Technical Report for the SAM (Student Achievement Monitoring) **project**

Table of contents

1. SAM Project—development objective, theory, results

The project is dedicated to the development of the toolkit for student achievement monitoring, or SAM, to be used at the level of a separate school or a municipal school system. The objective of the project development was to create a set of pedagogical tests that would help combine the measurement of academic achievement with its qualitative characteristic on the basis of its reference to a certain taxonomy of educational objectives.

The SAM model was elaborated within the framework of the cultural development theory as postulated by L. S. Vygotsky and his followers [1; 2; 3; 4; 5; 10]. The key element of this model is three-level taxonomy of developing subject competence, i.e. such cultural mode of operation that is inherent in some school subject [4; 6; 7; 9]. As per this taxonomy, three qualitative levels exist for the assimilation of the cultural pattern of action, and, for the purposes of brevity, they can be defined as formal, reflective and functional (Fig. 1.1). Each of these levels is linked to a certain type of orientation within the subject content and manifests itself in an ability to solve the relative class of tasks.

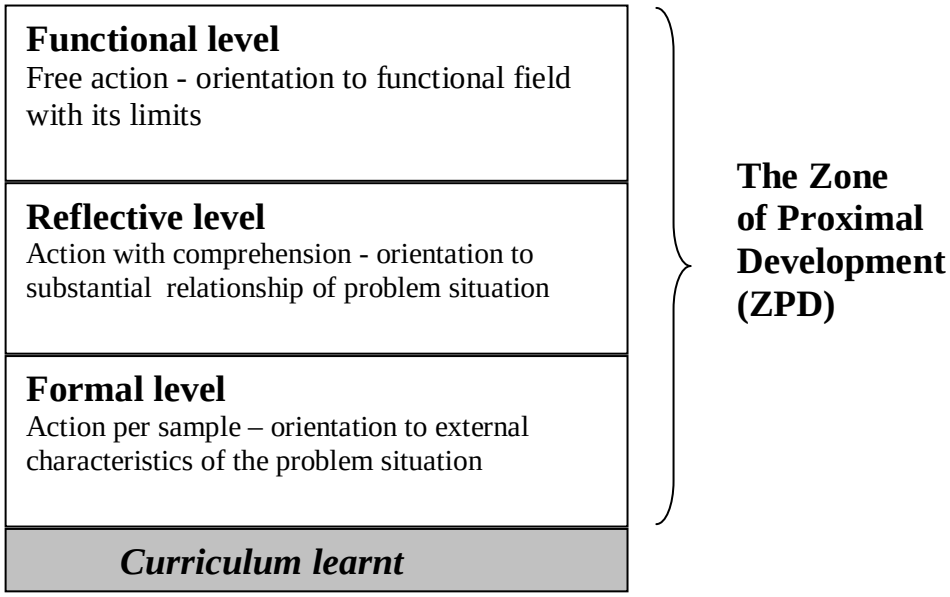


Fig. 1.1. Multi-level model for assimilating subject content

The described multi-level pattern was taken as a basis for creating SAM tests.

Proficiency levels of the mode of action were put into correspondence with indicators, i.e. with types of tasks. When designing a content-related test for each relatively integral content area, blocks of test items are developed. Each block includes three test items assigned to levels 1, 2, and 3, providing a natural difficulty-based hierarchy. Each block functions as a detector that identifies the level of assimilation of the respective content area (the level is identified by the most difficult test item in the block solved by the test taker). The set of blocks covering major curriculum areas enables to obtain a structural picture of the curriculum assimilation by the student.)

Blocks of items in mathematics and Russian language are presented below to illustrate the point made:

Block of math items

1 What number will be obtained if 10472 is divided by 34? Answer: _____
2 Absentminded Peter copied an exercise on multiplication of two numbers from the textbook. He wrote down the first multiplier correctly: 7. But in the second multiplier, he displaced two figures. As a result, he got 147. What answer should Peter have got if he had copied the exercise correctly? Answer: _____
3 What is the largest result that can be obtained if letters in the expression AB5 + BC2 are substituted with figures (different letters should be replaced with different figures)? Answer: _____

Note. All three items refer to the content area Numbers and Calculations. The first item suggests simple application of the rule (algorithm) of calculation. The second item requires analysis of the wrong arithmetic operation (with due regard for positional principle) and designing a plan for its correction. And finally, the third item assumes "playing" with the positional principle to obtain a specific value that meets the condition needed to get the largest value.

Block of Russian language items

1 Mark the sentence in which NOT ALL necessary commas were entered. 1. Осенний листопад. Листья летят, скачут, плывут. 2. Мальчишки просидели в засаде до вечера, но ушли почти ни с чем. 3. Мама сидела за компьютером и писала какой-то доклад. 4. На массивном столе с зелёным сукном лежал ноутбук папки с бумагами, калькулятор.
2 Put in punctuation marks in accordance with described situations. 1. One student put punctuation marks incorrectly so that the meaning of the sentence was that the sea waves washed ashore both objects and people. Один ученик неверно расставил знаки препинания, и у него получилось, что море выбросило на берег предметы и людей. **Прибоем выбросило на берег корабль__испанцев__лодку__ рыбака__катер.** 2. The other student put punctuation marks correctly, so that the sea waves washed ashore only objects. **Прибоем выбросило на берег корабль__испанцев __лодку__ рыбака__катер.**
3 Formulate the below sentence in a new manner, so that its sense will remain the same, but it will become a simple sentence with uniform parts. **Вода непрестанно трудится, и время от времени края горных уступов обрушиваются.** _____ _____

Note. Items in this block are related to punctuation in simple sentences with uniform parts. The first item includes several simple sentences assuming direct application of a relevant rule. The second item requires an ability to isolate uniform items in the conceptually ambiguous sentences, that is isolating and taking into account significant semantic relations shall be necessary. The third item is based on the

assumption that the examinee has an active ability to handle the concept of the simple sentence with uniform parts and is thus capable of transforming a complex statement into a simple one while keeping its meaning unchanged, that is it tests the ability to analyze options and to choose an adequate option.

Such blocks created using the contents of the main parts of the curriculum are combined into test forms. Accordingly, the test can be regarded as a system consisting of three subtests, each of which has a set of test items of the same level from different curriculum content areas (Fig. 1.2.).

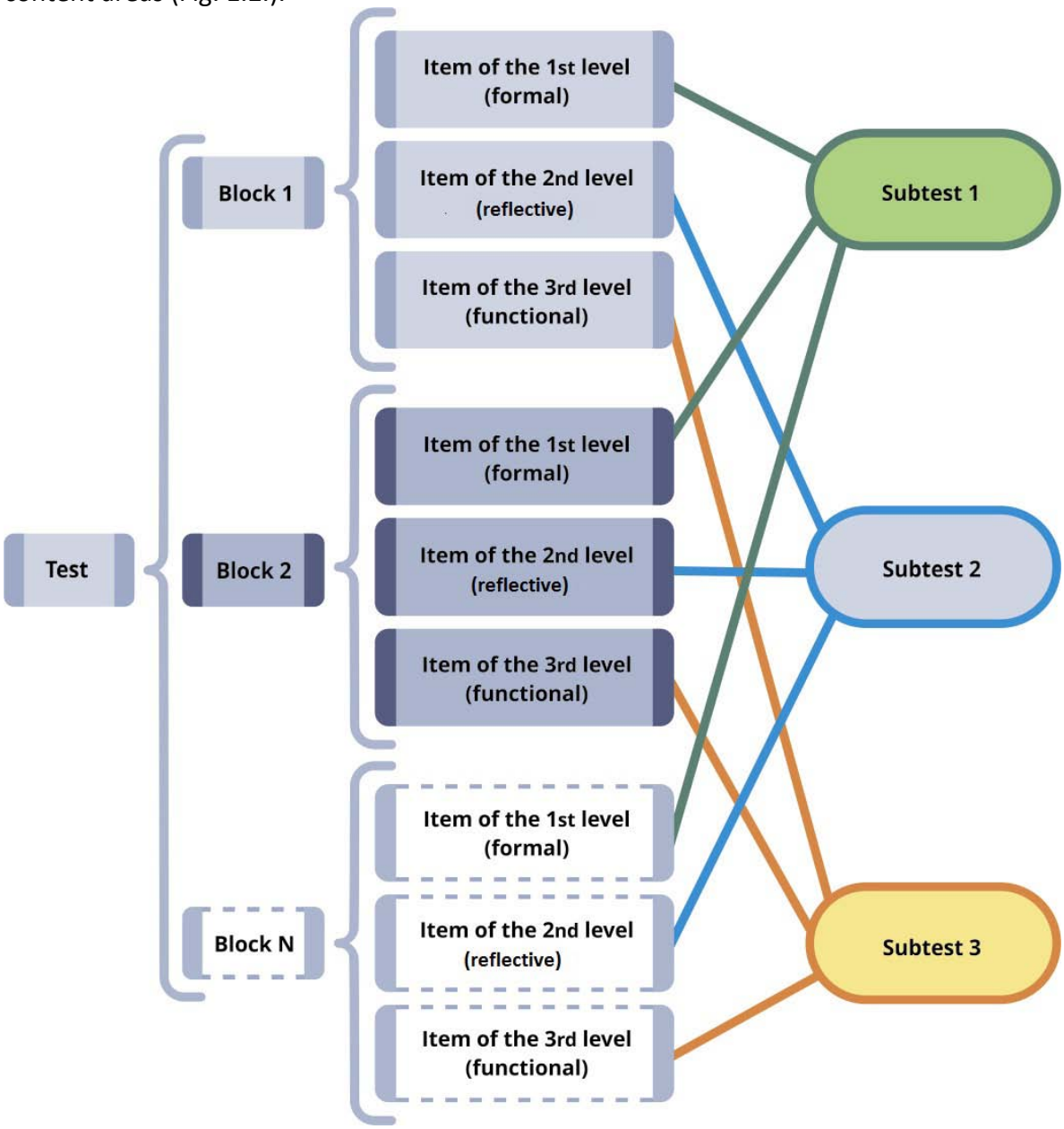


Fig. 1.2. Test structure

Thus the test can address a double objective: a) it works as a tool for integrative measurement of learning achievements (with its set of 45 items covering the main subject areas); and b) it provides diagnostic methodologies for defining the level of material assimilation (15 blocks of items).

SAM tests include items of different types: completion ones having a brief answer, multiple-choice ones having a choice of 4-5 offered options, also matching items, items that

require constructions, etc. (Test items were presented in blocks, the sequence of blocks made no difference.)

Score procedure uses a dichotomous approach: students get 1 point for a correct answer and 0 for incorrect (or absent) answer. The highest raw score that the test taker can achieve for completing the test equals 45. The highest raw score that each test taker can get for each of the three subtests (i.e., for each level) equals 15.

Each subject test is designed in several versions having similar statistical characteristics. All versions include common items enabling to perform equating and get the scores of all test takers on a common scale. The number of common items in different versions makes up at least 6 (at least two common blocks). Tests can be offered in paper-based or computer-based forms.

Within the SAM model, the tests having common structure were developed for two subjects: mathematics and Russian language.

SAM model and its tool system was fully described in the book: «SAM: a Tool for Student Achievement Monitoring» (2011) (edited by: P.G. Nezhnov, Ye.Yu. Kardanova).

Literature

1. Выготский Л.С. История развития высших психических функций //Собр. соч.: В 6 т., т. 3. М.: Педагогика, 1983.
2. Выготский Л.С. Мышление и речь //Собр. соч.: В 6 т., т. 2. М.: Педагогика, 1982.
3. Гальперин П.Я. Психология как объективная наука. М., 1998.
4. Давыдов В.В. Теория развивающего обучения. М.:ИНТОР, 1996.
5. Диагностика учебной успешности в начальной школе. М.: ОИРО, 2009.
6. Запорожец А.В. Психология действия. Москва-Воронеж: НПО «МОДЭК», 2000.
7. Исаев Е.И. Психологическая характеристика способов планирований у младших школьников // Вопр. психологии. 1983. № 2.
8. Лернер И.Я. Процесс обучения и его закономерности. М.: Знание, 1980Магкаев В.Х. Экспериментальное изучение планирующей функции мышления в младшем школьном возрасте // Вопросы психологии, 1974, № 5.
9. Магкаев В.Х. Экспериментальное изучение планирующей функции мышления в младшем школьном возрасте // Вопросы психологии, 1974, № 5.
10. Микулина Г.Г., Савельева О.В. К психологической оценке качества знаний у младших школьников // Психологическая наука и образование. 1997. № 2.
11. Нежнов П.Г. Опосредствование и спонтанность в модели «культурного развития» // Вестн.Моск.Ун-та. Серия 14. Психология. 2007, № 1.
12. Нежнов П.Г., Карданова Е.Ю., Эльконин Б.Д. Оценка результатов школьного образования: структурный подход // Вопросы образования, 2011, № 1, стр. 26-43.
13. Нежнов П.Г., Медведев А.М. Исследование теоретического анализа у школьников // Вопросы психологии, 1990, № 5.
14. Савельева О.В. Психологические критерии качества знаний младших школьников // Автореф… дисс. канд. пед. наук. М., 1989.
15. Симонов В.П. Диагностика степени обученности учащихся: Учебно-справочное пособие. М.: МПА, 1999
16. Эльконин Б.Д. Введение в психологию развития. М., 1994.
17. Эльконин Д.Б. Избранные психологические труды. М.: Педагогика, 1989.
18. Карданова Е.Ю. Моделирование и параметризация тестов: основы теории и приложения. – М.: Федеральный центр тестирования, 2008, 304 с.

19. 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 IAEA 2010 “Assessment for the Future Generations”. Bangkok
20. Карданова Е.Ю., Нейман Ю.М. Проблема выравнивания в современной теории тестирования. Вопросы тестирования в образовании, 2003, № 8, с. 21-40.
21. Kardanova, E., Nezhnov, P. (2011): School Achievements Monitoring Toolkit: Assessment Framework. The paper presented at the 37-th Annual Conference IAEA 2011. Manila.
22. Bloom, B., Englehart, M. Furst, E., Hill, W., & Krathwohl, D. (1956), *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*, New York, Toronto, Longmans, Green.
23. Mullis, I. V.S., Kennedy, A. M., Martin, M.O., Sainsbury, M. (2006), PIRLS 2006 *Assessment Framework and Specifications, 2nd Edition*, TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
24. Nezhnov, P. (2011). SAM – toolkit to assess primary school students’ academic achievements. CADMO. Innovations in assessment to meet changing needs. ANNO XIX, 1, pp.85-98.
25. SAM (School Achievement Monitoring): Инструмент мониторинга учебных достижений школьников // *под ред. Нежнова П.Г., Кардановой Е.Ю.*, 2011, 104 с.

2. Psychometric Characteristics of the SAM Tests

This report shows the results of the statistical analysis conducted for the SAM pilot testing data in mathematics and in the Russian language, which was run using P&P form in various regions of the Russian Federation in spring 2012. The total number of examinees was over 5 thousand fourth-graders with the secondary school.

The data were analyzed within the framework of classical test theory as well as modern test theory IRT.

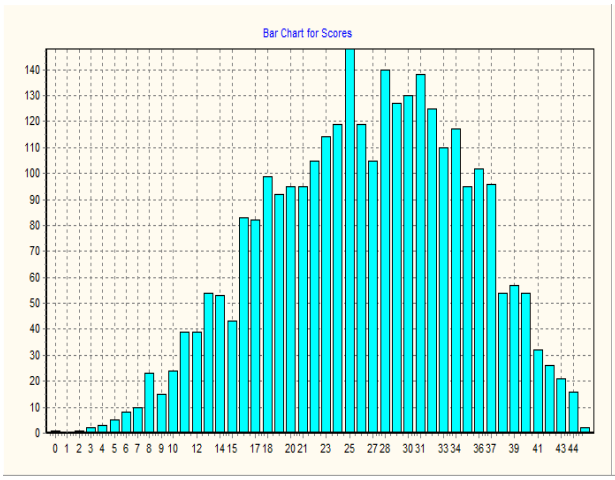
2.1. Results of the pilot testing statistical analysis for test items in mathematics

Two test forms in mathematics were used during pilot testing. The test consisted of 45 items grouped in 15 blocks. Table 2.1 shows the summary of statistical indices for two test forms **within the framework of the classical test theory**. Figure 2.1 presents raw score distribution histograms of test participants.

Table 2.1. Summary of test results (mathematics)

	Test form 1	Test form 2
Number of examinees	3018	2941
Raw score average	26	27
Standard deviation	8.37	8.55
Skewness	-0.21	-0.37
Kurtosis	-0.56	-0.36
Average difficulty level	0.59	0.61
Average discrimination 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

Test form 1



Test form 2

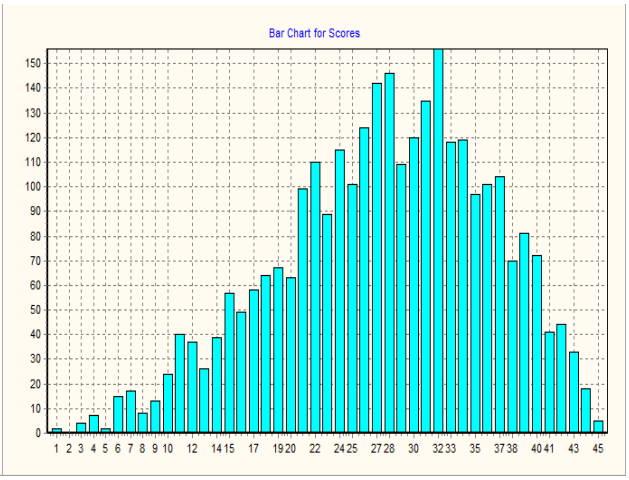


Figure 2.1. Raw score distribution histograms of test participants

Table analysis shows that statistical indices of the two test forms are quite close. Reliability indices are quite high: not lower than 0.9 for both test forms. Both average difficulty level and average discrimination indices are close to optimal values.

Table 2.2 shows detailed information on items in **test form 1.** When conducting classical analysis, an item's major characteristics are its difficulty and its discrimination. Item difficulty is defined by the percentage of the test participants who were able to complete this item correctly. The higher the value of this index, the easier is the item.

Discrimination characterizes the differentiative capacity of an item, that is its capability of making a distinction between test participants with differing levels of preparation. This report makes use of two discrimination indexes: the classical one (which is the difference between item difficulty values for two groups of test participants: 27% of the best ones, having higher scores, and 27% of the weakest ones, having lowest scores) and the adjusted point-biserial correlation coefficient (correlation between an item score and the overall test score, after the results for this specific item are removed). The value of 0.2 was chosen to be the critical value for discrimination indexes.

The analysis of Table 2.2 may lead to a conclusion that all items of the test under review function well. For the observed group of test participants, four Level 1 items were very easy – their item difficulty values were over 0.9, which means that these items were completed successfully by over 90% of examinees (in the table, these items are highlighted in green). These same four items are characterized by low discriminativity (in the table, low-discriminativity items are highlighted in pink). Thus, the extreme ease of completing these items manifested itself in their low level of discriminating fineness: an item which everyone can deal with successfully does not show any differentiation quality.

The rest of the test items show good statistical indices. Let us point out that difficulty-based hierarchy is pronounced within each block of Level 1, 2, and 3 items.

Table 2.2. Statistical indices of items functioning (mathematics, test form 1)

Item number and item index		Difficulty level	Discrimination index	Adjusted point-biserial correlation
1	M-C-01-1-1	0.93	0.13	0.21
2	M-C-01-1-2	0.63	0.62	0.49
3	M-C-01-1-3	0.17	0.37	0.38
4	M-C-03-1-1	0.98	0.06	0.15
5	M-C-03-1-2	0.71	0.47	0.39
6	M-C-03-1-3	0.29	0.37	0.30
7	M-M-02-1-1	0.88	0.31	0.39
8	M-M-02-1-2	0.73	0.51	0.43
9	M-M-02-1-3	0.33	0.53	0.42
10	M-M-03-1-1	0.87	0.28	0.32
11	M-M-03-1-2	0.66	0.63	0.50
12	M-M-03-1-3	0.45	0.60	0.44
13	M-M-06-1-1	0.74	0.49	0.44
14	M-M-06-1-2	0.49	0.68	0.51
15	M-M-06-1-3	0.43	0.54	0.40
16	M-M-11-1-1	0.89	0.20	0.27

17	M-M-11-1-2	0.56	0.72	0.54
18	M-M-11-1-3	0.21	0.42	0.39
19	M-R-02-1-1	0.73	0.60	0.51
20	M-R-02-1-2	0.52	0.67	0.50
21	M-R-02-1-3	0.36	0.51	0.40
22	M-R-05-1-1	0.97	0.05	0.14
23	M-R-05-1-2	0.68	0.49	0.39
24	M-R-05-1-3	0.16	0.34	0.35
25	M-G-01-1-1	0.82	0.27	0.28
26	M-G-01-1-2	0.49	0.56	0.41
27	M-G-01-1-3	0.20	0.32	0.29
28	M-D-03-1-1	0.81	0.28	0.28
29	M-D-03-1-2	0.54	0.61	0.45
30	M-D-03-1-3	0.31	0.54	0.45
31	M-D-05-1-1	0.81	0.40	0.42
32	M-D-05-1-2	0.70	0.50	0.40
33	M-D-05-1-3	0.25	0.42	0.35
34	M-D-08-1-1	0.93	0.15	0.28
35	M-D-08-1-2	0.66	0.54	0.44
36	M-D-08-1-3	0.19	0.26	0.25
37	M-R-03-1-1	0.87	0.28	0.35
38	M-R-03-1-2	0.62	0.47	0.35
39	M-R-03-1-3	0.35	0.49	0.38
40	M-C-05-1-1	0.82	0.35	0.35
41	M-C-05-1-2	0.59	0.61	0.46
42	M-C-05-1-3	0.40	0.72	0.54
43	M-M-08-1-1	0.88	0.29	0.37
44	M-M-08-1-2	0.46	0.71	0.53
45	M-M-08-1-3	0.35	0.67	0.52

Figures below display graphic representations of the indices from Table 2.1.

Figure 2.2 is a bar chart presenting the distribution of difficulty values (or p-values) for the items of the test under review. Item indexes are shown on the horizontal axis in the order of their appearance during the test while p-values are shown at the vertical axis. Level 1 items are shown in blue, Level 2 items in red and Level 3 items in green. This diagram provides a clear picture of the difficulty-based hierarchy within each block. Figure 2.3 also shows the distribution of difficulty values, this time as a curve diagram.

Figures 2.4 and 2.5 show the distribution of the discrimination indexes for the test items. And finally, Figure 2.6 displays a joint distribution of both p-values and discrimination indexes for the test items.

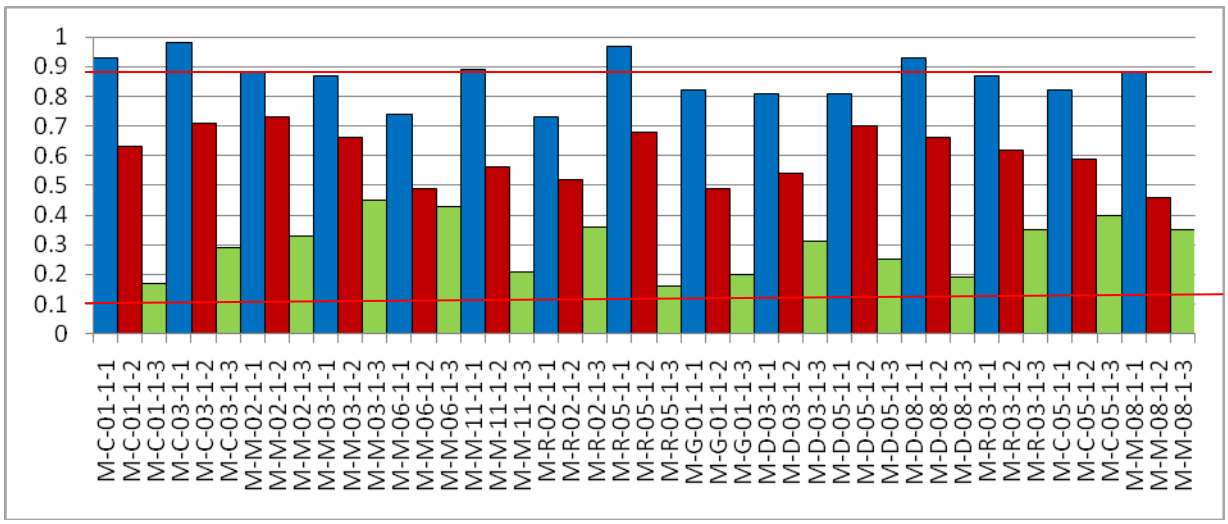


Figure 2.2. Diagram of p-values distribution (mathematics, test form 1)

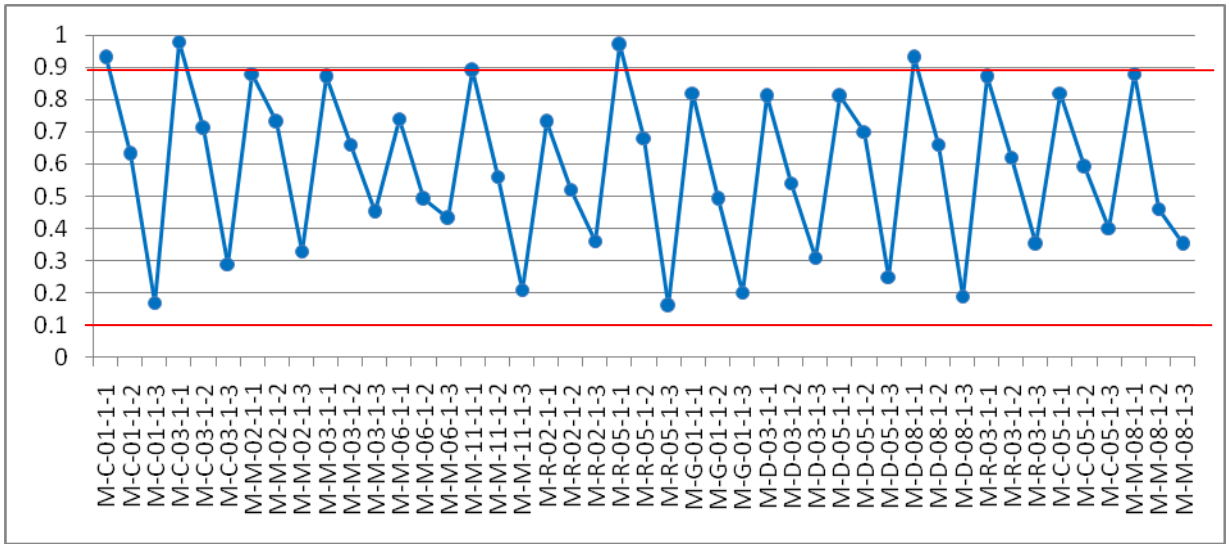


Figure 2.3. Difficulty values distribution diagram (mathematics, test form 1)

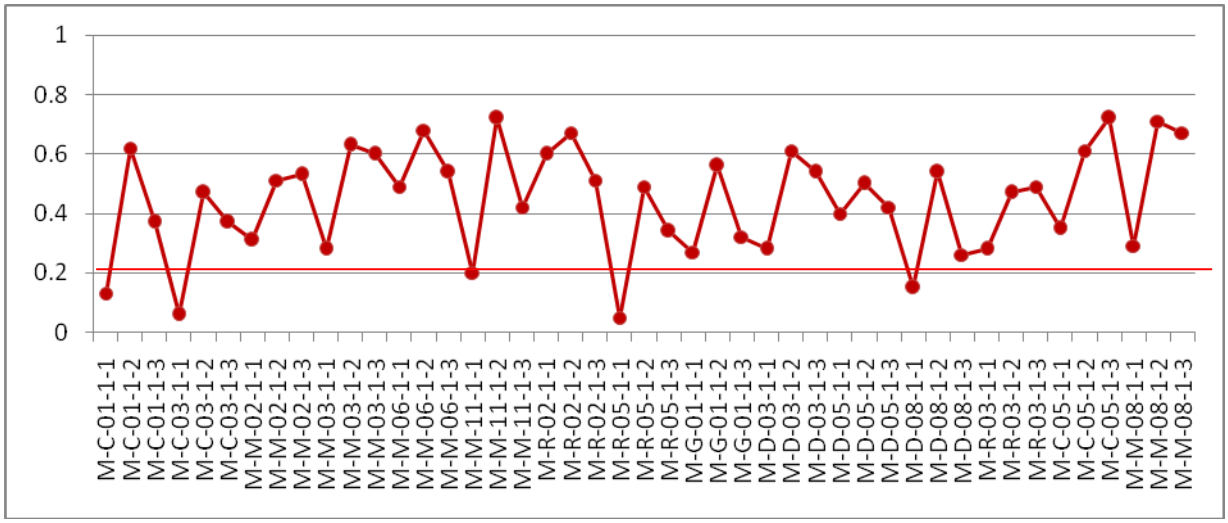


Figure 2.4. Diagram of discrimination indexes distribution for test items (mathematics, test form 1)

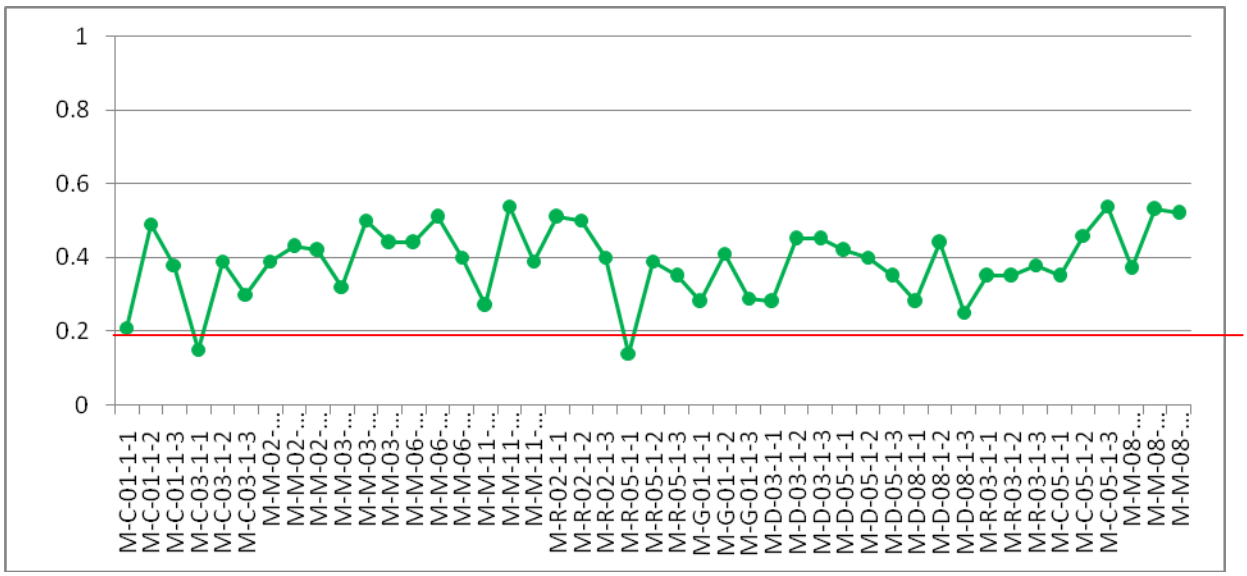


Figure 2.5. Point-biserial coefficient distribution diagram (mathematics, test form 1)

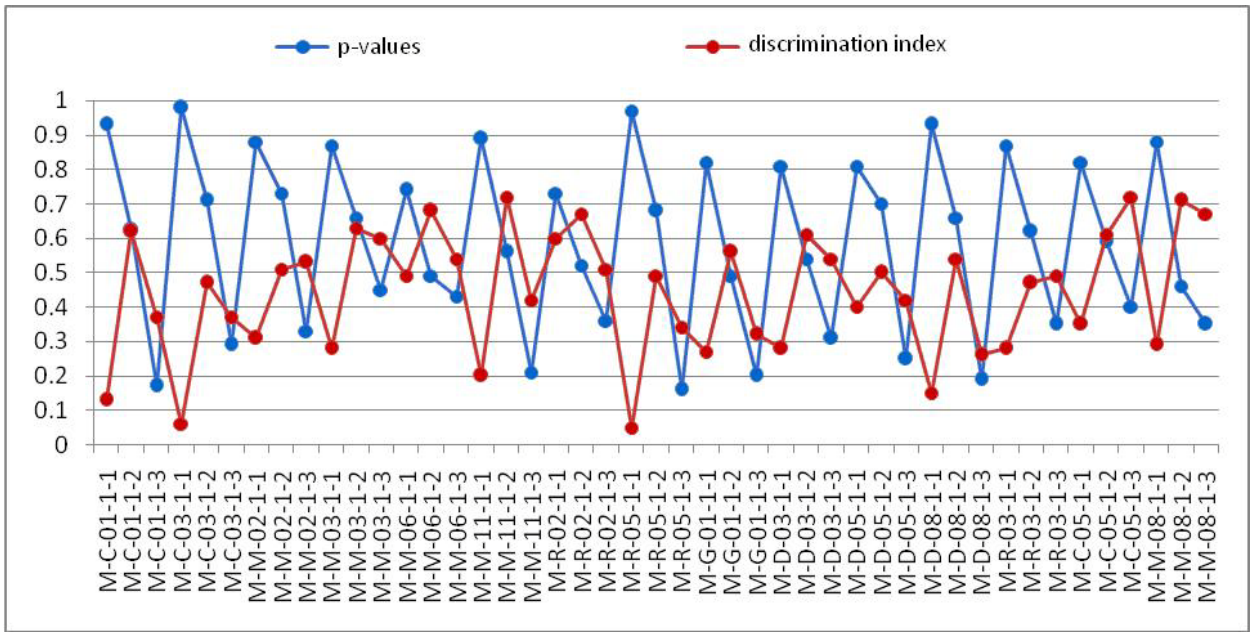


Figure 2.6. Joint representation of both p-values and discrimination indexes distribution (mathematics, test form 1)

The above figures show well that practically all items have good indices, i.e. optimal level of difficulty and high discrimination. Only four extremely easy Level 1 items have discrimination indexes below normal values.

Additionally, an **analysis of distractor functioning** was undertaken, relative to multiple-choice options. In this test, there were seven such items.

To analyze distractors, firstly, the distribution of examinee responses was reviewed over all response alternatives for multiple-choice items, the goal being the determination of non-functioning distractors. Secondly, correlation coefficients (point-biserial) were calculated between distractors and the overall grade point for the test. A distractor functions properly, if high ability examinees do not choose it as a correct answer. In this case, the correlation will have a negative value (it is desirable that it would be less than -0.2). And vice versa, the

correlation for a correct response option must have a positive value (the correlation factor for the correct answer option coincides with the discriminativity index for items that were dealt with earlier).

No serious problems were found for distractor functioning of the multiple-choice items with the test under consideration.

Tables 2.3 and 2.4 show examples of distractor analysis for two items. The correct answer option is marked by a star and is highlighted in colour. For item M-M-03-1-2 all indices are within norm: 66% of test participants completed the item correctly. The high value of the positive correlation factor is a witness to the fact that these examinees were good students with high test scores. The rest of the students who could not complete the item are distributed over three distractors. All distractors were, in fact, functioning: each was chosen by at least 5% of the test participants. Negative values of the correlation factors indicate that these were weak students with low test scores.

Table 2.3. Distractor analysis for item M-M-03-1-2

Item M-M-03-1-2				
Answer options	**А**	**Б***	**В**	**Г**
Examinee choice distribution	11%	66%	7%	6%
Correlation coefficient	-0,22	0,54	-0,29	-0,23

Table 2.4. Distractor analysis for item M-M-03-1-3

Item M-M-03-1-3					
Answer options	**a + b**	**a * b**	**a + 2b**	**2a + b***	**2a + 2b**
Examinee choice distribution	11%	9%	9%	44%	22%
Correlation coefficient	-0,20	-0,31	-0,10	0,49	-0,05

Item M-M-03-1-3 was more difficult: only 44% of the examinees completed it and 5% passed up on it. The high value of the positive correlation factor for the correct answer is a witness to the fact that these examinees were good students with high test scores. The rest of the students who could not complete the item are distributed over four distractors. All distractors were, in fact, functioning: each was chosen by at least 5% of the test participants. Negative values of the correlation factors indicate that these were weak students with low test scores. In this item, however, the 2a + 2b distractor is preferable: it was chosen by 22% of the examinees. At the same time, its correlation with the overall score is insignificant, which means that this distractor was attractive for both weak examinees and for those with a high level of preparation. Such a situation is characteristic for items having a certain catch that good students may not notice.

The analysis of test form 1 items (in mathematics) has thus shown within the framework of the classical test theory that all items exhibit good characteristics.

A more thorough analysis of the test was conducted **within the framework of the modern test theory IRT (or Item Response Theory)**. Please find below main results of this analysis.

The unidimensional dichotomous Rasch model was used as test model. As per this model, each test item is characterized by one parameter, which is difficulty, and each test participant is also characterized by one parameter, which is ability level. Estimates of all parameters (both of test participants and test items) are located on the common metric scale (the unit of measurement on this scale is called 'logit'), and they are provided with characteristics of estimation precision. The starting point of this scale is not defined and can be chosen arbitrarily. Aiming to define the starting point, the average value difficulty estimates for all items was selected to be equal to 0.

First of all, the study of test dimensionality was undertaken. This was done by factor analysis of standardized residuals of the examinee response to the item as compared to the statistical expectation, as per model [1]. It was shown that the **test can be acknowledged as significantly unidimensional**. This means that the test measures the only latent characteristic of those under test: the mathematics competence of test participants.

Further on, the fit of experimental test data with the measurement model used was analyzed. Table 2.5 represents statistical data across the test items. The items are lined up as per their order of appearance during the test.

Table 2.5. Statistical indices for test items

Item number	Item type	Difficulty estimate	Measurement error	Correlation coefficient	Fit statistics	
					weighted	*unweighted*
1	M-C-01-1-1	-2.36	0.07	0.24	1.06	1.32
2	M-C-01-1-2	0.01	0.04	0.51	0.94	0.90
3	M-C-01-1-3	2.73	0.06	0.45	0.95	0.87
4	M-C-03-1-1	-3.80	0.13	0.15	1.03	1.09
5	M-C-03-1-2	-0.58	0.05	0.40	1.05	1.10
6	M-C-03-1-3	1.92	0.05	0.36	1.14	1.24
7	M-M-02-1-1	-1.84	0.06	0.38	0.93	0.79
8	M-M-02-1-2	-0.61	0.05	0.46	0.97	0.91
9	M-M-02-1-3	1.36	0.05	0.48	1.00	1.02
10	M-M-03-1-1	-1.68	0.06	0.34	1.02	0.96
11	M-M-03-1-2	-0.15	0.05	0.52	0.93	0.85
12	M-M-03-1-3	1.01	0.04	0.49	1.00	1.00
13	M-M-06-1-1	-0.62	0.05	0.47	0.94	0.91
14	M-M-06-1-2	0.84	0.04	0.55	0.90	0.87
15	M-M-06-1-3	1.01	0.04	0.43	1.07	1.09
16	M-M-11-1-1	-1.87	0.06	0.29	1.04	1.28
17	M-M-11-1-2	0.43	0.04	0.57	0.86	0.80
18	M-M-11-1-3	2.41	0.05	0.45	0.98	0.93
19	M-R-02-1-1	-0.52	0.05	0.54	0.86	0.76
20	M-R-02-1-2	0.65	0.04	0.55	0.91	0.86
21	M-R-02-1-3	1.31	0.05	0.44	1.05	1.14
22	M-R-05-1-1	-4.17	0.16	0.10	1.03	1.41

23	M-R-05-1-2	-0.30	0.05	0.42	1.05	1.04
24	M-R-05-1-3	2.74	0.06	0.42	0.98	1.07
25	M-G-01-1-1	-1.34	0.06	0.29	1.11	1.43
26	M-G-01-1-2	0.84	0.04	0.46	1.03	1.07
27	M-G-01-1-3	2.56	0.05	0.34	1.08	1.53
28	M-D-03-1-1	-1.14	0.05	0.31	1.10	1.46
29	M-D-03-1-2	0.47	0.04	0.48	1.01	1.01
30	M-D-03-1-3	1.54	0.05	0.51	0.96	1.00
31	M-D-05-1-1	-1.20	0.05	0.43	0.93	0.96
32	M-D-05-1-2	-0.76	0.05	0.40	1.06	1.04
33	M-D-05-1-3	1.79	0.05	0.43	1.05	1.28
34	M-D-08-1-1	-2.85	0.09	0.26	0.95	1.53
35	M-D-08-1-2	-0.19	0.05	0.46	1.00	0.97
36	M-D-08-1-3	2.63	0.05	0.31	1.11	1.52
37	M-R-03-1-1	-1.80	0.06	0.34	0.97	1.32
38	M-R-03-1-2	0.01	0.04	0.39	1.12	1.15
39	M-R-03-1-3	1.40	0.05	0.41	1.08	1.18
40	M-C-05-1-1	-1.28	0.06	0.36	1.03	1.10
41	M-C-05-1-2	0.26	0.04	0.49	0.97	1.00
42	M-C-05-1-3	0.84	0.05	0.58	0.86	0.82
43	M-M-08-1-1	-1.96	0.07	0.36	0.94	0.90
44	M-M-08-1-2	0.87	0.04	0.57	0.88	0.83
45	M-M-08-1-3	1.38	0.05	0.57	0.87	0.81

Table review results in a conclusion that there are no items in the test which the participants could not achieve (that is, they were not completed due to lack of time). Columns 3 and 4 of the table display item difficulty assessment values as well as corresponding measurement errors (in logits). Next column presents values of the correlation factor, for correlation between item scores and ability levels of test participants. This index shall be interpreted in the same manner as discriminativity indexes in the classical test theory. Only two items show the correlation factor value to be under 0.2 (items 4 and 22), these are Level 1 items, they have a lower difficulty level.

