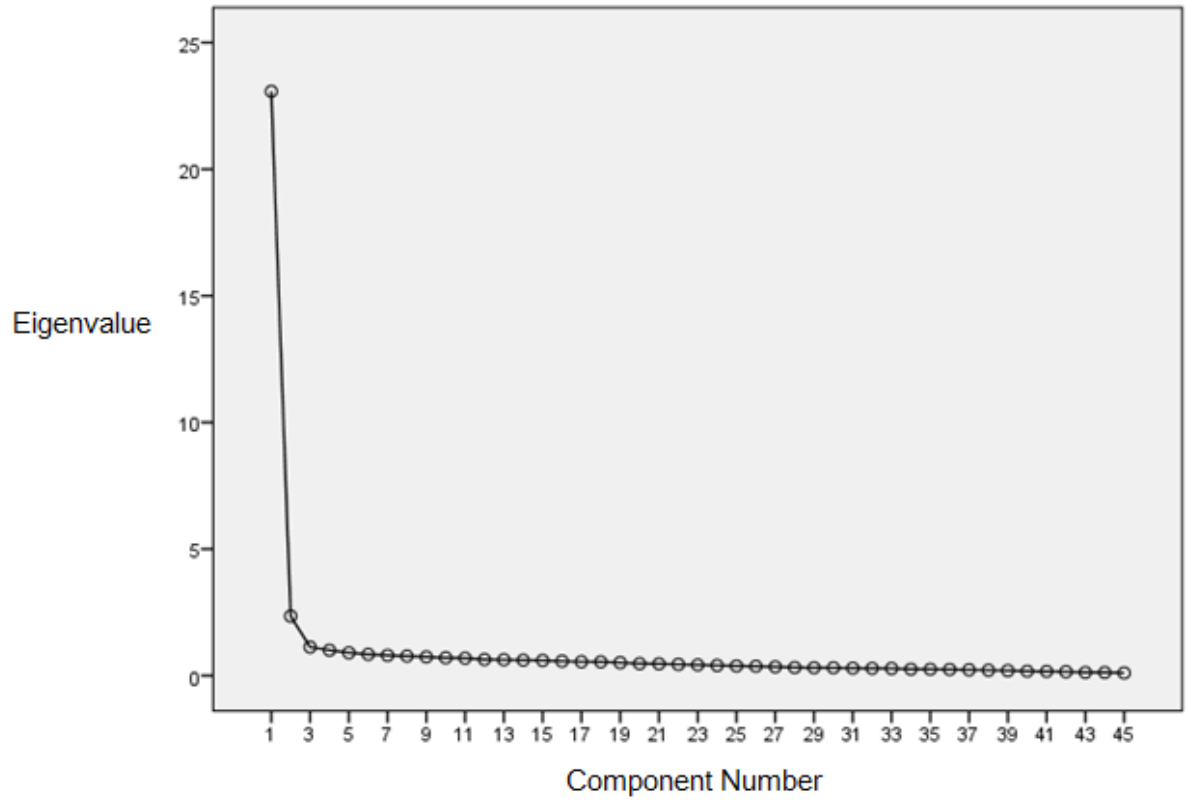

Figure 2. Screen plot of the eigenvalues

2.2. IRT approach

In addition, principal component analysis of the standardized residuals (Linacre, 1998; Smith, 2002) was used to confirm the unidimensionality of data. The results are presented in Table 2.

Table 2. *Table of Standardized Residual Variance (in Eigenvalue Units)*

	-- Empirical --			Modeled
Total raw variance in observations =	77.8	100.0%		100.0%
Raw variance explained by measures =	32.8	42.1%		41.6%
Raw variance explained by persons =	13.9	17.9%		17.6%
Raw Variance explained by items =	18.9	24.3%		24.0%
Raw unexplained variance (total) =	45.0	57.9%	100.0%	58.4%
Unexplned variance in 1st contrast =	1.7	2.2%	3.9%	
Unexplned variance in 2nd contrast =	1.5	1.9%	3.3%	
Unexplned variance in 3rd contrast =	1.4	1.9%	3.2%	
Unexplned variance in 4th contrast =	1.4	1.8%	3.1%	
Unexplned variance in 5th contrast =	1.3	1.7%	2.9%	

The most important information for dimensionality study is contained at the bottom of the table. Unexplained variance in 1st, 2nd, ... contrast is size of the first, second, ... contrast (component) in the principal component decomposition of standardized residuals, i.e., variance that is not explained by the Rasch measures, but that is explained by the contrast. We are trying to explain the data by the estimated Rasch measures: the person abilities and the item difficulties. The Rasch model also predicts random statistically-unexplained variance in the data. This unexplained variance should not be explained by any systematic effects.

According to Rasch model simulations (www.rasch.org/rmt/rmt191h.htm), it is unlikely that the 1st contrast in the "unexplained variance" (residual variance) will have a size larger than 2.0. If so, a secondary dimension in the data appears to explain more variance than is explained by the Rasch item difficulties. Here unexplained variance in the 1st contrast is 1.7. Also the variance explained by the 1st contrast is 2.2%. This is much smaller than the variance explained by the item difficulties (24.3%). So we can conclude that there are no reasons to claim about the second dimension in the data.

Conclusion: the SAM tests can be considered as essentially unidimensional. It means that only one latent trait explains students' responses to SAM test items.

3. Certification of quality and reliability of SAM tests

The research on verification of SAM tests psychometric quality was run during a special research in spring 2012. The total number of examinees was about 6000 fourth-grade students from different regions of the Russian Federation.

The data were analysed within the framework of classical test theory (including distractor analysis) as well as modern test theory IRT.

The full version of SAM psychometric quality verification study is included in the SAM Technical Report. Only main results are presented here.

3.1. Overall analysis under CTT

Summary of results for SAM test in mathematics are presented in Table 3. Distribution of test items difficulty levels and discrimination indices for test form 1 are shown in Figure 1.

Here are the main results of item analysis:

- average difficulty levels and discrimination indices are close to optimal values;
- item difficulties (p-values) are within 0.16 - 0.98 range;
- four items of the 1st level feature law discrimination that can be accounted for by their simplicity (over 90% of examinees have completed these items);
- there is a difficulty-based hierarchy within each block of three items of the 1st, 2nd and 3rd levels.
- all items have satisfactory psychometric characteristics;
- the distractors of MC items function correct;
- the test forms statistical characteristics are very close for both test forms, that allows to suggest that the test forms are parallel. (It should be note, that it is not important for SAM, because test forms had common items to make it possible to equate different test forms).

Table 3. *Summary of results for test in mathematics*

	Test form 1	Test form 2
Number of examinees	3018	2941
Raw score average	26	27
Standard deviation	8.37	8.55
Average difficulty level	0.59	0.61
Avegare dicrimination index	0.44	0.46
Average point-biserial coefficient	0.39	0.39
Reliability index (KR20)	0.90	0.91
Standard error of measurement	2.61	2.61

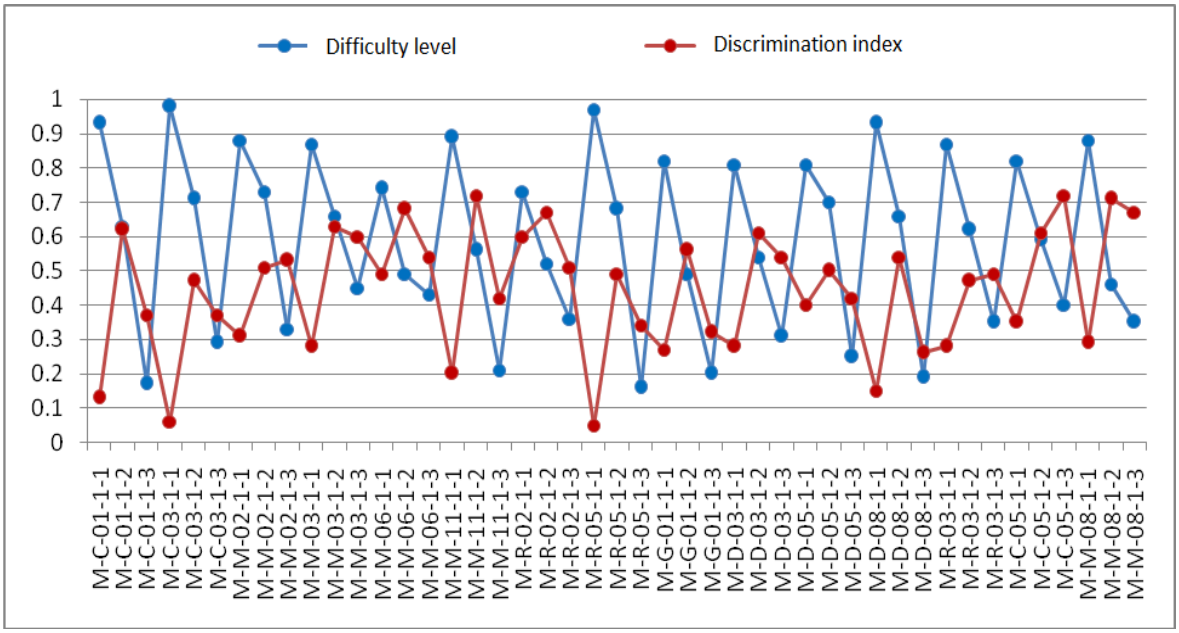

Figure 3. Distribution of difficulty levels and discrimination indices (Mathematics, test form 1)

Reliability. Classical test theory assumes that a test score additively consists of a reliable component (also called true score) and a component caused by random measurement error. The objective of the reliability analysis is to estimate the degree to which test-score variance is due to true-score variance. For tests intended for making important decisions, reliability lower than 0.8 is considered "insufficient", between 0.8 and 0.9 "sufficient" and above 0.9 "good" (Evers, A., 2001). In our case the reliability coefficient (KR20) is not less than 0.9 for both test forms, that is good. In the section 3.2 below methods from IRT reliability analysis will be also provided.

.

3.2. Analysis under IRT

The one-parameter dichotomous Rasch model was selected as a model for test data modeling and students scaling. The possibility of its use was proved in special study, devoted to this issue (Kardanova, 2010). The WINSTEPS software was used for data treatment under the Rasch model. Table 4 provides summary statistics of all measured students.

Table 4. *Summary statistics of students measures (math, test form 1)*

SUMMARY OF 3015 MEASURED (NON-EXTREME) PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	26.4	45.0	.65	.41	.99	.0	1.06	.1
S.D.	8.3	.0	1.31	.07	.20	1.0	.72	1.0
MAX.	44.0	45.0	4.82	1.03	2.11	4.3	9.90	4.4
MIN.	2.0	45.0	-4.26	.37	.48	-3.3	.20	-2.3

REAL RMSE	.43	TRUE SD	1.24	SEPARATION 2.91	PERSON RELIABILITY .89
MODEL RMSE	.41	TRUE SD	1.25	SEPARATION 3.03	PERSON RELIABILITY .90

S.E. OF PERSON MEAN = .02

Reliability. Rasch analysis provides person and item reliability and separation indices (Stone, 2004) that indicate in Table 4. The person reliability is 0.89. This means that the proportion of observed student variance considered true is 89% and the errors of SAM measures (about 0.41 logits) are small in relation to the spread of SAM students measures along the scale (about 9 logits). The person separation index is a measure of the reliability with which persons can be grouped into distinct trait levels on the basis of their performance on the measure. We observed a person separation index of almost 3 (2.91 more exactly), which indicates that test can divide the students into three distinct groups of proficiency.

Similar analysis can be conducted to evaluate item reliability and separation.

Figure 4 shows the variable map, in which the relative distribution of test takers and of items is shown across the common metric scale. On the left of the figure, there is the logit scale. On the map, test takers are represented on the left side and the items on the right side. More difficult items and high ability students are located in the upper part of the map, while easier items and low ability students are placed in the lower part of the map.

Main results of SAM tests analysis under Rasch model are:

- there are no items in the test which the participants could not achieve (that is, they were not completed due to lack of time);

- the difficulties of items are well targeted against the students and cover the range of students measures adequately;

- the majority of items demonstrate satisfactory psychometric characteristics;

- the fit of SAM individual items to the measurement model is reasonably good (INFIT and OUTFIT mean-square and standardized fit statistics provided by Winsteps were used to evaluate model fit).

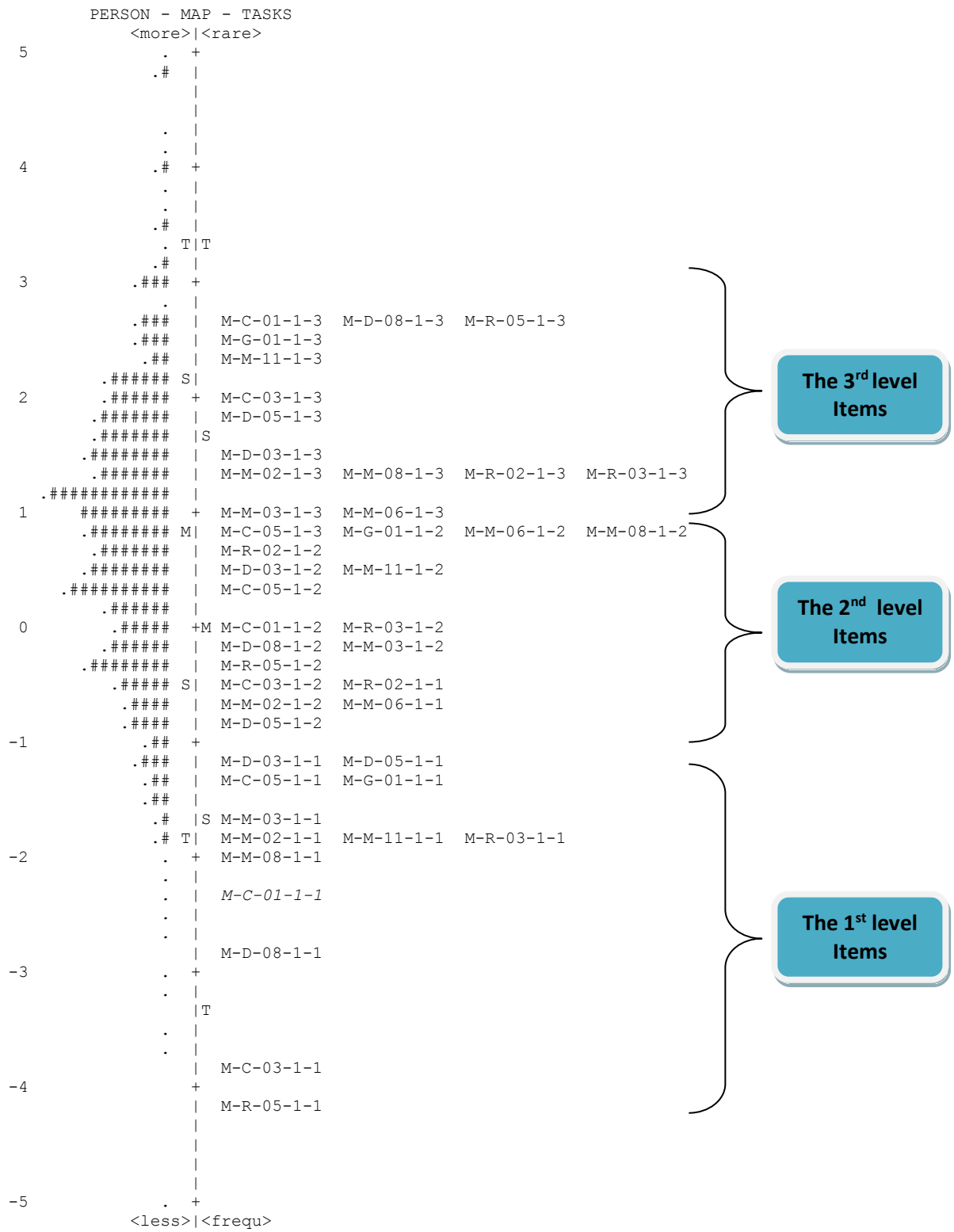

Figure 4. Variable map (mathematics, test form1)

Wright and Masters (1982) recommend the examination of clusters of items on the logit scale as a basis for interpretation of the latent variable. SAM items can be clustered into three groups in accordance with the items level. All the 1st level items are located in the lower part of the map, further there are the 2nd level items and, still further on, the 3rd level items. It means that the 3rd level items are the most difficult ones in this test and the 1st level items are the easiest ones.