The last two columns of Table 2.5 show fit statistics values that characterize the fit of the data to the used measurement model. In this report two versions of fit statistics are used: unweighted and weighted. The unweighted fit statistics is more sensitive to extreme, unexpected responses – when, for example, a well-prepared test participant will suddenly provide a wrong answer to an easy item, or, vice versa, when a weak test participant suddenly handles correctly some difficult item. The weighted version of fit statistics helps lower the impact of extreme and unexpected responses. Thus, these two statistics modes belong to different parts of the distribution of test participants' ability levels. The fit statistics values interval of (0,5; 1,5) is cited in the literature to be most productive for conducting measurements, and the interval (0,8; 1,2) is considered the same in the case of large-scale testing with far reaching consequences. The last interval was taken in this study to be acceptable for establishing fit statistics.

Analysis of Table 2.5 leads to a conclusion that values of weighted fit statistics are within norm for all test items. For some items, values of unweighted fit statistics exceed the right critical limit of 1.2 (such items are highlighted pink in the table). This means that some test participants gave surprising answers to these items.

In other words, practically all test items show a satisfactory fit to the measurement model used. Items having unweighted fit statistics values higher than critical will be dealt with below. First of all, however, we shall show an example of a good item which was functioning with an ideal fit to the model: this is item 2. This is a Level 2, average-difficulty item. Figure 2.7 shows the characteristics curve for this item, which is the probability curve for giving a correct response to an item, depending on ability levels of test takers (red line). Little crosses in this figure pinpoint empirical distribution of test takers' responses to this item. They represent average score for this particular item across groups of examinees (the whole sample was divided into ten parts, depending on the test score). Limits of the 95% confidence range are also shown in this figure for points of empirical distribution.

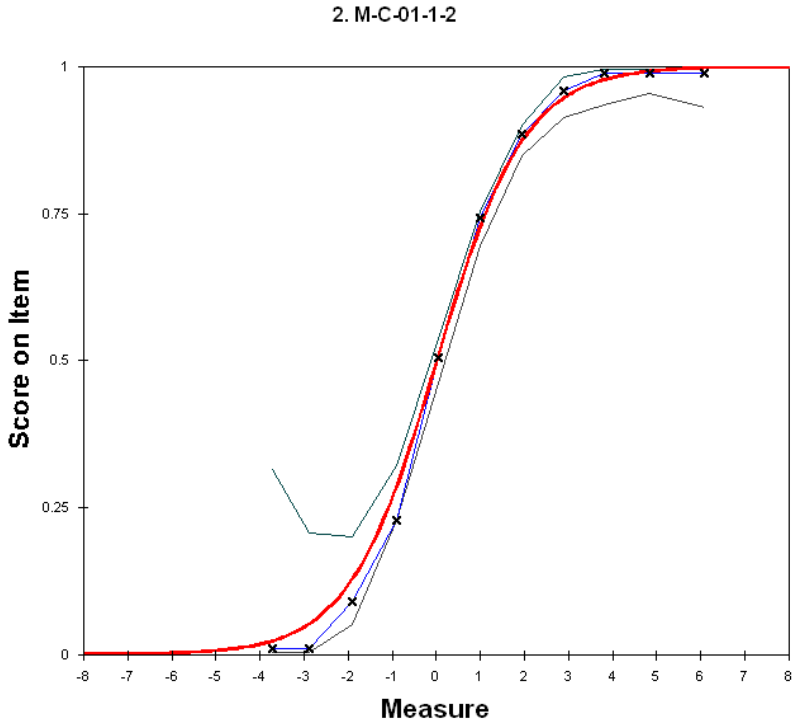


Figure 2.7. Item M-C-01-1-2 characteristics curve

We can see for this item that all empirical distribution points do not get beyond the limits of the confidence range, and that means that the misfit between expected model values and empirical data is not significant or, in other words, that the test takers' responses to this item correlate well with the model used. The same is true for the absolute majority of test items.

To compare, we shall review in more details one of the problematic items – for example, item 36 for which the value of unweighted fit statistics is equal to 1.52 (one of the highest among the rest of the items). The characteristics curve for this item is shown in Figure 2.8. This item is a difficult one and of Level 3. Only 19% of test participants could complete it correctly. And, as the figure shows, there were some test takers who completed this item correctly, although there was a very low probability of that. There are not too many of those, however, so in general the item in a satisfactory fit with the model.

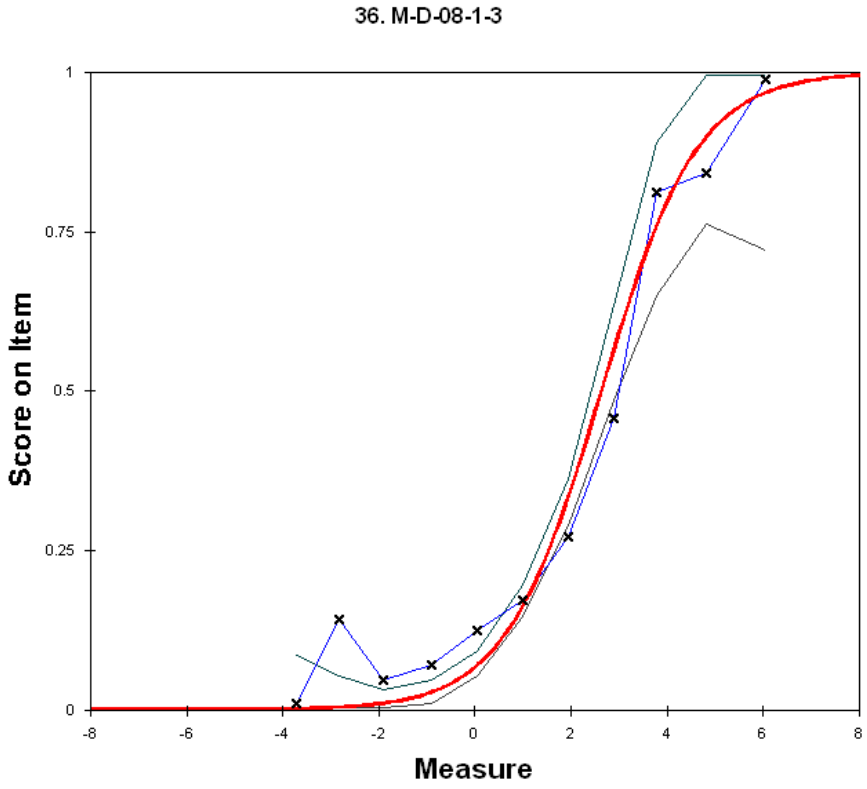


Figure 2.8. Item M-C-01-1-3 characteristics curve

Figure 2.9 shows the variable map, in which the relative distribution of test takers and of items is shown across the general 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. Items with more difficulty and those test takers who are better prepared are located in the upper part of the map, while easier items and insufficiently well prepared test takers are concentrated in the lower part of the map. The map will help analyze the joint distribution of items in relation to this group of test takers and to diagnose and pinpoint test problems.

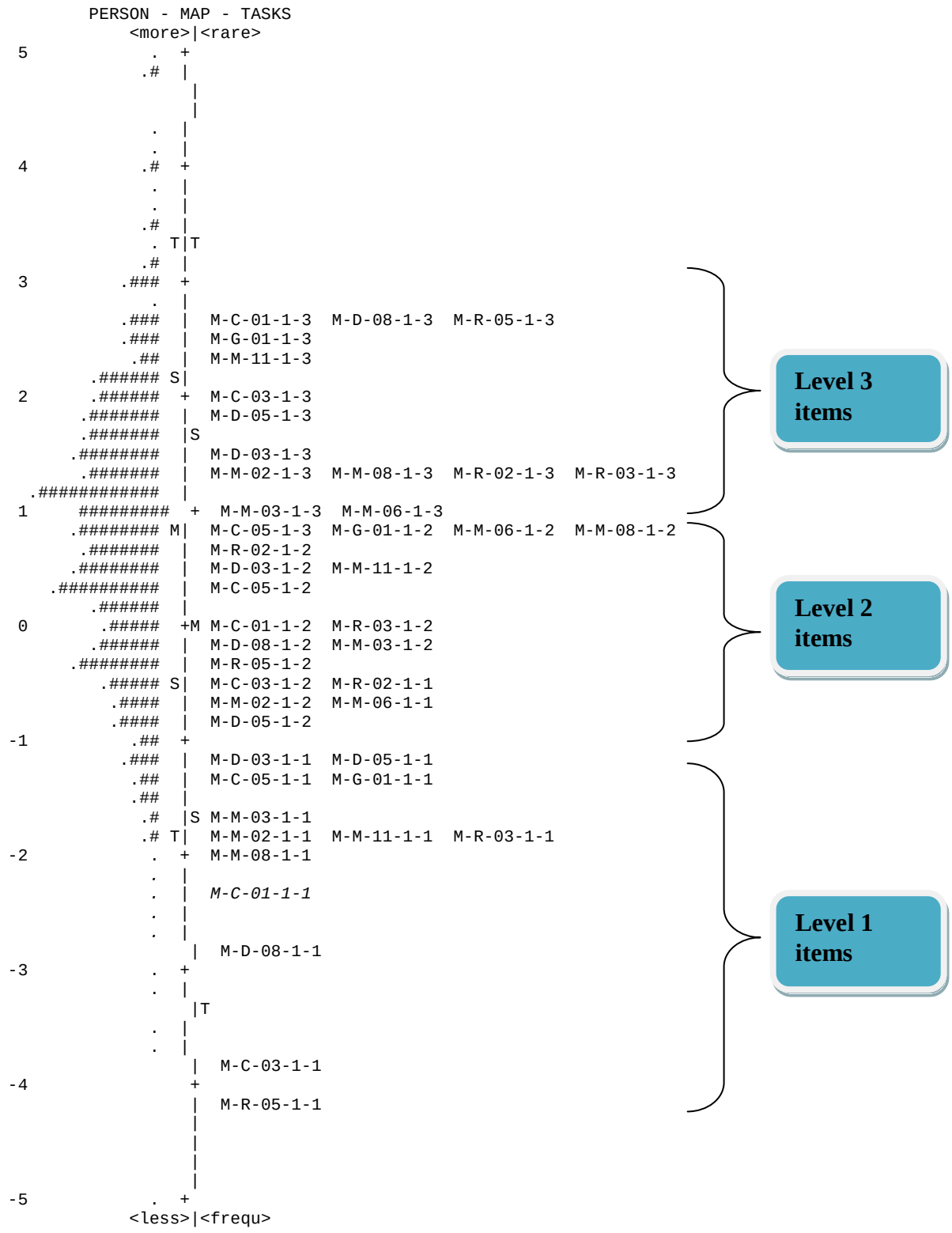


Figure 2.9. Variables map (mathematics, test form 1)

As can be noted, the distribution of examinee measures is close to normal. The examinees sample is located higher up with regard to the items sample, which means that the test proved to be easy for this group of test takers. It is noteworthy how relatively wide is the span of the examinee measures. Some Level 1 items turned out to be very easy while at the upper part of the map there is a lack of difficult Level 3 items for test takers with high ability level.

All level 1 items are in the lower part of the map, further there are level 2 items and, still further on, level 3 items, the most difficult ones in this test.

In general one could recommend making some Level 1 and 3 items more complex (they are in the lower and the upper areas of the map), for this would help center the test with regard to the group of test takers and also to reduce the measurement error for both weak and strong examinees.

A similar analysis was done for **test form 2 in mathematics** (analysis data are not shown in this report); all indexes are within norm, all conclusions are similar.

Lastly **parameter equating** (for examinee measures and item difficulty) was undertaken, and all parameters were transferred to a common scale. To conduct the equating procedure, common items were introduced into both test forms, six items in total (2 common blocks, in test items 7-9 and 22-24), which amounts to 14% of the total number of items. Table 2.6 shows characteristics of common items within the framework of the classical test theory.

Table 2.6. Common item characteristics for two test forms (mathematics)

Test item number	Item index	**Test form 1**		**Test form 2**	
		Difficulty level (p-value)	Discriminativity index	Difficulty level (p-value)	Discriminativit y index
7	M-M-02-1-1	0.88	0.31	0.87	0.32
8	M-M-02-1-2	0.73	0.51	0.76	0.51
9	M-M-02-1-3	0.33	0.53	0.35	0.55
22	M-R-05-1-1	0.97	0.05	0.97	0.06
23	M-R-05-1-3	0.68	0.49	0.66	0.50
24	M-R-05-1-2	0.16	0.34	0.13	0.26

Test results for general tests in two test forms are very close and they satisfy quality criteria as well as demonstrate a good fit with the model. Test item #22 is an exception: as was pointed out earlier, it is an extremely easy test item so that 97% of the students could complete it in both test forms. As a result, this test item has a very low discrimination rate.

To create a common scale the method of separate calibration was chosen with common parameters fixed which makes it possible to display all parameters on the common scale [3]. The scale for test form 1 was chosen to be the common scale.

Figure 2.10 shows characteristics curves for two test forms in mathematics. We can see that these characteristics curves practically coincide, which tells us about the equivalence of both test forms, i.e. that the equating procedure was successful. The equating error can be statistically estimated, although there is no need for it in this particular case.

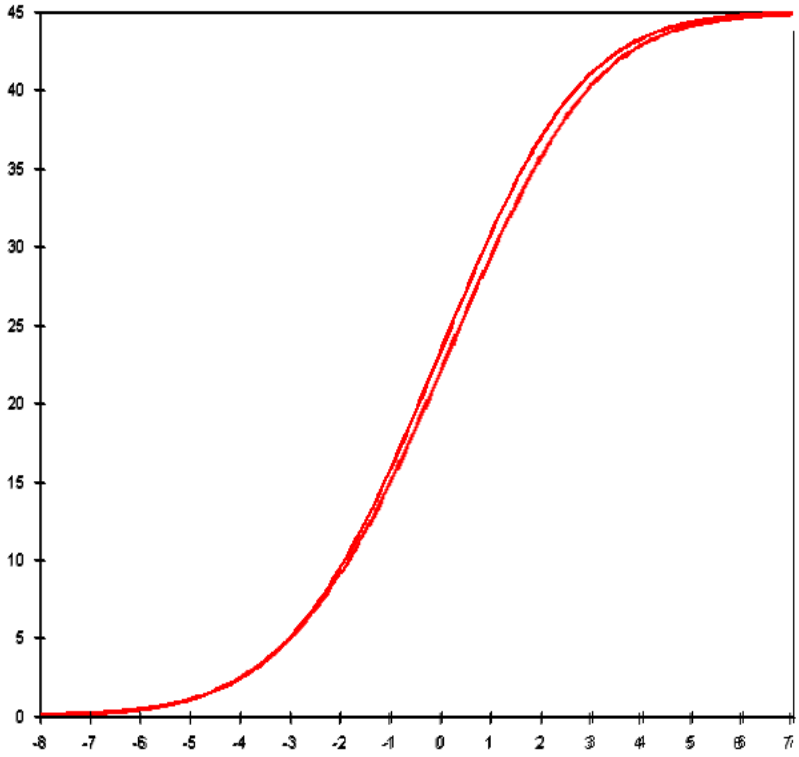


Figure 2.10. Characteristics curves for two test forms after the equating procedure (mathematics)

Consequently, the analysis of the mathematics test within the framework of both classical and modern test theory brings us to the conclusion that this **test represents a quality measuring tool and can be used for the evaluation of mathematics proficiency of testing participants**.

2.2. Results of the pilot testing statistical analysis for test items in language competence

Two test forms in Russian language were used during pilot testing. The test consisted of 45 items grouped in 15 blocks. Table 2.7 shows the summary of statistical indices for two test forms **within the framework of the classical test theory**. Figure 2.11 presents raw score distribution histograms of testing procedure participants.

Table 2.7. Summary of test results (Russian language)

	Test form 1	Test form 2
Number of examinees	2980	2967
Raw score average	23,35	23,13
Standard deviation	8,74	8,70
Skewness	-0,14	-0,19
Kurtosis	-0,68	-0,69
Average difficulty level	0,52	0,51
Average discrimination index	0,46	0,47
Average point-biserial coefficient	0,39	0,39
Reliability index (KR20)	0,90	0,90
Standard error of measurement	2,70	2,70

Test form 1 Test form 2

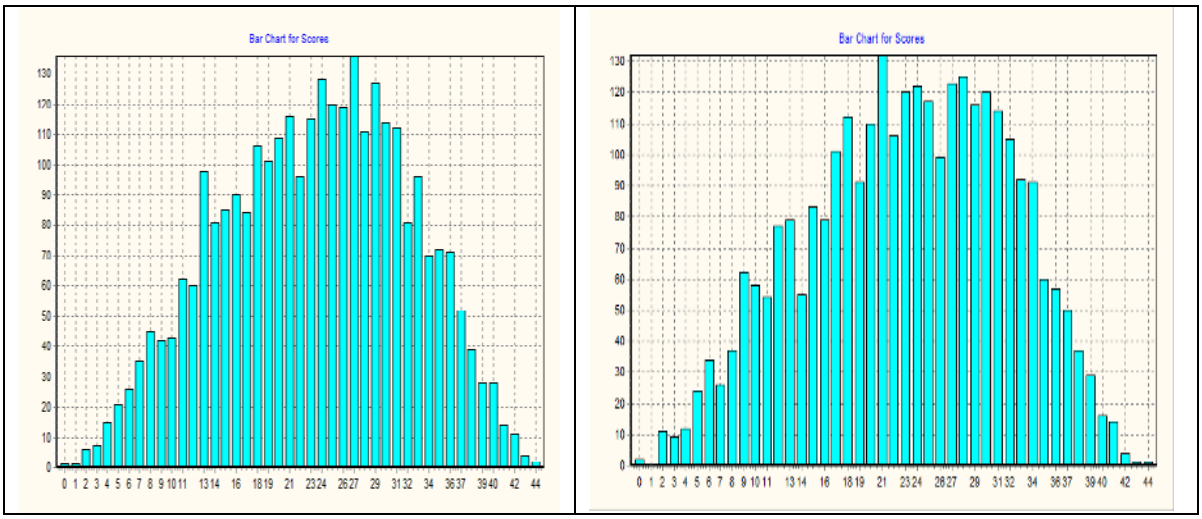


Figure 2.11. Raw score distribution histograms of testing participants

Table analysis shows that statistical indices of the two test forms are very close. Additionally, the hypothesis that these were parallel test forms was examined (using the chi-square method) and it was shown that these test forms can be indeed acknowledged as parallel. Reliability indices are quite high: 0.9 for both test forms. Both average difficulty level and average discrimination indices were close to optimal values.

Table 2.8 shows detailed information on items in **test form 1.** When conducting classical analysis, an item's major characteristics are its difficulty and its discriminative ability. Item difficulty (its difficulty level) is defined by the percentage of the test participants who were able to complete that item correctly. The higher the value of this index, the easier is the item.

Discrimination characterizes the discrimination fineness of an item, that is its capability of making a distinction between test participants with differing levels of preparation. This report makes use of two discriminativity indexes: the classical one (which is the difference of item difficulty values for two groups of test participants: 27% of the best ones, having higher scores, and 27% of the weakest ones, having lowest scores) and the adjusted point-biserial correlation coefficient (correlation factor between an item score and the overall test score, after the results

for this specific item are removed). The value of 0.2 was chosen to be the critical value for discriminativity indexes.

The analysis of Table 2.8 will lead to a conclusion that all items of the test under review function well. For the observed group of test participants, one Level 1 items was very easy (its item difficulty value was 0.96 which means that it was completed successfully by 96% of examinees) and one item was very difficult (its item difficulty value was 0.07 which means that it was completed successfully by only 7% of examinees); in the table, these items are highlighted in green. These same two items are characterized by low discriminativity (in the table, low-discriminativity items are highlighted in pink). Thus, the extreme ease and the extreme difficulty of completing these items manifested itself in their low level of discriminating fineness.

The remaining test items show good statistical indices. Let us point out that difficulty-based hierarchy is pronounced within each block of Level 1, 2, and 3 items. The block of items 14-16 is, however, an exception: in this block, Level 3 item turned out to be somewhat easier than Level 2 item (these items are highlighted blue in the table).

Table 2.8. Statistical indices of items functioning (Russian language, test form 1)

Item number and item index		Difficulty level	Discrimination index	Adjusted point-biserial coefficient
1	L-O-1-01-1-1	0.85	0.33	0.39
2	L-O-1-01-1-2	0.80	0.33	0.32
3	L-O-1-01-1-3	0.36	0.56	0.44
4	L-O-2-01-1-1	0.96	0.10	0.23
5	L-O-2-01-2-2	0.68	0.53	0.43
6	L-O-2-01-1-3	0.36	0.48	0.38
7	L-M-3-01-1-1	0.75	0.39	0.33
8	L-M-3-01-1-2	0.39	0.53	0.41
9	L-M-3-01-1-3	0.28	0.49	0.41
10	L-M-2-01-1-1	0.74	0.35	0.29
11	L-M-2-01-1-2	0.68	0.56	0.48
12	L-M-2-01-1-3	0.24	0.37	0.33
13	L-M-1-02-1-1	0.80	0.37	0.36
14	L-M-1-02-1-2	0.33	0.53	0.42
15	L-M-1-02-1-3	0.43	0.56	0.43
16	L-F-1-02-1-1	0.82	0.36	0.39
17	L-F-1-02-1-2	0.42	0.58	0.45
18	L-F-1-02-1-3	0.32	0.60	0.48
19	L-F-2-02-1-1	0.79	0.46	0.45
20	L-F-2-02-1-2	0.61	0.33	0.25
21	L-F-2-02-1-3	0.35	0.35	0.27
22	L-F-2-01-1-1	0.85	0.31	0.35
23	L-F-2-01-1-2	0.50	0.6	0.46
24	L-F-2-01-1-3	0.28	0.55	0.47
25	L-F-2-03-1-1	0.67	0.59	0.5
26	L-F-2-03-1-2	0.57	0.76	0.58
27	L-F-2-03-1-3	0.21	0.38	0.37

28	L-L-3-01-1-1	0.88	0.28	0.37
29	L-L-3-01-1-2	0.65	0.5	0.4
30	L-L-4-01-1-3	0.29	0.42	0.35
31	L-L-2-01-1-1	0.56	0.58	0.45
32	L-L-2-01-1-2	0.38	0.48	0.35
33	L-L-2-01-1-3	0.23	0.32	0.29
34	L-L-1-01-1-1	0.63	0.45	0.35
35	L- L-1-01-1-2	0.30	0.44	0.36
36	L-L-1-01-1-3	0.07	0.11	0.19
37	L-L-5-01-1-1	0.75	0.46	0.41
38	L-L-5-01-1-2	0.38	0.56	0.43
39	L-L-5-01-1-3	0.20	0.41	0.39
40	L-S-2-01-1-1	0.77	0.47	0.45
41	L-S-2-01-1-2	0.54	0.64	0.5
42	L-S-2-01-1-3	0.26	0.53	0.46
43	L-S-1-01-1-1	0.75	0.49	0.46
44	L-S-1-01-1-2	0.34	0.47	0.37
45	L-S-1-01-1-3	0.33	0.61	0.5

Figures below display graphic representations of the indices from Table 2.8.

Figure 2.12 is a bar chart presenting the distribution of difficulty values (p-values) for the items of the test under review. Item indexes are shown on the horizontal axis in the order of their appearance during the test while p-values are shown on the vertical axis. Level 1 items are shown in blue, Level 2 items in red and Level 3 items in green. This diagram provides a clear picture of the difficulty-based hierarchy within each block. Figure 2.13 also shows the distribution of difficulty values, this time as a curve diagram.

Figures 2.14 and 2.15 show the distribution of the discrimination indexes for the test items. And finally, Figure 2.16 displays a joint distribution of both p-values and discrimination indexes for the test items.

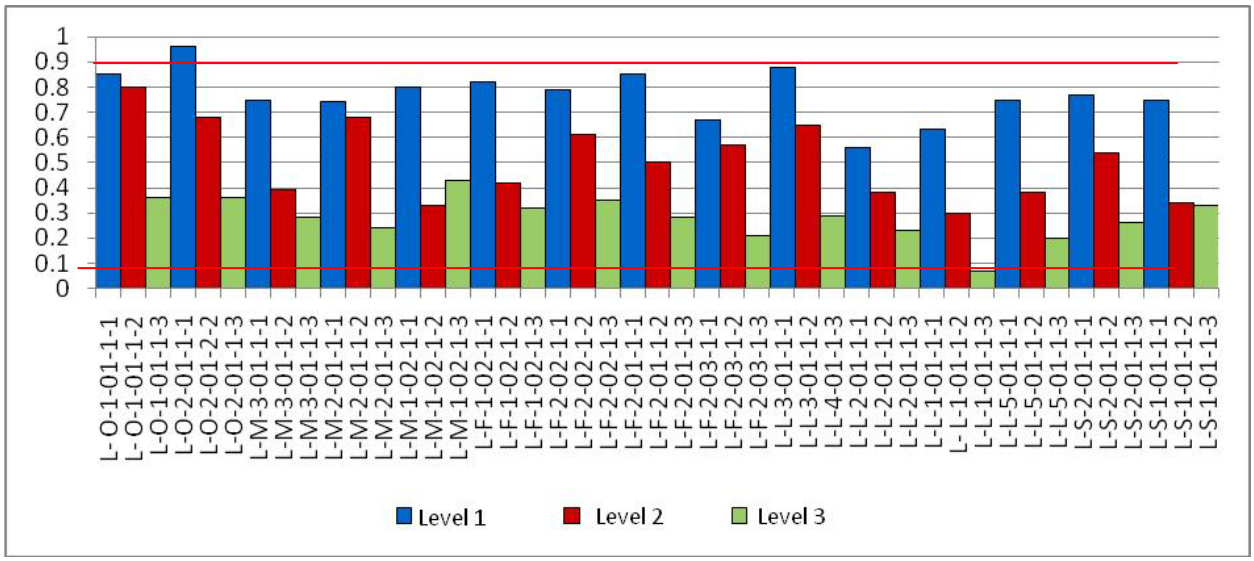


Figure 2.12. Diagram of p-values distribution (Russian language, test form 1)

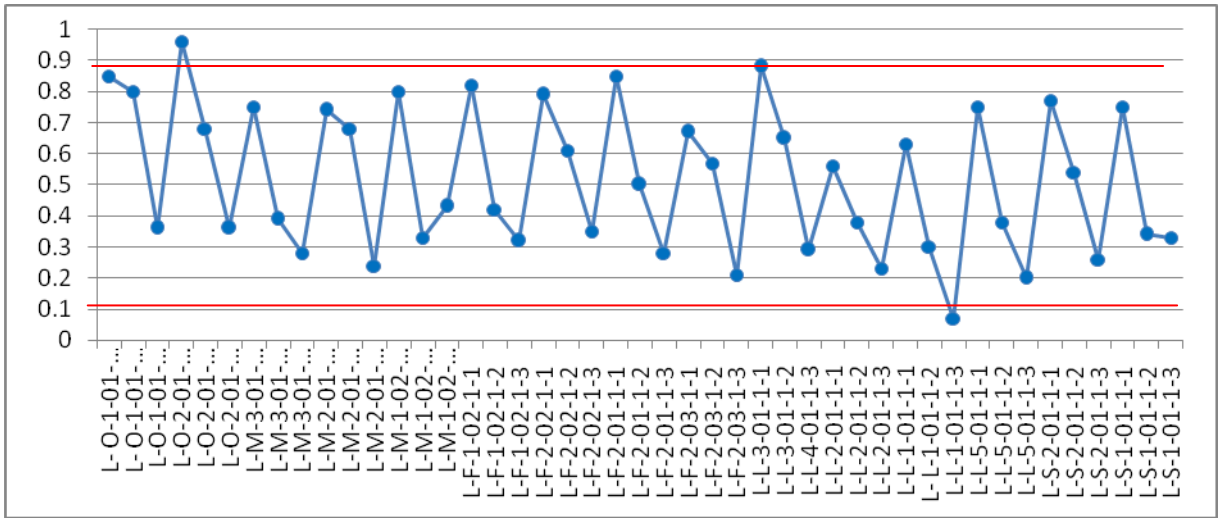


Figure 2.13. Difficulty values distribution diagram (Russian language, test form 1)

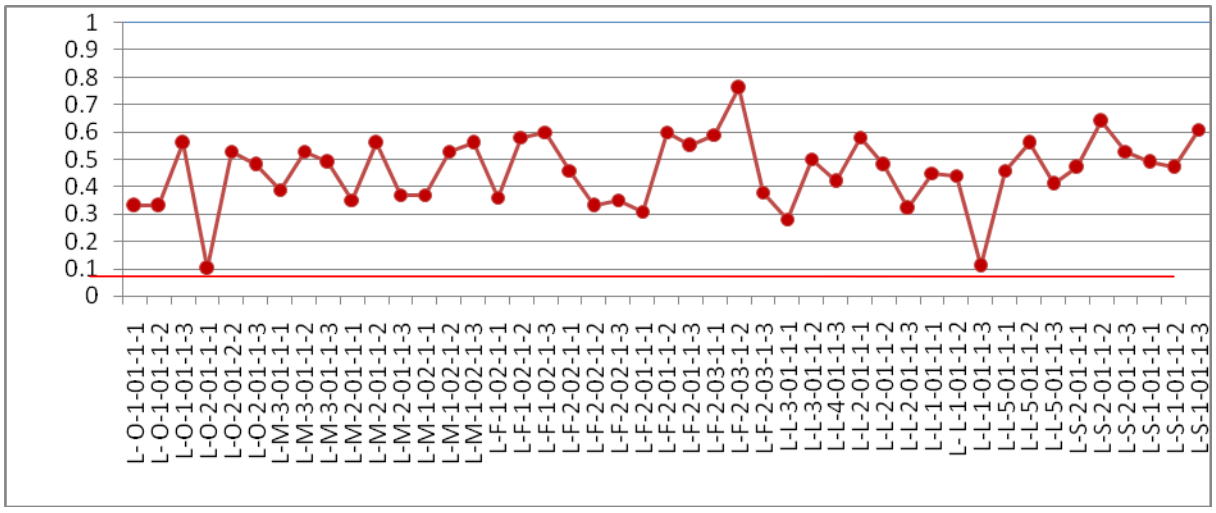


Figure 2.14. Diagram of discrimination indexes distribution for test items (Russian language, test form 1)

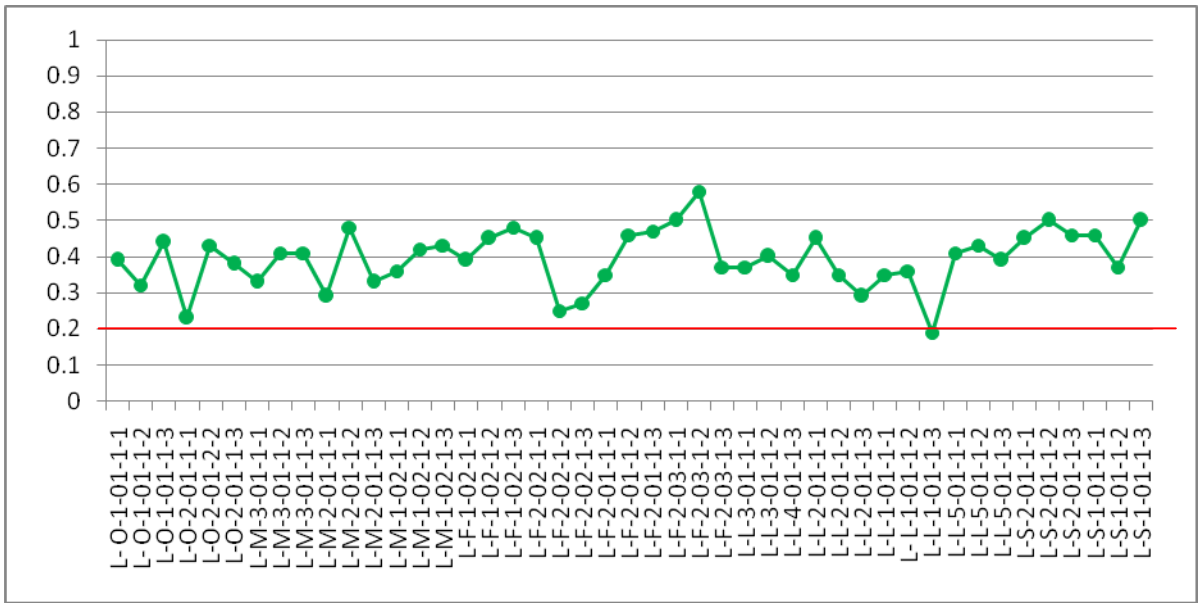


Figure 2.15. Point-biserial coefficient distribution diagram (Russian language, test form 1)

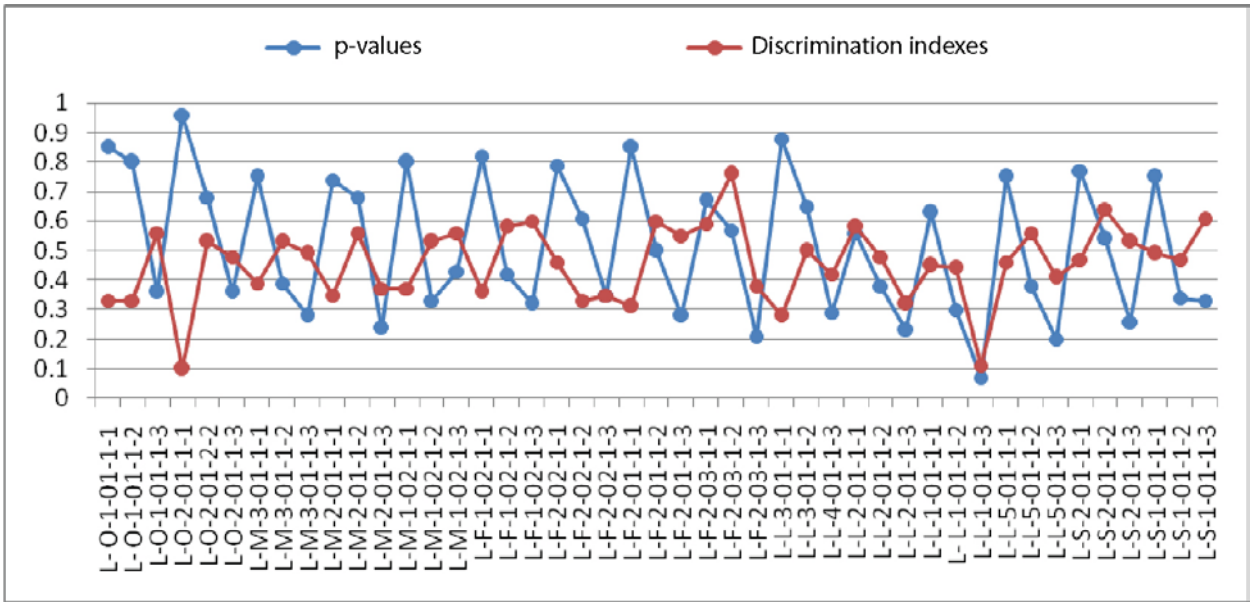


Figure 2.16. Joint representation of both p-values and discrimination indexes distribution (Russian language, test form 1)

The above figures show well that practically all items have good indices, i.e. optimal level of difficulty and high discrimination. Discrimination index values for only two items are below normal: they were, as was noted above, the easiest and the most difficult items.

Additionally, an **analysis of distractor functioning** was undertaken for closed-form items (with multiple-choice options). In this test, there were 14 such items, and in some of them more than one correct answer was possible.

To analyze distractors, firstly, the distribution of examinee responses was reviewed over all response alternatives for closed-type items, the goal being the determination of non-functioning distractors. Secondly, correlation factors (point-biserial correlation coefficients) were calculated for distractors and the overall score of the test. A distractor functions properly, if high ability examinees do not choose it as a correct answer. In this case, the correlation factor will have a negative value (it is desirable that it would be less than -0.2). And vice versa, the correlation factor for a correct response option must have a positive value (the correlation factor for the correct answer option coincides with the discriminativity index for items that were dealt with earlier).

No serious problems were found for distractor functioning of the closed-type items with the test under consideration.

Tables 2.8 and 2.9 show examples of distractor analysis for two items. The correct answer option is marked by a star and is highlighted in colour. For item L-M-3-01-1-1 all indices are within norm: 75% of test participants completed the item correctly. The high positive value of the correlation factor means that these examinees were good students with high test points. The rest of the students who could not complete the item are distributed over three distractors (3% of examinees passed up this item). All distractors were, in fact, functioning: each was chosen by at least 5% of the test participants. Negative values of the correlation factors indicate that these were weak students with low test points.

Table 2.8. Distractor analysis for item L-M-3-01-1-1

Item L-M-3-01-1-1				
Answer options	**А**	**Б**	**В***	**Г**
Examinee choice distribution	8%	7%	75%	7%
Correlation coefficient	-0,15	-0,15	0,33	-0,19

Table 2.9. Distractor analysis for item L-F-2-02-1-2

Item L-F-2-02-1-2				
Answer options	**А**	**Б**	**В***	**Г**
Examinee choice distribution	17%	8%	61%	12%
Correlation coefficient	-0,09	-0,29	0,25	0,01

Item L-F-2-02-1-2 was more difficult: 61% of the examinees completed it and 2% passed up on it. The high value of the positive correlation factor for the correct answer is a witness to the fact that these examinees were good students with high test points. The students who could not complete the item were distributed over three distractors. All distractors were, in fact, functioning: each was chosen by at least 5% of test participants. Distractors А and Г were more preferable for test takers, and its correlation with the overall grade is close to zero, which means that these distractors were attractive for both weak examinees and for those with a good level of preparation. Such a situation is characteristic for items having a certain catch that good students may not notice.

The analysis of test form 1 items (in Russian language) has thus shown within the framework of the classical test theory that all items exhibit good characteristics.

A more thorough analysis of the test was conducted **within the framework of the modern test theory IRT (Item Response Theory)**. Please find below main results of this analysis.