Figure 5 shows test information function for test form 1 in mathematics. It is the Fisher information for the test on each point along the latent variable. The test information function reports the "statistical information" in the data corresponding to each score or measure on the complete test. Since the standard error of the Rasch person measure (in logits) is 1/square-root (information), the units of test information would be in inverse-square-logits (Stone, 2008). We see that the test information function peaks in the interval (-1, +2), where the majority of examinees are located (that follows from the analysis of students distribution – Figure 4). It means that the test is well centered regarding the sample of examinees and allows to measure students ability with a minimal error of measurement.

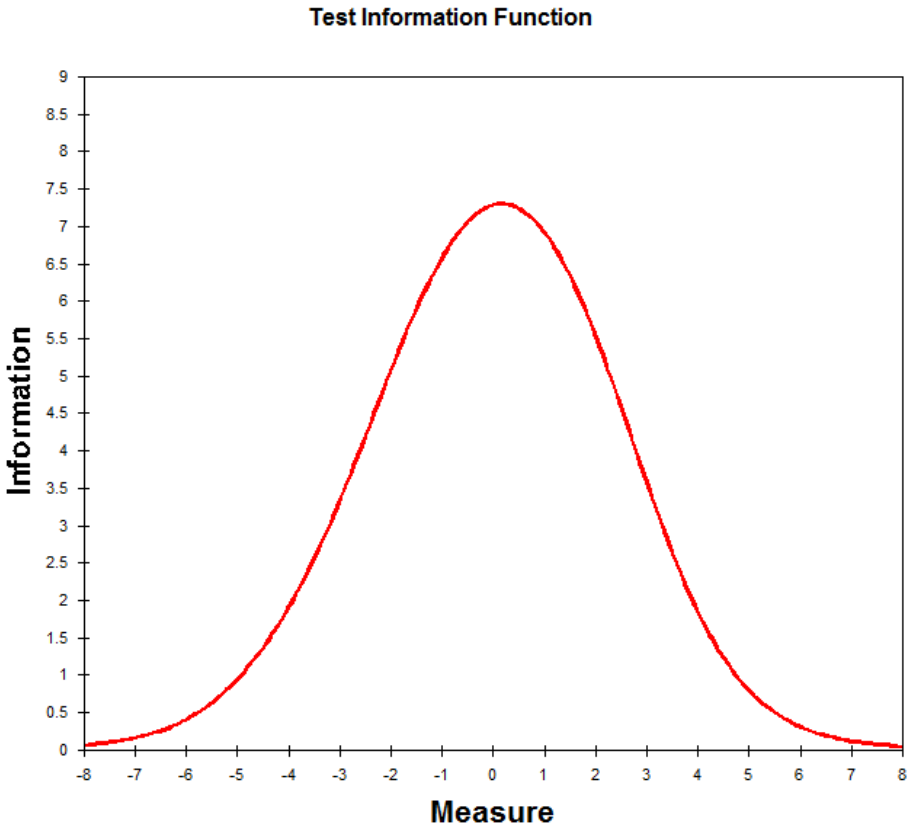

Figure 5. Test information function (mathematics, test form1)

Conclusion. The SAM tests can be regarded as a high quality measurement tool.

4. Evidence for fair test use (DIF analysis)

An item demonstrates DIF (Differential Item Functioning) if test participants with the same ability level who belong to different groups have varying chances to complete an item correctly.

In other words, an item functions in a different manner for different groups of test takers, and representatives of one of the groups can be evaluated unfairly. DIF analysis is designed to help to detect the items which will show different functioning with regard to different groups of examinees; it also establishes the degree of impact that this phenomenon may have on test participants scores.

Regarding SAM test, the basis for creating various special groups of test participants can be participant gender, region of residency (various constituent territories of the Russian Federation), country of residency (when using these tests in the Russian Federation and in CIS countries), testing language (when the tests are translated into other languages), test form (paper- or computer-based). This report presents DIF analysis results with regard to participants of different gender (boys / girls).

The full version of SAM items DIF analysis is included in the SAM Technical Report. Only conclusions are presented here.

DIF-analysis of SAM test items was conducted with Winsteps software (Linacre, 2011). Two statistics– Student's t-test and Mantel-Haenzel statistics - were used to check DIF.

On the whole girls' results are somewhat higher than boys' results (Table 5).

Table 5. *Test results for both genders (mathematics, test form 1)*

	Females	Males
Sample size	1471	1545
Observed raw score: average (SD)	26.7 (8.4)	26.2 (8.3)
Ability estimate: average (SD)	0.76 (1.15)	0.69 (1.11)

Additionally Figure 6 presents item difficulty distributions separately for samples of boys and girls.

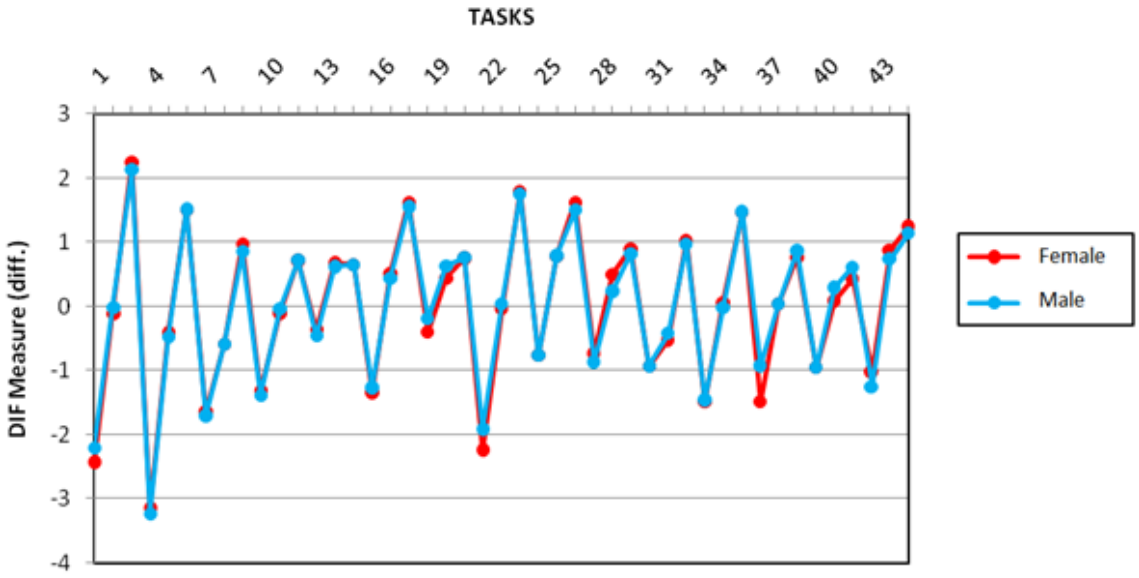

Figure 6. Items difficulties for boys and girls

Main results of DIF analysis are:

- 7 items of test form 1 in mathematics can be considered as demonstrating DIF based on gender: 5 in favour of girls and 2 in favour of boys;

- none of these item difficulty differences between groups of girls and boys exceeds the accepted threshold of practical significance (0.5 logits), which suggests that the divergence in item difficulty for student groups under review is insignificant. (Wang, 2008);

- an additional study (comparing of the test characterictic curves) has shown that the fact that there is insignificant DIF doesn't tell on examinee's test score and so on his/her proficiency level.

Conclusion. SAM items do not demonstrate different functioning with regard to representatives of different gender.

5. Testing of hypotheses that follow from the theoretical foundation of the test construct

Verification of the theoretical model that was realized in SAM tests allows to suggest two hypotheses as minimum that can be tested empirically.

The *first hypothesis* follows from the three-faceted taxonomy and can be formulated in the following way:

If three levels of the syllabus acquisition described theoretically don't contradict the logic of functional genesis, then the test items related to the same content area and meeting the criteria of three levels, should demonstrate a difficulty-based hierarchy. More exactly, the items of three levels related to the same block and meeting the theoretically-grounded criteria of three levels should be built into a difficulty-based hierarchy.

The *second hypothesis* which has been verified in the research relates the time frame for the syllabus acquisition process. The hypothesis follows from periodization of cultural development and can be formulated in the following way:

Towards the end of the primary school the syllabus is expected to be acquired on the 2nd, reflexive, level. Acquiring this syllabus on the 3rd, functional, level is expected to happen towards the end of the middle school.