The unidimensional dichotomous Rasch model was used as testing model. According to this model, each test item is characterized by one parameter, which is difficulty, and each test participant is also characterized by one parameter, which is ability level. Estimates of all parameters (both of test participants and test items) are located on the common metric scale (the unit of measurement on this scale is called ‘logit’), and they are provided with evaluation precision characteristics. The starting point of this scale is not defined and can be chosen arbitrarily. Aiming to define the reference point, the average value difficulty estimates for all items was selected to be equal to 0.

First of all, the study of test dimension was undertaken. This was done by factor analysis of standardised residuals of the examinee response to the item as compared to the statistical expectation, in accordance with the model [1]. It was shown that the **test can be acknowledged as significantly unidimensional**. This means that the test measures the only latent characteristic of those under test: language competence of test participants.

Further on, the fit of empirical test data with the measurement model used was analyzed. Table 2.10 represents statistical data across all test items. The items are lined up as per their order of appearance during the test.

Table 2.10. Statistical indices for test items

Item number	Item type	Difficulty estimate	Measurement error	Correlation coefficient	Fit statistics	
					weighted	*unweighted*
1	L-O-1-01-1-1	-1.93	0.06	0.42	0.94	0.82
2	L-O-1-01-1-2	-1.57	0.05	0.37	1.05	1.07
3	L-O-1-01-1-3	0.97	0.04	0.47	0.97	0.94
4	L-O-2-01-1-1	-3.54	0.10	0.28	0.95	0.76
5	L-O-2-01-2-2	-0.72	0.05	0.47	0.97	0.96
6	L-O-2-01-1-3	0.98	0.04	0.42	1.03	1.05
7	L-M-3-01-1-1	-1.19	0.05	0.37	1.09	1.07
8	L-M-3-01-1-2	0.87	0.04	0.46	0.99	0.97
9	L-M-3-01-1-3	1.31	0.05	0.44	0.98	0.97
10	L-M-2-01-1-1	-1.05	0.05	0.33	1.13	1.22
11	L-M-2-01-1-2	-0.71	0.05	0.51	0.91	0.87
12	L-M-2-01-1-3	1.70	0.05	0.37	1.05	1.07
13	L-M-1-02-1-1	-1.66	0.05	0.40	0.99	0.96
14	L-M-1-02-1-2	1.21	0.04	0.46	0.96	0.92
15	L-M-1-02-1-3	0.40	0.04	0.45	1.01	0.99
16	L-F-1-02-1-1	-1.70	0.06	0.42	0.96	0.91
17	L-F-1-02-1-2	0.72	0.04	0.49	0.95	0.94
18	L-F-1-02-1-3	1.17	0.04	0.53	0.88	0.82
19	L-F-2-02-1-1	-1.44	0.05	0.48	0.92	0.79
20	L-F-2-02-1-2	-0.34	0.04	0.31	1.19	1.38
21	L-F-2-02-1-3	0.70	0.05	0.27	1.24	1.33
22	L-F-2-01-1-1	-2.11	0.06	0.38	0.97	0.92
23	L-F-2-01-1-2	0.20	0.04	0.49	0.96	0.93
24	L-F-2-01-1-3	1.49	0.05	0.51	0.88	0.84
25	L-F-2-03-1-1	-0.95	0.05	0.50	0.91	0.81
26	L-F-2-03-1-2	-0.07	0.04	0.60	0.82	0.76
27	L-F-2-03-1-3	1.85	0.05	0.40	0.98	1.00
28	L-L-3-01-1-1	-2.59	0.07	0.36	0.93	0.85
29	L-L-3-01-1-2	-0.59	0.04	0.42	1.04	1.06
30	L-L-4-01-1-3	1.26	0.05	0.37	1.08	1.25
31	L-L-2-01-1-1	-0.11	0.04	0.47	0.99	0.98
32	L-L-2-01-1-2	0.84	0.04	0.37	1.10	1.16
33	L-L-2-01-1-3	1.54	0.05	0.32	1.11	1.23
34	L-L-1-01-1-1	-0.89	0.05	0.33	1.14	1.20
35	L-L-1-01-1-2	1.31	0.05	0.40	1.04	1.02
36	L-L-1-01-1-3	3.29	0.08	0.22	1.04	1.26
37	L-L-5-01-1-1	-1.40	0.05	0.39	1.03	1.01
38	L-L-5-01-1-2	0.85	0.04	0.46	0.98	0.97
39	L-L-5-01-1-3	1.81	0.05	0.42	0.97	0.89
40	L-S-2-01-1-1	-1.56	0.05	0.44	0.95	0.83
41	L-S-2-01-1-2	-0.11	0.04	0.49	0.96	0.94
42	L-S-2-01-1-3	1.32	0.05	0.48	0.93	0.90
43	L-S-1-01-1-1	-1.64	0.06	0.39	0.98	0.97

44	L-S-1-01-1-2	1.05	0.04	0.38	1.07	1.09
45	L-S-1-01-1-3	0.99	0.05	0.51	0.92	0.87

Table review results in a conclusion that there are no items in the test which the participants could not reach (in other words, that they were not completed due to lack of time). Columns 3 and 4 of the table display item difficulty estimation values as well as corresponding measurement errors (in logits). Next column presents values of the correlation factor, for correlation between item scores and ability levels of test participants. This index shall be interpreted in the same manner as discriminativity indexes in the classical test theory. Only two items show the correlation factor value to be under 0.2 (items 4 and 22), these are Level 1 items, they have a lower difficulty level.

The last two columns of Table 2.10 show fit statistics values that characterize the fit of the data to the used measurement model. Two versions of fit statistics are used in this report: unweighted and weighted. The unweighted fit statistics is more sensitive to extreme, unexpected responses when, for example, a high-ability test participant will suddenly provide a wrong answer to an easy item, or, vice versa, when a low-ability test participant suddenly handles correctly some difficult item. The weighted version of fit statistics helps lower the impact of extreme and unexpected responses. Thus, these two statistics modes belong to different parts of the distribution of test participants' ability levels. The fit statistics values interval of (0,5; 1,5) is cited in the literature to be most productive for conducting measurements, and the interval (0,8; 1,2) is considered the same in the case of large-scale testing with far reaching consequences. The latter interval was taken in this study to be acceptable for establishing fit statistics.

Analysis of Table 2.10 leads to a conclusion that values of weighted fit statistics are within norm for all test items except one. For some items, values of unweighted fit statistics exceed the right-hand critical limit of 1.2 (such items are highlighted pink in the table). This means that some test participants gave unexpected answers to these items.

In other words, practically all test items show a satisfactory fit to the measurement model used. Items having unweighted fit statistics values higher than critical will be dealt with below. First of all, however, we shall show an example of a good item which was functioning with an ideal fit to the model: this is item 3. This is a Level 3 item, its difficulty is higher than average. Figure 2.17 shows the characteristics curve for this item, which is the probability curve for giving a correct response to an item, depending on ability levels of test takers (red line). Little crosses in this figure pinpoint empirical distribution of test takers' responses to this item. They represent the average score for this particular item across groups of examinees (the whole sample was divided into ten parts, depending on the test score). Limits of the 95% confidence range are also shown in this figure for points of empirical distribution.

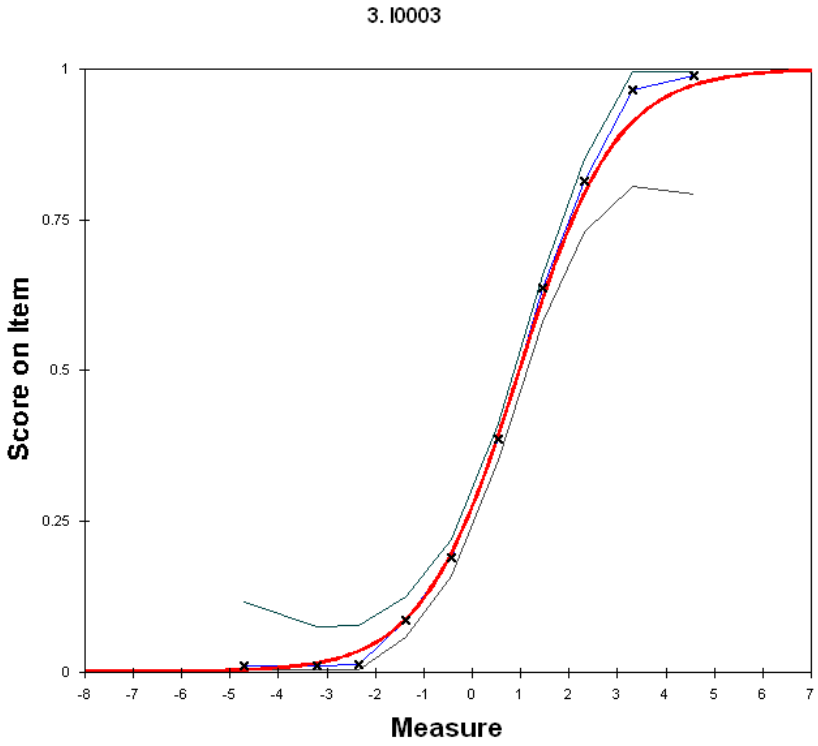


Figure 2.17. Item L-O-1-01-1-3characteristics curve

We can see for this item that all empirical distribution points do not get beyond the limits of the confidence range, and that means that the misfit between expected model values and empirical data is not significant or, in other words, that the test takers' responses to this item correlate well with the model used. The same is true for the absolute majority of test items.

To compare, we shall review in more details one of the problematic items – for example, item 10 for which the value of unweighted fit statistics is equal to 1.22. The characteristics curve for this item is shown in Figure 2.18. This is Level 1 item and quite an easy one: 74% of test participants could complete it correctly. And, as the figure shows, there were some weak test takers who completed this item correctly, although there was a very low probability of that. This was a multiple-choice item (4 answers were offered, while only one of them was correct), so we could assume that this may have been the result of accidental guessing. Aiming to verify this assumption, let us analyze distractors for this item in Table 2.11.

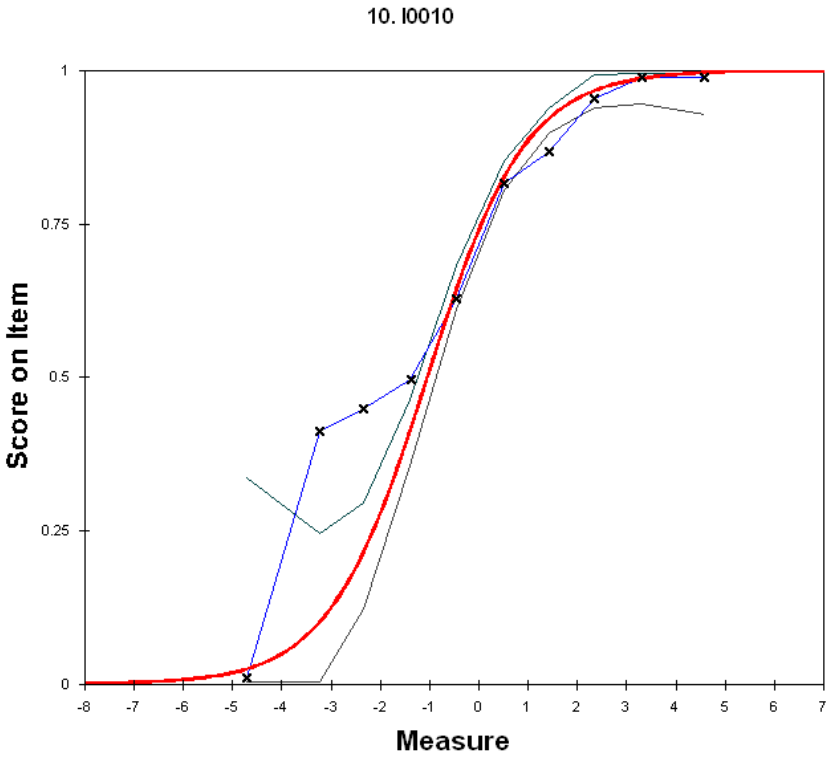


Figure 2.18. Item L-M-2-01-1-1 characteristics curve

Table 2.11. Distractor analysis for item L-M-2-01-1-1

Item L-M-2-01-1-1				
Answer options	**А**	**Б***	**В**	**Г**
Examinee choice distribution	10%	74%	2%	12%
Correlation coefficient	-0,02	0,29	-0,1	-0,28

We can see that distractor В is not functioning: only 2% of test takers chose it. The relatively high positive correlation of the correct answer means that well-prepared students made this choice. Those who did not know the correct answer would guess between two options, А and Г. Hence the result: the probability of guessing the answer correctly is relatively high for weak students, and this is reflected in Figure 2.19 and the statistics of this item.

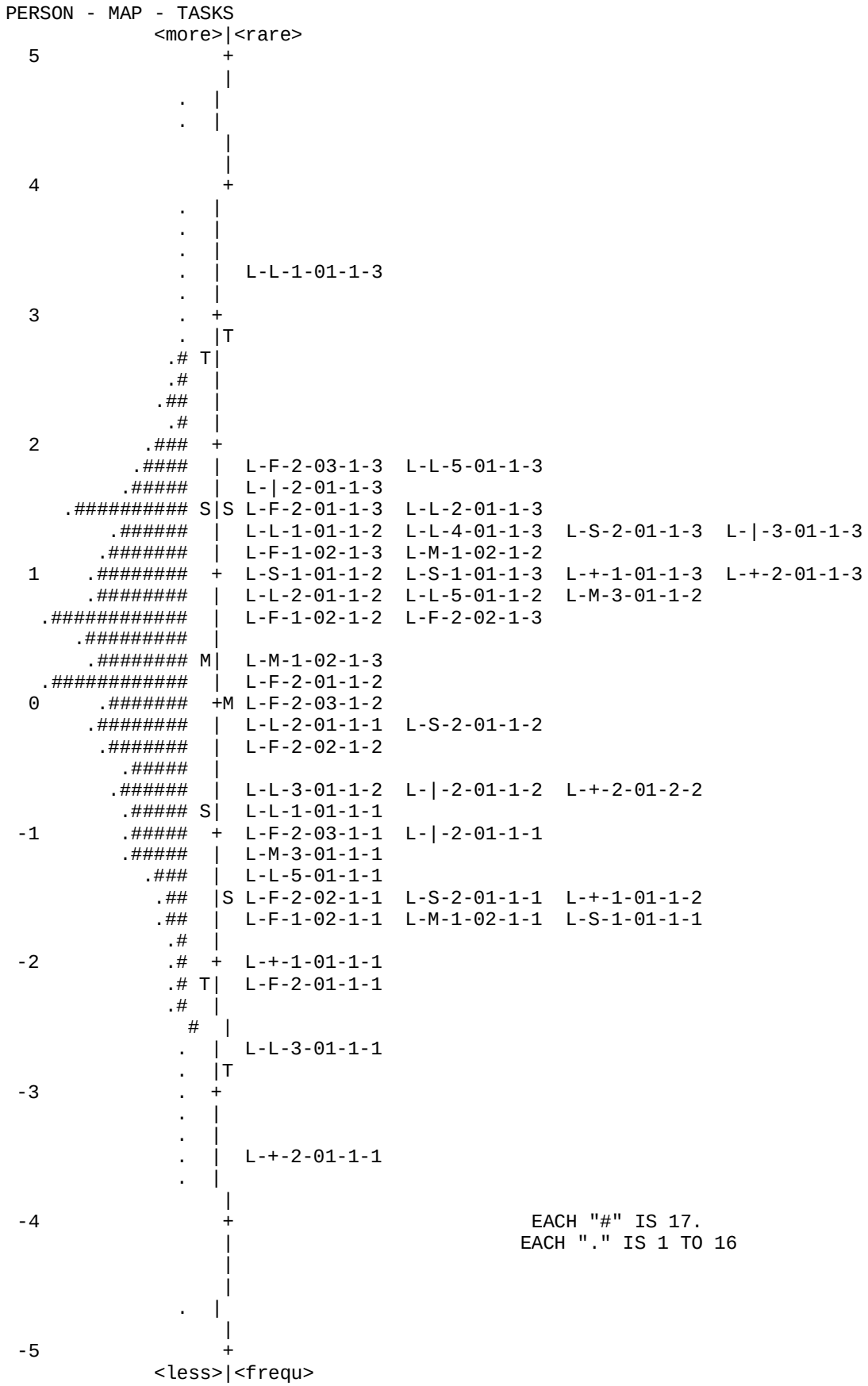


Figure 2.19. Variable map (Russian language, test form 1)

Figure 2.19 shows the variable map, in which the relative distribution of test takers and of items is shown across the general 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. Items with more difficulty and those test takers who are better prepared are located in the upper part of the map, while easier items and insufficiently well prepared test takers are concentrated in the lower part of the map. The map will help analyze the joint distribution of items in relation to this group of test takers and to diagnose and pinpoint test problems.

As can be noted, the distribution of examinee measures is close to normal. The examinees sample is located somewhat higher up with regard to the items sample, which means that the test proved to be easy for this group of test takers. It is noteworthy how relatively wide is the span of the examinee measures. Some Level 1 items turned out to be very easy while at the upper part of the map there is a lack of difficult Level 3 items for test takers with high ability level.

All level 1 items are in the lower part of the map, further there are level 2 items and, still further on, level 3 items, the most difficult ones in this test.

In general one could recommend making some Level 1 and 3 items more complex (they are in the lower and the upper areas of the map), for this would help center the test with regard to the group of test takers and also to reduce the measurement error for both strong examinees.

A similar analysis was done for **test form 2 in Russian language** (analysis data are not shown in this report); all indexes are within norm, all conclusions are similar.

Lastly **parameter equating** (for examinee measures and item difficulty) was undertaken, and all parameters were transferred to a common scale. To conduct the equating procedure, common items were introduced into both test forms, six items in total (2 common blocks, in test items 4-6 and 34-36), which amounts to 14% of the total number of items. Table 2.12 shows characteristics of common items within the framework of the classical test theory.

Table 2.12. Common item characteristics for two test forms (Russian language)

Test item number	Item index	**Test form 1**		**Test form 2**	
		Difficulty level (p-value)	Discriminativity index	Difficulty level (p-value)	Discriminativity index
7	L-O-2-01-1-1	0.96	0.10	0.96	0.10
8	L-O-2-01-2-2	0.68	0.53	0.70	0.52
9	L-O-2-01-1-3	0.36	0.48	0.33	0.50
22	L-L-1-01-1-1	0.63	0.45	0.57	0.51
23	L-L-1-01-1-2	0.30	0.44	0.29	0.43
24	L-L-1-01-1-3	0.07	0.11	0.07	0.12

Test results for general tests in two test forms are very close and they satisfy quality criteria as well as demonstrate a good fit with the model. Test items #7 and #22 are an exception: as was pointed out earlier, they are, respectively, the easiest and the most difficult test items. Psychometric characteristics of these items, however, (specifically, the fit with model) are satisfactory within the IRT framework.

To create a common scale the method of separate calibration was chosen with common parameters fixed which makes it possible to display all parameters on the common scale [3]. The scale for test form 1 was chosen to be the common scale.

Figure 2.20 shows characteristics curves for two test forms in Russian language . We can see that these characteristics curves practically coincide, which tells us about the equivalence of both test forms, i.e. that the equating procedure was successful. The equating error can be statistically estimated, although there is no need for it in this particular case.

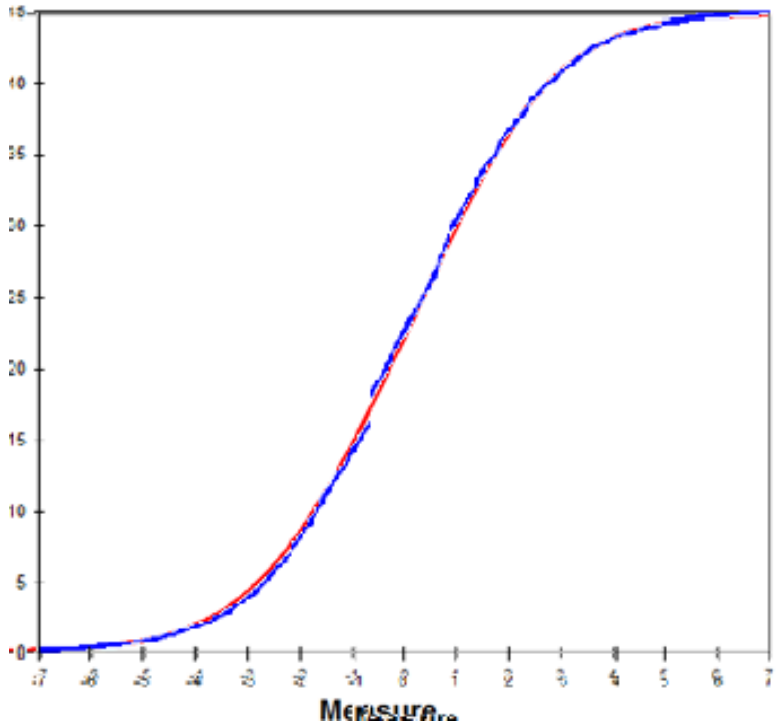


Figure 2.20. Characterstics curves for two test forms after the equating procedure (Russian language)

Consequently, the analysis of the Russian language test within the framework of both classical and modern test theory brings us to the conclusion that this **test represents a quality measuring tool and can be used for the evaluation of language prociency of testing participants**.

Literature

1. *Smith Jr. E. V.* Detecting and Evaluating the Impact of Multidimensionality using Item Fit Statistics and Principal Component Analysis of Residuals // Journal of Applied Measurement. – 2002,Vol. 3, №2, p.205-231
2. *Smith R.M.* Applications of Rasch Measurement. – Chicago: Mesa Press, 1992
3. Карданова Е.Ю., Нейман Ю.М. Проблема выравнивания в современной теории тестирования. *Вопросы тестирования в образовании*, 2003, № 8, с. 21-40.

3. SAM Validization

The development of an educational test incorporates validization, that is evaluating its adequacy to the subject and to the goal of the testing. As a rule, there are three main validity characteristics: content validity, construct (conceptual) validity and criterion validity (current and predictive). Each of these characteristics tackles its own area of issues. Thus, content validity characterizes the degree of adequacy that test materials have with regard to the competence being tested. Construct validity establishes the measure for the coordination between testing results and the psychological model of the ability being tested. Lastly, criterion validity pinpoints the measure of consistency between testing results and known objective criteria.

It is worth considering the validization of benchmarks that were established for the interpretation of SAM results. This study is described in the chapter called Estimation of Examinees.

3.1. Content validity

Obviously, no test can contain the content of the whole curriculum. Content validity for a test will only characterize a degree of optimum selection of curriculum content. The study of a test's content validity aims to examine the quality of the items sample that were included in the test, in terms of their being adequate with regard to the competence being tested. It is not possible to evaluate validity using some quantitative measure. The main method for the study of a test's content validity is doing expert evaluation of the content of each item as well as of the content of the test as a whole, which should be conducted by specialists, that is to get under way the peer review of the content.

The principles of selecting SAM test content were outlined and justified in test specifications, and they were presented in detail in [1].

The test in mathematics includes main content areas of mathematics as they are represented in the primary school curriculum.

In the process of content selection Russian Federation State Standard of Primary General Education (Russian Ministry of Education and Science order "Concerning the Approval and Implementation of the Russian Federation State Standard for Primary General Education" of October 06, 2009, #373) was used.

The subject scope of the test was divided into five sections: Numbers and Calculations, Measurement of Values, Regularities, Dependences, Elements of Geometry. The content basis of the test can be represented as a matrix (Table 3.1), which includes, as follows:

- Curriculum sections (5 sections);
- Mathematics tools (concepts, representations, principles, rules, formulae, schemes, etc.), mastering which is at the basis of mathematical competence.

Table 3.1. Mathematics test content

Content sections	Orientation points in mathematics
Numbers and Calculations	Sequence of natural numbers Number scale Positional principle Arithmetic operations properties Order of operations
Measurement of Values	Relation between number, value and unit Relation between the whole and its parts Formula for rectangle area
Regularities	"Induction step" Recurrence (periodicity)
Dependences	Relations between uniform quantities (equality, inequality, multiplicity, difference, "whole and parts") Direct proportionality between values Derived quantities: velocity, labor productivity, etc. Relationship between units
Elements of Geometry	Form and other qualities of figures (main types of geometrical figures) Spatial relationship between geometric figures – Symmetry

The structural unit of the test is a block of three units (Level 1, Level 2, and Level 3) corresponding to one curriculum section. Table 3.2 shows an approximate unit proportion for various sections.

Table 3.2. Tentative section representation in the test book

Content areas	Number of blocks	Number of items
Numbers and Calculations	4	12
Measurement of Values	5	15
Regularities	2	6
Dependences between Values	2	6
Elements of Geometry	2	6
TOTAL	15	45

The test in Russian language includes main content areas of the Russian grammar as they are represented in the primary school curriculum.

The following regulatory documents were used in the process of content selection:

- Russian Federation State Standard of Primary General Education (Russian Ministry of Education and Science order "Concerning the Approval and Implementation of the Russian Federation State Standard for Primary General Education" of October 06, 2009, #373)
- Tentative basic primary education curriculum recommended for use at educational institutions by decision of the Coordination Council with the General Education Department of the Russian Ministry of Education and Science, dealing with issues of implementing the state

educational standard (see Coordination Council minutes #1 of July 27-28, 2010).

The subject of testing corresponds to the Russian Federation State Educational Standard (see Results of Program Acquisition: FSES, pp. 10-11) in the following part:

- Mastering initial understanding regarding the norms of the Russian literary language (orthoepical, lexical and grammatical norms) as well as the rules of polite speech; developing a sense of direction in the goals, tasks, devices and conditions of communication; making the right selection for successful achievement of communicative tasks by turning to adequate linguistic means;
- Mastering educational actions with linguistic units and developing an ability of using the working knowledge for achieving cognitive, practical and communicative tasks.

The content area of the Russian language test was divided into two sections – "Word, its meaning and its spelling" and "Utterance and its appearance in written language"; this corresponds to the two aspects of speech: nominative and communicative.

The first content area assumes that word is to be acquired first and foremost from the viewpoint of the form and meaning correlation and that means that linguistic mechanisms of forming and expressing ideas and notions should be used. The test items in this section of the test are built upon the material which corresponds to such parts of the "Russian language" school subject as "Phonetics", "Word composition", "Morphology" and "Vocabulary".

The second content area is guided by the impact of sequential deployment of various meanings in language and in speech. Thus this area embraces types of syntagmatic links between words, phrases and parts of sentence, that is it is connected to the acquisition of syntactic tools of language. The test items in this section of the test are built upon the material of the section "Syntax and Punctuation".

The content basis of the test in Russian language can be represented as a matrix (Table 3.3), which includes, as follows:

- Curriculum sections;
- Language tools (concepts, representations, rules, diagrams, etc.), mastering which is at the basis of language competence.

Table 3.3. Russian language test content

Content areas and sections	**Orientation points in the language material**
Word, its meaning and its spelling	
Phonetic and written form of words (phonetics, graphics and orthography)	– "Sound vs. letter" relation – Principles of Russian writing – Orthographic rules – Principle of syllabic division
Composition of a word (morphemics and word formation)	– "Form vs. meaning" relation – Derivation relation

Word meaning (vocabulary)	– Semantic relations (synonymic, antonymic, genus-species relations) – Polysemy – Transfer of meaning models
Word forms (morphology)	– Allotment principles for parts of speech – Paradygmatic relations (types of inflexion: declension, conjugation) – Grammatical categories (gender, number, case, person, tense)
Utterance and its appearance in written language	
Sentence (syntax and punctuation)	– Syntactic relations: compounding, hypotaxis – Grammatic basis of a sentence – Principles of Russian punctuation

The structural unit of the test is a block of three units (Level 1, Level 2, and Level 3) corresponding to one curriculum section. Table 3.4 shows an approximate unit proportion for various sections and levels.

Table 3.4. Tentative section representation in the test book

Content areas	**Number of blocks**	**Number of items**
Phonetics, graphics and ortghography	2	6
Morphemics and word formation	3	9
Morphology	4	12
Vocabulary	4	12
Syntax and punctuation	2	6
TOTAL	15	45

All test items in Mathematics and Russian Language underwent internal expert evaluation: correctness of expression was verified as well as whether test form was adequate to the item content.

Development and pilot testing of test items included consultations with primary school teachers and subject teachers with regard to test content.

At the final stage expert evaluation was conducted for the content of both tests which was made by external experts who were subject teachers. Expert opinions are presented in the appendixes.

3.2. Construct validity

According to Anastasi, “construct validity of a test shows to what degree its results could be seen as a measure of some theoretical construct” [2]. To substantiate construct validity various hypotheses were formulated regarding relationships between testing results and the psychological model of the competence being tested.

In the SAM model the key construct to be verified is a three-level taxonomy of the syllabus acquisition which is incorporated into age-related context. This construct is linked, as a minimum, to two hypothetical propositions which require empirical verification:

1. The items of three levels related to the same block and meeting the theoretically-grounded criteria of three levels must reveal a corresponding difficulty-based hierarchy.

2. According Elconin's age periodization in the primary school the syllabus is expected to be acquired reflectively. In other words, this is meant to be done through comprehension. Towards the end of the primary school the third level is still rudimentary. Acquiring this syllabus functionally (at level 3) is expected to happen in middle school.

An additional approach to construct validation calls for a comparison between item difficulty predicted by test developers and the empirical difficulty of these same items.

Below study results are shown for the substantiation of construct validity.

Verification of the first hypothesis

Pilot testing of SAM tests gave a positive answer to the first point. The p-values of different level items of the mathematics test are featured in Figure 3.1. (Blue columns present level 1 items, red columns – level 2 items and green columns – level 3 items). The items of each block demonstrate a correct difficulty-based hierarchy: in each block, p-values decrease from the first level to the third one.

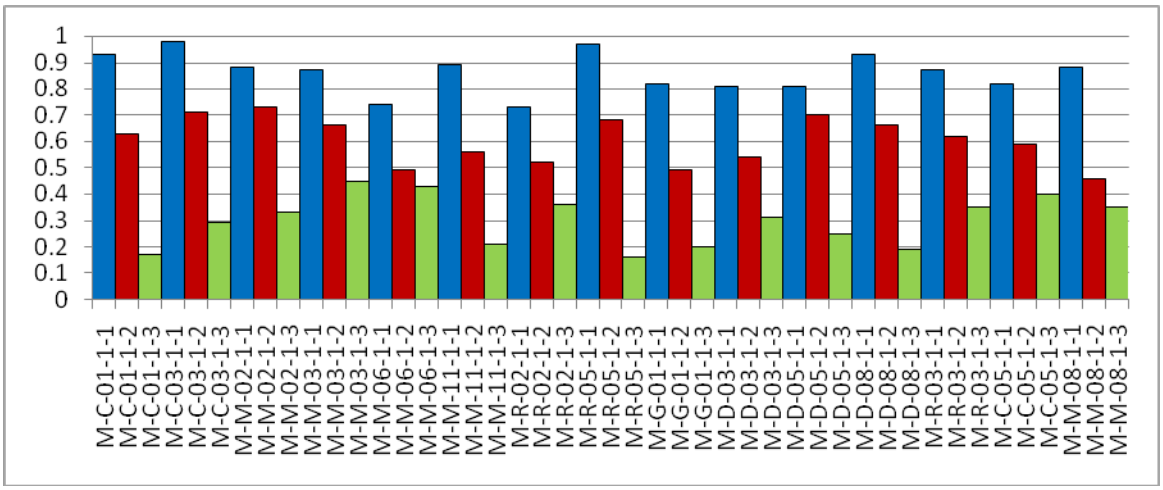


Figure 3.1. Distribution of difficulty levels (mathematics, test form 1)

Additionally, Table 3.5 shows p-values of items depending on the level. On the whole items of level 1 are easier than items of level 2, and items of level 2 are easier than items of level 3. This is also confirmed in separate content areas (see Table 3.6, in which average p-values of items are presented as a function of content areas).

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.

Table 3.5. The p-values of test items depending on their level

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

Table 3.6. Average p-values of items as a function of content areas

Content area	Test total	Level 1	Level 2	Level 3
Numbers and Calculations	0,61	0,91	0,64	0,29
Measurement of values	0,59	0,85	0,58	0,35
Regularities	0,59	0,86	0,61	0,29
Dependences	0,58	0,85	0,64	0,25
Elements of geometry	0,51	0,82	0,49	0,20

Verification of the second hypothesis

Verification of the second hypothesis was done in a special study conducted in the years 2011-2012. In 2011 the tests in mathematics and in the Russian language were administered on four age groups – students of the 4th, 6th, 8th and 10th grades. One year later, in 2012, the same tests were administered on the same students who were studying at that time in the 5th, 7th, 9th and 11th grades. Testing was done in spring, at the end of the academic year.

Further on, test results in mathematics shall be presented. The first mathematics cross section was done by 396 students and the second one by 412 students. The total of students who completed both cross sections was 374. Table 3.7 shows data on parallel distribution of the students.

Table 3.7. Student distributions in parallels (mathematics)

	Cross section 1	Cross section 2	Both cross sections
Grades 4-5	104	102	93
Grades 6-7	103	108	99
Grades 8-9	104	111	100
Grades 10-11	85	91	82
TOTAL	396	412	374

It is noteworthy that in this study no goal was set to achieve a representativeness of a sample with regard to the total statistical universe. Students in all grades were relatively strong students.

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 separate all participants into four groups according to the level of their achievement. The procedure of benchmarking will be described in Chapter 5.

Four proficiency levels were identified that correspond to the following content 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.

Table 3.8 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.) This table demonstrates that the least changes in the number of students who showed a certain degree of improvement can be seen for the transfer from the 10^{th} to the 11^{th} grade. In these grades changes in the percentages for each proficiency level were under 2%. By this time results become stable: most student participants are at proficiency level 3. The largest changes in the percentage of students who belong to different levels can be observed between Grade 6 and 7 and also between Grade 7 and 8, that is in middle school. Here the difference in proficiency level 2 and level 3 are between 8% and 16%. This fact confirms our hypothesis that the acquisition of curriculum at the functional level must and will happen within the framework of middle school.

Table 3.8. Distribution of students of different grades according to the proficiency level (mathematics)

	Proficiency level 1	Proficiency level 2	Proficiency level 3
Grade 4	16%	64%	18%
Grade 5	10%	60%	30%
Grade 6	7%	55%	38%
Grade 7	4%	43%	53%
Grade 8	1%	29%	70%
Grade 9	2%	24%	74%
Grade 10	1%	17%	82%
Grade 11	1%	15%	84%

Towards the end of primary school (Grade 4) the subject material of the syllabus is acquired (at a general education school) at the reflective level: most students (64%) are at proficiency level 2. Proficiency level 3 is just starting to be built up: only 18% of the students are at proficiency level 3 (and that in classes with well-prepared, strong learners). This confirms

our hypothesis that normally the subject matter of the primary school syllabus must be acquired at the reflective level, that is at the level of understanding.

Figure 3.2 demonstrates in a graphic form data from Table 3.8. It is possible to see that the percentage of students who show achievement at the second proficiency level steadily declines as the student age gets higher. The percentage of students who show achievement at proficiency level 3 keeps increasing.

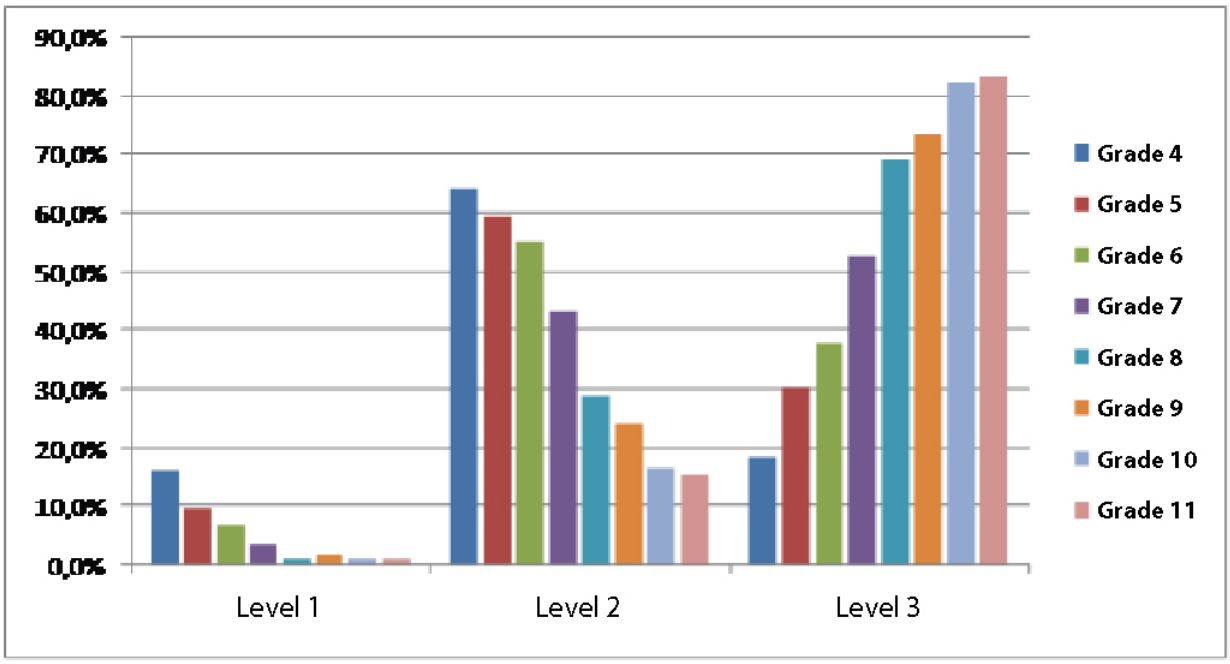


Figure 3.2. Distribution of test participants in proficiency levels as a function of grade (mathematics)

Figure 3.3 shows the distribution of students from different classes who could achieve various proficiency levels in mathematics. The diagram shows well the age dynamics in content acquisition. By the end of primary school (Grade 4) proficiency level 2 is dominant. Starting at Grade 8 (when students leave general school), proficiency level 3 is dominant. This provides good support for the SAM theoretical model, i.e. provides evidence to its validity.

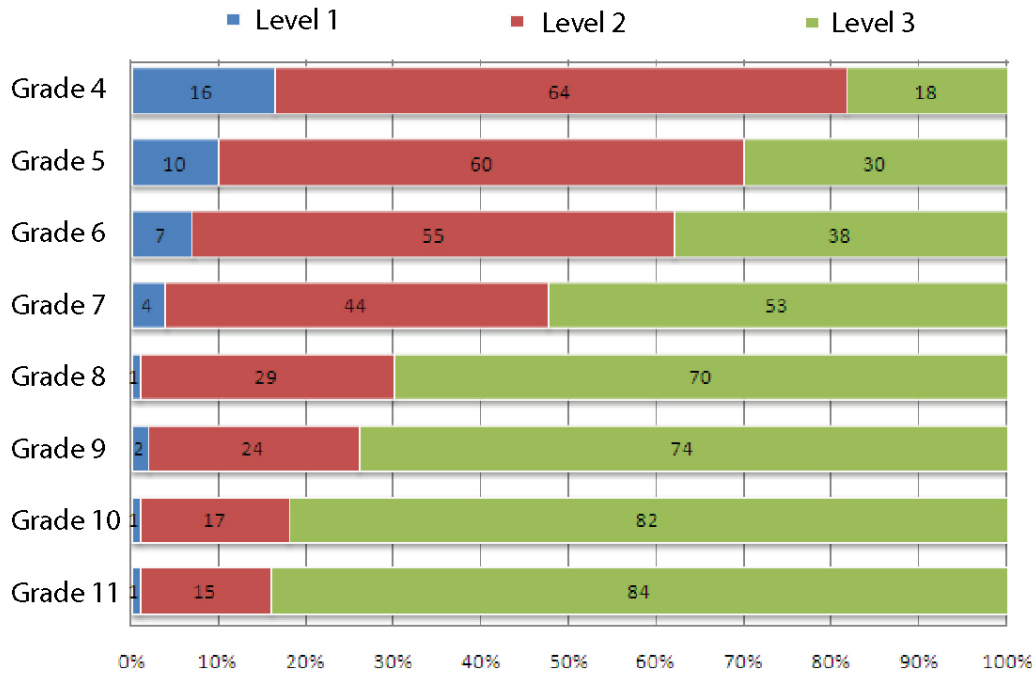


Figure 3.3. Students distribution of different grades depending on proficiency level (mathematics)

Let us point out that the results of this research can be more widely-used, namely for the estimation of student individual progress over time. It is possible to do so because the same students were tested twice with a one-year gap between the tests. Student progress can be monitored using two characteristics: test score and proficiency level which the student has. When a student gets better or worse proficiency level, it is the result of his/her getting a higher or lower test score. The change in a test score will not necessarily lead to changing to a proficiency level.