5.1. Verification of the first hypothesis

The first hypothesis was partly supported in the section 3.2. The same sample was used for the research, 3018 students of the fourth grade.

The p-values of different level items of the test in mathematics are featured in Figure 7 (blue columns present the 1st level items, red columns – the 2nd level items and green columns – 3rd level items). The items of each block demonstrate a difficulty-based hierarchy: p-values decrease from the first level to the third one in each block.

Additionally p-values of items depending on the level are shown in Table 6. On the whole items of the 1st level are easier than items of the 2nd level, and items of the 2nd level are easier than items of the 3rd level. This is also seen in every content area for both subjects.

Significance of differences in items difficulty for all tests was confirmed statistically: differences are significant on the 0.05 level.

Thus, the items of the same block, which correspond to the theoretically given criteria for the three levels, feature a corresponding difficulty-based hierarchy, that is they reflect the logic of functional genesis according to the accepted taxonomy and this is a sign of validity for this construct.

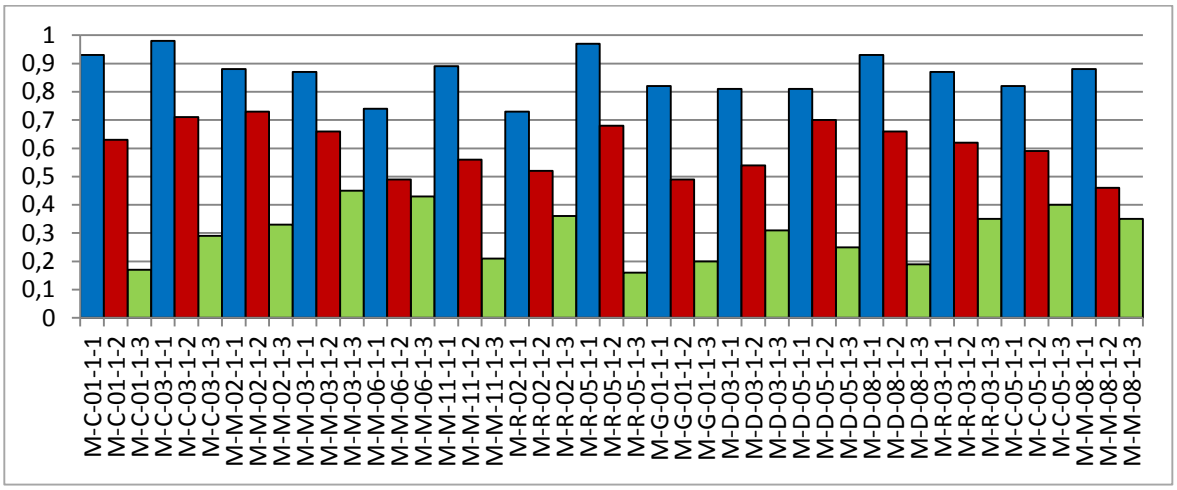

Figure 7. Distribution of difficulty levels (Mathematics, test form 1)

Table 6. *Difficulties of test items depending on their level*

		Difficulty			
	Number of items	Average	Standard deviation	Minimum value	Maximum value
The 1st level items	15	0,86	0,07	0,73	0,98
The 2nd level items	15	0,60	0,09	0,46	0,73
The 3rd level items	15	0,30	0,09	0,16	0,45
Total	45	0,59	0,25	0,16	0,98

Conclusion. The first hypothesis is confirmed.

5.2.Verification of the second hypothesis

Verification of the second hypothesis was done in a special study conducted in years 2011-2012. The study had two stages. In 2011 the SAM tests in mathematics and the Russian language were administered to four age groups – students of the 4th, 6th, 8th and 10th grades. One year later the same tests were administered to the same students who were studying at the moment in the 5th, 7th, 9th and 11th grades. Testing was done in spring, at the end of academic year. Sample included about 100 examinees in each grade from two schools that can be considered as good ones.

The main results of this research (for SAM mathematics test) will be presented below.

On the first stage the test results of the 4th, 6th, 8th and 10th grades students were analyzed. Figure 8 shows changes of test scores averages for students of different grades. Figure 9 shows the success profile in mathematics for the given student samples (average percentages of successfully completed items as a function of level for each grade). We can see the differences in syllabus acquisition, which relate to the 2nd and 3rd levels mainly.

It is important to note, that in this study no goal was set to achieve a representativeness of a sample with regard to the total population of these grades students. Students in all grades were relatively strong students.

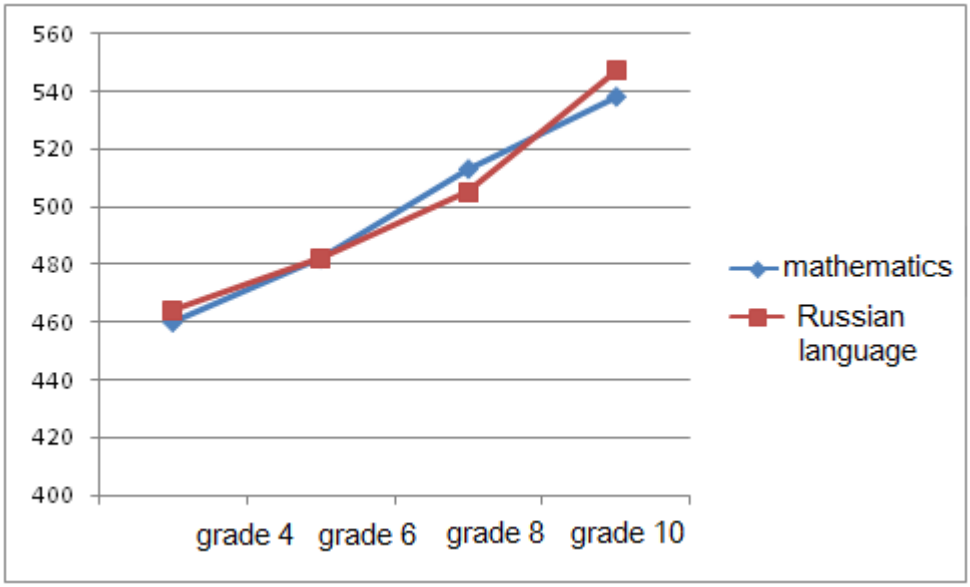

Figure 8. Average test scores for different grades

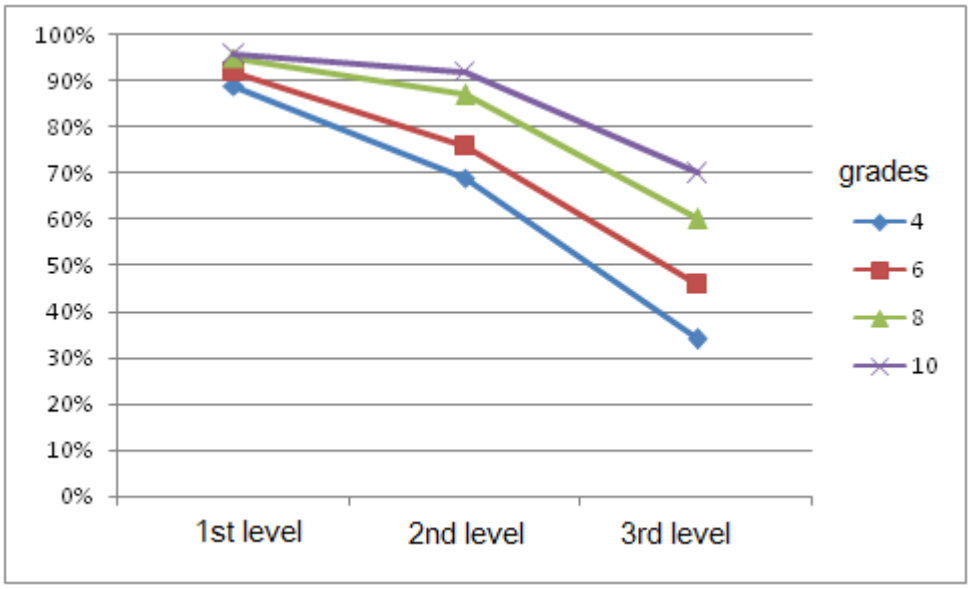

Figure 9. Success profile for the given student samples

To make the interpretation of the results for the SAM test participants possible on the basis of the three-level testing model, benchmarks were set that helped to separate all participants into four groups according to the level of their achievement. The procedure of benchmarking is described in SAM Framework and Technical Report.

Four proficiency levels were identified that correspond to the following criteria:

Proficiency level 0 (below level 1) – the student completes less than 50% of level 1 items ;

Proficiency level 1 – the student completes at least 50% of level 1 items;

Proficiency level 2 - the student completes at least 50% of level 2 items;

Proficiency level 3 - the student completes at least 50% of level 3 items.

Figure 10 shows the distribution of test participants of different age groups depending on the level of their achievement, depending on their grade. (The table does not show level 0, since the number of testing participants who qualified for this level was negligible. It is the result of a relatively strong sample, as was explained earlier.)

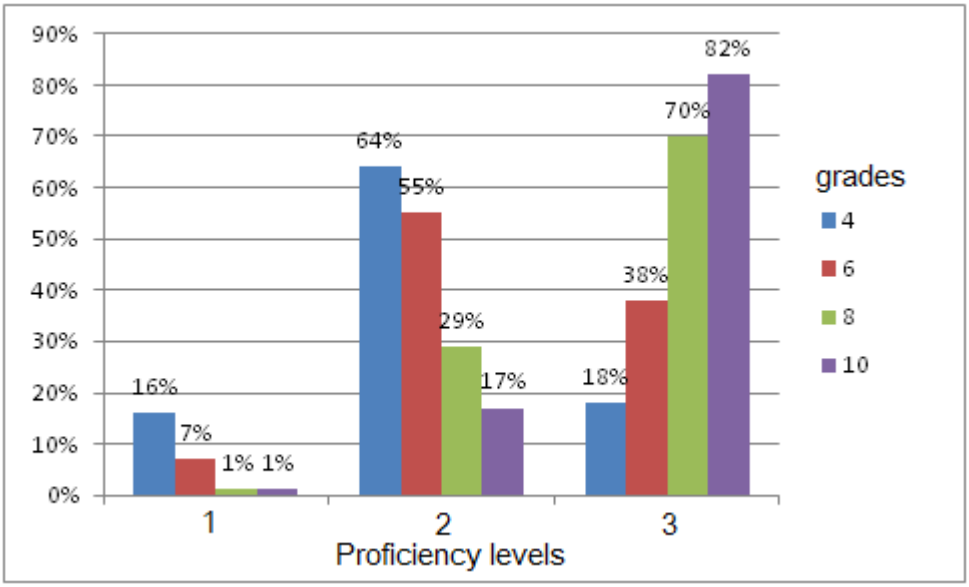

Figure 10. Distribution of students of different grades according to the proficiency level (mathematics)

From the Figure 10 we see that towards the end of primary school (the fourth grade) the majority of students demonstrate the 2nd proficiency level, that means that the reflexive level of syllabus acquisition is dominant. The percentage of students who demonstrates the 3rd level is increasing with the student age and begins to dominate at grade 8. These facts support our suggestion that in good situation towards the end of primary school syllabus is acquired on the 2nd, reflexive, level, that is comprehensive level. Acquiring this syllabus on the 3rd, functional, level is achieved in the middle school.

The purpose of the second stage of the research was to estimate the one year change in students achievement, because the same tests were administered to the same students who were studying at the moment in the 5th, 7th, 9th and 11th grades.

The diagram in the Figure 11 shows students distribution on proficiency levels depending on grade for both stages. Each initial sample (from the first stage of the research) is presented twice in Figure 11. For example, data for 4th and 5th grades relate to the same sample of students, and so on for each pair of grades – 6th and 7th , 8th and 9th, 10th and 11th.

The diagram details and supports the picture of syllabus acquisition received at the first stage. So, the percentage of students who demonstrate the 1st level is sinking to almost 0 with the increasing the school grade, and the percentage of students who demonstrate the 3rd (functional) level is growing up from the grade 4 to grade 11.

Thus, from the research we can conclude, that towards the end of the primary school (the 4th grade) the subject matter is acquired on comprehension level – most students (64%) stand at the 2nd proficiency level. The 3rd proficiency level is still rudimentary. Starting with the 8-9 grades (upon finishing the middle school) the 3rd proficiency level becomes dominant.

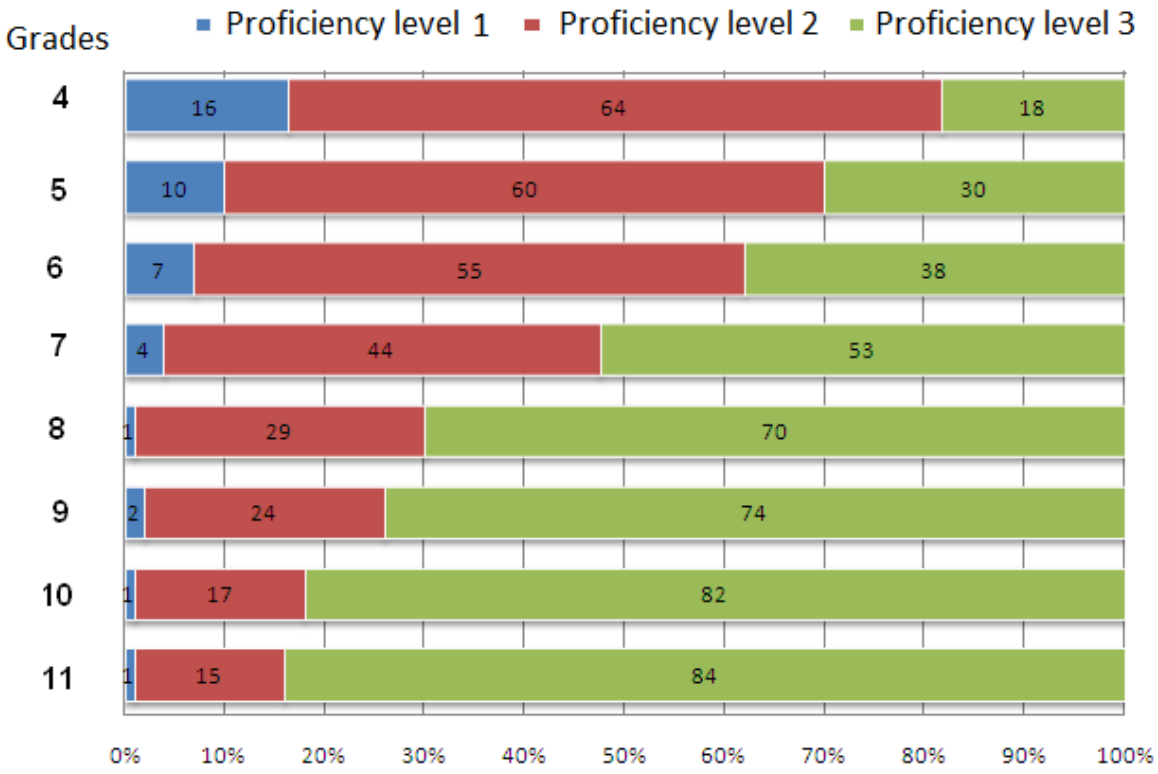

Figure 11. Students distribution of different grades depending on proficiency level in mathematics

Note. The results of this research can be more widely-used, namely for the estimation of student individual progress in time. It is possible to do because the same students were tested twice with a one year gap. Improving the proficiency level of a student is evidence of his individual progress in subject acquisition. In the 4-5 grades 61% of examinees have retained their proficiency level (4% - the 1st level, 44% - the 2nd and 13% - the 3rd); 29% of examinees have improved their proficiency level: 10% have passed from the 1st level to the 2nd one, 1% - from the 1st to the 3rd and 18% examinees have passed from the 2nd to the 3rd level.

Conclusion. The second hypothesis is confirmed.

6. Criterion validity

Research on criterion-related evidence should demonstrate that a test score is a good predictor of non-test behavior or outcome criteria (Evers, A. et al., 2010). Prediction can focus on the future (predictive validity), the same moment in time (concurrent validity), or sometimes even on the past (retrospective validity). So, criterion validity implies that in order to establish the validity of test results and their interpretation, it is necessary to compare test results with some external criterion related with the measured construct.

Usually criterion validity study includes predictive and concurrent validity. Predictive validity shows how well a test can predict future criterion scores. Concurrent criterion validity answers the question how test results are related to a criterion at present.