Improving the proficiency level of a student is evidence of his individual progress in subject acquisition. Changes in test score will show whether the student is going towards a higher level or if he/she is "stalled".

Table 3.9 presents data regarding the percentage of students in two parallels (Grades 4-5 and 6-7) who demonstrated a certain combination of levels at the first and second cross-section. We can see in this table that 9% of Grade 5 students got a lower level as compared with their own results in Grade 4: 2% went down, from proficiency level 2 to level 1 and 7% went from proficiency level 3 to level 2. 29% of Grade 5 students improved their level: 10% went up from proficiency level 1 to level 2; 1% went up from level 1 to level 3; and 18% of students went up from proficiency level 2 to level 3. Finally, 61% of the testing participants kept their presence at their previous level: 4% at proficiency level 1, 44% at level 2 and 13% at level 3. In the same manner we can interpret the data for the Grade 6-7 parallel. In this case, the majority of participants – 23% - improved their proficiency going up from level 2 to level 3.

Table 3.9. Percentage of students who showed a certain combination of levels at the first and second cross-section (mathematics)

Proficiency level combinations Cross-section 1 / cross-section 2	Grade 4-5	Grade 6-7
1/ 0	0	0
1/ 1	4	2
1/ 2	10	4
1/ 3	1	0
2/ 0	0	0
2/ 1	2	1
2/ 2	44	31
2/ 3	18	23
3/ 0	0	0
3/ 1	0	0
3/ 2	7	9
3/ 3	13	29

Let us make a note that students of other parallels (Grades 8-9 and 10-11) do not show a significant change in their levels, because in these parallels the dominating number of participants are at level 3.

Further on, there are results of a similar analysis for the Russian language test.

The first Russian language cross section was done by 382 students. The second Russian language cross section was done by 409 students. The total of students who completed both cross sections was 363. Table 3.10 shows data on parallel distribution of the students.

Table 3.10. Student distributions in parallels (Russian language)

	Cross section 1	Cross section 2	Both cross sections
Grades 4-5	102	102	92
Grades 6-7	102	107	99
Grades 8-9	98	111	95
Grades 10-11	80	89	77
TOTAL	382	409	363

Table 3.11 shows the distribution of test participants of different age groups depending on the proficiency level. This table demonstrates that the least changes in the number of students who showed a certain degree of improvement can be seen for the transfer from the 10^{th} to the 11^{th} grade. In these grades changes in the percentages for each proficiency level were minimal. By this time results become stable: most student participants were at level 3. The largest changes in the percentage of students who belong to different proficiency levels

can be observed in middle school. This fact confirms our hypothesis that the acquisition of curriculum at the functional level must and will happen within the framework of middle school.

Table 3.11. Distribution of students of different school grades according to the proficiency level (Russian language)

	Proficiency level 1	Proficiency level 2	Proficiency level 3	Proficiency level 0
Grade 4	37%	52%	10%	1%
Grade 5	24%	51%	20%	5%
Grade 6	21%	55%	22%	2%
Grade 7	13%	54%	31%	2%
Grade 8	16%	40%	43%	1%
Grade 9	7%	41%	52%	0%
Grade 10	1%	24%	75%	0%
Grade 11	3%	26%	71%	0%

Towards the end of primary school (Grade 4) the subject material of the syllabus is acquired (at a general education school) at the reflective level: most students (52%) were already at level 2. Level 3 is just starting to be built up: only 10% of the students are at Level 3 (and that in classes with well-prepared, strong learners). This confirms our hypothesis that normally the subject matter of the primary school syllabus must be acquired at the reflective level, that is at the level of understanding.

Figure 3.4 demonstrates in a graphic form the data from Table 3.11. It is possible to see that the percentage of students who show the first level steadily declines as the student age gets higher. At the same time, the percentage of students who show achievement at Level 3 keeps increasing from Grade 4 to Grade 10 (with insignificant fluctuations in the area for Grade 7-8 for level 1 and in the area for Grade 10-11 for level 3). As far as level 2 is concerned, the percentage of students who stay at this level (that is those showing achievements of level 2) is relatively stable for Grades 4-7 (at the level of 50%-55%) and later on it keeps decreasing: to the level of 40% in Grades 8-9 and to 25% in Grades 10-11.

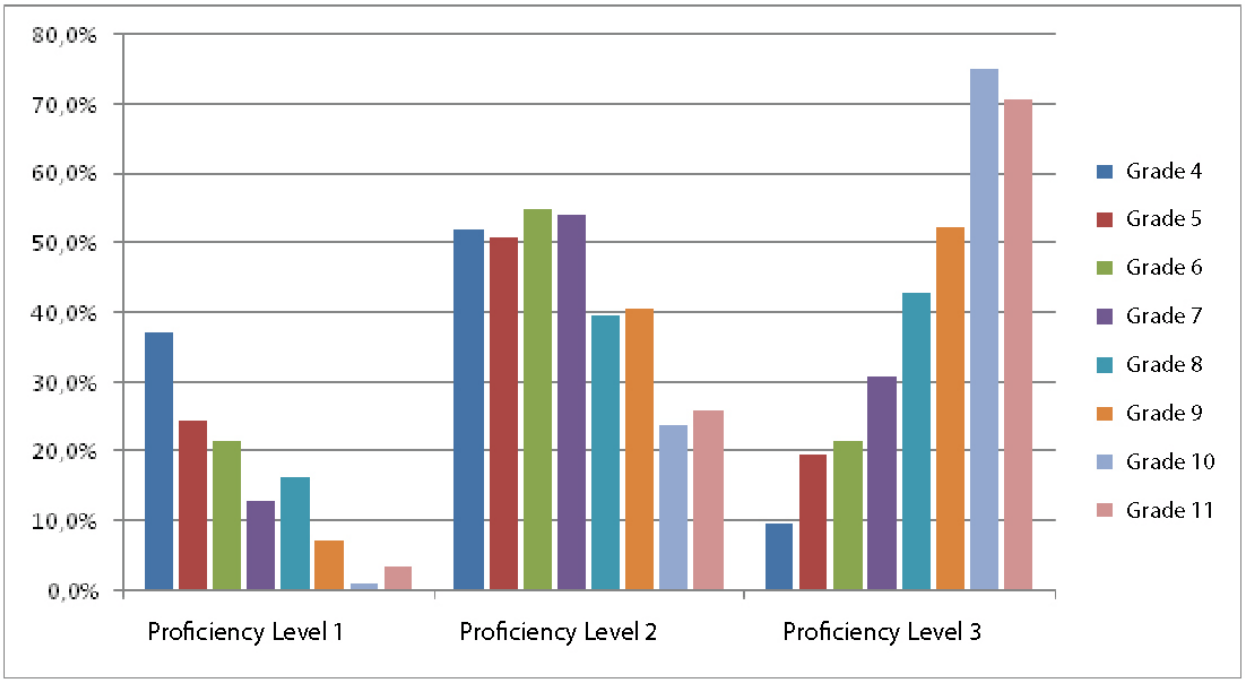


Figure 3.4. Distribution of test participants in proficiency *levels as a function of grade (Russian language)*

It is interesting to note that the least changes in the percentage of students who demonstrated a certain level of achievement can be observed for the transfer from Grade 5 to 6 and also from Grade 10 to 11. In these grades the percentage differences for each level are not greater than 4%. Greatest changes in the percentage of students can be observed, as follows:

- For level 1 – from Grade 4 to 5 (13%);
- For level 2 – from Grade 7 to 8 (14%) and from Grade 9 to 10 (17%);
- For level 3 – from Grade 9 to 10 (23%).

Figure 3.5 shows the distribution of students from different grades who could achieve various levels in the Russian language. The diagram shows well the age dynamics in content acquisition. By the end of primary school (Grade 4) Level 2 is dominant. Starting at Grade 9 (when students leave general school), level 3 is dominant. This provides good support for the SAM theoretical model, i.e. provides evidence to its validity.

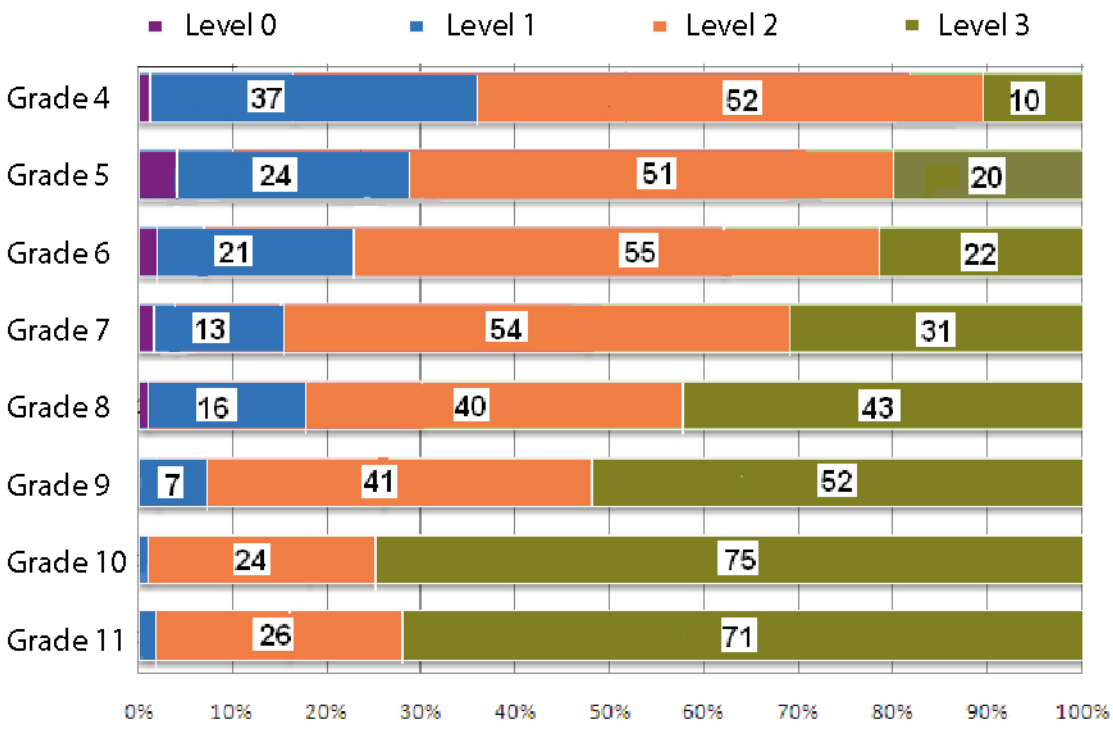


Figure 3.5. Students distribution of different grades depending on proficiency level (Russian language)

Let us point out that, just as with mathematics, the results of this research with the Russian language proficiency can be more widely-used, namely for the estimation of student individual progress over time. It is possible to do so because the same students were tested twice with a one-year gap between the tests. Student progress can be monitored using two characteristics: test score and proficiency level which the student has achieved. When a student achieves better or worse proficiency level, it is the result of his/her getting a higher or lower test score. The change in a test score will not necessarily lead to changing to a different level.

Improving the proficiency level of a student is evidence of his individual progress in subject acquisition. Changes in test score will show whether the student is going towards a higher level or if he/she is "stalled".

Table 3.12 presents data regarding the percentage of students in two parallels (Grades 4-5 and 6-7) who demonstrated a certain combination of levels at the first and second cross-section. We can see in this table that 9% of Grade 5 students got a lower level as compared with their own results in Grade 4: 2% went down, from level 1 to level 0; 3% went down from level 2 to level 1, and 4% went from level 3 to level 2. 27% of Grade 5 students improved their level: 13% went up from level 1 to level 2 and 14% of students went up from level 2 to level 3. Finally, 63% of the testing participants kept their presence at their previous level: 20% at level 1, 36% at level 2, and 7% at level 3. In the same manner we can interpret the data for the Grade 6-7 parallel. In this case, the majority of participants – 34% - kept their level 2 and 17% improved their proficiency going up from level 2 to level 3.

Table 3.12. Percentage of students who showed a certain combination of proficiency *levels at the first and second cross-section*

Proficiency level combinations Cross-section 1 / cross-section 2	Grade 4-5	Grade 6-7
1/ 0	2	0
1/ 1	20	10
1/ 2	13	11
1/ 3	0	0
2/ 0	0	0
2/ 1	3	3
2/ 2	36	34
2/ 3	14	17
3/ 0	0	0
3/ 1	0	0
3/ 2	4	6
3/ 3	7	16

Let us make a note that students of other parallels (Grades 8-9 and 10-11) do not show a significant change in their levels, because in these parallels the dominating number of participants are at level 3.

Comparing difficulties of test items predicted by test developers and those observed after testing

While developing each test item, its authors must have an idea regarding the cognitive basis of an action for its completion and thus have an opportunity to predict the level of difficulty for a potential sample of test participants. If the predictions of test developers are close to empirical values, there is an evidence that their understanding of the actions by the examinees correspond to the facts.

To conduct a relevant study SAM test in mathematics proficiency was selected. Test developers were asked to estimate expected difficulty of all items prior to their pilot testing. Interjudge consistency was estimated, and consistency index value was found to be 0,831. After the pilot testing of the items we estimated the consistency index as a correlation between expected and empirical p-values for the whole test as well as for each subtest with items at Level 1, 2 and 3. Table 3.13 presents correlation indexes.

Table 3. Correlation indexes for expected and empirical p-values

Level 1 items	0,882
Level 2 items	0,973
Level 3 items	0,928
The whole test	0,931

Figure 3.6. shows the relation between expert expectations and empirical difficulty characteristics for the whole test. These results show quite high level of consistency between experts as well as a high correlation between their expectations and empirical results, which is one more count of evidence for the construct validity.

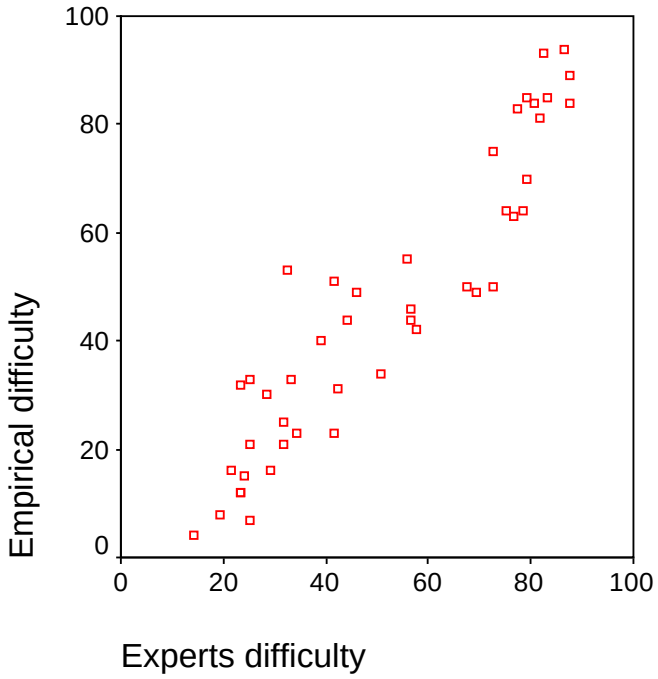


Figure 3.6. Correlation between expert-estimated and empirical difficulty

3.3. Criterion validity

As noted above, 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. Criterion validity can be divided into predictive and current validity. Predictive validity shows how well a test can predict future criterion scores . Current criterion validity answers the question how test results are related to a criterion at present.

Below there are results of studying SAM criterion validity.

SAM predictive validity study

This study was based on SAM pilot testing in one of the regions of the Russian Federation in spring 2011. Total sample was 941 students from 12 schools. This sample was compiled as a representative sample stratified on the basis of 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 value about 500 and standard deviation of 50 (the scaling procedure will be described in detail in Chapter 5). Additionally, as was mentioned above, to make the interpretation of the results for the SAM test participants possible on the basis of the three-level test model, benchmarks were set that helped separate all participants into four groups according to the level of their achievement. Figure 3.7 shows the distribution of test participants over proficiency levels in mathematics.

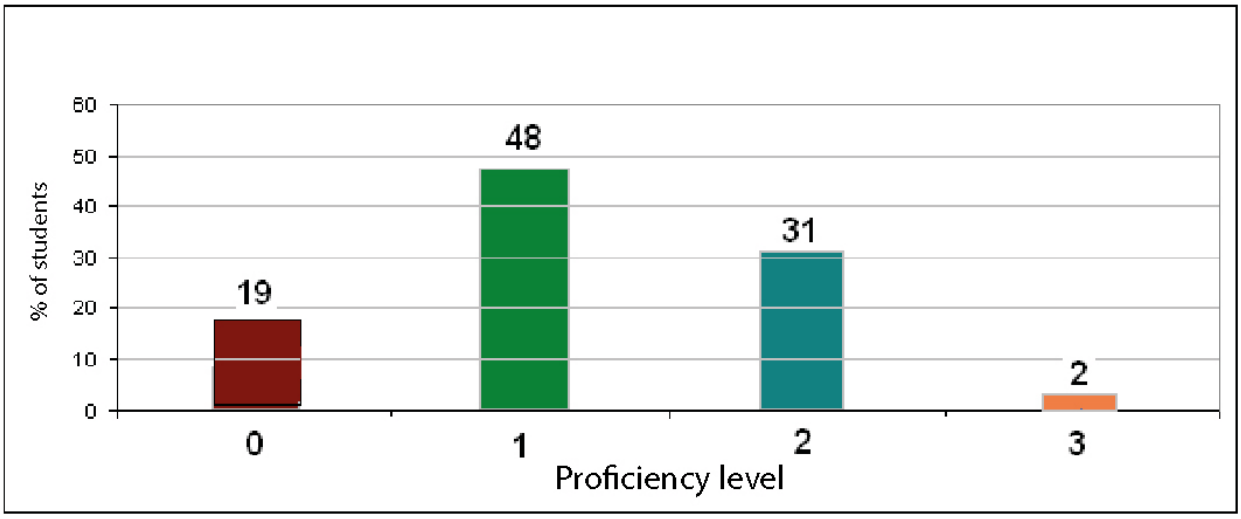


Figure 3.7. Test participant distribution over proficiency levels (mathematics)

We can observe that on the whole, all over the region, 19% of all examinees were at proficiency level 0 (i.e. not even level 1 was acquired); 48% were at proficiency level 1 (only level 1 was acquired); 31% - at proficiency level 2 (level 2 was acquired) and only 2% - at proficiency level 3 (level 3 was acquired).

Figure 3.8 shows the distribution of test participants across levels of proficiency depending on the school (school indexes are shown left, on the vertical axis. Schools were arranged in the descending order of their average test score. It can observed that the number of students at proficiency level 0 fluctuates from 8% to 48%, depending on the particular school, and the number of students at proficiency level 3 does not exceed 7%. Proficiency level 1 dominates in all schools.

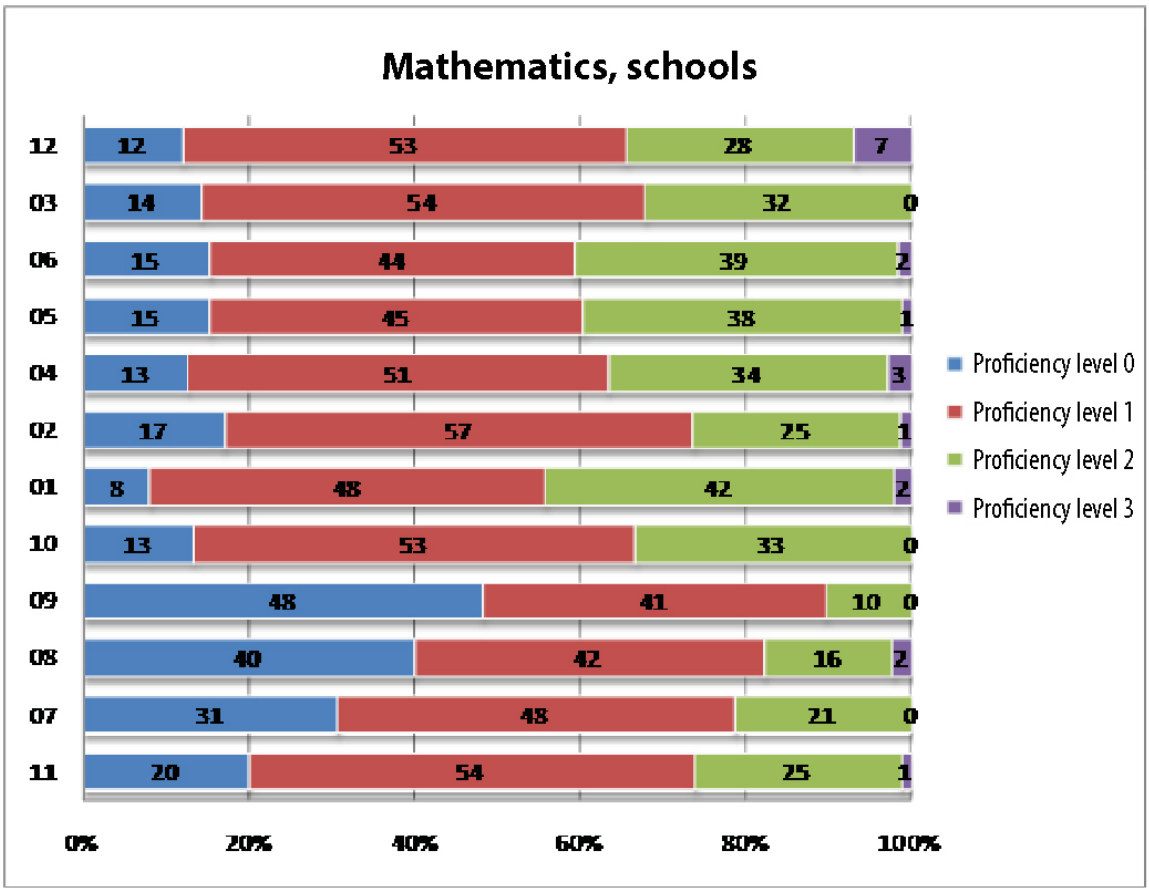


Figure 3.8. Test participant distribution over proficiency levels depending on the school (mathematics)

To study predictive validity of the SAM test, marks in mathematics were gathered with the same students one year later (at that time, they were studying in the fifth grade). Those were school marks, of course, and thus subjective, plus the grading scale in schools has a very low discriminativity. There exists, however, no other universal criterion showing student progress in mathematics in the fifth grade. It was possible to collect school grades from 649 students, which makes 67% of the total test sample. Consequently, for the purposes of our study only test results of those students were selected whose achievement results in mathematics were known for Grade 5. 110 students (17%) from the selection were at proficiency level 0; 311 students (48%) were at proficiency level 1; 219 students (34%) were at proficiency level 2 (they acquired level 2); and 10 (1%) were at proficiency level 3. It is worth noting that the proficiency level distribution of selected students is very close to the distribution across the whole region shown in Figure 3.7.

One student of 649 got mark 2 [meaning: Failing]. This student was removed from subsequent analysis. The remaining students got the following grades distribution: 218 (34%) got mark 3 (satisfactory); 275 (42%) got mark 4 (Good); and 155 students (24%) got mark 5 (Excellent).

Table 3.14 shows mark distribution for students at different proficiency levels; Figure 3.9 shows the same distribution in a diagram. We can see that all students who were put into proficiency level 3 according to SAM test results got mark 5 (Excellent). Students who were put into proficiency level 2 were mainly distributed between marks 4 (Good) and 5 (Excellent). One

half of the students who were put into proficiency level 1 got mark 4 (Good) and about one third from level 1 got mark 3 (Satisfactory). Finally, for students who were put into proficiency level 0, the dominant mark was 3.

Table 3.14. Student marks distribution (for students at different proficiency levels), % (mathematics)

Proficiency levels	Mark		
	3	**4**	**5**
Proficiency level 0	82	17	1
Proficiency level 1	35	49	16
Proficiency level 2	9	48	43
Proficiency level 3			100

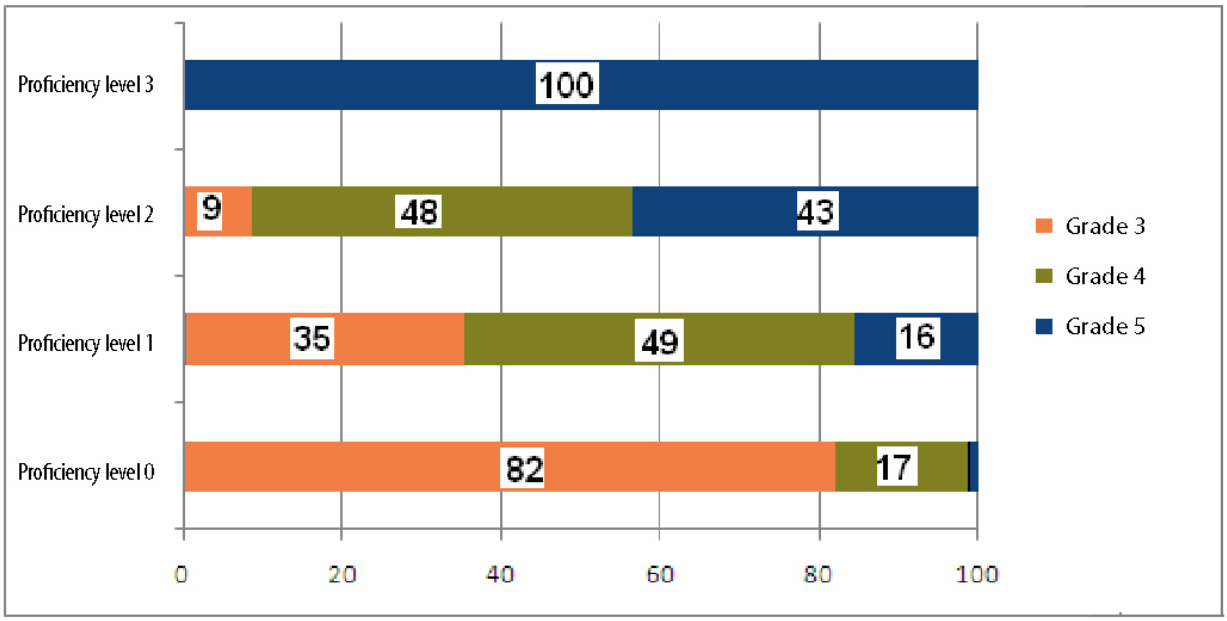


Figure 3.9. Distribution of student marks depending on student proficiency level (mathematics)

Table 3.15 shows the student distribution across proficiency levels, for students who were awarded different marks for their academic achievement; Figure 3.10 shows the same distribution as a diagram. We can see that 62% of students who achieved mark 5 (Excellent) were put into proficiency level 2; 55% of students awarded with mark 4 (Good) were put into proficiency level 1 and 38% of mark 4 students were put into proficiency level 2. Finally, 41% of mark 3 (Satisfactory) students were put into proficiency level 0 and 50% into proficiency level 1. No student who was put into proficiency level 3 achieved other marks than 5 (Excellent).

Table 3.15. Proficiency level distribution of students having different marks, in % (mathematics)

Mark	Proficiency level			
	0	**1**	**2**	**3**
3	41	50	9	
4	7	55	38	
5		31	62	7

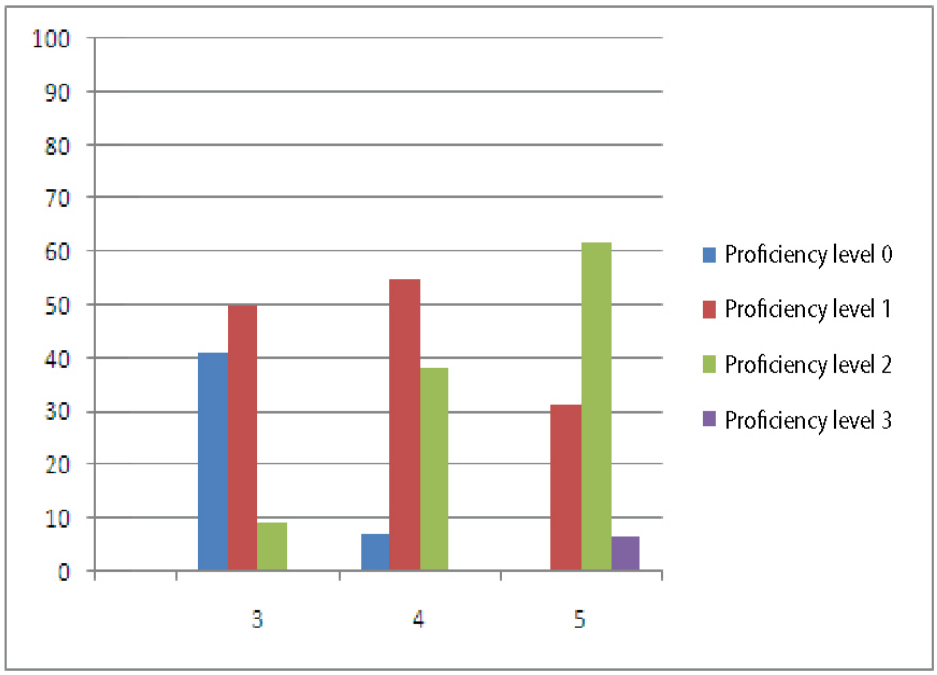


Figure 3.10. Student distribution across proficiency levels, depending on their marks (mathematics)

In addition, the correlation index for student ability score and their school marks was calculated, and also the correlation index for student proficiency level and their school marks. These correlation indexes were, respectively, 0.6 and 0.56. These values are quite high, thus providing a favorable result for predictive validity of SAM test in mathematics.

A similar study was done for the SAM test in Russian language. Overall situation is similar. Figure 3.11 shows the distribution of test participants over proficiency levels in Russian language for the whole region. Please note that only one student was put into proficiency level 3. This student was removed from subsequent analysis and this is why level 3 will not be represented here.

We can observe that on the whole, all over the region, 15% of all examinees were at proficiency level 0 (i.e. not even level 1 was acquired); 64% were at proficiency level 1 (only level 1 was acquired); and 21% - at proficiency level 2 (level 2 was acquired).

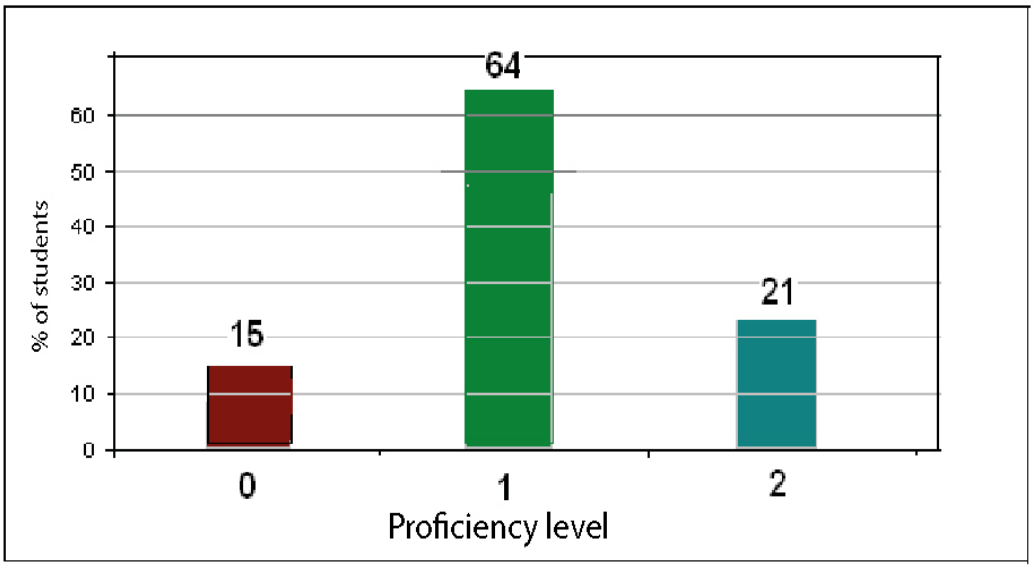


Figure 3.11. Test participant distribution over proficiency levels (Russian language)

To study predictive validity of the SAM test, test marks in Russian language were gathered with the same students one year later (at that time, they were studying in the fifth grade). The sample of students whose school grades were accessible included 638 students, which makes 67% of the total test sample. Consequently, for the purposes of our study only test results of those students were selected whose marks in Russian language were known for Grade 5. 159 students (25%) from the selection were at proficiency level 0; 348 students (55%) were at proficiency level 1; 130 students (20%) were at proficiency level 2 (they acquired level 2); and 1 student was at proficiency level 3 (he was removed from the study analysis). It is worth noting that the proficiency level distribution of selected students is different from the distribution for the whole region shown in Figure 3.13.

Not one of the 638 students got mark 2 (Failing). The students got the following marks distribution: 228 (36%) got mark 3 (Average); 316 students (50%) got mark 4 (Good); and 94 students (14%) got mark 5 (Excellent).

Table 3.16 shows student mark distribution for students at different proficiency levels; Figure 3.12 shows the same distribution in a diagram. We can see that 82% of the students who were put into proficiency level 0 according to SAM test results in Russian language got mark 3 (Satisfactory). Most of the students who were put into proficiency level 1 (62%) got mark 4 (Good). Students who were put into proficiency level 2 were mainly distributed between marks 4 (Good) and 5 (Excellent).

Table 3.16. Student marks distribution (for students at different proficiency levels), % (Russian language)

Proficiency levels	Mark		
	3	**4**	**5**
Proficiency level 0	82	17	1
Proficiency level 1	28	62	10
Proficiency level 2	7	49	43

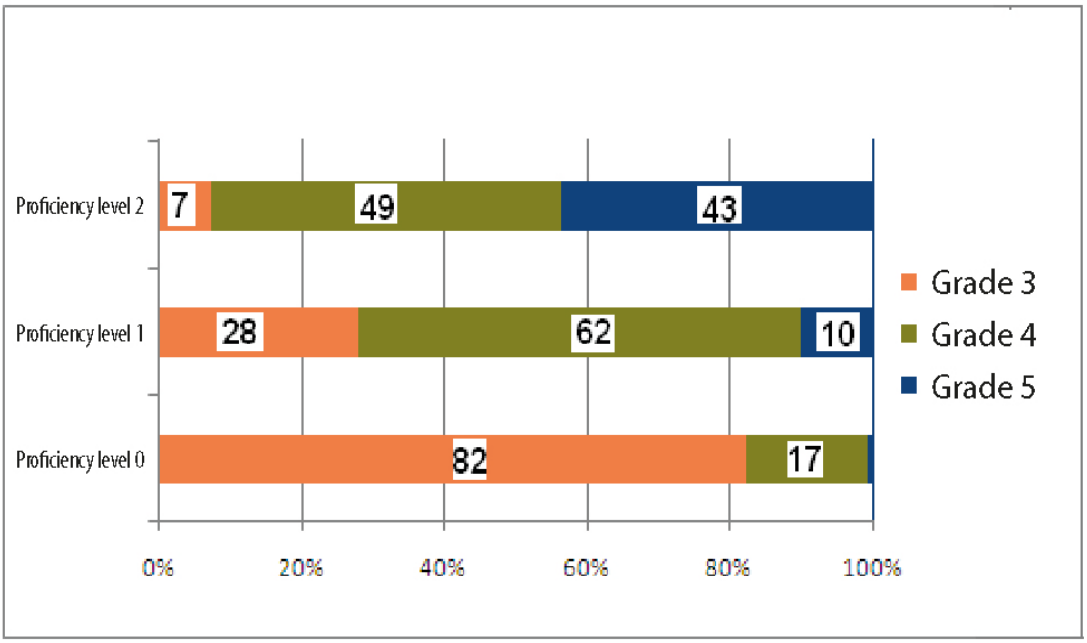


Fig. 3.12. Distribution of student marks depending on student proficiency level (Russian language)

Table 3.17 show the student distribution across proficiency levels, for students who were awarded different marks for their academic achievement; Figure 3.13 shows the same distribution as a diagram. We can see that 61% of students who got mark 5 (Excellent) were put into proficiency level 2; 70% of students awarded with mark 4 (Good) were put into proficiency level 1 and 21% of mark 4 students were put into proficiency level 2. Finally, 55% of mark 3 (Satisfactory) students were put into proficiency level 0 and 41% into proficiency level 1.

Table 3.17. Proficiency level distribution of students having different marks, in % (Russian language)

Mark	Proficiency level		
	0	**1**	**2**
3	55	41	4
4	9	70	21
5	1	37	61

In addition, the correlation index for student ability score in Russian language and their school mark was calculated, plus the correlation index for student proficiency level and their school mark. These correlation indexes were, respectively, 0.64 and 0.58. These values are close to corresponding values for the mathematics test. As noted above, they are quite high, thus providing a favorable result for predictive validity of SAM test in Russian language.

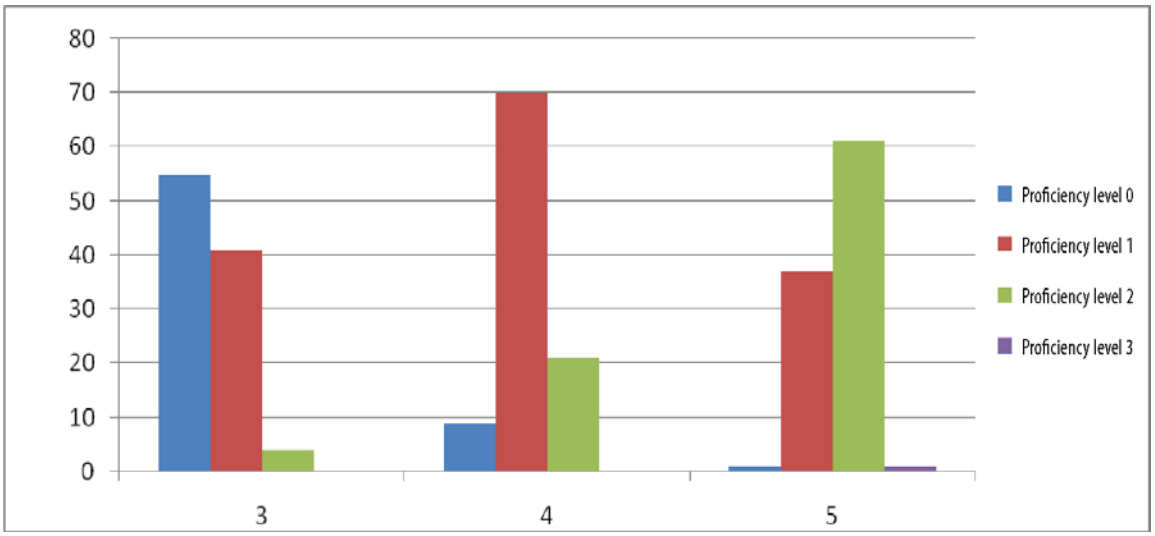


Fig. 3.13. Student distribution across proficiency levels, depending on their marks (Russian language)

Consequently, this study leads to the conclusion that SAM tests in mathematics and in Russian language feature high predictive validity.