Below there are results of studying SAM criterion validity (the full version is presented in Technical Report).

6.1. SAM Predictive validity study

The SAM predictive validity study was based on SAM pilot testing in one of the regions of the Russian Federation in spring 2011. The total sample was 941 primary school students from 12 schools. This sample was compiled as a representative sample stratified on two parameters: school type (general education school vs. gymnasium, lyceum) and school location (city vs. village). The testing was conducted at the end of Grade 4, which is the end of primary school education.

SAM testing results were transferred to the 1000-point scale with the average about 500 and standard deviation of 50 (the scaling procedure is described in detail in Technical Report). Additionally, as was mentioned in the section 5.2, benchmarks were set that helped separate all participants into four groups according to the level of their achievement. Figure 12 shows the distribution of test participants over proficiency levels for this research.

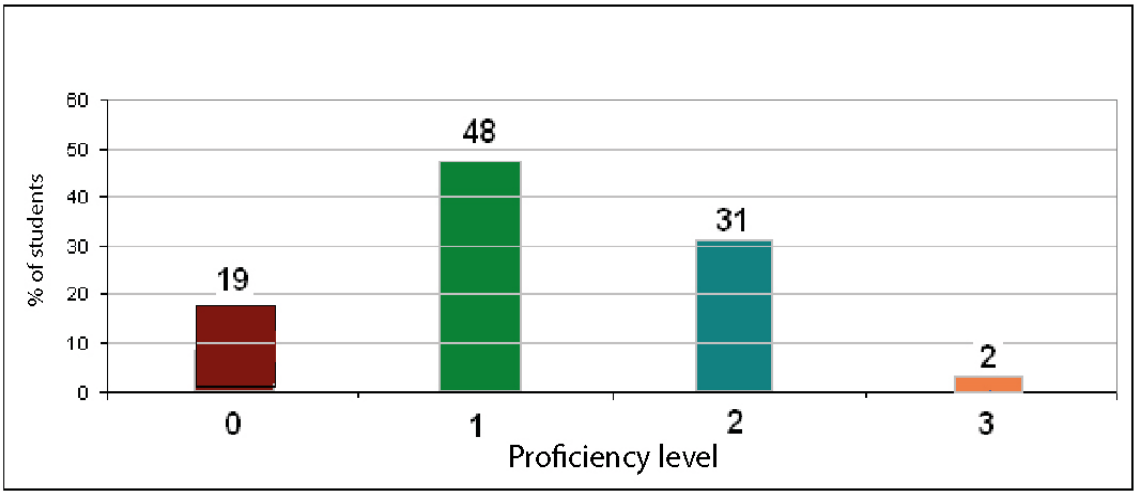

Figure 12. Test participant distribution over proficiency levels (mathematics)

To study predictive validity of SAM the same students' marks in mathematics were gathered one year later (they were studying in the 5th grade at the moment).

The main conclusions are (Figure 13):

- all students who were put into the 3rd proficiency level had a 5 (excellent) mark;
- students who were put into the 2nd level were mainly distributed between 4 (good) and 5 marks;
- among the students who were put into the 1st level one half had a 4 mark and about one third had a 3 (satisfactory) mark;
- and finally, for students who were put into the 0 level, the dominant mark is 3.

The correlation between the students' ability score and their school marks is 0.6 and the correlation between their proficiency level and the school mark is 0.56. These values are quite high, statistically significant and speak in favour of predictive validity of SAM test in mathematics.

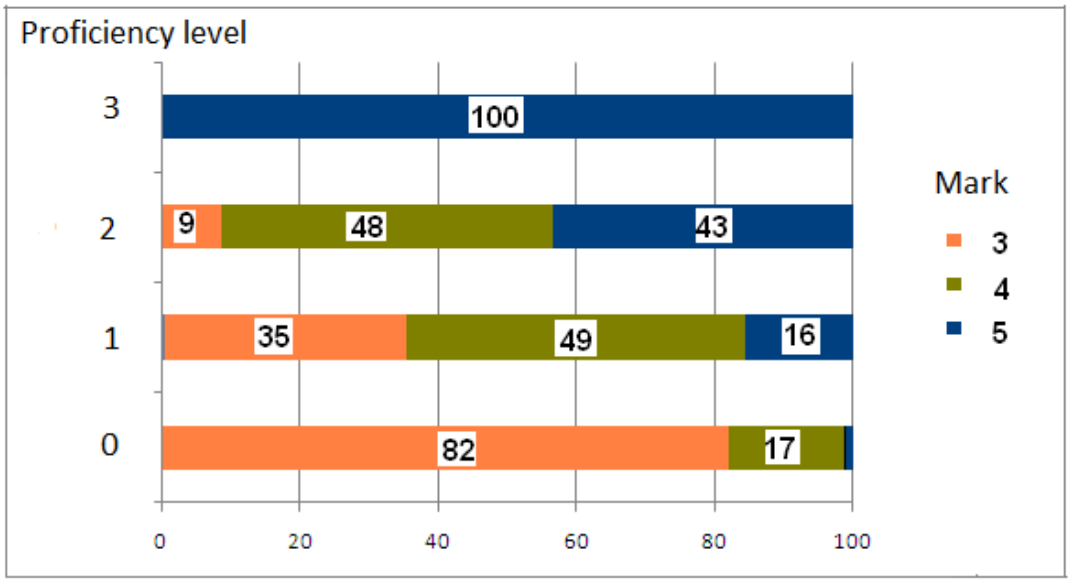

Figure 13. Distribution of student marks depending on student proficience level (mathematics)

6.2. SAM concurrent validity study

This study was based on SAM pilot testing in one of the regions of Russian Federation in spring 2012. The total sample in mathematics was 4406 examinees. All examinees were primary school graduates. A special feature of this particular pilot study was that practically all Grade 4 students from primary schools of the whole region were tested. In mathematics on the whole 2% of examinees stand at the proficiency level 0; 27% - at the 1st level; 54% - at the 2nd and 17% - at the 3rd one (Figure 14).

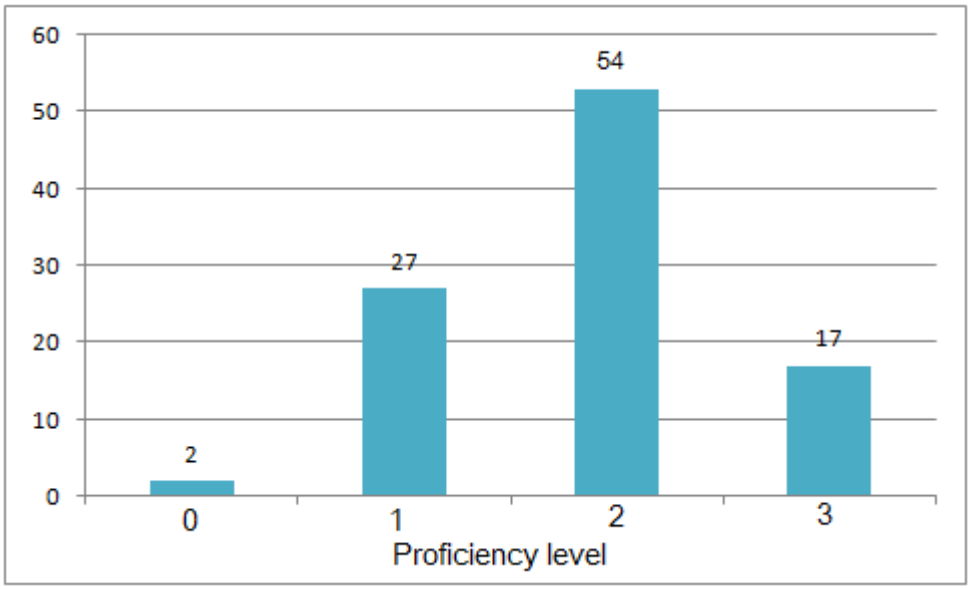

Figure 14. Test participant distribution over proficiency levels (mathematics)

Concurrent validity is studied using an external criterion for which the data are collected simultaneously with conducting experiments on the method tested. Students marks in mathematics that were expected by the end of primary school were chosen as a criterion for

current validity study. These estimates were collected from school teachers during the testing. It was possible to gather estimated marks on 3955 students. 39% of students were expected to have mark 3, 52% - mark 4 and 9% - mark 5.