SAM current validity study

This study was based on SAM pilot testing in one of the regions of the Russian Federation in spring 2012. Total sample amounted to 4406 students in the mathematics study and 4385 students in the Russian language study. A special feature of this particular pilot study was that practically all Grade 4 students from primary schools of the whole region were tested. Current validity is studied using an external criterion for which the data are collected simultaneously with conducting experiments on the method tested. Subject marks that were expected for students at their completion of primary school were taken as a criterion for the current validity study. These estimates were collected from school teachers during the testing. It was possible to gather estimated marks on 3955 students for mathematics and on 3893 students for Russian language. Table 3.18 shows the distribution of marks for mathematics and for Russian language.

Table 3.18. Distribution of expected student marks, %

Expected annual subject mark	Mathematics	Russian language
3	39	41
4	52	51
5	9	8

Table 3.19 presents the distribution of test participants over proficiency levels. We can see that in mathematics (for the whole region) 2% of test takers were at proficiency level 0 (even level 1 was not acquired yet); 27% were at proficiency level 1 (only level 1 was acquired); 54% were at proficiency level 2 (level 2 was acquired), and 17% were at proficiency level 3 (level 3 was also acquired). In the Russian language the distribution over proficiency levels was as follows: 4%, 39%, 39% and 12% for proficiency levels 0, 1, 2, and 3, respectively.

Table 3.19. Distribution of test participants over proficiency levels, %

Proficiency levels	Mathematics	Russian language
0	2	4
1	27	39
2	54	39
3	17	12

Table 3.20 shows the distribution of marks in mathematics for students from different proficiency levels, and in Figure 3.14 the same distribution is shown in a diagram. We can see that 89% of students who were put into Proficiency level 0 in mathematics after SAM test results have mark 3 (Satisfactory). 64% of students who were put into Proficiency level 1 have mark 3 (Satisfactory) and 34% have mark 4 (Good). Most students who were put into Proficiency level 2 (59%) have mark 4 (Good). The same is true for students who were put into Proficiency level 3, however 25% of them have mark 5 (Excellent).

Table 3.20. Distribution of marks for students from various proficiency levels, % (mathematics)

Proficiency levels	Mark		
	3	**4**	**5**
Proficiency level 0	89	9	2
Proficiency level 1	64	34	2
Proficiency level 2	33	59	8
Proficiency level 3	13	62	25

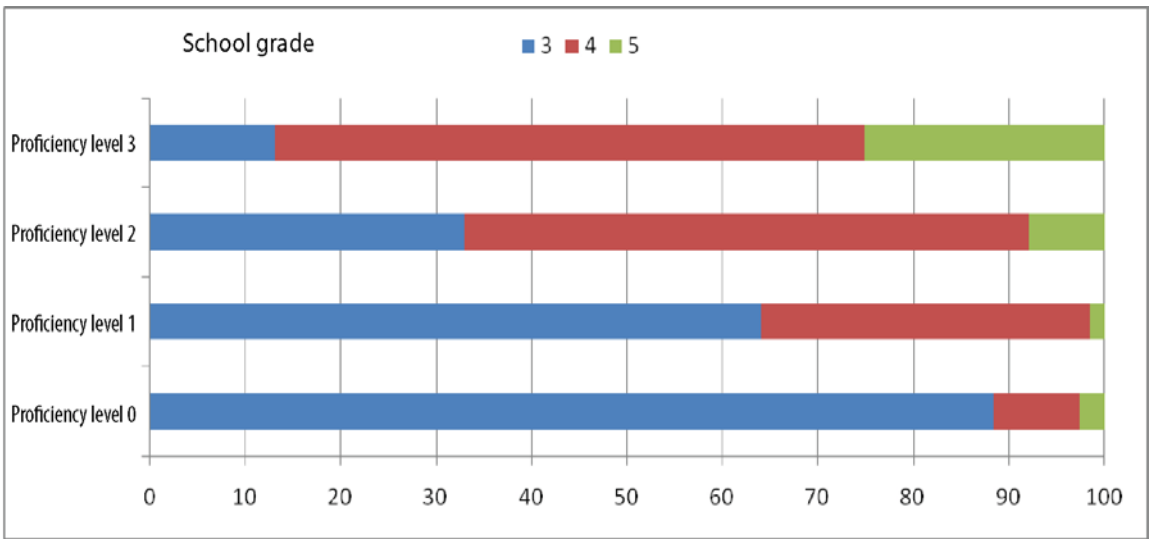


Figure 3.14. Mark distribution of students depending on proficiency level (mathematics)

Table 3.21 shows the proficiency level distribution of students having different marks, and in Figure 3.15 the same distribution is shown in a diagram. We can see that of the students who got mark 5 (Excellent) 47% were put into proficiency level 2 and 48% into proficiency level 3; of the students who got mark 4 (Good) 61% were put into proficiency level 2 and 21% into proficiency level 3; and finally, of the students who got mark 3 (Satisfactory) 4% were put into proficiency level 0, 44% into proficiency level 1 and 46% into proficiency level 2.

Table 3.21. Proficiency level distribution of students having different marks, % (mathematics)

Grade		Proficiency level		
	0	**1**	**2**	**3**
3	4	44	46	6
4	0	18	61	21
5	1	4	47	48

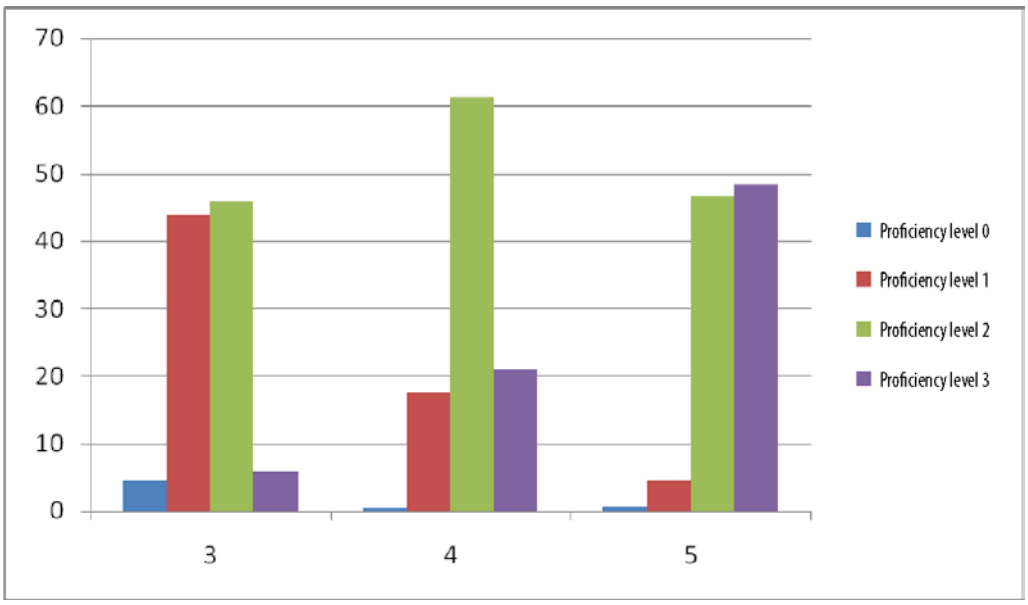


Figure 3.15. Student distribution over proficiency levels depending on awarded marks (mathematics)

In addition, the correlation index for student ability score in mathematics and their expected school mark was calculated, and also the correlation index for proficiency level into which a student was put and his school mark. These correlation indexes were, respectively, 0.46 and 0.41. These values are lower than the corresponding indexes for the predictive validity, which can be explained in this manner: teachers may have been unwilling to give higher expected scores to students. This is supported by the fact that the percentage of expected 5 (Excellent) grades stayed under 10% (see Table 3.18), which does not correspond, in our opinion, to the actual number of students who got 5 in mathematics and in Russian language at their completion of the primary school.

Further, results of a similar study for Russian language are presented. Table 3.22 shows the distribution of marks in Russian language for students from different proficiency levels, and in Figure 3.16 the same distribution is shown in a diagram. We can see that 84% of students who were put into Proficiency level 0 in Russian language after SAM test results have mark 3 (Satisfactory). 53% of students who were put into Proficiency level 1 have mark 3 (Satisfactory) and 45% have mark 4 (Good). Most students who were put into Proficiency level 2 (65%) have mark 4 (Good). The same is true for students who were put into Proficiency level 3, however 28% of them have mark 5 (Excellent).

Table 3.22. Distribution of marks for students from various proficiency levels, % (Russian language)

Proficiency levels	Grade		
	3	**4**	**5**
Proficiency level 0	84	15	1
Proficiency level 1	53	45	2
Proficiency level 2	27	65	9
Proficiency level 3	11	61	28

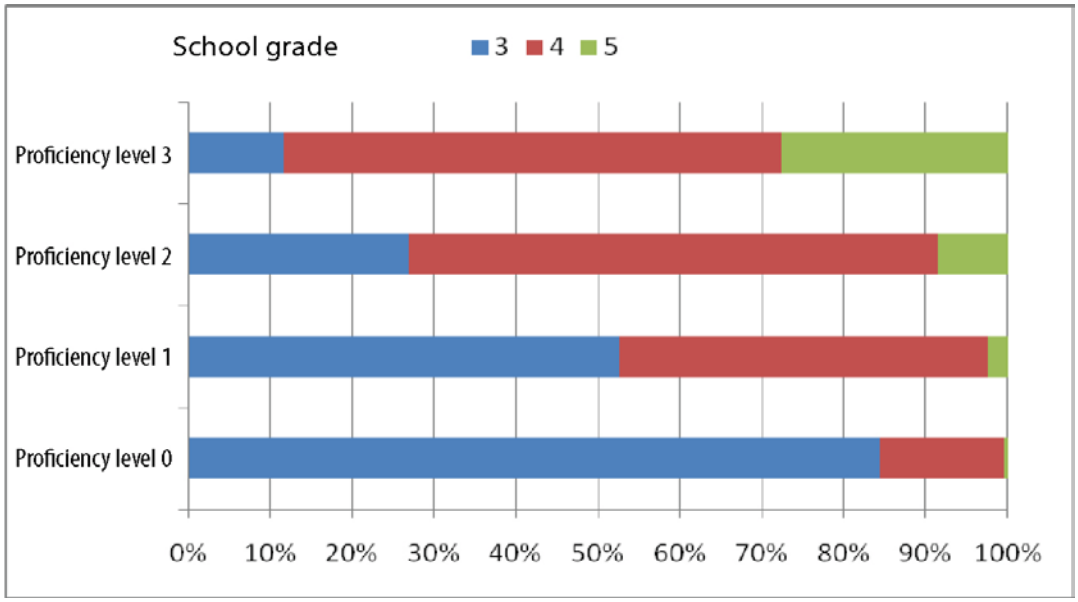


Figure 3.16. Mark distribution of students depending on proficiency level (Russian language)

Table 3.23 shows the proficiency level distribution of students having different marks, and in Figure 3.17 the same distribution is shown in a diagram. We can see that of the students who achieved mark 5 (Excellent) 43% were put into proficiency level 2 and 44% into proficiency level 3; of the students who achieved mark 4 (Good) 49% were put into proficiency level 2 and 14% into proficiency level 3; and finally, of the students who achieved mark 3 (Satisfactory) 21% were put into proficiency level 0, 51% into proficiency level 1 and 25% into proficiency level 2.

Table 3.23. Proficiency level distribution of students having different marks, % (Russian language)

Mark		Proficiency level		
	0	**1**	**2**	**3**
3	21	51	25	3
4	3	34	49	14
5	1	12	43	44

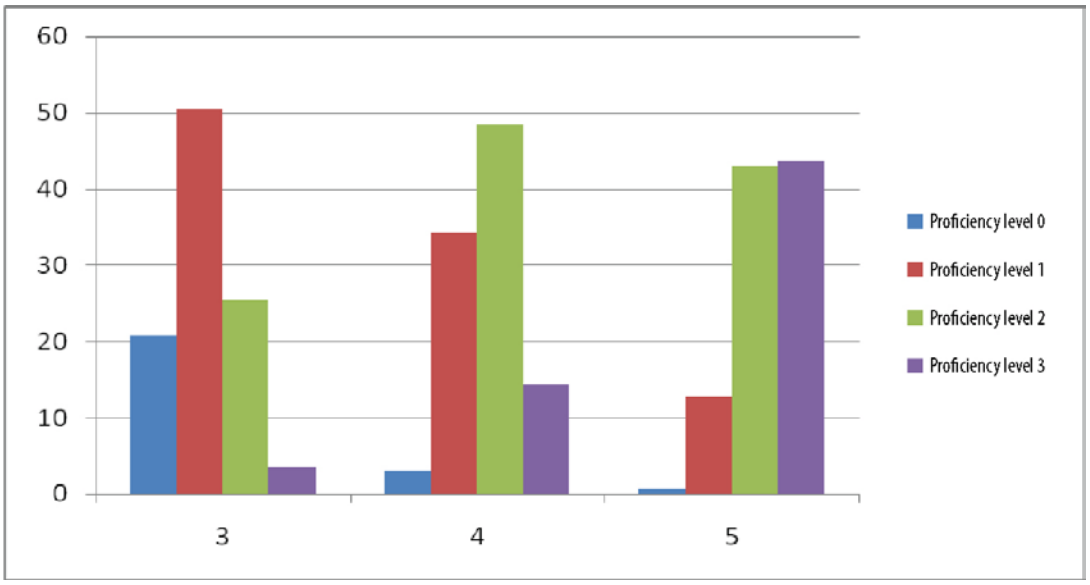


Fig. 3.17. Student distribution over proficiency levels depending on awarded marks (Russian language)

In addition, the correlation index for student ability score in Russian language and their expected school mark was calculated, and also the correlation index for proficiency level into which a student was put and his school mark. These correlation indexes were, respectively, 0.48 and 0.45. These values are close to similar indexes in mathematics.

Literature

1. SAM (School Achievement Monitoring): Инструмент мониторинга учебных достижений школьников // *под ред. Нежнова П.Г., Кардановой Е.Ю.*, 2011, 104 с.
2. *Анастази А.*, Урбина С. Психологическое тестирование. СПб.: Питер, 2003. 674 с.

4. Differential item functioning with regard to different examinee groups (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. For example, a test item in mathematics may contain words which examinees from a certain group do not know or comprehend poorly. In this case the probability of a successful completion of this item is lower for participants from this particular group, even though their mathematics proficiency is the same as with other participants. If the test incorporates several of such items, this will influence the test scores of participants from this group: they will be sufficiently lower. DIF analysis is designed to help discover such 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 for various SAM test item functioning with regard to participants of different gender (boys / girls).

A unidimensional dichotomous Rasch model was selected for SAM test modeling. As per conditions of this model, item parameter estimates must be stable (invariable) relative to examinee groups which provided these estimates (on the hypothesis that the item fits to the model) [1]. If an item demonstrates DIF with regard to some participant groups, the invariant property will not be fulfilled: difficulty estimates for this item which were evaluated following test results of these groups, will show statistically significant difference. One of the DIF methods of detection (that we were using in this report) is separate calibration t-test [2]. The statistics is calculated separately for each item:

$$t_i = \frac{\delta_i^{(1)} - \delta_i^{(2)}}{\sqrt{(\sigma_i^{(1)})^2 + (\sigma_i^{(2)})^2}},$$

where $\delta_i^{(1)}$ is difficulty estimate for the test item i based on the first examinee group test completion; $\delta_i^{(2)}$ is difficulty estimate for the - test item *i* based on the second examinee group test completion; $\sigma_i^{(1)}$, $\sigma_i^{(2)}$ are standard errors of measuring $\delta_i^{(1)}$ and $\delta_i^{(2)}$. The t_i statistics shows asymptotically normal distribution with zero mathematical expectation and unit variance. When statistics values exist with modulus greater than 2 (i.e. $|t_i| > 2$), this is an indication of exceedingly large degrees of deviation for item difficulty scores, that is an indication of DIF.

To compare item difficulty it is necessary for item difficulty values that were arrived at after two calibrations to be on the same scale. For the creation of a common scale this study

used the method of “Constant Items” [3], in which a set of DIF-free items serves as anchor link for the establishment of the common scale and all other items are tested for DIF. Both non-statistical procedures (item content analysis) and statistical procedures (comparing item characteristics for the whole data population as well as for each group separately) were used for the selection of such items.

It is noteworthy that when sample size is large, the t-statistics (just as any other type of statistics) may provide statistically significant values even with small DIF in an item. This is why when conducting item DIF analysis a minimum difference in item difficulty for two examinee groups is set as a certain threshold value defining its practical importance. We consider the DIF amount equal to 0.5 logit to be just this kind of threshold value [4].

On top of t-statistics this study was using one of the best known DIF study methods within the classical test theory – the Mantel-Haenzel statistics method (MH) which involves a direct comparison of the chances for completing an item correctly by examinees with the same ability score, but coming from different groups [5]. MH statistics shows a chi-square distribution with one degree of freedom if null hypothesis of DIF nonoccurrence is correct.

A real-life test will never be ideal and always has some degree of DIF. It is necessary to find out whether the DIF value is relatively small and thus the test can be considered stable (invariable) in practical terms (i.e. it does not distort scores of examinees from different groups) or whether items with DIF introduce significant distortions in the measuring of examinees. To this end this study compared test characteristics curves for separate groups of examinees[1]. Tests were preliminarily equated using the method of anchoring item parameters, with items having no DIF. If characteristics curves practically coincide, the presence of DIF items will not impact significantly the test score of the test participant, and the results can be acknowledged as fair relative to various groups of participants. Otherwise, if the curves do not coincide, the presence of items with DIF shall exert an impact onto participant scores and thus must be additionally studied.

To conclude, in this study when conducting SAM test item analysis to find out the presence of gender DIF a procedure was used consisting of the following stages:

1) setting up a common scale for the two groups;

2) selecting items that may include DIF through using t-test method with separate calibration for each group as well as Mantel-Haenszel (MH) method;

3) selecting item with DIF on the basis of the results from step 2 and taking into account threshold value of practical significance (the minimum threshold in the difference of item difficulty for various examinee groups is taken to be equal to 0.5 logit);

4) determining DIF significance at test level: comparing test characteristic curves of two groups of participants.

DIF-analysis of SAM test items was conducted with Winsteps software.

4.1. DIF analysis of mathematics item test

[1] Test characteristics curve represents a plot of mathematical expectation of the test score depending on a student’s ability score.

Table 4.1 incorporates data on the number of test participants of both sexes and on their average score for the mathematics test (test form 1). Data of analysis for test form 2 are similar, so they are not shown in this report. Figure 4.1 presents histograms of test score distributions for participants from groups under review: pink is for girls and blue is for boys.

Table 4.1. Test results for both genders (mathematics, test form 1)

Indicator	Females	Males
Sample	1471	1546
Observed raw score: average (SD)	26.7 (8.4)	26.2 (8.3)
Ability estimate: average (SD)	0.69 (1.32)	0.61 (1.11)

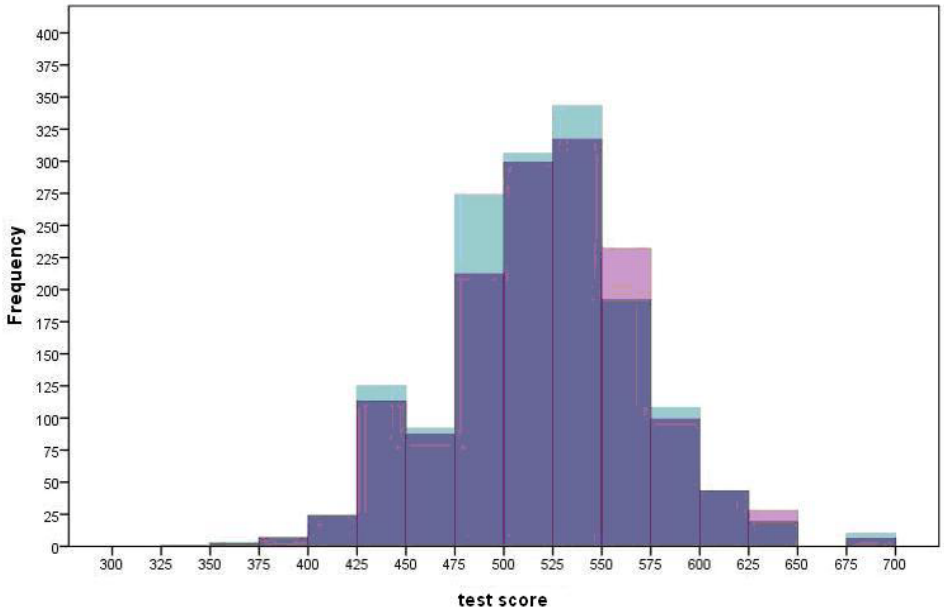


Figure 4.1. Test score distribution for test participants from groups under review (mathematics, test form 1)

Thus results for girls and boys are, on the whole, very close: girls did the test a tiny bit better than boys, but these differences are hardly significant (the significance of differences will be dealt with in Chapter 6).

DIF analysis results for SAM item tests are shown in Table 4.2. Items appear in the same order as they were coming up in the test. Columns 2 and 3 of the table show item difficulty estimates as well as corresponding measurement errors that resulted from the testing of girls (all data are in logits). The next two columns show the same data that were arrived at from the testing of boys. Columns 6 and 7 display the difference in difficulty estimates for the two participant groups and the corresponding measurement error. Next column presents values of t-statistics. Values from the (-2,+2) interval are acceptable for this statistics if the item doesn't demonstrate DIF. Finally, the last two columns show values of Mantel-Haenzel (MH) statistics as well as its significance level (that is, the probability of appearance for such data, under the hypothesis that the item does not demonstrate DIF). Minimum threshold of such probability is

taken to be equal to 0.05. DIF statistics values that go beyond its critical values are highlighted pink in the table.

Table 4.2. DIF analysis results (mathematics, test for 1)

#	Girls		Boys		Difference in diff.	S.E. of difference	t-statistics	Mantel-Hanzel	
	Difficulty	Meas.err.	Difficulty	Meas.err.				statistics	Prob.
1	-2.58	0.11	-2.34	0.10	-0.25	0.15	-1.66	3.47	0.06
2	-0.1	0.06	-0.01	0.06	-0.09	0.09	-1.05	1.44	0.23
3	2.37	0.07	2.37	0.07	0	0.1	0	0.1	0.75
4	-3.37	0.15	-3.57	0.16	0.2	0.22	0.93	0.33	0.56
5	-0.4	0.06	-0.55	0.06	0.15	0.09	1.7	1.85	0.17
6	1.57	0.06	1.62	0.06	-0.05	0.09	-0.55	0.26	0.61
7	-1.7	0.08	-1.78	0.08	0.08	0.12	0.67	0.11	0.74
8	-0.65	0.07	-0.65	0.06	0	0.09	0	0.02	0.89
9	1.22	0.06	1.16	0.06	0.06	0.08	0.75	0.94	0.33
10	-1.48	0.08	-1.56	0.08	0.08	0.11	0.75	0.33	0.57
11	-0.17	0.06	-0.12	0.06	-0.05	0.09	-0.58	0.05	0.82
12	0.79	0.06	0.79	0.06	0	0.08	0	0.06	0.81
13	-0.5	0.07	-0.65	0.06	0.16	0.09	1.72	3.32	0.07
14	0.67	0.06	0.62	0.06	0.05	0.08	0.56	1.03	0.31
15	0.81	0.06	0.81	0.06	0	0.08	0	0.52	0.47
16	-1.54	0.08	-1.6	0.08	0.06	0.11	0.49	0	0.95
17	0.43	0.06	0.36	0.06	0.07	0.08	0.8	2.11	0.15
18	1.81	0.06	1.92	0.06	-0.11	0.09	-1.18	1.11	0.29
19	-0.52	0.07	-0.42	0.06	-0.09	0.09	-1.04	1.03	0.31
20	0.45	0.06	0.64	0.06	-0.19	0.08	-2.24	4.47	0.03
21	1.02	0.06	1.06	0.06	-0.04	0.08	-0.54	0.46	0.5
22	-2.58	0.11	-2.46	0.1	-0.13	0.15	-0.83	0.92	0.34
23	-0.19	0.06	-0.07	0.06	-0.12	0.09	-1.42	1.92	0.17
24	2.1	0.07	2.19	0.07	-0.09	0.1	-0.9	1.43	0.23
25	-0.94	0.07	-1.09	0.07	0.15	0.1	1.48	0.97	0.33
26	0.71	0.06	0.71	0.06	0	0.08	0	0.17	0.68
27	1.91	0.06	1.79	0.06	0.12	0.09	1.31	1.14	0.29
28	-0.95	0.07	-1.13	0.07	0.18	0.1	1.82	2.16	0.14
29	0.51	0.06	0.24	0.06	0.27	0.08	3.19	12.07	0
30	1.26	0.06	1.21	0.06	0.04	0.08	0.5	0.82	0.37
31	-1.01	0.07	-1.11	0.07	0.1	0.1	1.01	0.33	0.57
32	-0.5	0.07	-0.44	0.06	-0.06	0.09	-0.68	0.2	0.66
33	1.51	0.06	1.51	0.06	0	0.09	0	0	0.98
34	-1.93	0.09	-1.86	0.08	-0.08	0.12	-0.61	1.58	0.21
35	-0.03	0.06	-0.2	0.06	0.18	0.09	2.06	3.66	0.06
36	1.93	0.07	1.93	0.07	0	0.09	0	0.04	0.85
37	-1.68	0.08	-1.26	0.07	-0.42	0.11	-3.79	16.19	0
38	0	0.06	0	0.06	0	0.09	0	0.07	0.8
39	1.05	0.06	1.18	0.06	-0.13	0.09	-1.55	2.88	0.09
40	-1.09	0.07	-1.09	0.07	0	0.1	0	0	0.97

41	0.07	0.06	0.28	0.06	-0.21	0.09	-2.45	6.67	0.01
42	0.7	0.06	0.96	0.06	-0.27	0.09	-3.1	8.37	0
43	-1.24	0.08	-1.6	0.08	0.37	0.11	3.3	11.5	0
44	0.91	0.06	0.8	0.06	0.11	0.09	1.28	3	0.08
45	1.38	0.07	1.25	0.07	0.14	0.09	1.43	3.22	0.07

Seven items of test form 1 in mathematics are highlighted because at least one statistics value went beyond the critical point: five of them are in favour of girls and two are in favor of boys. Values of item difficulty difference between groups of girls and boys are highlighted in green. None of these values 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. Some insignificant exceedance of critical values can be explained by the large sample size and by random factors. This conclusion is also confirmed by content analysis of items noted as functioning for boys or for girls.

To confirm this statement Figure 4.2 presents item difficulty distributions separately for samples of boys and girls.

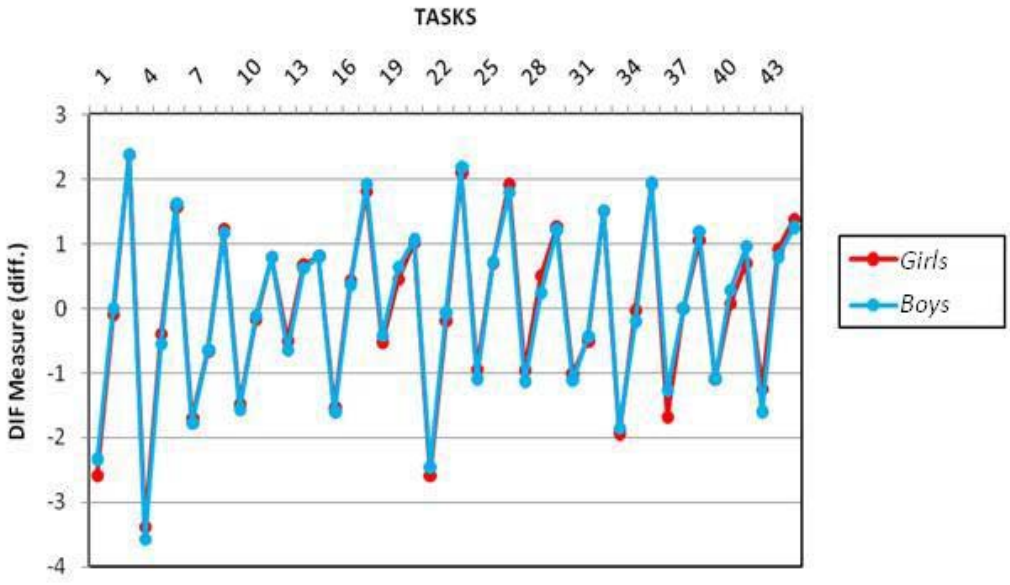


Figure 4.2. Item difficulty distribution diagram for various groups of students (Mathematics, test form 1)

Figure 4.3 shows the characteristic curve of item 37 that has the largest t-statistics value in absolute magnitude (-3.79) and also the largest value of divergence between item difficulty for two groups (-0.42). This item is in favour of girls. In other words, this is the worst item in terms of demonstrating DIF. The characteristic curve of an item demonstrates the probability (in conformity with the model used) of completing this item correctly, depending on examinee ability levels (green line of the diagram). Little crosses on the figure indicate the points of empirical distribution for examinee answers to this item: red crosses show boys' answers and blue crosses are for those of girls. They represent average score for this item in examinee groups with different ability levels.

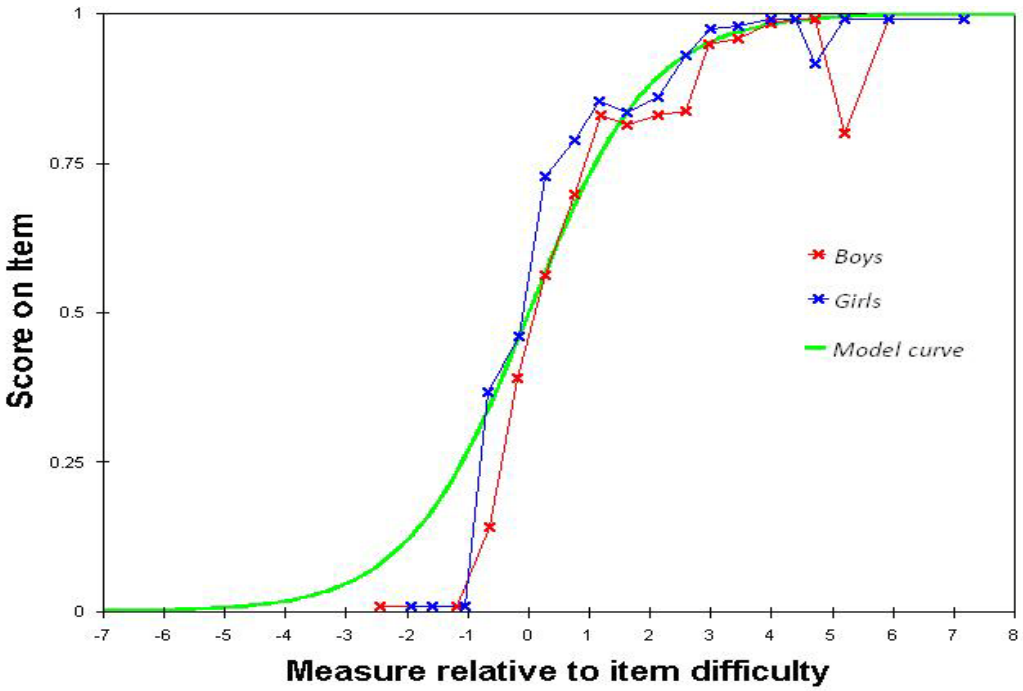


Figure 4.3. Characteristic curve and points of empirical distribution for two participant groups with item 37 (mathematics, test form 1)

As a comparison, Figure 4.4 shows characteristic curve for item 12 for which difficulty estimates in two participant groups are the same and thus this item is completely free of DIF.

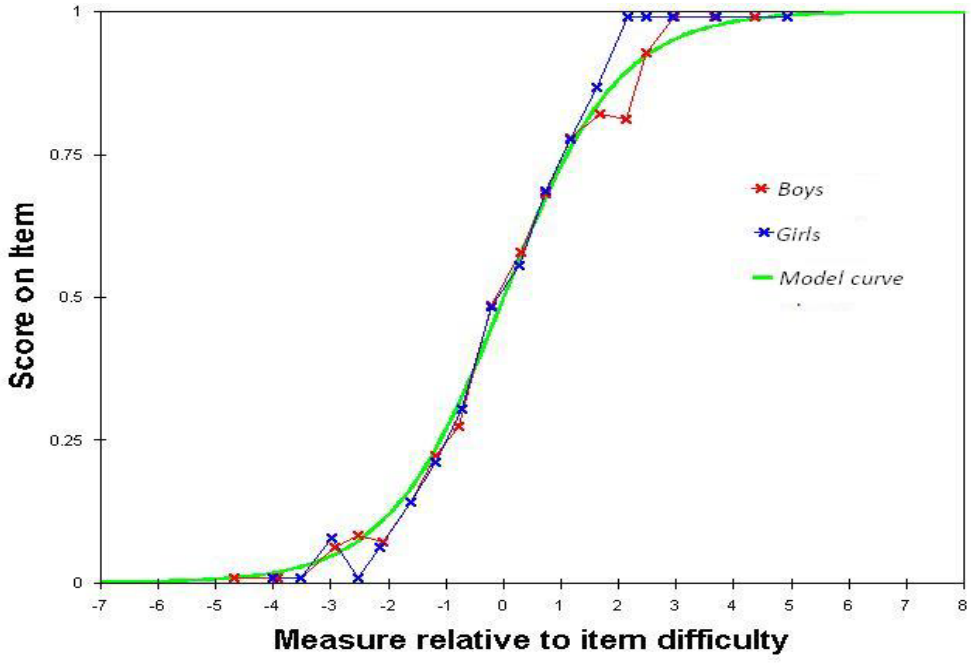


Figure 4.4. Characteristic curve and points of empirical distribution for two participant groups with item 12 (mathematics, test form 1)

As we can see, insignificant differences in the empirical item difficulty shall always be observed. However, study results have shown that in this case such differences are not significant in practical sense.

To confirm the suggestion that small differences in item difficulties will not impact test results for representatives of both groups under review test characteristic curves for both groups are shown in Figure 4.5 (blue line is for boys and red is for girls). Items 3, 8, 12, 15, 26, and 31 were chosen as anchor items between two samples: results of preliminary study have shown them to be free of DIF. The test characteristic curves for boys and for girls practically coincide, which means that the presence of DIF did not impact the examinee test scores and, consequently, his of her proficiency level.

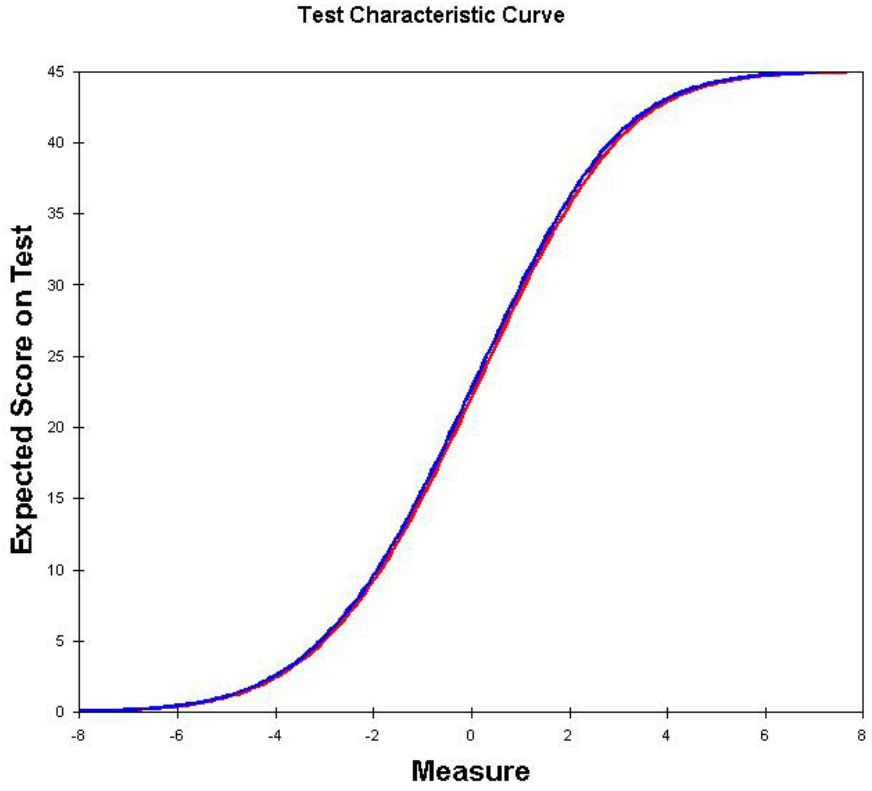


Figure 4.5. Test characteristic curves for the two reviewed student groups (mathematics, test form 1)

Thus results of analysis for test items in mathematics clearly state that these items do not demonstrate different functioning with regard to representatives of different gender.

4.2. DIF analysis of language competence item test

Table 4.3 shows data on the number of test participants of both sexes and on their average score for the language competence test (test form 1). Data of analysis for test form 2 are similar, so they are not shown in this report. Figure 4.6 presents histograms of test score distributions for participants from groups under review: pink is for girls and blue is for boys.

Table 4.1. Test results for both genders (Russian language, test form 1)

Indicator	Females	Males
Sample	1440	1537
Observed raw score: average (SD)	24.7 (8.5)	22.1 (8.8)
Ability estimate: average (SD)	0.34 (1.22)	-0.04 (1.26)

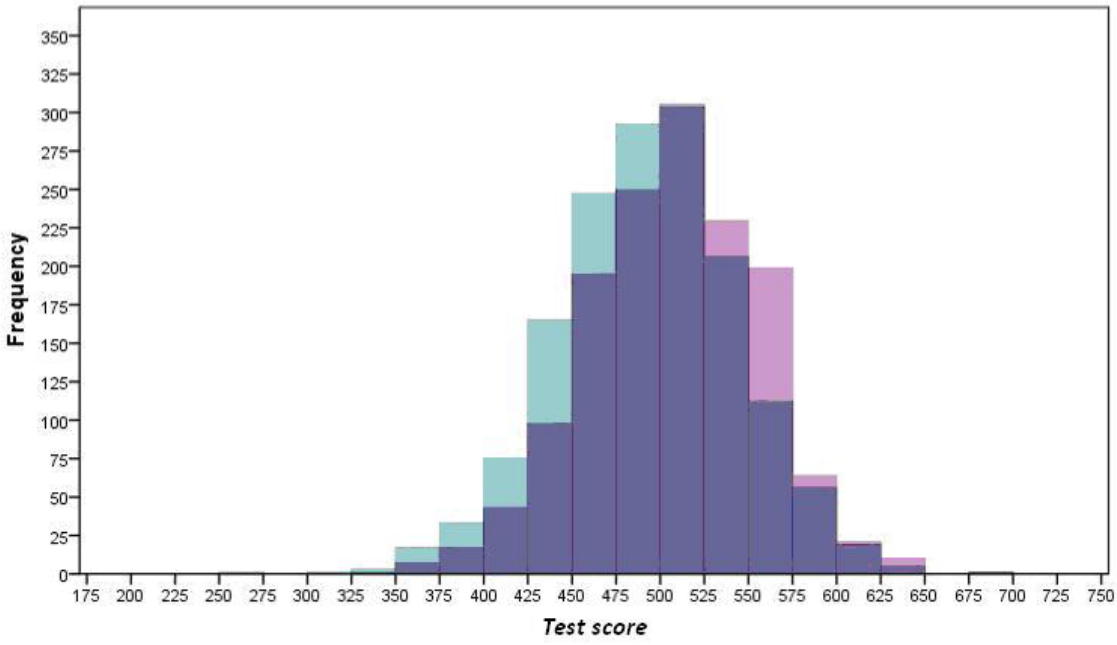


Figure 4.6. Test score distribution for test participants from groups under review (Russian language, test form 1)

In other words, girls demonstrate somewhat better results in Russian language than boys do (the significance of differences will be dealt with in Chapter 6).