The main conclusions are (Figure 15):

- among the students who were expected to have mark 5, 47% were put into the 2nd proficiency level and 48% - to the 3rd one;
- among the students who were expected to have mark 4, 61% were put into the 2nd level and 21% - to the 3rd one.
- among the students who were expected to have mark 3, 4% were put into the 0 proficiency level, 44% - to the 1st level and 46% - to the 2nd one.

The correlation between the students ability score and their expected school marks is 0.46 and the correlation between the proficiency level and the school mark is 0.41. All coefficients are statistically significant on the 0.05 level.

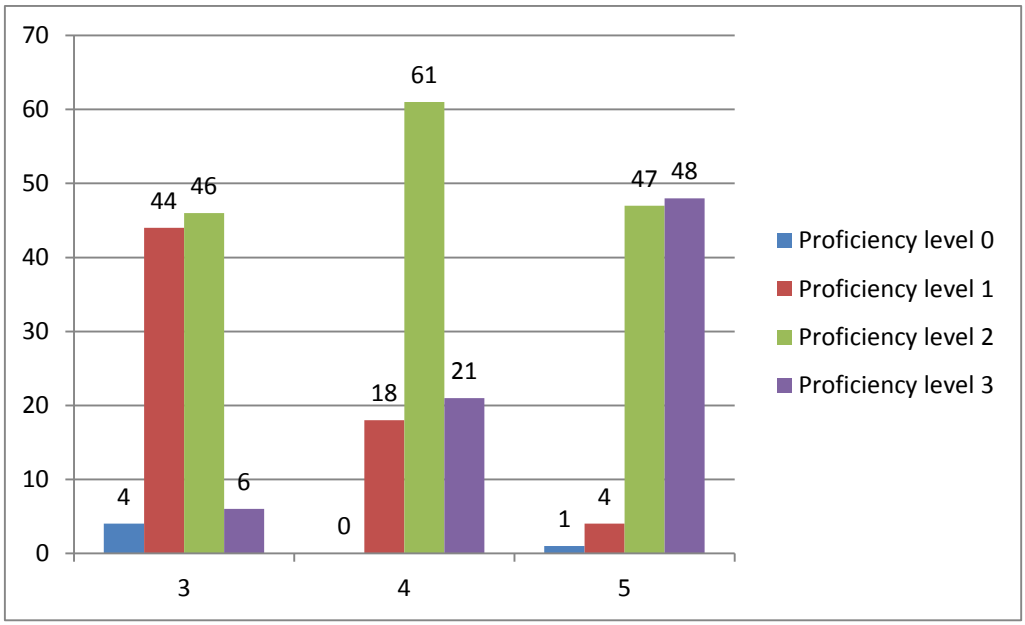

Figure 15. Student distribution into proficiency levels depending on their mark (mathematics)

7. Convergent validity

Convergent validity refers to the degree to which two measures of constructs that theoretically should be related, are in fact related. Convergent validity, along with discriminant validity, is considered a subtype of construct validity. Two these types of validity are traditionally used in sociology, psychology, and other behavioral sciences. Convergent validity can be established if two similar constructs correspond with one another, while discriminant validity applies to two dissimilar constructs that are easily differentiated.

To establish SAM convergent validity, we need to show that measures that should be related are in reality related. So we need to find instrument aimed measuring related construct.

We decided to use for this purpose an instrument of monitoring of educational achievements in mathematics of primary school students that was developed by the Center of Quality Assessment of Russian Academy of Education (http://www.centeroko.ru/fgos/fgos.htm). This test is curriculum-based and serves the purpose to evaluate the level of primary school

students' preparation in according with the Federal state educational standard for primary school. This test will be called AT test (Achievement Test) in this chapter.

A special research was conducted to compare test results on these two tests. Among students who completed AT test, students with high test scores were selected. The maximum possible score for AT test was 24. Students with test scores not less than 20 were selected for our research.

These students can be identified as having high level of mathematical preparation in according to the Federal standard. Then SAM math test was administered to these students.

Sample of students who completed both tests included 1785 students. All of them were the fourth grade students, the last grade of primary school. Testing was done in spring 2012, at the end of academic year.

The hypothesis tested was: the results of these students on SAM tests should be high, most of them should be put into 2nd and 3rd proficiency levels.

The results of the research are presented shortly below.

Figure 16 shows histogram of raw scores of the students on SAM test. The results of these students on SAM test are very high, average score is 33, standard deviation is 6.5, average p-value is 0.74 (see table 3 for comparison with the SAM testing sample).

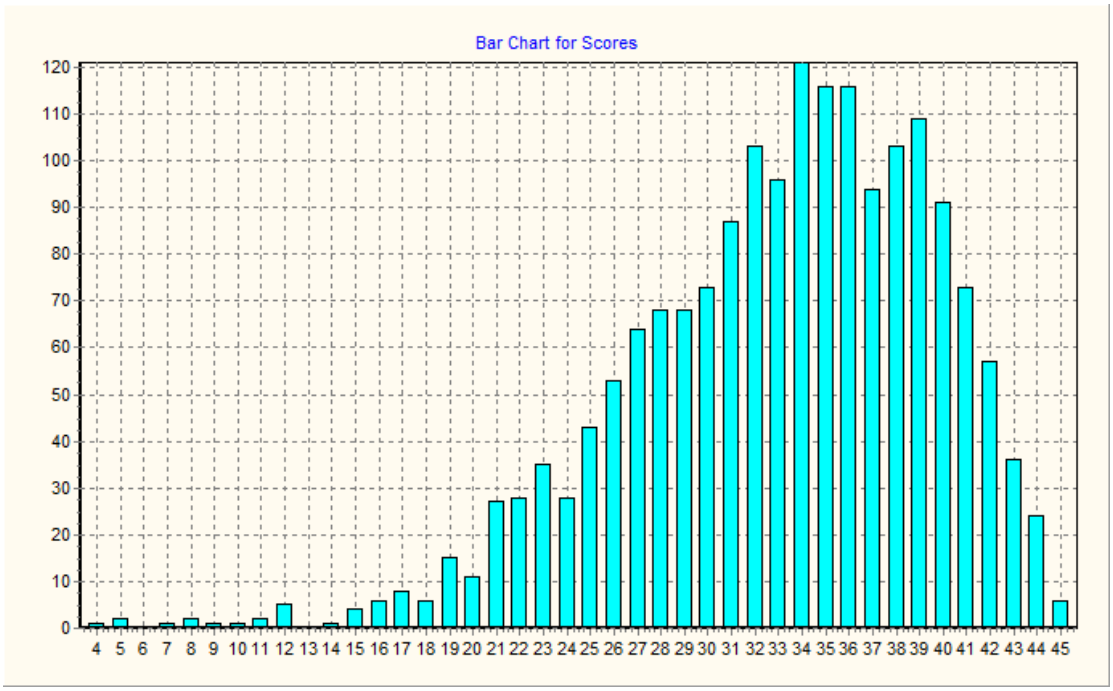

Figure 16. Histogram of raw scores

Figure 17 shows the distribution of participants of this research over proficiency levels: 52% of all participants stay at level 2 and 39% - at level 3. For comparison, Figure 14 shows the distribution of test participants over proficiency levels for the sample of SAM pilot testing in 2012 (4406 students).

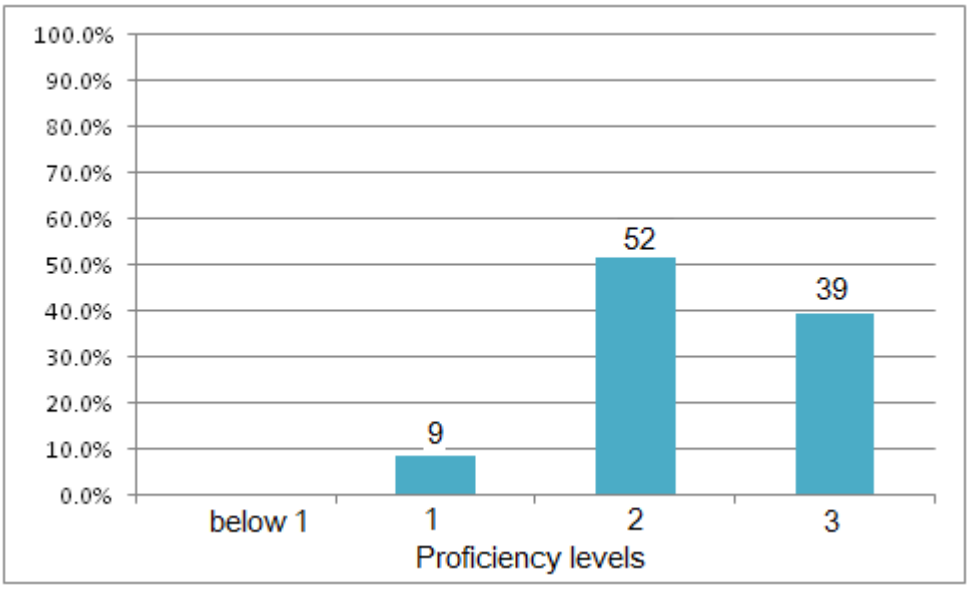

Figure 17. Student distribution into proficiency levels

Thus, 91% of the students stay at levels 2 and 3, that confirms our hypothesis.