DIF analysis results for SAM item tests are shown in Table 4.4. Items appear in the same order as they were coming up in the test. Columns 2 and 3 of the table show item difficulty estimates as well as corresponding measurement errors that resulted from the testing of girls (all data are in logits). The next two columns show same data that were arrived at from the testing of boys. Columns 6 and 7 display the difference in difficulty estimates for the two participant groups and the corresponding measurement error. Next column presents values of t-statistics. Values from the (-2,+2) interval are acceptable for this statistic if the item doesn't demonstrate DIF. Finally, the last two columns show values of Mantel-Haenzel (MH) statistics as well as its significance level (that is, the probability of appearance for such data, under the hypothesis that the item does not demonstrate DIF). Minimum threshold of such probability is taken to be equal to 0.05. DIF statistics values that go beyond its critical values are highlighted pink in the table.

Table 4.4. DIF analysis results (Russian language, test for 1)

#	Girls		Boys		Difference in difficulty	S.E. of difference	t-statistics	Mantel-Hanzel	
	Difficulty	Meas.err.	Difficulty	Meas.err.				Chi-square statistics	Prob.
1	-2,26	0,09	-2,01	0,07	-0.25	0.12	-2.14	3.93	0.05
2	-1,74	0,08	-1,68	0,07	-0.06	0.11	-0.56	0.87	0.35
3	0,86	0,06	0,94	0,06	-0.08	0.09	-0.9	0.33	0.57
4	-3,79	0,16	-3,51	0,12	-0.28	0.2	-1.45	1.92	0.17
5	-0,81	0,07	-0,81	0,06	0	0.09	0	0.22	0.64
6	0,92	0,06	0,92	0,06	0	0.09	0	0.32	0.57
7	-1,27	0,07	-1,34	0,07	0.07	0.1	0.68	0.00	0.99
8	0,84	0,06	0,65	0,06	0.19	0.09	2.21	4.34	0.04
9	1,33	0,06	1,39	0,07	-0.06	0.09	-0.67	0.23	0.63
10	-1,04	0,07	-1,33	0,07	0.29	0.09	3.07	3.71	0.05
11	-0,87	0,07	-0,76	0,06	-0.11	0.09	-1.22	0.46	0.50
12	1,64	0,07	1,57	0,07	0.07	0.09	0.78	0.11	0.74
13	-1,7	0,08	-1,59	0,07	-0.11	0.1	-1.03	1.36	0.24
14	1,01	0,06	1,14	0,06	-0.14	0.09	-1.55	1.90	0.17
15	0,57	0,06	0,48	0,06	0.08	0.08	1	1.27	0.26
16	-1,88	0,08	-1,67	0,07	-0.21	0.11	-1.98	2.73	0.10
17	0,65	0,06	0,56	0,06	0.09	0.08	1.05	1.52	0.22
18	1,06	0,06	1,16	0,06	-0.1	0.09	-1.08	0.51	0.48
19	-1,69	0,08	-1,49	0,07	-0.2	0.1	-1.95	1.57	0.21
20	-0,32	0,06	-0,6	0,06	0.27	0.09	3.19	1.79	0.18
21	0,96	0,06	0,96	0,06	0	0.09	0	1.17	0.28
22	-2,31	0,09	-1,95	0,07	-0.36	0.12	-3.07	9.91	0.002
23	0,15	0,06	0,15	0,06	0	0.08	0	0.59	0.44
24	1,42	0,06	1,35	0,07	0.07	0.09	0.73	2.25	0.13
25	-0,71	0,06	-0,84	0,06	0.13	0.09	1.46	4.82	0.03
26	-0,16	0,06	-0,24	0,06	0.08	0.08	0.89	5.76	0.02
27	1,87	0,07	1,83	0,07	0.04	0.1	0.43	0.10	0.75
28	-2,49	0,1	-2,23	0,08	-0.26	0.13	-2.06	3.06	0.08
29	-0,56	0,06	-0,8	0,06	0.24	0.09	2.73	5.84	0.02
30	1,36	0,06	1,24	0,06	0.12	0.09	1.32	0.41	0.52
31	-0,14	0,06	-0,14	0,06	0	0.08	0	0.05	0.82
32	0,91	0,06	0,63	0,06	0.27	0.09	3.19	7.11	0.01
33	1,68	0,07	1,68	0,07	0	0.1	0	0.92	0.34
34	-0,56	0,06	-0,56	0,06	0	0.09	0	0.42	0.52
35	1,24	0,06	1,24	0,06	0	0.09	0	0.15	0.70
36	3,47	0,11	3,17	0,11	0.3	0.15	1.96	2.26	0.13
37	-1,4	0,07	-1,18	0,06	-0.23	0.1	-2.33	5.44	0.02
38	0,77	0,06	0,82	0,06	-0.05	0.09	-0.57	0.23	0.63
39	1,8	0,07	2,01	0,07	-0.21	0.1	-2.07	3.59	0.06
40	-1,53	0,08	-1,3	0,06	-0.23	0.1	-2.3	2.67	0.10
41	-0,08	0,06	-0,01	0,06	-0.07	0.08	-0.87	0.07	0.80
42	1,47	0,06	1,55	0,07	-0.08	0.09	-0.84	0.02	0.88

43	-1,33	0,07	-1,21	0,06	-0.12	0.1	-1.26	0.67	0.42
44	1,06	0,06	0,97	0,06	0.09	0.09	1.02	0.47	0.49
45	1,05	0,06	1,05	0,06	0	0.09	0	1.53	0.22

Eleven items of test form 1 are highlighted because at least one statistics value went beyond the critical point: six of them are in favor of girls and five are in favor of boys. Values of item difficulty difference between groups of girls and boys are highlighted in green. None of these values 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. Some insignificant exceedance of critical values can be explained by the large sample size and by random factors. This conclusion is also confirmed by content analysis of the items noted as functioning for boys or for girls.

To confirm this statement Figure 4.7 presents item difficulty distributions separately for samples of boys and girls.

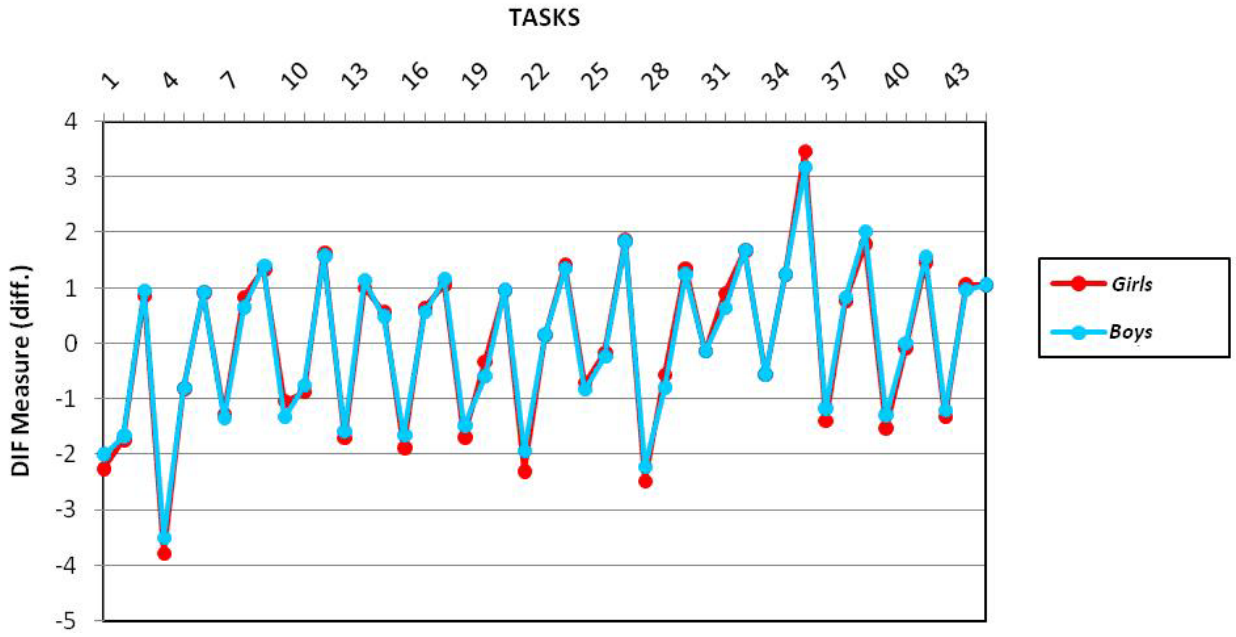


Figure 4.7. Item difficulty distribution diagram for various groups of students (Russian language, test form 1)

Figure 4.8 shows the characteristic curve of item 32 that has the largest t-statistics value in absolute magnitude (3.19) and also the largest value of difference between item difficulty estimates for two groups (0.27). This item is in favor of boys. In other words, this is the worst item in terms of demonstrating DIF. The characteristic curve of an item demonstrates the probability (in conformity with the model used) of completing this item correctly, depending on examinee ability levels (green line of the diagram). Little crosses on the figure indicate the points of empirical distribution for examinee answers to this item: red crosses show boys' answers and blue crosses are for those of girls. They represent average score for this item in examinee groups with different ability levels.

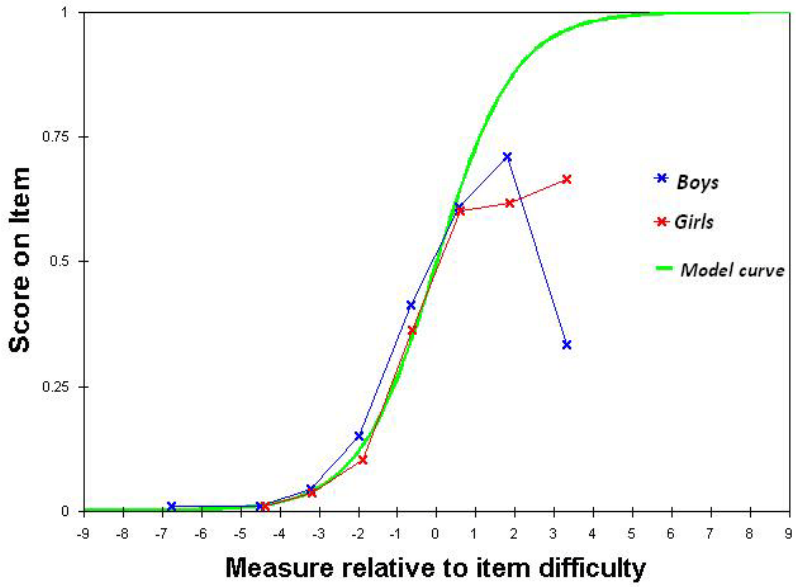


Figure 4.8. Characteristic curve and points of empirical distribution for two participant groups with item 32 (Russian language, test form 1)

As a comparison, Figure 4.9 shows characteristic curve for item 5 for which difficulty estimates in two participant groups are the same and thus this item is completely free of DIF.

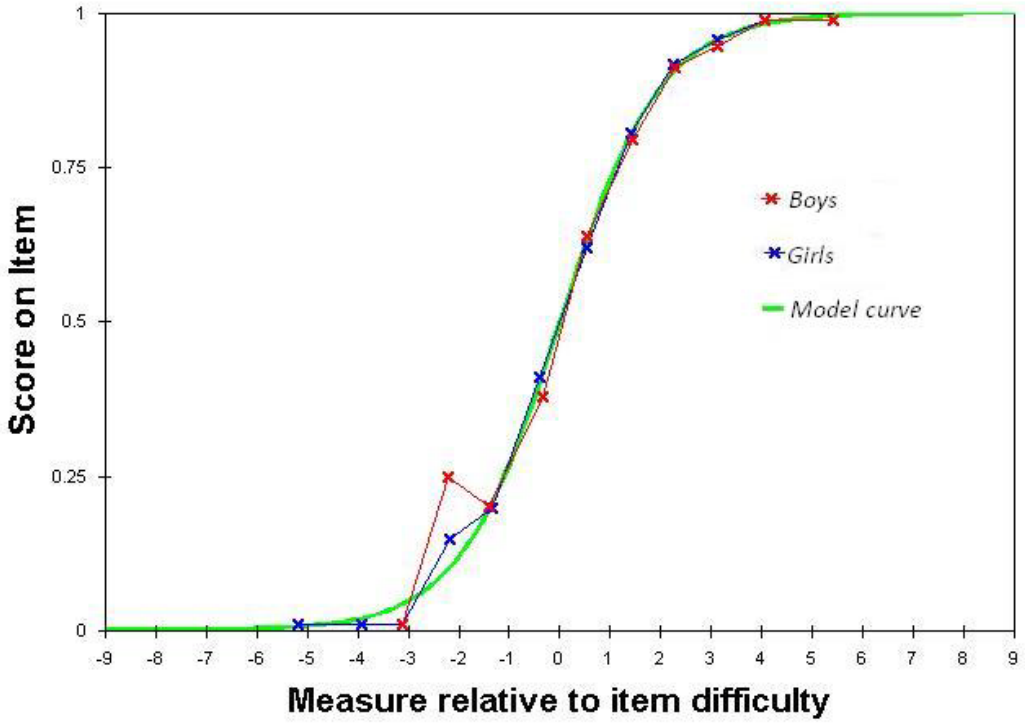


Figure 4.9. Characteristic curve and points of empirical distribution for two participant groups with item 5 (Russian language, test form 1)

As we can see, insignificant differences in the empirical item difficulty shall always be observed. More so, for item 32 a runout is observed at higher score values when students with high ability levels completed the item worse than it was expected. Notably the runout is observed both for girls and for boys, i.e. irrelevant of participant gender. However, there were few of such participants (otherwise, the item would not have been in good agreement with the measurement model and that would have been noted earlier, in Chapter 2). Consequently, this item must be analyzed in terms of clearness of its wording, but it functions satisfactorily from the DIF point of view: as study results have shown, differences in difficulty are not significant in practical sense.

To confirm the suggestion that small differences in item difficulties will not impact test results for representatives of both groups under review test characteristic curves for both groups are shown in Figure 4.10 (blue line is for boys and red is for girls). Items 12, 13, 17, 35, 41, and 42 were chosen as anchor items between two samples: results of preliminary study have shown them to be free of DIF. The test characteristic curves for boys and for girls practically coincide, which means that the presence of DIF did not impact the examinee test scores and, consequently, his of her proficiency level.

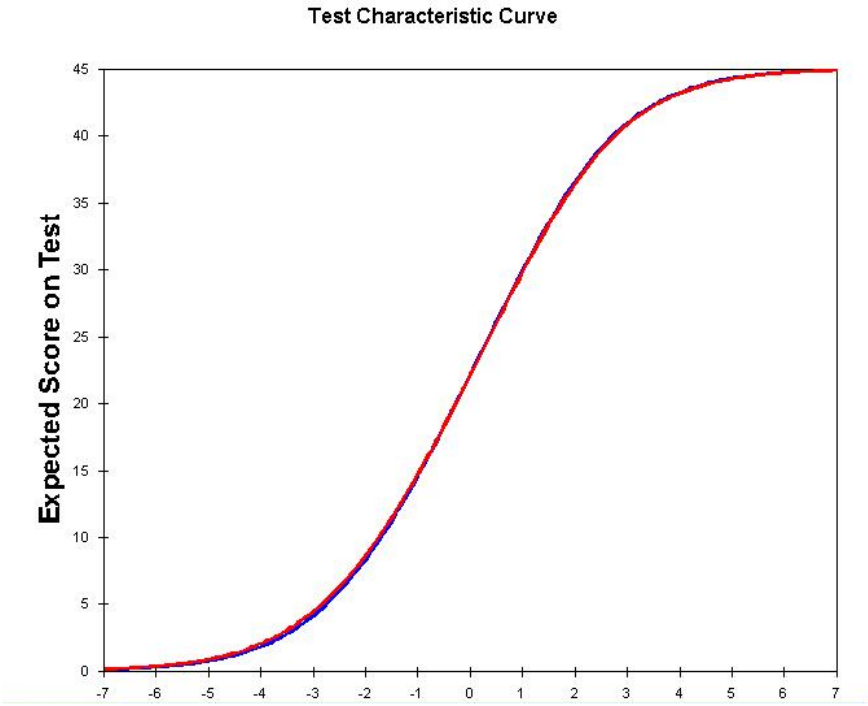


Fig. 4.10. Test characteristic curves for the two reviewed student groups (Russian language, test form 1)

Thus results of analysis for test items in Russian language state that these items do not demonstrate different functioning with regard to representatives of different gender.

Literature

1. Smith, R. M., and Suh, K. K. (2003). Rasch fit statistics as a test of the invariance of item parameter estimates. Journal of Applied Measurement, 4, 153-163.
2. Smith. R. M. (2004). Detecting item bias with the Rasch model. Journal of Applied Measurement, 5(4), 430-449
3. Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. Journal of Experimental Education, 72, 221-261.
4. Wang, W.-C. (2008). Assessment of differential item functioning. Journal of Applied Measurement, 9(4), 387-408
5. Dorans N.J. (1989). Two New Approaches to Assessing Differential Item Functioning: Standardization and the Mantel-Haenszel Method. Applied Measurement in Education, 2(3), 217-233.
6. Linacre J. M. (2011). A User's Guide to WINSTEPS. Program Manual 3.71.0. (http://www.winsteps.com/a/winsteps.pdf).

5. Estimation of examinees

Estimation of examinees shall happen by using two approaches – norm-referenced and criterion-referenced – that are combined in keeping with the modern test theory.

Within the framework of the norm-referenced approach each test participant is assigned a test score after mathematical treatment of the test results (the dichotomous Rasch model is used as a test model). Test scores of all test participants are on the same metric scale (), regardless of the time of test administration and specific set of test items completed. For reporting test results a 1000-point scale is used with a mean at about 500 and standard deviation of 50. In the future all scores arrived at by practical use of the SAM tool shall also be put onto this scale, and that would provide an opportunity of achievement comparison for each participant over time, i.e. monitor the academic progress of students.

To implement criterion-referenced approach a graded version of achievement scale was developed based on integral scores of test participants and on threshold values which separate all participants into groups corresponding to various proficiency levels of achievement. This provides an opportunity for qualitative assessment of curriculum acquisition via indicating the leading type of orientation for solving tasks of different levels. Methods of creating a common scale and establishing benchmarks are specified in [1].

It was accepted by test developers that a level can be considered as acquired if at least 50% of items at this level are completed correctly. (This refers, of course, to probability estimates.)

Four proficiency levels were identified which correspond to the following content criteria:

Proficiency level lower than level 1 – even level 1 was not acquired: students belonging to this group can complete fewer than 50% of level 1 items.

Proficiency level 1 – level 1 is acquired: students belonging to this group can complete at least 50% of the level 1 items.

Proficiency level 2 – level 2 is acquired: students belonging to this group can complete at least 50% of level 2 items.

Proficiency level 3 – level 3 is acquired: students belonging to this group can complete at least 50% of the level 3 items.

For mathematics, the following benchmarks were established: transfer to proficiency level 1 – 430 points; transfer to proficiency level 2 – 500 points; transfer to proficiency level 3 – 570 points. A benchmark indicates the lower limit of a corresponding proficiency level. If a participant test score went over this benchmark, the corresponding proficiency level is regarded as acquired. This means that there is a 50% probability that a given test participant will be able to complete more than 50% of items in this level. All participants whose test results are under a given value are considered as someone who did not acquire this proficiency level, not to speak of all the subsequent ones. Interpretation of proficiency levels provided through the example of mathematics test is shown in Figure5.1.

Mathematical competence scale

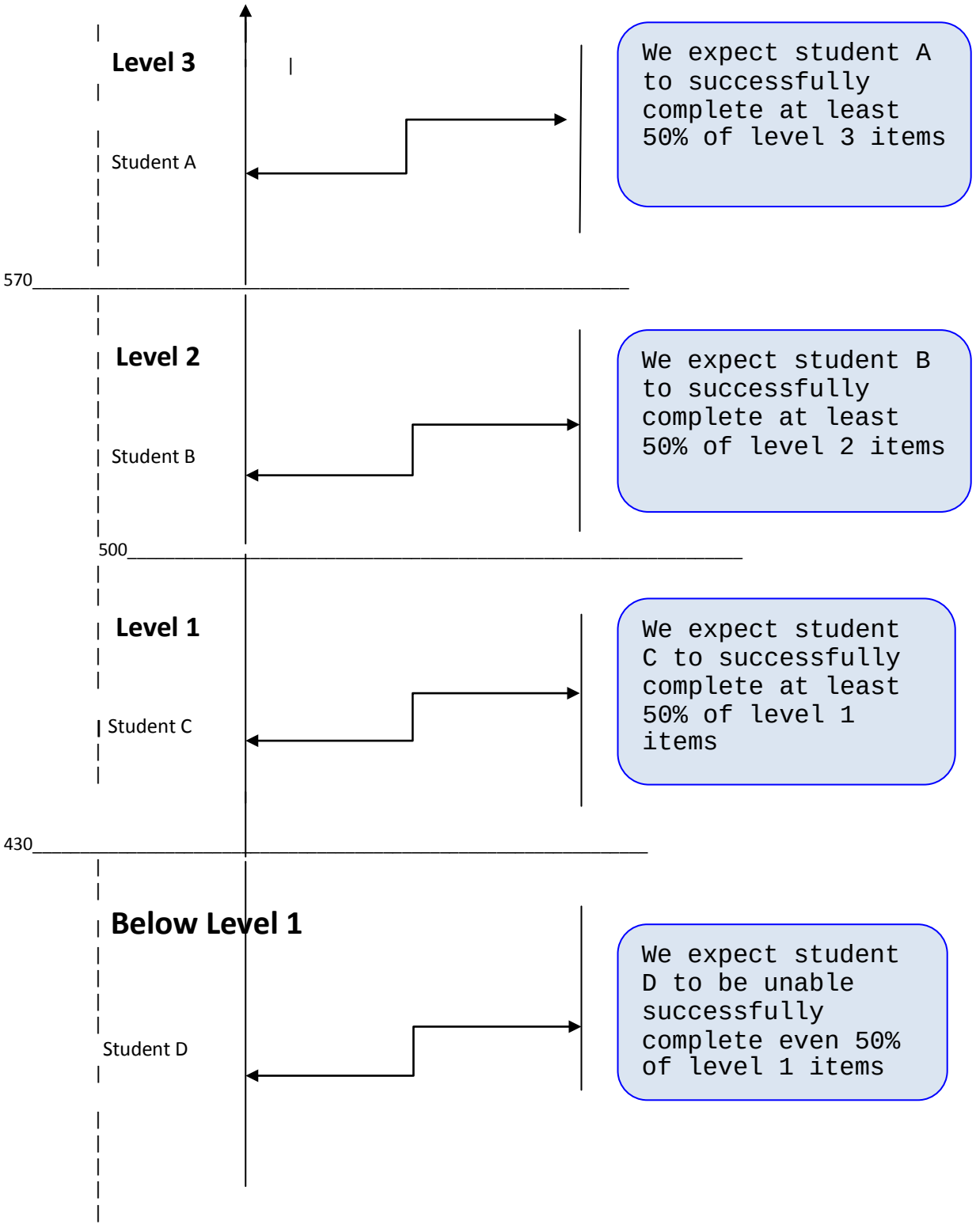


Figure5.1. Categorization of the competence scale in mathematics

Consequently, each test participant gets not only an integral overall test score, but also a proficiency level which he or she could achieve. If a student did not reach even proficiency level 1, that means that he (or she) was able to complete fewer than 50% of level 1 items. In this case there is practically zero probability that he would be able to complete items from levels 2 and 3. If a student did reach proficiency level 1 (but did not reach higher proficiency levels),

that means that he (or she) was able to complete over 50% of level 1 items. In this case the student can complete fewer than 50% of level 2 items and there exists a very insignificant probability that he would be able to complete some level 3 items.

Further on, if a student could reach proficiency level 2 (but did not reach proficiency level 3), that means that he (or she) was able to complete not fewer than 50% of level 2 items. In this case the student can complete most of level 1 items, but fewer than 50% of level 3 items.

And, finally, if a student could reach proficiency level 3, he (or she) could complete over 50% of level 3 items. In this case this student would be able to complete any level 1 item and most of level 2 items.

Such a separation of test participants according to proficiency levels allows an interpretation of results in terms of level-sensitive assessment model, i.e. it can provide qualitative interpretation to the overall (integral) score.

Figure 5.2 represents variable map for one of test forms in mathematics (the structure of variable map was described in Chapter 2).

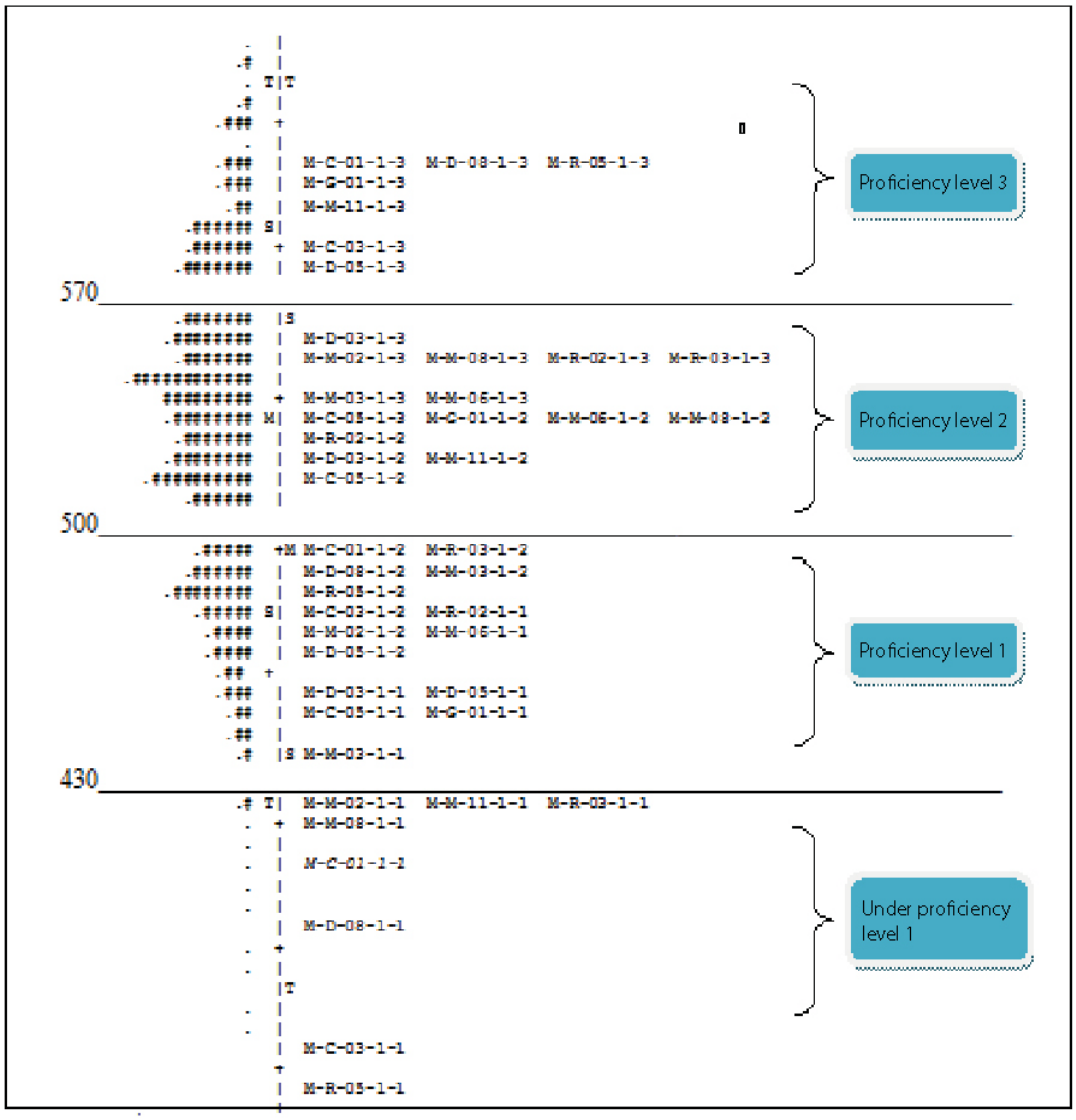


Figure 5.2. Variable map (mathematics, test form 1)

Horizontal lines on the map indicate boundaries between proficiency levels corresponding to benchmarks. Thus most of test participants are at proficiency levels 1 and 2,

with level 2 dominating. This corresponds to the expectations of the assessment model developers that by the end of the primary school the curriculum is acquired at the second (reflective) level.

Test participants in Russian language testing will be dealt with in a similar manner.

Establishing benchmarks and separating test participants into groups will help provide a qualitative interpretation to the overall (integral) score. Aiming to confirm the resulting conclusions regarding ability levels of test participants, a study was undertaken to validate the reported benchmarks. There were following directions of study:

- Checking whether the test model in use corresponds separately to each participant group (at each proficiency level);
- Establishing benchmarks by some other method and comparing results (to this end, benchmarks were established the Angoff method which is often used in traditional testing);
- Comparing item difficulty indexes of tests as expected according to the model and empirical item difficulty of these items for the test participants who belong to various proficiency levels;
- Assessing benchmark measurement errors.

Study results are presented in [2]. The main conclusion of the study is that benchmarks were set valid and can be used for qualitative interpretation of test results.

Established benchmarks are based on the three-level SAM testing model, and they allow each test participant to identify the acquired proficiency level. There are three of such proficiency levels, corresponding to three model levels. However, the degree of the proficiency can be different within a proficiency level: a student may have just acquired a certain level, to barely “hang” to it, or he may acquire it, i.e. feel confident about having had mastered it. Let us remind ourselves that a proficiency level is considered acquired when at least 50% of the corresponding level items were completed correctly. Let us agree that a proficiency level should be considered as not only acquired, but also mastered (assimilated) if at least 75% of the current level items were completed correctly. Such an interpretation of test results will help identify students who are already confident about their position at the current proficiency level and ready to transfer to the next level of proficiency. The illustration of the process of achieving and acquiring proficiency levels is shown in Figure. 5.3.

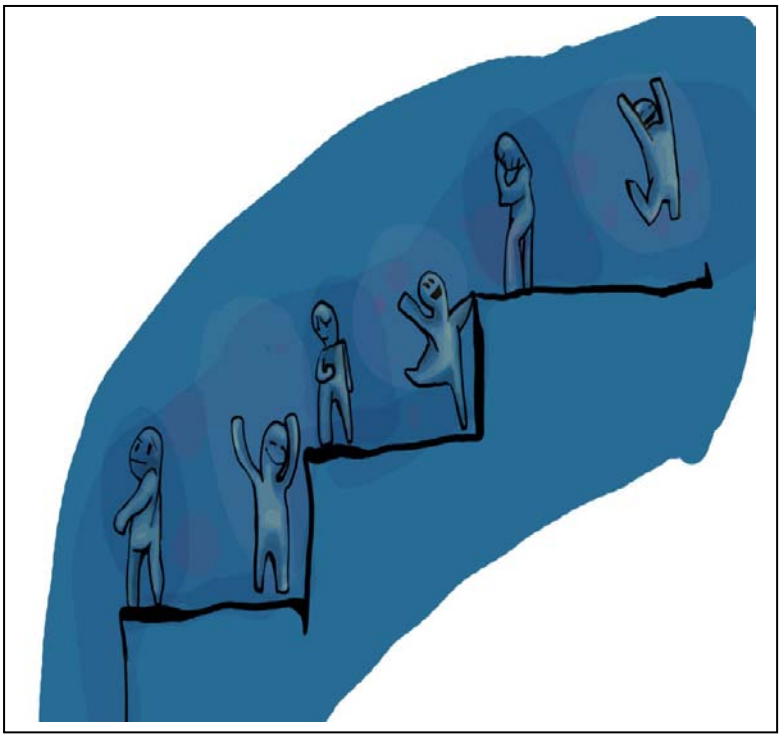


Figure 5.3. Illustration of the various degrees of reaching proficiency levels

Figure 5.4 shows the distribution of the participants of 2012 pilot testing in mathematics (a sample containing around 6000 students) over proficiency levels: 2% could not reach level 1, and half of all participants stay at level 2.

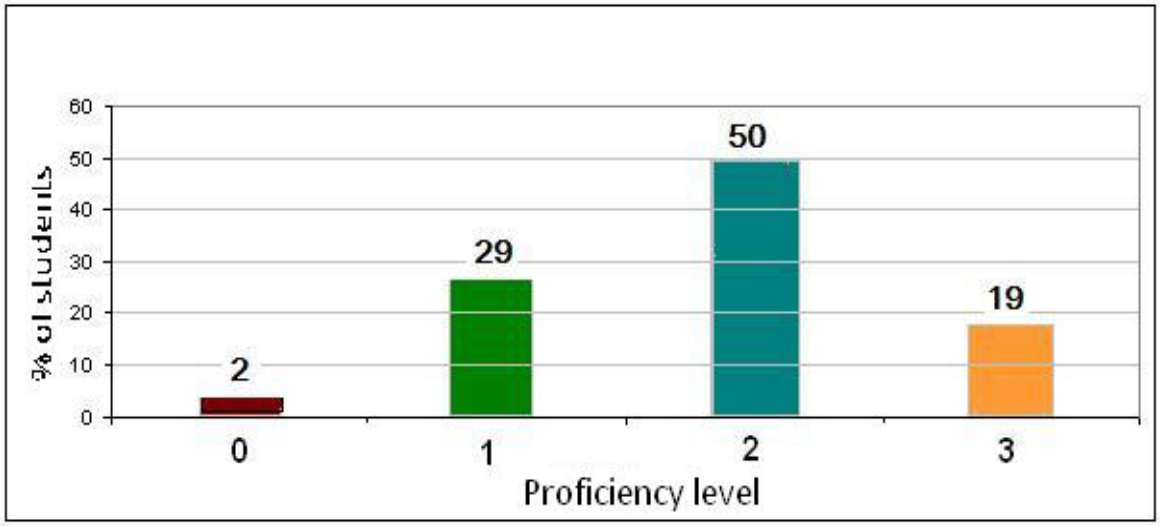


Figure 5.4. Distribution of the test participants over proficiency levels (mathematics)

Table 5.1 shows the distribution of examinees over proficiency levels taking into account the degree of proficiency level assimilation.

Table 5.1. Test examinee distribution over proficiency levels (mathematics)

Proficiency level	Total (students)	%	Degree of assimilation	Total (students)	% of the total	% of proficiency level
1	1725	29	Acquired	665	11	39
			Assimilated	1060	18	61
2	2974	50	Acquired	1864	32	63
			Assimilated	1110	18	37
3	1124	19	Acquired	760	13	68
			Assimilated	364	6	32

Thus, we can see that of 2974 students who are at proficiency level 2 only 37% could assimilate it. Similarly, of 1124 students who are at proficiency level 3 only 32% could assimilate it. As far as level 1 is concerned, the majority of students (61%) who are already on it did assimilate it.

Figure 5.5 shows in a diagram the student distribution over proficiency levels taking in to account the degree of assimilation of proficiency levels. This interpretation of test results can be useful for a more detailed assessment of student achievement.

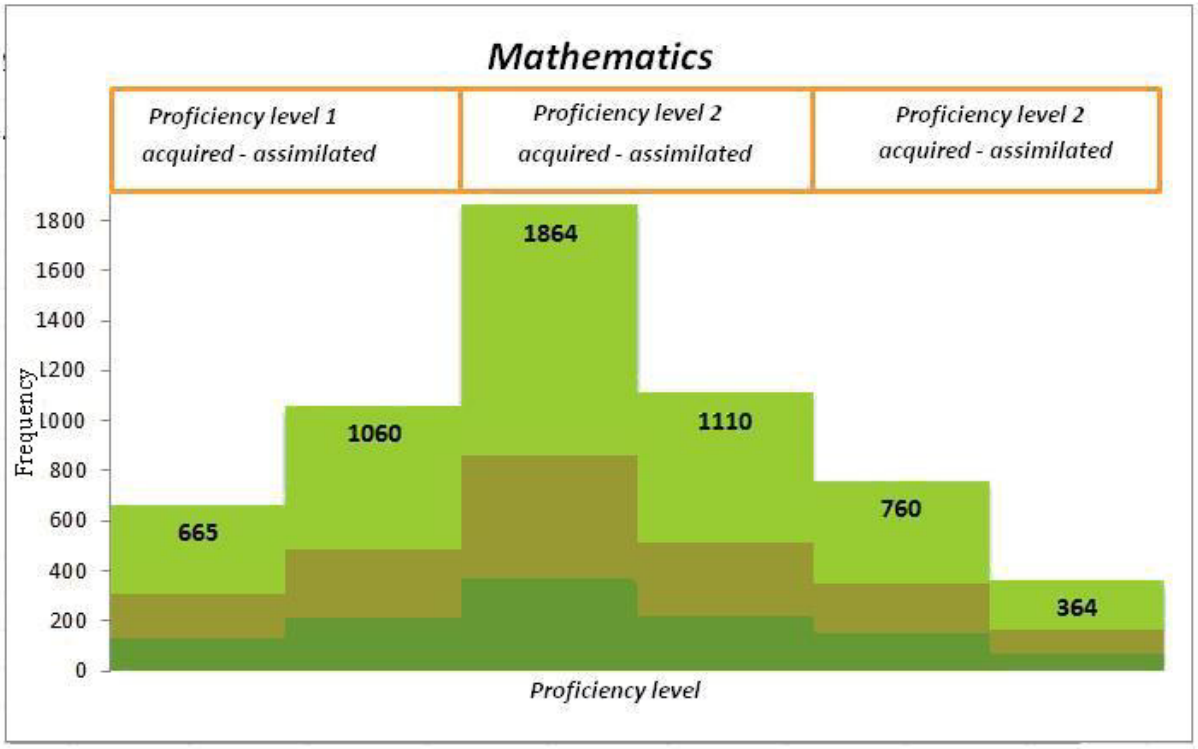


Figure 5.5. Examinee distribution over proficiency levels taking into account the degree of proficiency level assimilation (mathematics)

Figure 5.6 shows the distribution of 2012 pilot testing in Russian language (a sample containing around 6000 students) over proficiency levels. Consequently, 11% could not reach level 1, and approximately equal numbers of study participants stay at level 1 and 2.

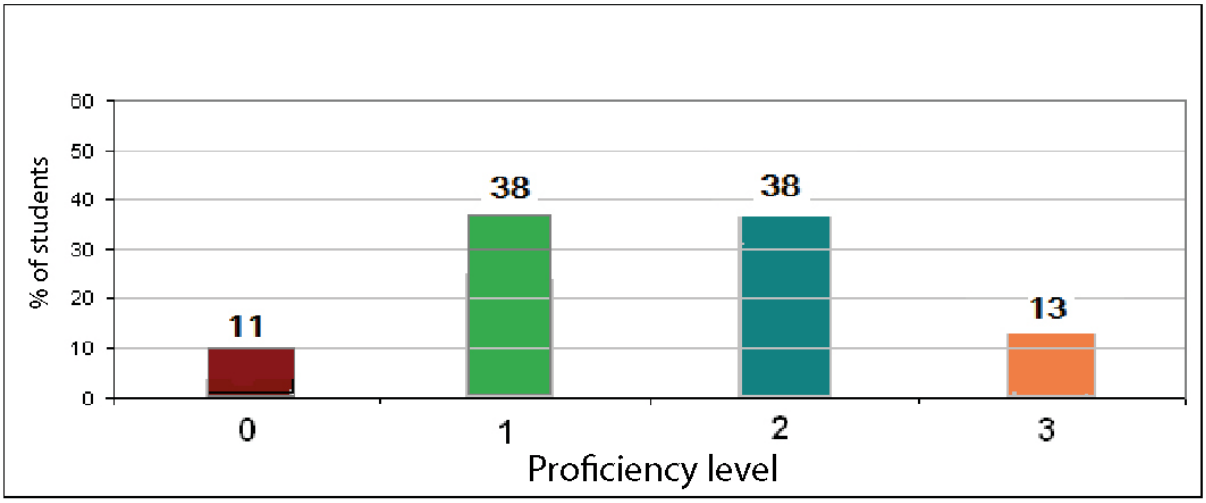


Fig. 5.6. Distribution of the test participants over proficiency levels (Russian language)

Table 5.2 shows data of the distribution of examinees over proficiency levels taking into account the degree of proficiency level assimilation. The table does not include 631 students (11% of the total sample) who could not reach proficiency level 1.