Figure 18 shows the success profiles in mathematics for the given student sample in comparing with the sample of SAM pilot testing in 2012. There are shown average percentages of successfully completed items as a function of item level for each sample. We can see the differences in syllabus acquisition between two samples, which relate to the 2nd and 3rd levels mainly. Students of the sample analyzed can complete in average 79% of the 2nd level items and 50% of the 3rd level items.

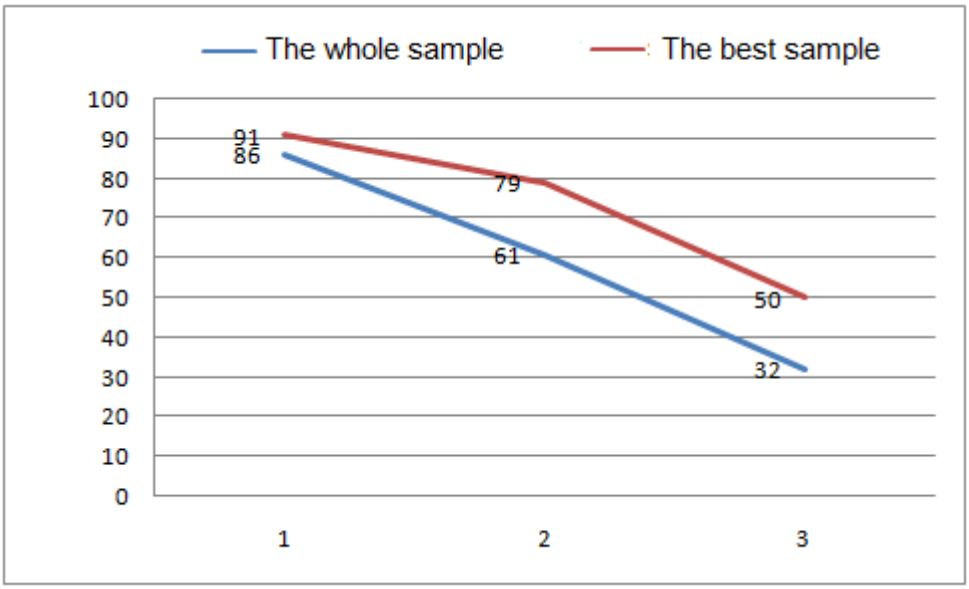

Figure 18. Success profiles for two student samples

Average scores of students analyzed on SAM test and AT test depending on proficiency level are presented in Table 7. We see that test scores of these students on SAM test vary significantly, that means that SAM test possesses high differentiative power.

Table 7. *Average scores of students on SAM test depending on proficiency level*

	Average AT test score	*Average SAM test score*
Proficiency level 1	20	478
Proficiency level 2	21	536
Proficiency level 3	23	598
Total		555

The correlation between the students ability scores on two tests is about 0.3. The low value can be explained by lack of variance among test scores on AT test: test scores vary only in range from 20 to 24. So correlation coefficient is not an appropriate measure of convergent validity for our research. Evidences presented above support the claim that measures on AT test and SAM test that should be related are in reality related. It is a support of convergent validity of SAM test.

8. Other validity evidence research

8.1. Prediction of items difficulty

It is important that people who developed the test have a common understanding of what constitutes each level of students' thinking. To support this issue a special research was conducted aimed at confirmation of theoretical item difficulty hierarchy (predicted by test developers) with empirical data.

5 test developers (in mathematics) were asked to estimate expected difficulty of all items for population, the test is designed for. Inter-judge consistency was estimated: coefficient of consistency is 0.83. After test administration we estimated correlation between expected and empirical p-values for the whole test and for each subtest of items of the 1st, 2nd and 3rd levels. Table 7 presents coefficients of correlation. These results show quite high level of consistency between experts and high correlation between their expectations and empirical results, that is one more evidence of construct validity.

Table 7. *Coefficients of correlation between expected and empirical p-values*

	Correlation
The 1st level items	0.882
The 2nd level items	0.973
The 3rd level items	0.928
The whole test	0.931

8.2. The potential of SAM three-faceted taxonomy to be communicated to other researchers

The research question is whether the understanding of what constitutes each level of students' thinking can be communicated to other researchers who then should be able to use the definitions provided by the test developers to categorize problems into different levels in a consistent way.

Or, in other words, can a high level of agreement be obtained from experts in education outside the team of SAM developers?

To test it a special study was conducted. Below is the description of the research.

A short one-hour lecture about SAM theoretical model and three-faceted taxonomy was given to 22 school teachers - participants of a seminar on improving teaching. The sample included 15 primary school teachers and 7 others who were teachers in history, literature, chemistry, mathematics and physics. After the lecture, the teachers were asked to distribute 30 SAM math items in accordance with their level (formal, reflexive and functional). The items were chosen from SAM tests in mathematics, two random items from each block. Further, all items were given in random order. In a similar way 30 items from a SAM test in the Russian language were chosen for the second stage of the experiment.

So, the teachers were asked to define the level of each item (in math and the Russian language separately) in accordance with the three-faceted taxonomy. Inter-rater consistency was estimated: the coefficient of consistency equaled 0.63. The consistency between the teachers' ratings and the items real levels was 0.62 for mathematics items and 0.73 for Russian language items.

The findings of the research suggest that SAM theoretical foundation is understandable for teachers and the three-faceted taxonomy of SAM test items can be communicated to other researchers.

9. Ongoing research

As it was mentioned in the beginning of this paper, construct validation research is never completed. Below there are a few directions of ongoing research.

1. Connection of SAM test results and educational program that this school is realized. There are more than 10 different educational programs for primary school in Russia, and schools have a choice what program to use. The results of preliminary research allow to suggest that educational program has an impact on the students achievement in primary school.

2. Connection of SAM test results and teachers characteristics. In Russia, in the primary school there is a single teacher for both core subjects - Russian language and mathematics. The learning outcomes of the class are basically associated with this person, so our intention is to explore teachers' beliefs and practices as well as their connection to quality of education.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Evers, A. (2001). The revised Dutch rating system for test quality. *International Journal of Testing*, 1:2, 155–182.

Evers, A., Sijtsma, K., Lucassen, W. and Meijer, R. R.(2010). The Dutch Review Process for Evaluating the Quality of Psychological Tests: History, Procedure, and Results. *The International Journal of Testing*, 10:4, 295 — 317.

Kardanova, E. (2010). The development of the toolkit for assessment of subject competences of primary school students. The paper presented at the 36-th Annual Conference of the International Association for Educational Assessment IAEA 2010. Bangkok

Linacre, J.M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome measurement*, 2, 266-283.

Linacre J. M. (2011). A User's Guide to WINSTEPS. Program Manual 3.71.0. (http://www.winsteps.com/a/winsteps.pdf).

Smith, Jr. E. V. (2002). Detecting and Evaluating the Impact of Multidimensionality using Item Fit Statistics and Principal Component Analysis of Residuals. *Journal of Applied Measurement*, 3:2, 205-231.

Stone, M.H. (2004). Substantive scale construction. In E.V.Smith, R.M.Smith (eds), *Introduction to Rasch measurement* (pp.201-225). Maple Grove, MN: JAM Press.

Stone, M.H. (2008). Fisher's Information Function and Rasch Measurement. *Journal of Applied Measurement*, 9:2, 125-135.

Wang, W.-C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement*, 9(4), 387-408

Wright, B.D., Masters, G.N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.