Table 5.2. Test examinee distribution over proficiency levels (Russian language)

Proficiency level	Total (students)	%	Degree of assimilation	Total (students)	% of the total	% of proficiency level
1	2287	38	Acquired	1278	21	56
			Assimilated	1013	17	44
2	2271	38	Acquired	1898	33	84
			Assimilated	373	6	16
3	757	13	Acquired	664	11	88
			Assimilated	93	2	12

Thus, we can see that of 2271 students who are at proficiency level 2 only 16% could assimilate it (that is, they deal confidently with it); this is a much smaller number than in mathematics. Similarly, of 757 students who are at proficiency level 3 only 12% could assimilate it. As far as level 1 is concerned, there are 2287 students (38%) at it. Most of them (56%) could not assimilate it yet.

Figure 5.7 shows in a diagram the student distribution over proficiency levels taking in to account the degree of assimilation of proficiency levels.

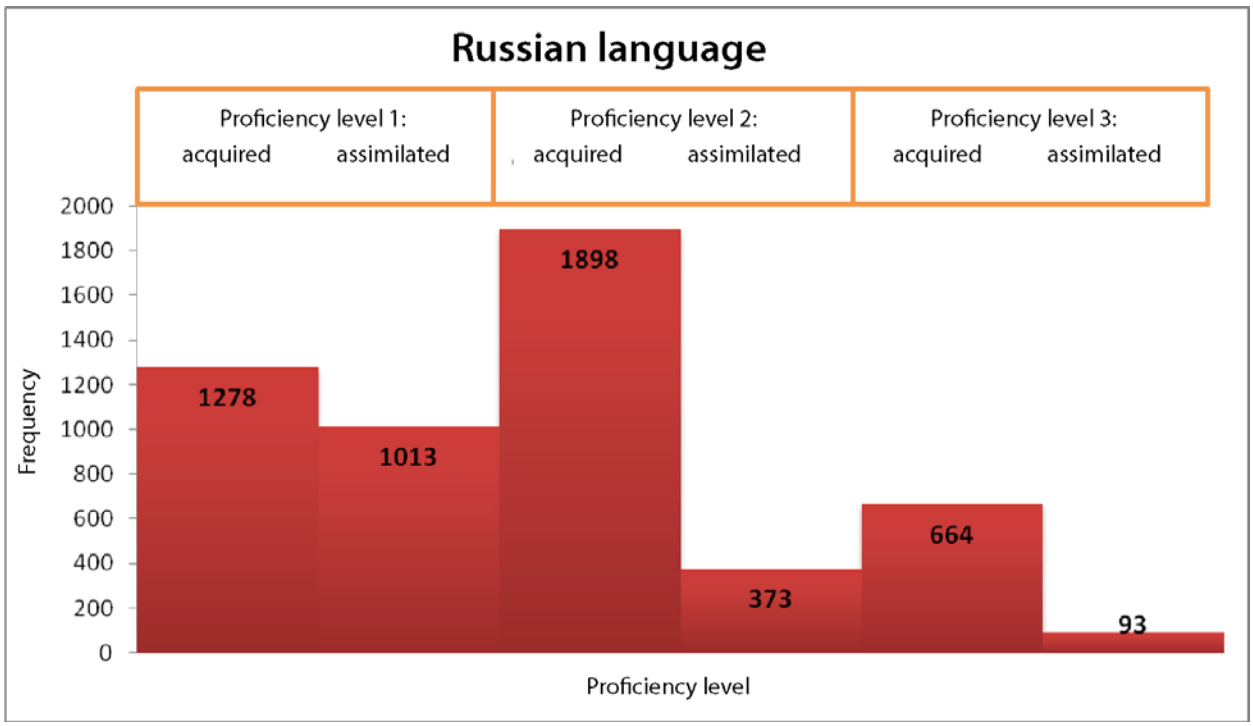


Figure 5.7. Examinee distribution over proficiency levels taking into account the degree of proficiency level assimilation (Russian language)

Literature

1. SAM (School Achievement Monitoring): Инструмент мониторинга учебных достижений школьников // *под ред. Нежнова П.Г., Кардановой Е.Ю.*, 2011, 104 с.

2. *Kardanova E., Gaponova N.* New technologies of assessment: School Achievements Monitoring Toolkit // The paper presented at EDULEARN12, Barcelona, Spain, 2012. http://library.iated.org/view/KARDANOVA2012NEW

6. Preliminary analysis of test results

SAM test results can be represented using three key groups of indicators: integral scores (raw score and test score), proficiency levels, and three-dimensional profiles.

Raw score is the sum of points achieved by a test participant across all items. Test score is the final score result that a student gets after the mathematical treatment of the raw score aimed at obtaining estimates on the common metric scale. This scale is common for all test takers irrespective of the time frame of testing and a specific set of items that were dealt with. Test results under this indicator are presented on a 1000-point scale designed for each school subject test via special studies of the basic sample of students.

Beside integral test score, each test participant gets a proficiency level which this student could acquire. A total of 4 categories of achievement was allocated: Under Level 1 (for students who could not acquire even Level 1 of mediation), Proficiency Level 1 (for those who acquired level 1), Proficiency Level 2 (for those who acquired level 2), and Proficiency Level 3 (for those who acquired level 3). The separation of test participants into proficiency levels makes it possible to interpret the results in terms of level-sensitive assessment model, i.e. it can provide qualitative interpretation to the overall (integral) score.

Additionally, the SAM test enables to obtain a structural characteristic of the assessed competency: its three-dimensional profile. The profile will be set up using raw (percentage) scores that a student got for each level and it shows the share of material assimilated at each of the three levels, i.e. constituents of the integral score that was taken into three subscales. It is exactly the profiles and their changes during the monitoring process that provides most consistent and meaningful picture of the process subject content assimilation.

Preliminary analysis of test results was run on the basis of SAM pilot testing in one of the regions of the Russian Federation in spring 2012. This pilot testing was characterized by the fact that almost all primary school leavers (4th-graders) of this region were tested. Table 6.1 represents general data regarding this SAM pilot testing.

Table 6.1. General testing information

	Number of students	Number of schools	Number of classes	Number of communities
Mathematics	4639	192	293	136
Russian language	4649	192	293	136

Table 6.2 shows data on the distribution of students in the region over proficiency levels depending on the school subject tested. Figure 6.1 shows the same distribution as a diagram. We can see that mathematics results are somewhat better than Russian language results. Proficiency level 2 dominates in mathematics (over 50% of students could acquire it) while in Russian language an approximately equal number of students (some 40%) remain in proficiency levels 1 and 2. Proficiency level 3 does not show too many students: 12% in Russian language and 18% in mathematics.

Table 6.2. Student distribution in the region over proficiency levels

		Under level 1	Level 1	Level 2	Level 3	Total
Mathematics	Number	106	1305	2390	838	4639
	%	2%	28%	52%	18%	100%
Russian language	Number	494	1801	1791	563	4649
	%	11%	39%	38%	12%	100%

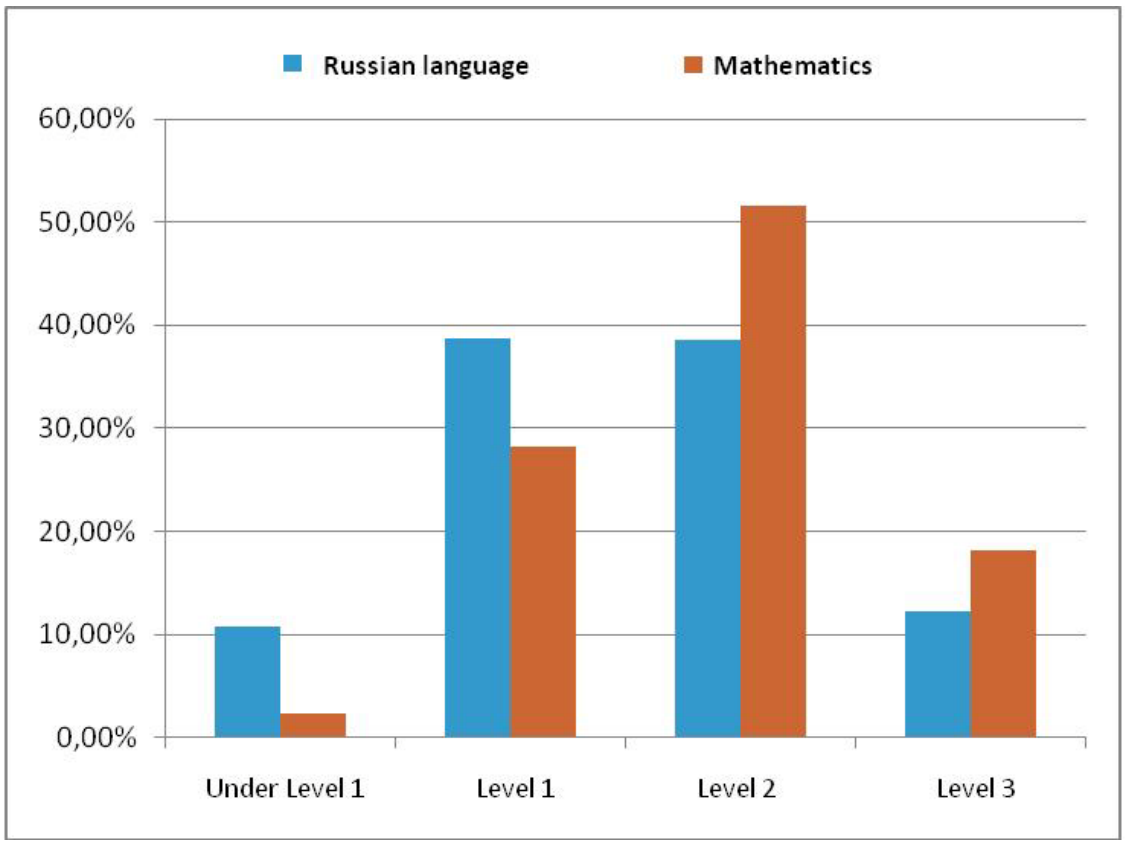


Figure 6.1. Student distribution in the region over proficiency levels depending on the school subject tested

Table 6.3 shows average percentages of correctly completed items for different levels, depending on the school subject. In Figure 6.2 the same information is represented in a graphic manner, as a chart.

Table 6.3. Average percentage of completed items for different levels

	Level 1 items	Level 2 items	Level 3 items
Mathematics	86	61	32
Russian language	76	49	28

We can see that mathematics results are better than the Russian language results for all levels. For level 3 items, however, this advantage is not significant while for level 1 and 2 items it is quite sizeable. In general, primary school leavers can handle level 1 items well (the material was acquired formally), but only half of the material was acquired on the second, reflective

level (with comprehension) and only about one third of the material was acquired at the functional level.

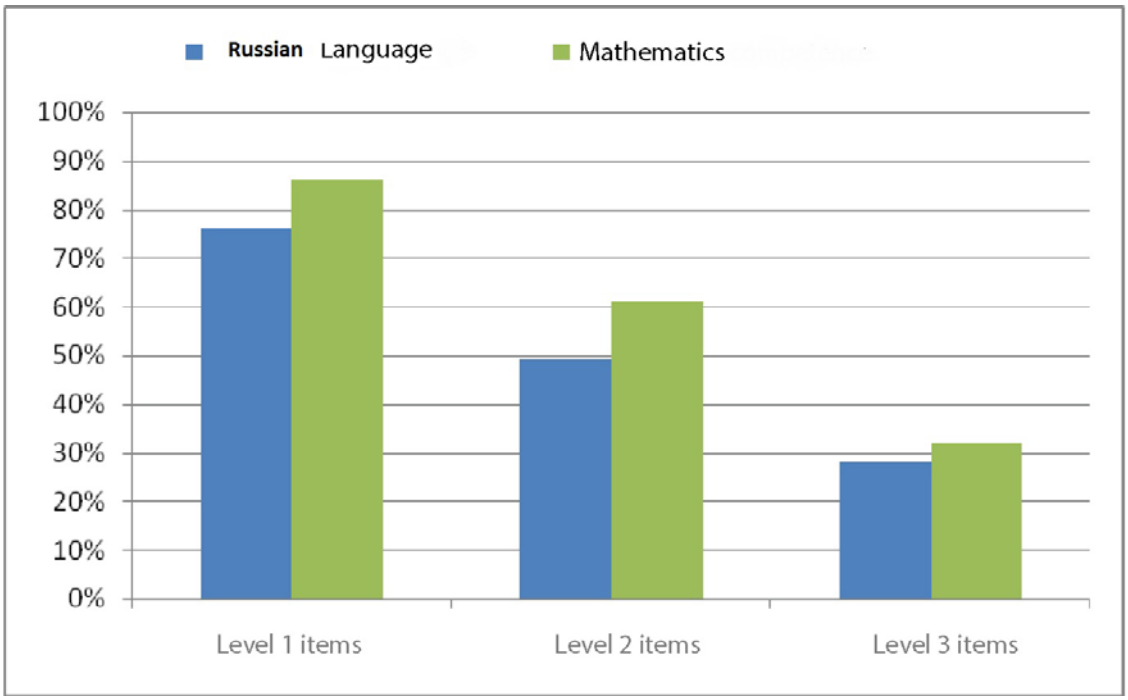


Figure 6.2. Percentage of correctly completed items from different levels depending on the school subject

For the purpose of comparative judgment on academic progress it is possible to use norms of test completion. They are a set of indicators which were established empirically following the testing results of a clearly defined sample of examinees. This sample must be representative with regard to the entire assembly of persons for whom the test was designed, and it should also be quite large in its size. In this case testing was encompassing all students who were leaving primary schools in the region under study, so this allows for establishing test completion norms for this region (i.e. regional norms).

Norms can be established both on the individual and on group levels. At the individual level the following indicators can be observed:

- Statistically average individual norm is average indicator and standard deviation of test completion by sample students (Table 6.4).
- Percentile norms are indicators based on the percentage of test participants who completed their test at a certain level.

Table 6.4. Statistically average individual norms

	Average mean	Standard deviation
Mathematics	522	49
Russian language	499	50

Offering percentile norms can boil down to, for example, displaying the test completion indicator that corresponds to 90th percentile (which means that 90% of students completed the test worse) and to 10th percentile (90% of students completed the test better) – see Table 6.5.

Table 6.5. Percentile individual norms

	10^{th} percentile	90^{th} percentile
Mathematics	459	581
Russian language	433	557

That is to say, if the test score of a mathematics test participant is higher than 581, this student finds himself in the of the 10% of the best ones; and if participant score is under 459, the student is in the group of 10% of the worst ones. Percentile norms in Russian language are interpreted in a similar manner. Let us note that the norms in mathematics (both statistical average and percentile norms) are higher than the corresponding indicators for Russian language tests, and this means that mathematics results are higher than those in Russian language in this specific sample of students.

Additionally, on top of statistically average data on individual levels, it is possible to select average results across schools. At this level, the following indicators can be analyzed:

- Statistically average group norm is average indicator across schools and its standard deviation (Table 6.6).
- Socio-cultural norm is an average indicator for the group of leading schools (Table 6.7).

Table 6.6. Statistically average group norms

	Average mean	Standard deviation
Mathematics	517	34
Russian language	499	36

Leader group is a group of schools (for example, 25% of the total number of schools) which have the highest results. The leading group of schools serves as reference for the whole educational community, i.e. it provides a socio-cultural norm as a realistic norm of “tomorrow” and serves as an additional basis for the estimation of test results.

Table 6.7. Socio-cultural norms

	Test score
Mathematics	561
Russian language	543

These norms are statistical, they help to interpret test results from the viewpoint of their comparability among themselves.

6.1. Analysis of test results in mathematics

Figure 6.3 shows a histogram for the test score distribution of mathematics test participants. The average value of the score is 522, standard deviation is 41. In general, test score distribution is close to normal.

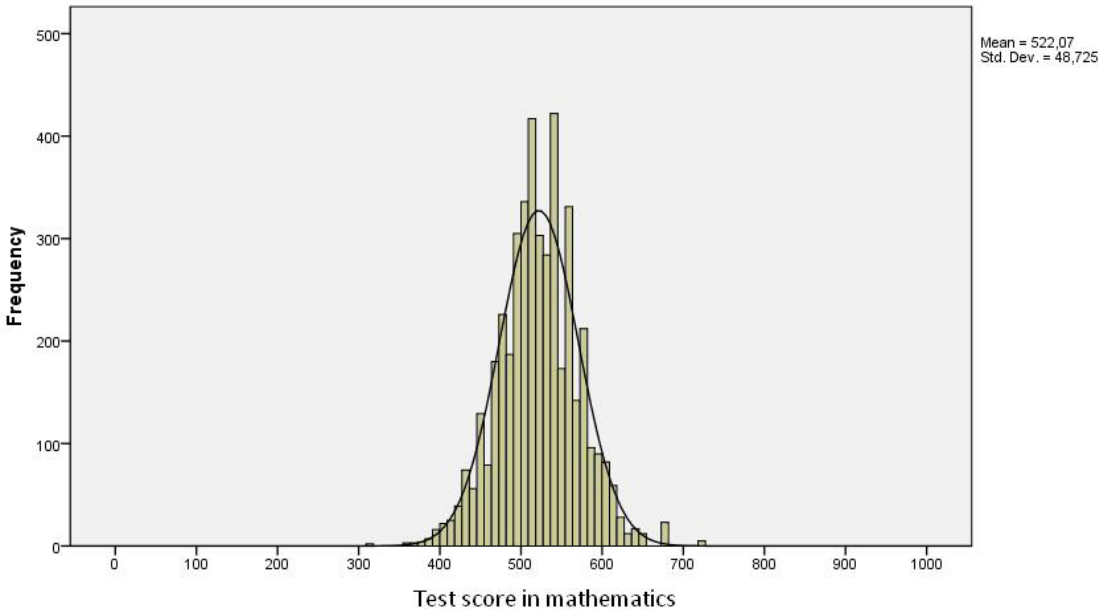


Figure 6.3. Distribution ot examinee test scores (mathematics)

Fig. 6.4 displays the distribution of mathematics test participants across proficiency levels.

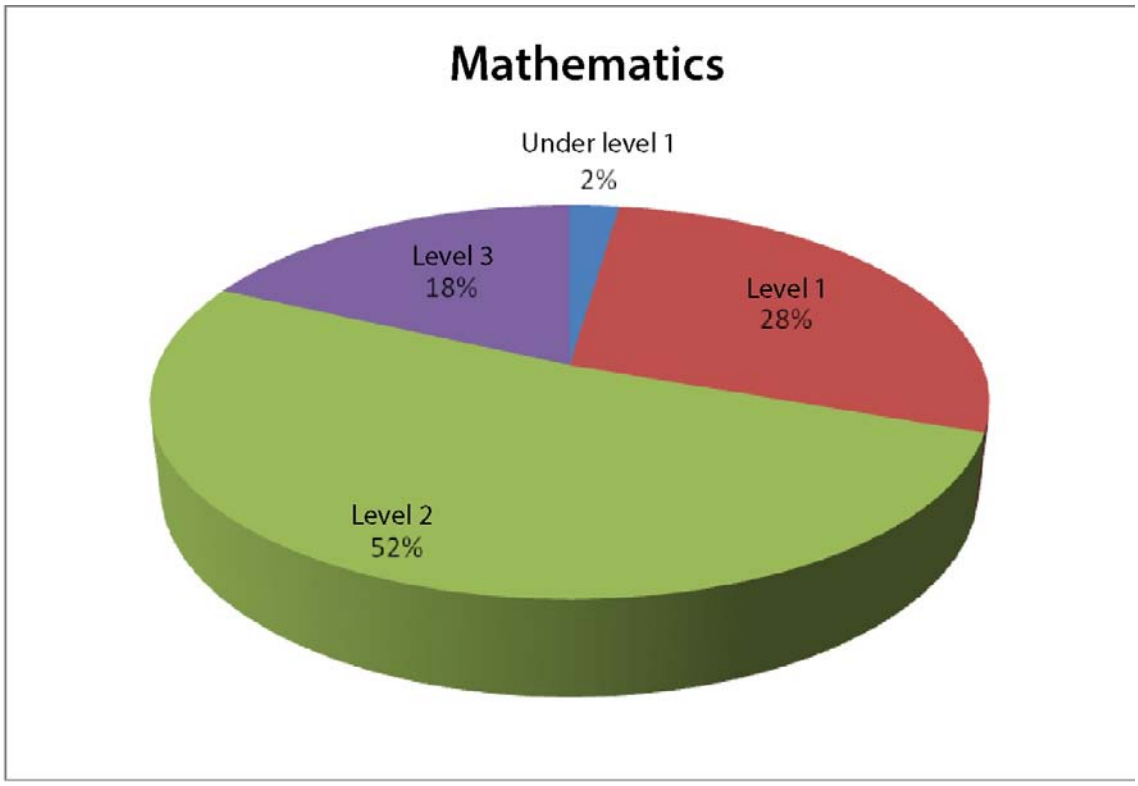


Fig. 6.4. Test participants distribution across proficiency levels (mathematics)

Fig. 6.5 shows the success profile in mathematics for the given student sample (average percentages of successfully completed items as a function of level).

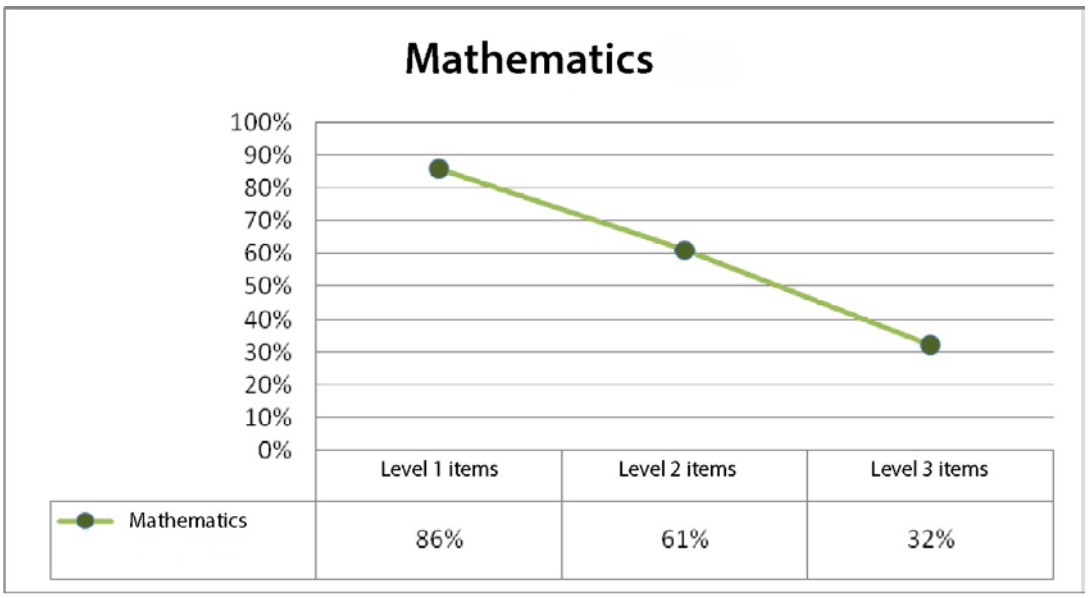


Fig. 6.5. Mathematics profile for the current student participant sample

The profile can be read quite easily. If we assume that the test covers, in due proportion, main curriculum sections, then level 1 scale tells us that the main part of that curriculum was formally acquired (over 80% of level 1 items were completed); the second scale shows that about 60% was acquired reflectively (with comprehension), and the last scale informs us that around one third was acquired functionally.

Figure 6.6 shows the test participant distribution across proficiency levels in various schools of a given region (on the diagram 20 schools are represented, that were randomly chosen from the total of 192). The horizontal axis features percentage of students in each proficiency level, and the vertical axis shows schools. For each school average test score of students from this school is shown in parentheses. Schools were ranged in order of increasing average mean for general test scores.

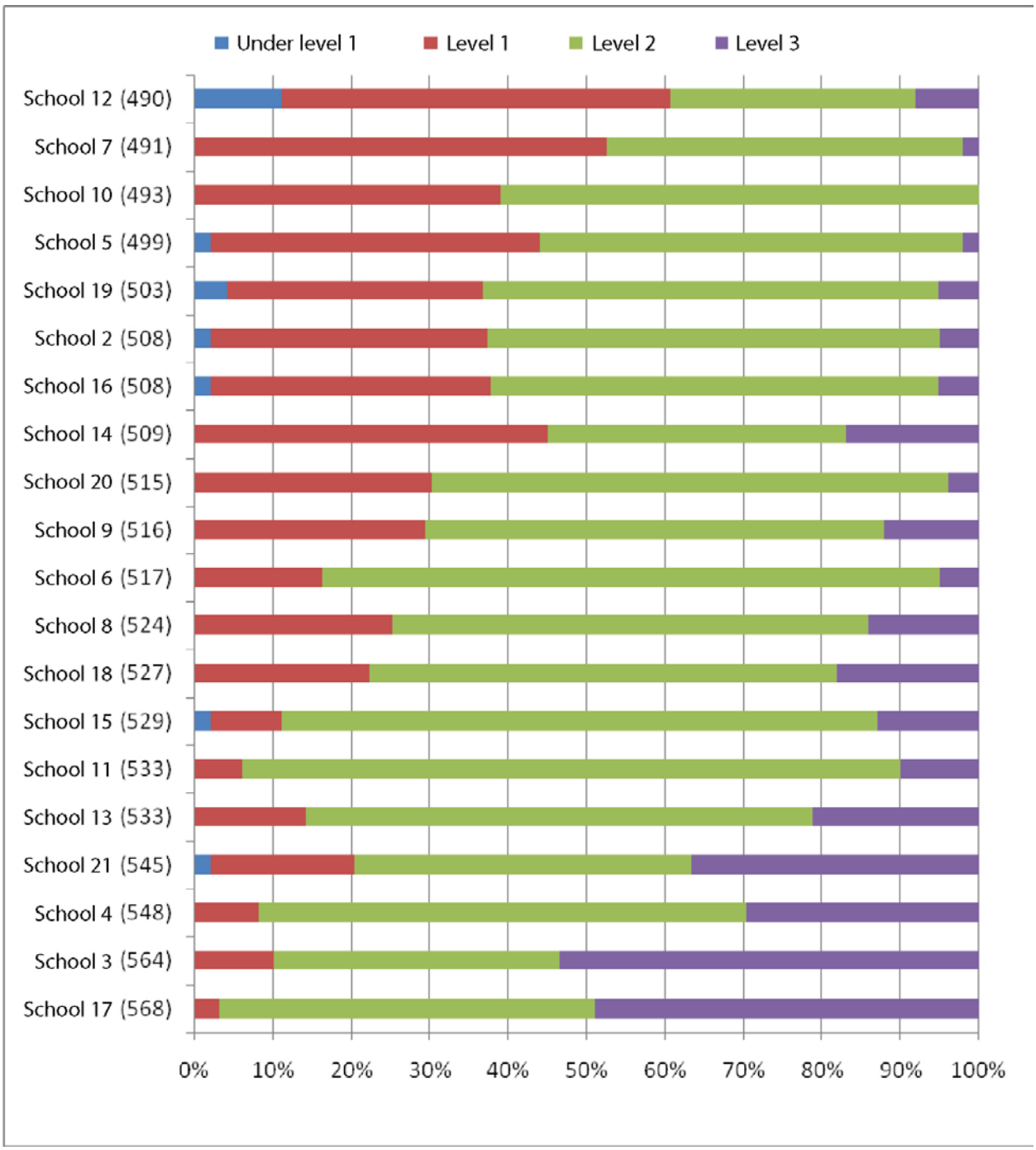


Figure 6.6. Distribution of students from various schools across proficiency levels (mathematics)

Thus, we can see that the number of students at various proficiency levels shows strong fluctuations, depending on the school. In schools located at the top of the diagram level 1 dominates (red), while in schools at the bottom of the diagram level 3 is dominant (purple). In most schools, however, level 2 dominates (green), and this means that the content was acquired at the reflective level (comprehension level). The number of participants who did not reach level 1 was not large even for schools with low results.

It is interesting to compare student distribution across proficiency levels in different classes of the same school. Figure 6.7 demonstrates such distribution for one of the schools in the given region, at which average indicators were lower than the average for the whole region.

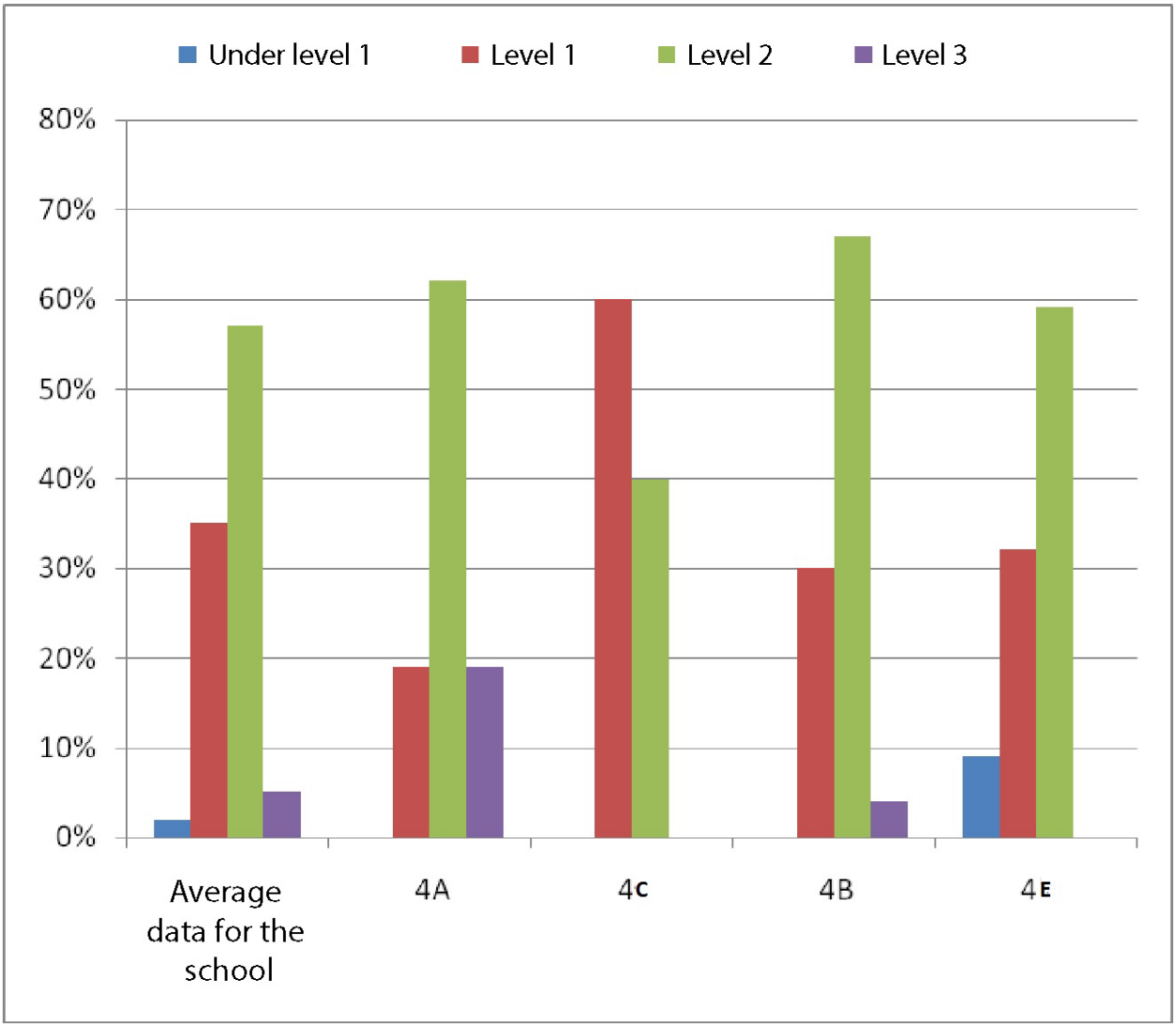


Fig. 6.7. Student distribution across proficiency levels in different classes of the same school (mathematics)

Comparing distributions across proficiency levels in different classes of the same school, we can see that in classes 4A, 4Б and 4B the proficiency level 2 dominates while in the 4Г class the proficiency level 1 dominates (this corresponds to level 1, the formal level of the curriculum acquisition). The best class is 4A because almost 20% of its students could reach the proficiency level 3 and only under 20% of its students still stay on the proficiency level 1. Class 4B follows, with 30% of its students at the proficiency level 1 and over 65% at the proficiency level 2. In class 4B there is almost no functional level of acquiring the curriculum: only fewer than 5% of the students are at the proficiency level 1. In class 4Б almost 10% of its students did not reach the proficiency level 1, and 30% of its students acquired the material at the formal level (with no comprehension). To know these relationships is very important when choosing strategies for working with problems in a particular class. It is even more important to know the causes for such a situation, and that would require additional study.

As part of this report preliminary analysis results were presented that feature links for the mathematics test results with various factors, i.e. participant gender, test form that the student had to complete, type of educational institution, location of the school etc. A more in-depth

study of the factors impacting the results of training outcome in the primary school is beyond the framework of this report.

Please find below the main study results.

1) SAM results in mathematics show statistical significance of examinee gender: girls completed the test better than boys.

Table 6.8 shows the ratio of boys and girls in this sample, and Table 6.9 demonstrates the distribution of different gender participants across proficiency levels. Figure 6.8 provides a graphic interpretation of this same distribution.

Table 6.8. Ratio of boys and girls (mathematics)

Gender	**Number**	**%**
Females	2078	47,2
Males	2328	52,8

Table 6.9. Gender distribution across proficiecny levels (mathematics)

			Proficiency level in Mathematics				
			Under level 1	1	2	3	Total
Gender	Female	Number	45	523	1100	408	2076
		%	2,2%	25,2%	53,0%	19,7%	100,0%
	Male	Number	58	664	1233	370	2325
		%	2,5%	28,6%	53,0%	15,9%	100,0%

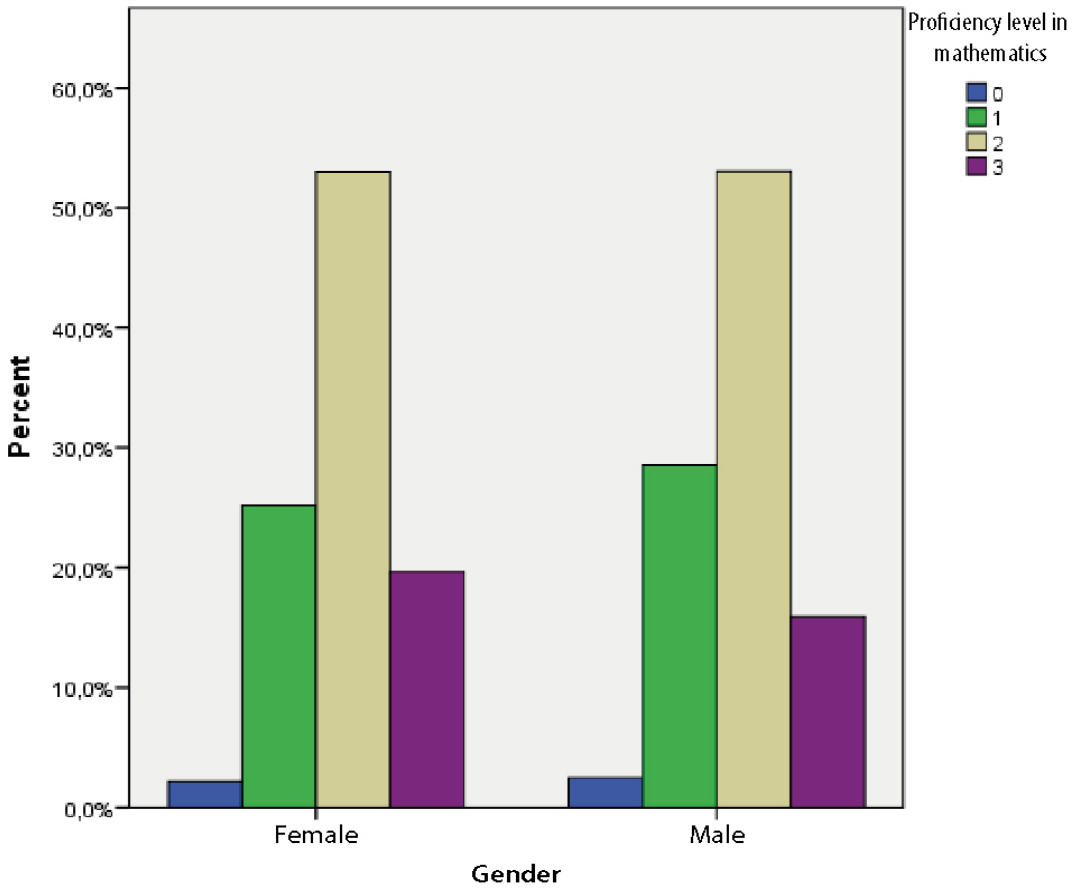


Figure 6.8. Distribution of examinees across proficiency levels depending on gender (mathematics)

It was statistically shown that there exists the dependence (even though a weak one) between the testing result in mathematics and participant gender: girls complete the test somewhat better than boys.

2) SAM results in mathematics depend significantly on the type of school: students of grammar schools complete the test better that students of comprehensive schools. Table 6.10 demonstrates the ratio of students at educational facilities of different type in this sample, and Table 6.11 shows the distribution of participants of different gender across proficiency levels. In Figure 6.9 you can see the same distribution as a graphic.

Tabl 6.10. Ratio of students from various types of educational facilities (mathematics)

Type of school	**Number**	**Percent**
General knowledge	3754	85,2
Gymnasium	647	14,7

Table 6.11. Test participant distribution across proficiency levels depending on the type educational facility (mathematics)

Type of educational facility * Proficiency level in mathematics

		Proficiency level in mathematics				
		Under 1	1	2	3	Total
General education	Number	98	1082	1986	588	3754
	%	2,6%	28,8%	52,9%	15,7%	100,0%
Gymnasium	Number	5	105	347	190	647
	%	,8%	16,2%	53,6%	29,4%	100,0%

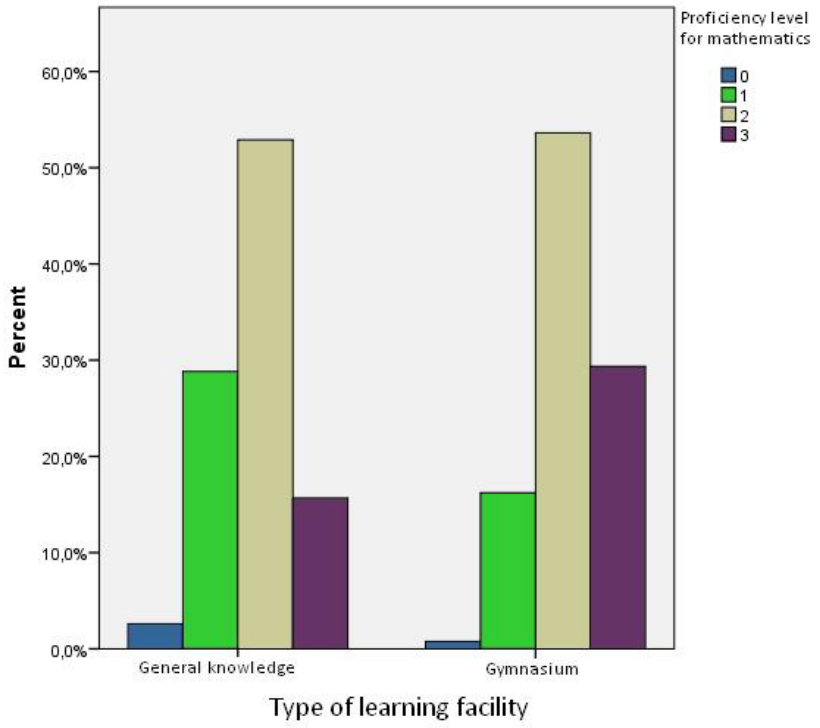


Figure 6.9. Distribution of examinees across proficiency levels depending on the type of school (mathematics)

It was statistically shown that there exists the dependence (even though it is weak) between the mathematics test result (test score and proficiency level) and the type of learning facility: students from specialized schools (gymnasiums) complete the test somewhat better than their counterparts from general knowledge schools.

3) SAM results in mathematics depend significantly on the school location: students of city schools demonstrate better results than students of schools in rural areas.

Table 6.12 shows the frequency distribution of students from communities of various types, and Table 6.13 shows the distribution of these test participants across proficiency levels. Figure 6.13 shows the correlation of testing results (test block) and the community type.

Table 6.12. Frequency distribution of students from communities of various types (mathematics)

		Frequency	Percent
Valid	City	3161	71,7
	Township, settlement	618	14,0
	Village	606	13,8
	Total	4385	99,5
Missing	System	21	,5
Total		4406	100,0

Table 6.13. Student distribution in different types of communities across proficiency levels (mathematics)

Type of community		Proficiency level in mathematics				Total
		Under level 1	Level 1	Level 2	Level 3	
City	Number	58	754	1725	624	3161
	%	1,8%	23,9%	54,6%	19,7%	100,0%
Tonwship	Number	16	188	312	102	618
	%	2,6%	30,4%	50,5%	16,5%	100,0%
Village	Number	28	237	290	51	606
	%	4,6%	39,1%	47,9%	8,4%	100,0%

It was statistically shown that there exists the dependence between the test result in mathematics (test score and proficiency level) and types of communities: city students perform the test better than students from townships and students from townships do it better than village students.

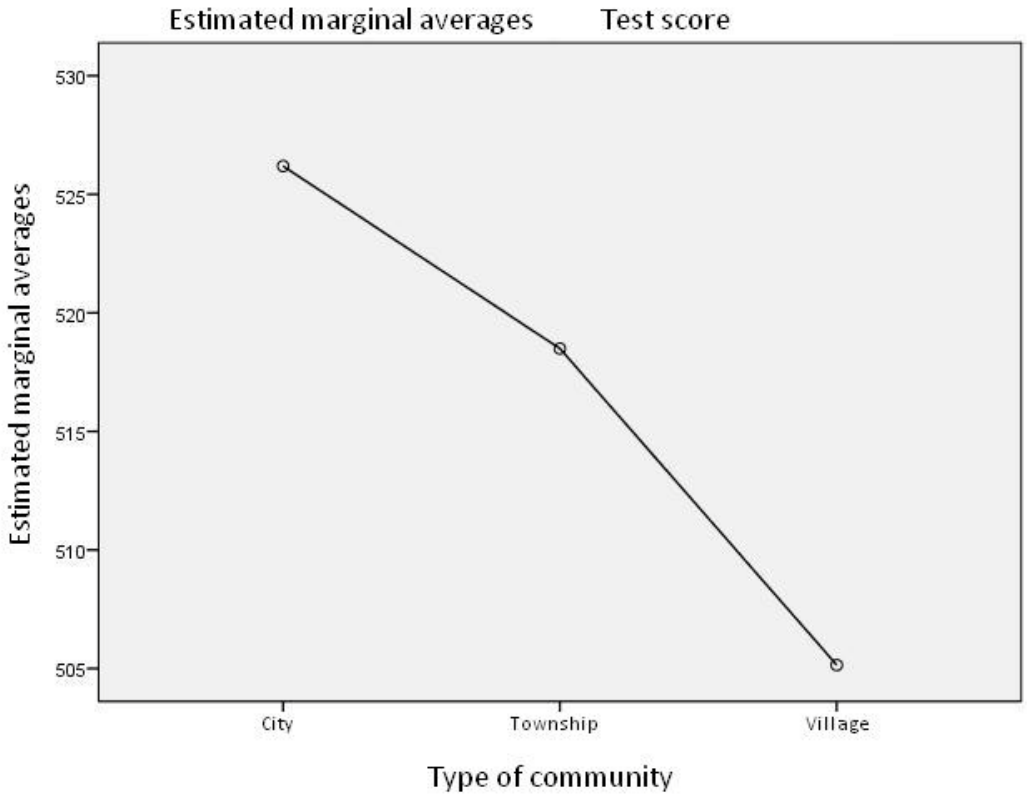


Figure 6.10. Relation between test results and school location (mathematics)

An additional study is necessary in order to get as clear picture as possible regarding factors impacting the results of teaching mathematics in primary schools.

6.2. Analysis of test results in Russian language

Figure 6.11 shows a histogram for the test score distribution of Russian language test participants.

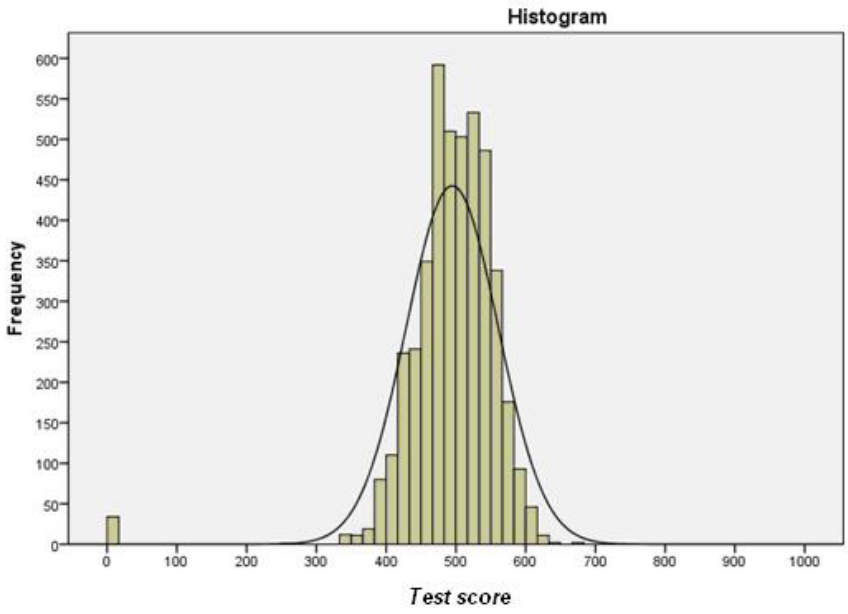


Figure 6.11. Distribution ot examinee test scores (Russian language)

Figure 6.12 displays the distribution of Russian language test participants across proficiency levels.

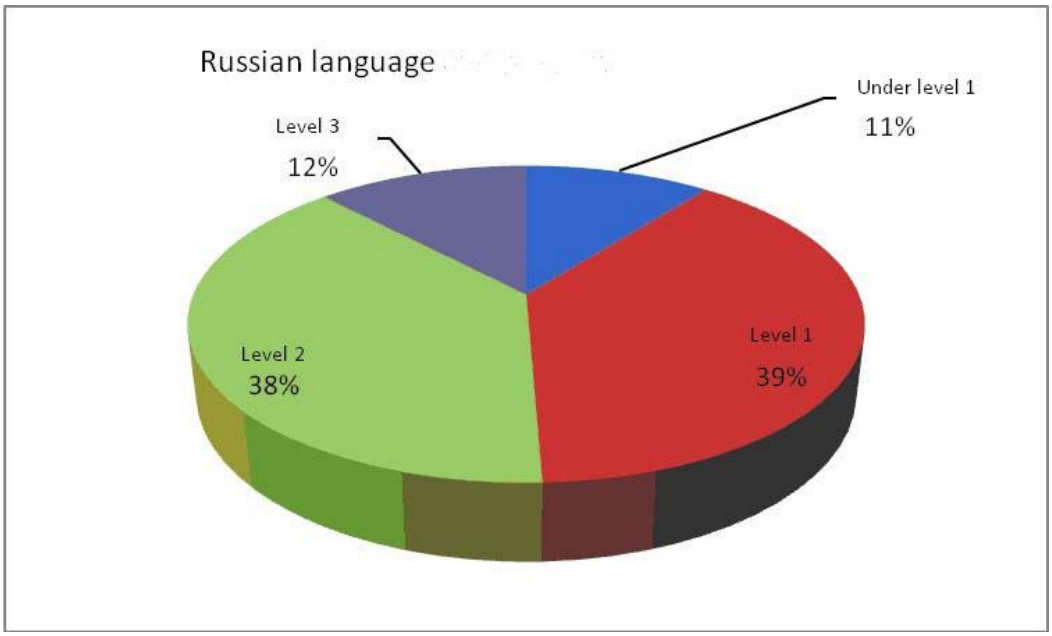


Figure 6.12. Test participants distribution across proficiency levels (Russian language)

Figure 6.13 shows the success profile in Russian language for the given student sample (average percentages of successfully completed items as a function of level).

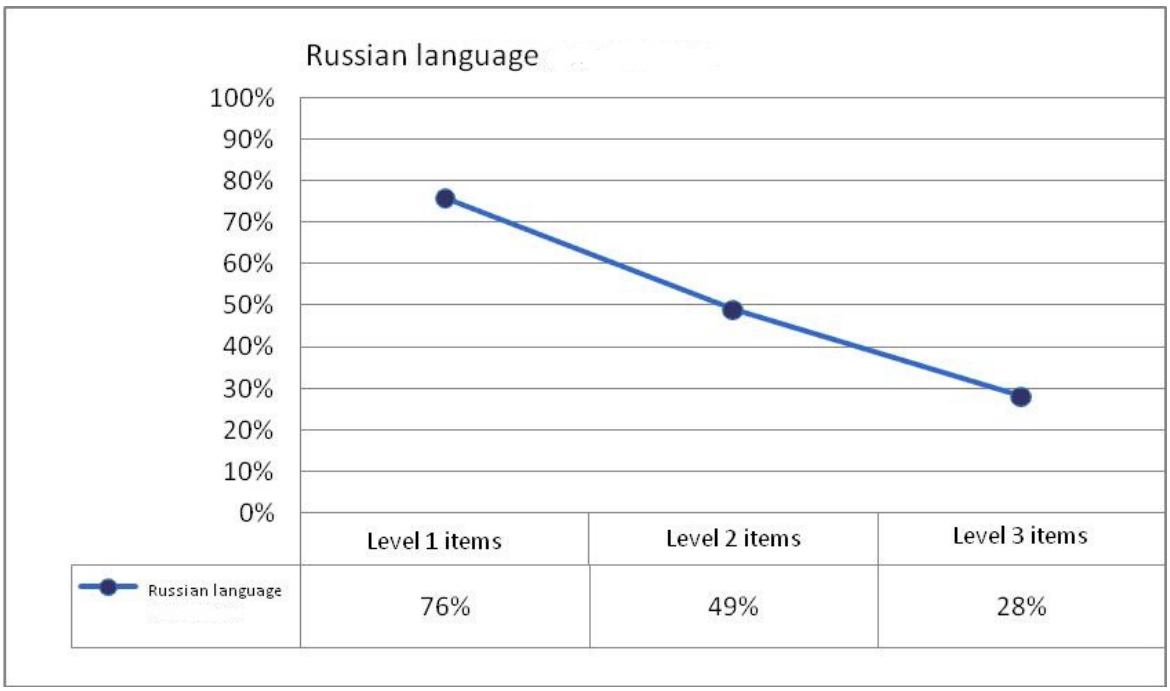


Figure 6.13. Russian language profile for the current student participant sample

The profile can be read quite easily. If we assume that the test covers, in due proportion, main curriculum sections, then level 1 scale tells us that the main part of that curriculum was formally acquired (almost 80% of level 1 items were completed); the second scale shows that only 50% was acquired reflectively (with comprehension), and the last scale informs us that less than one third was acquired functionally.

Figure 6.14 shows the test participant distribution across proficiency levels in various schools of a given region (on the diagram 20 schools are represented randomly chosen from the total of 192: these are the same schools as for mathematics in Figure 6.6). The horizontal axis features percentage of students in each proficiency level, and the vertical axis shows schools. For each school average test score of students from this school is shown in parentheses. Schools were ranged in order of increasing average mean for general test scores.

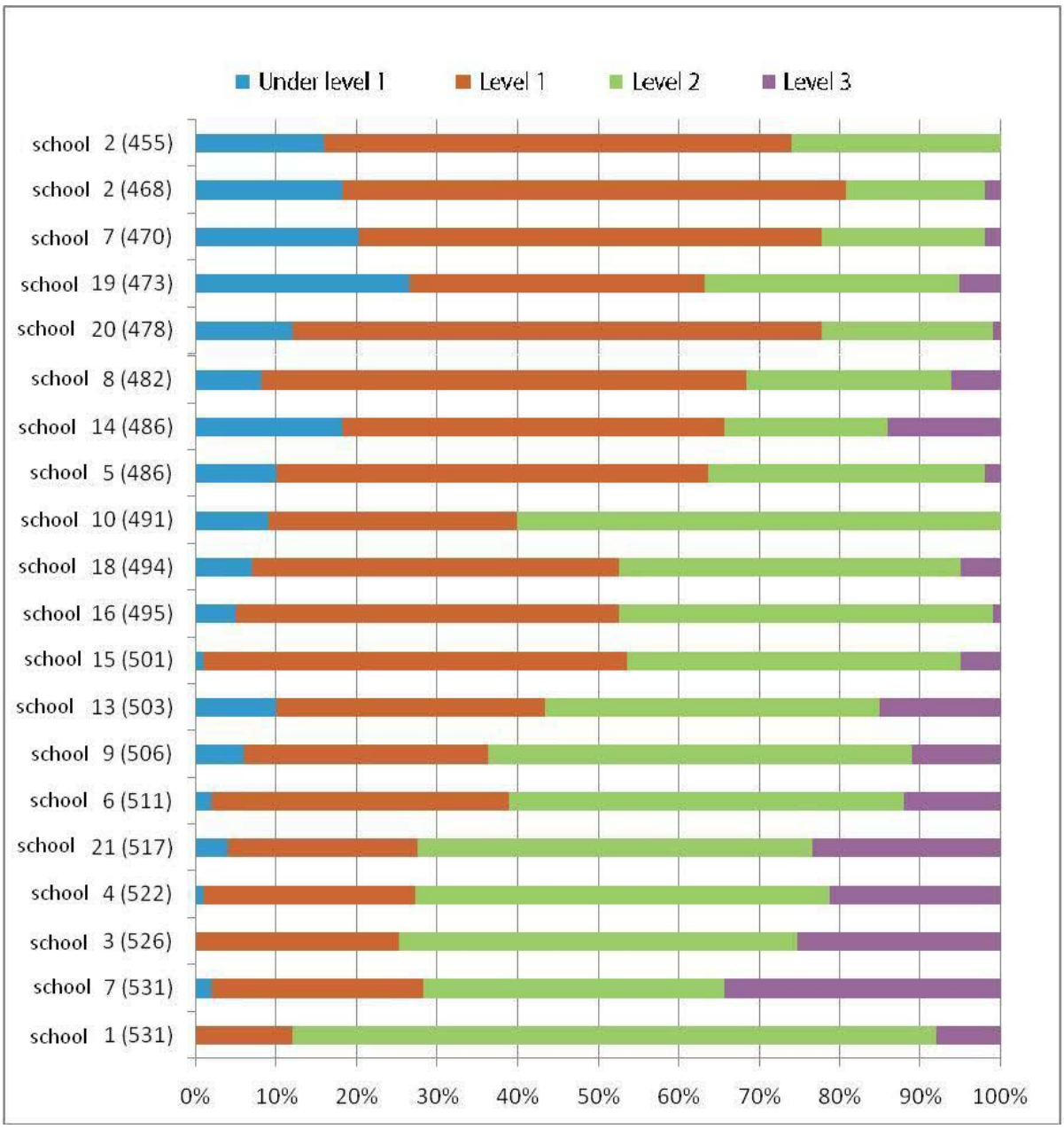


Figure 6.14. Distribution of students from various schools across proficiency levels (Russian language)

Thus, we can see that the number of students at various proficiency levels shows strong fluctuations, depending on the school. In schools located at the top of the diagram level 1 dominates (red), and a relatively large number of students stay under level 1 (blue) while level 3 is almost unavailable. In schools at the bottom of the diagram level 2 (green) is dominant and level 3 (purple) is also actively represented. In general, Russian language results are worse than the results in mathematics, in these same schools.

It is interesting to compare student distribution across proficiency levels in different classes of the same school. Figure 6.15 demonstrates such distribution for one of the schools in the given region, at which average indicators were lower than the average for the whole region. (We took the same school as in Figure 6.7.)

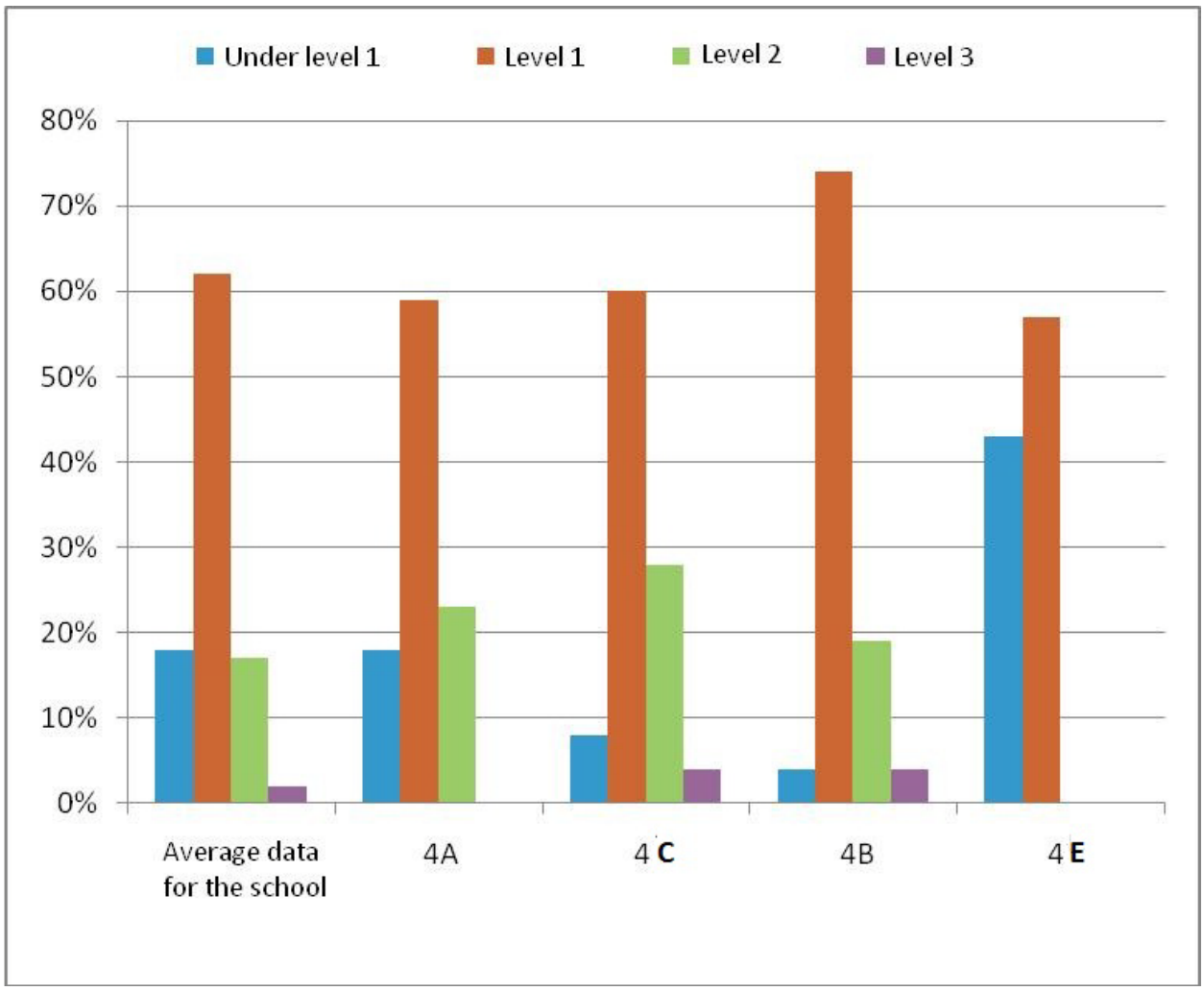


Figure6.15. Student distribution across proficiency levels in different classes of the same school (Russian language)

Comparing distributions across proficiency levels in different classes of the same school, we can see that in all classes the proficiency level 1 dominates. The 4Б class shows the weakest results: over 40% of students did not reach the proficiency level 1, the remaining students in this class could acquire the material at the formal level (with no comprehension). In this class, there are no students at higher levels. The best for now is 4Г class: around 30% of its students could acquire the material at level 2 (level of comprehension). However, even in this class almost 10% of students did not reach the proficiency level 1.

To compare let us analyze one more school (Figure 6.16). Its classes show even more variance in terms of quality of training. The best is 4A class: over 60% of students reached the proficiency level 2 and almost 10% are at the proficiency level 3. In 4B class, however, 55% of students could not reach the proficiency level 1 and the remaining 45% are at the proficiency level 1. That is, both reflective and functional levels of content acquisition are not represented in 4В class. 4Б class is very inconsistent in terms of training: 20% of students could not reach the proficiency level 1 while 10% did reach the proficiency level 3. The remaining students in this class are spread almost equally between the proficiency levels 1 and 2.

To know these relationships is very important when choosing strategies for working with problems in a particular class. It is even more important to know the causes for such a situation, and that would require additional study.

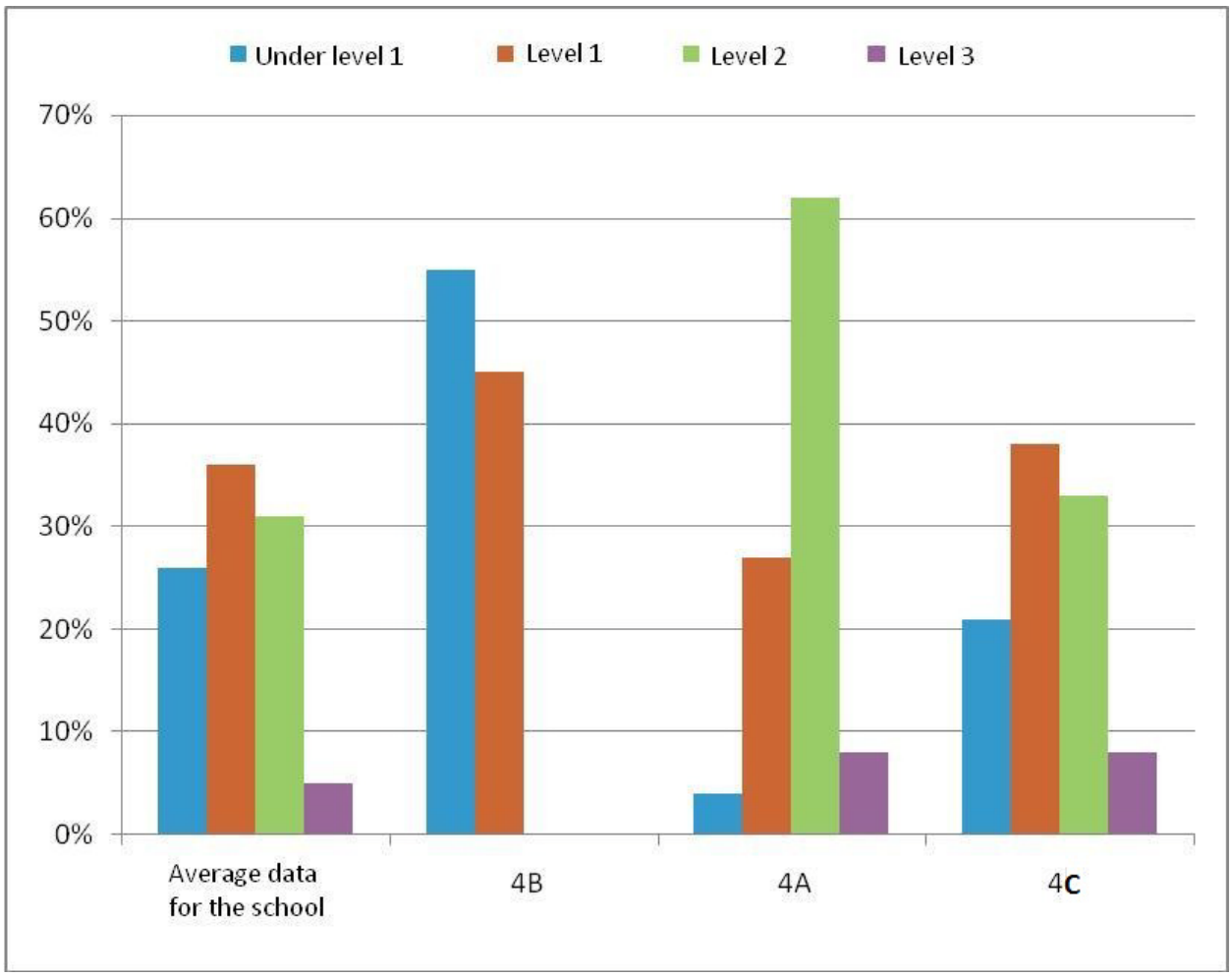


Figure 6.16. Student distribution across proficiency levels in different classes of school #2 (Russian language)

As part of this report the preliminary analysis results were presented that feature links for the Russian language test results with various factors, i.e. participant gender, test form that the student had to complete, type of educational insittution, location of the school etc. A more in-depth study of the factors impacting the results of training outcome in the primary school is beyond the framework of this report.

Please find below the main study results.

1) SAM results in Russian language show statistical significance of examinee gender: girls completed the test better than boys.

Table 6.14 shows the ratio of boys and girls in this sample, and Table 6.15 demonstrates the distribution of different gender participants across proficiency levels. Figure 6.17 provides a graphic interpretation of this same distribution.

Table 6.14. Ratio of boys and girls (Russian language)

Gender	**Number**	**%**
Females	2075	47,1
Males	2326	52,9

Table 6.15. Gender distribution across proficiecny levels (Russian language)

			Proficiency level in Russian language				
			Under level 1	1	2	3	Total
Gender	Female	Number	161	736	849	329	2075
		%	7,8%	35,5%	40,9%	15,9%	100,0%
	Male	Number	314	988	811	213	2326
		%	13,5%	42,5%	34,9%	9,2%	100,0%

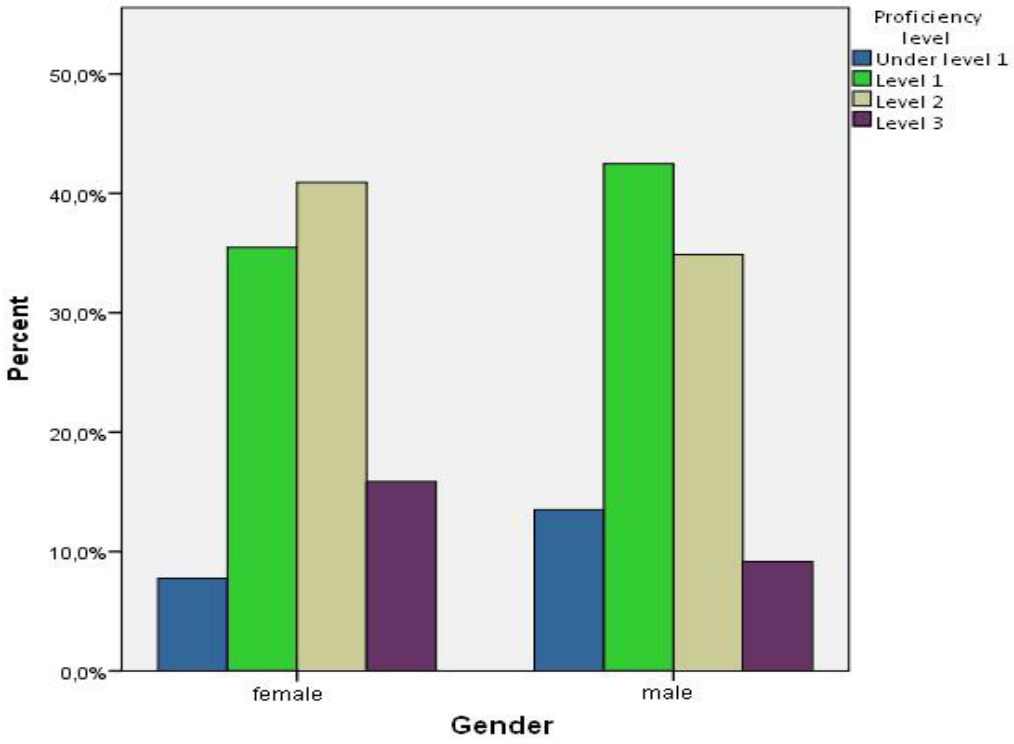


Figure 6.17. Distribution of examinees across proficiency levels depending on gender (Russian language)

It was statistically shown that there exists the dependence (even though a weak one) between the testing result in Russian language and participant gender: girls complete the test somewhat better than boys do.

2) SAM results in Russian language depend significantly on the type of school: students of grammar schools (gymnasiums) complete the test better that students of comprehensive (general education) schools.

Table 6.16 demonstrates the ratio of students at educational facilities of different type in this sample, and Table 6.17 shows the distribution of participants of different gender across proficiency levels. In Figure 6.18 you can see the same distribution as a graphic diagram.

Table 6.10. Ratio of students from various types of educational facilities (Russian language)

Type of school	**Number**	**Percent**
Comprehensive (general education)	3754	85,3
Gymnasium (grammar school)	647	14,7

Table 6.17. Test participant distribution across proficiency levels depending on the type educational facility (Russian language)

		Proficiency level in Russian language				Total
		Under 1	1	2	3	
General education	Number	444	1489	1406	415	3754
	%	11,8%	39,7%	37,5%	11,1%	100,0%
Gymnasium	Number	31	235	254	127	647
	%	4,8%	36,3%	39,3%	19,6%	100,0%

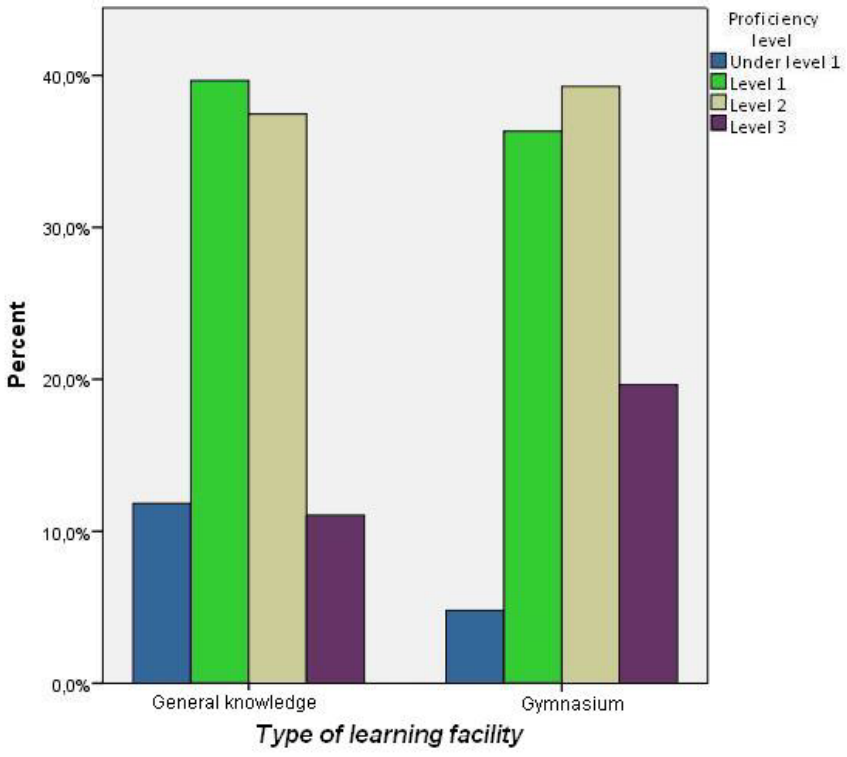


Figure 6.18. Distribution of examinees across proficiency levels depending on the type of school (Russian language)

It was statistically shown that there exists a dependence (even though it is very weak) between the Russian language test result (test score and proficiency level) and the type of learning facility: students from specialized schools (gymnasiums) complete the test somewhat better than their counterparts from general knowledge schools.

3) SAM test results in Russian language do not show dependence to the type of community.

Table 6.18 shows the frequency distribution of students from communities of various types, and Table 6.19 shows the distribution of these test participants across proficiency levels. Figure 6.19 shows the correlation of testing results (test block) and the community type.

Table 6.18. Frequency distribution of students from communities of various types (Russian language)

		Frequency	Percent
Valid	City	3161	71,8
	Township, settlement	618	14,0
	Village	606	13,8
	Total	4385	99,6
Missing	System	16	,4
Total		4401	100,0

Table 6.19. Student distribution in different types of communities across proficiency levels (Russian language)

Type of community		Proficiency level in Russian language				Total
		Under level 1	Level 1	Level 2	Level 3	
City	Number	281	1237	1228	415	3161
	%	8,9%	39,1%	38,8%	13,1%	100,0%
Township, settlement	Number	91	254	208	65	618
	%	14,7%	41,1%	33,7%	10,5%	100,0%
Village	Number	98	229	218	61	606
	%	16,2%	37,8%	36,0%	10,1%	100,0%

It was statistically shown that there exists no dependence between the test result in Russian language (test score and proficiency level) and types of communities: students results

are not significantly related to the type community. It can be seen in Figure 6.19 that average score for students from different community types differs by under 10 points.

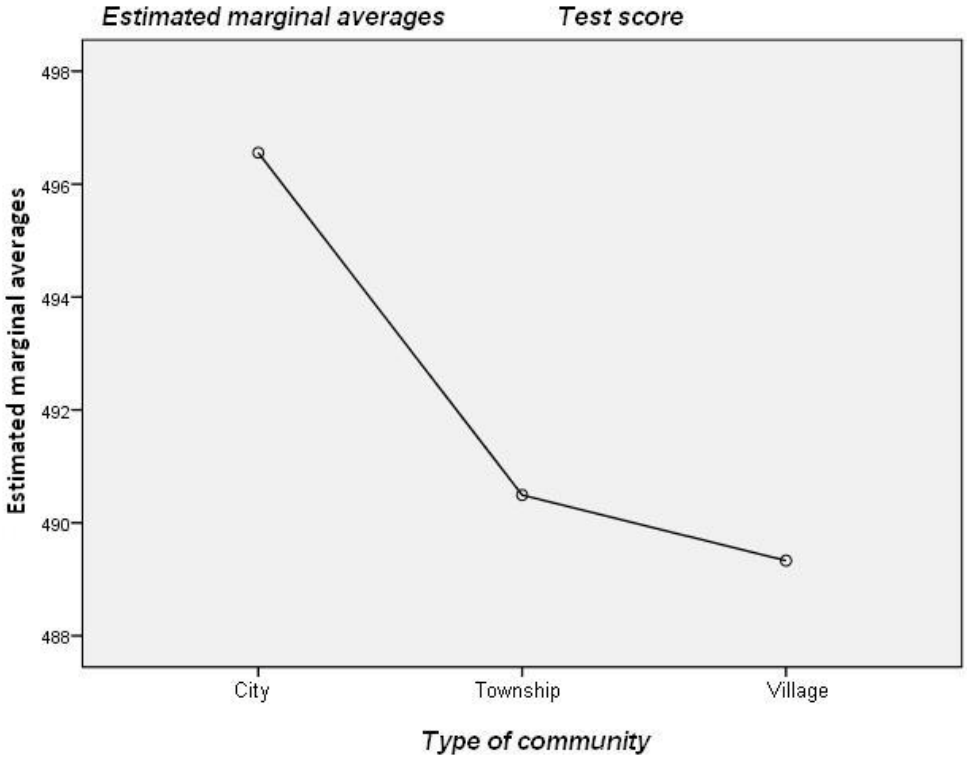


Figure 6.19. Relation between test results and school location (Russian language)

4) In conclusion, let us analyze the dependence between test results and test forms, using Russian language test as an example. In this region two test forms were utilized.

Table 6.20 shows the ratio of students who were completing different test forms in this sample, and Table 6.21 shows the distribution of participants for different test forms across proficiency levels. Figure 6.20 shows the same distribution as a diagram.

Table 6.20. Student ratio for different test forms (Russian language)

Test form	**Number**	**%**
First	2219	50,4
Second	2182	49,6

Table 6.21. Test participant distribution across proficiency levels depending on test from (Russian language)

.

			Proficiency level in the Russian language				Total
			Under level 1	Level 1	Level 2	Level 3	
Test form in Russian language	1	Number	231	871	832	285	2219
		%	10,4%	39,3%	37,5%	12,8%	100,0%
	2	Number	244	853	828	257	2182
		%	11,2%	39,1%	37,9%	11,8%	100,0%

.

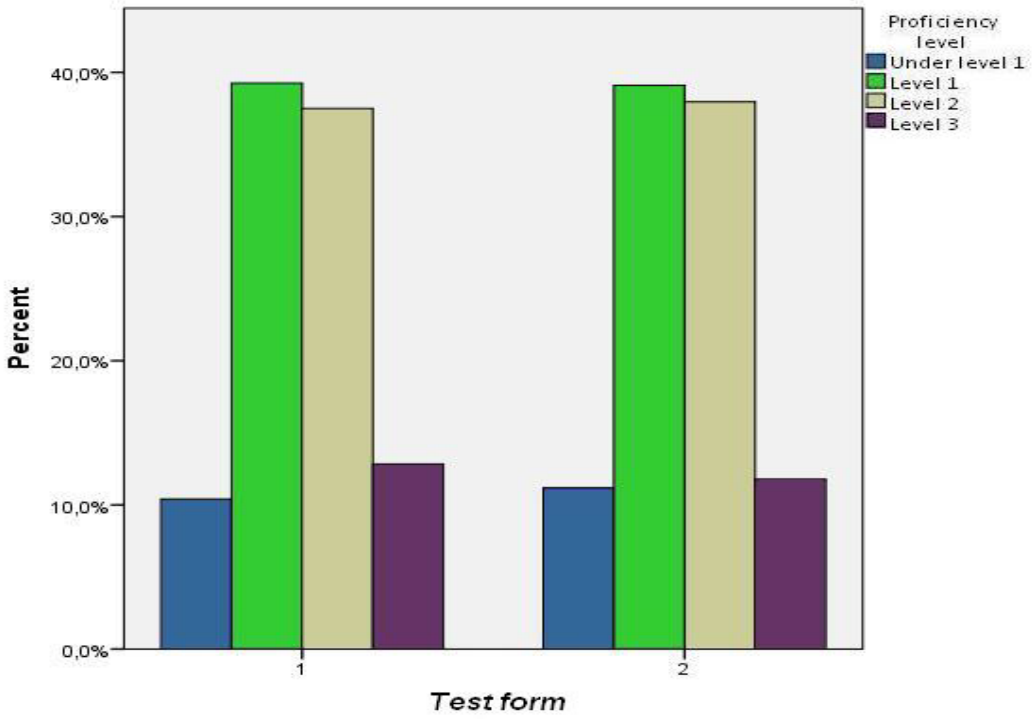


Fig. 6.20. Test participant distribution across proficiency levels depending on test forms (Russian language)

It was statistically shown that there exists no dependence between the Russian language test results (test score and proficiency level) and the test form: student results are not significantly related to the test form used.

An additional study is necessary in order to get as clear picture as possible regarding factors impacting the results of teaching Russian language in primary schools.