Ed. P. Nezhnov, E. Kardanova, B. Elkonin

Authors:
P. Nezhnov (p. 1); E. Kardanova (p. 3, p. 4); A. Vorontsov, V. Zaslavsky, O. Sviridova, O. Sokolova, L. Ryabinina (p. 5); S. Gorbov, V. Zaslavsky (p. 2; Annex 1, Annex 2)

STUDENT ACHIEVEMENTS MONITORING

SAM

SAM (Student Achievement's Monitoring) is a testing toolkit to assess subject competencies of school students. Theoretical framework of the toolkit relies upon the teaching/learning process concept based on L.S. Vygotsky's ideas. On the design side, SAM tests are noted for the embedded diagnostic mechanism providing information on the quality of content assimilation. The SAM model can be implemented for different school subjects. At the moment, mathematics and Russian language tests have been developed for primary school students. SAM is designed for the follow-up and improvement of the learning process in schools, and is intended for teachers, methodologists and education management authorities.

AKNOWLEDGMENTS

SAM – is a result of the creative efforts of a large number of professionals, without whom this project would not simply taken place.

We are particularly grateful to A.Volkov, V. Bolotov, Is. Froumin who initiated the search for new approaches to assessing the quality of school education, took part in the theoretical understanding of the problem, contributed to the creation of conditions for development and provision counseling and human support to the development team. The key contribution to the project from B. Hasan, who took on risky responsibility for organizing and conducting the initial level of development which was carried out by Institute of developmental psychology and pedagogy in Krasnoyarsk, should be also noted.

We are grateful to our colleagues who put their creativity to the conceptualization, design and implementation of the model SAM at different levels of development. This A. Aronov, O. Balandin, E. Vysotskaya, A. Gilyano, A. Dorohova, V. Zaytseva, O. Znamenskaya, I. Kim, C. Klevtsova, G. Kudina, Z. Novlyanskaya, O. Obuhova, O. Ostroverh, A. Skripka, T. Timkova, G. Zuckerman, T. Chaban, E. Chudinova, T. Yustus.

We would also like to note the great work of statisticians and psychometricians V. Ovchinnikov and V. Karpinskiy that provided debugging psychometric characteristics of test materials, and assisted in the development of methods of estimation of test participants, as well as students of MA program of Higher School of Economics "Measurement in Psychology and Education" who were involved in processing of approbations database.

At all levels of the project, we received good support from I. Efremova, Y. Efremova, Y. Koreshnikova, E. Shirshikova who helped us to solve urgent organizational problems.

Significant assistance and support of our work was provided by G. Kovaleva whose friendly and constructive criticism, as well as direct assistance in testing materials helped to formulate the basic provisions of the approach and to determine the potential of our tools. We express our

gratitude to the experts - M. Zelman and M. Vasileva, as well as K. Polivanova and A. Mayorov – for the examination of SAM, valuable comments and recommendations for improvement.

Finally, it is impossible to overestimate the multilateral support that was provided to the project by The Moscow office of the World Bank and the Center for International Cooperation in Education Development (CICED).

TABLE OF CONTENTS

INTRODUCTION

Assessment of school students' subject competences is a key issue in the educational practice. Assessment is a stimulating guidance for students and parents, and also a feedback reference for the teachers as well as an important indicator for school administrators. And finally, learning achievements make up a significant part of school reporting to education authorities. It is the evaluation practice with relevant toolkits that acts as a primary promoter of educational goals accepted by the community, demonstrates the degree of their achievement, and therefore provides the basis for making consistent teaching and managerial decisions.

To be able to successfully perform this important coordinating function, the evaluation system should match the specific objectives addressed by various actors of the educational practice. Thus, on the one hand, there are objectives related to the governance of the overall school system where the primary focus is placed on the accurate measurement of academic outcomes enabling to compare the performance of teaching teams, evaluate relative efficiency of different teaching approaches. On the other hand, there are objectives associated with the organization, improvement and follow-up of the teaching/learning process per se, where the qualitative characteristics of the process results are of special importance. That is the categorization of educational outcomes within the framework of a certain taxonomy of educational objectives differentiating specific levels of content acquisition. Therefore, there is a need for a readily available testing toolkit equally addressing both aspects of the assessment.

Possible approach to solving this problem is a model of teacher test SAM (Student Achievements Monitoring). At the moment, based on this model mathematics and Russian language tests have been developed for the primary school that can be used in grades 4-5. However, the tests can be also adapted for grade 3 to provide early diagnostics of learning outcomes.

The main users of SAM are the teachers, methodologists, school administrators and local education management authorities.

1. THEORETICAL BASIS AND DESIGN FEATURES OF THE SAM

SAM is based on the theory of cultural development, suggested by scientific school of Vygotsky. The main aspects of this theory, that become the basic inputs for SAM model, are the following.

Psychological maturing of child is a process of cultural development, that has cycled nature. Necessary precondition of each cycle of development is learning, i.e. transmission patterns of cultural experience (knowledge) to a child. Full cultural pattern is a system of signs, that sets a particular cultural ability, i.e. a generalized mode of action. For instance, in schools mathematical mode of action is presented via concepts, principles, ,schemes, mathematical tasks, solving algorithms, etc. This set of symbolic structures is being transmitted to a child within school learning process. Namely, a teacher explains and shows how and by what means certain mathematical operations are done. Then teacher works to ensure that child reproduced the main conclusions and solved routine tasks.

However, according to Vygotsky, the transmission of cultural pattern marks only the beginning of the educational process. In future the student should really assimilate this knowledge as a system, so to be able to solve all types of the tasks corresponding to this method.

The process of assimilating of cultural content has a hidden, long-term and spontaneous nature. The seeds of knowledge, sown in the pupil's minds, grow very gradually in each of them into the ability to act reasonably, i.e. a subject competence. Therefore, Vygotsky defined the process of assimilating of action modes as the functional development. According to his theory, the role of learning is to found the pupils' zone of proximal development - the possibility of the formation of a certain cultural ability. (The possibility that sometimes is not realized, and then the knowledge gained lies dormant in memory.) Vygotsky also believed that acquiring of the mode of action passes three levels. Further studies allowed to clarify what kind of levels and why there are three of them.

The important step was made by P. Galperin, who showed that action method is determined by the content of the orientation. And mastering of the action method is essentially the assimilation of an appropriate system of orientation points. Further, owing to the work of B. El'konin, V. Davydov, A. Zaporozhets, A. Leontiev and their co-workers there were three necessary ingredients in the orientation of action method identified:

1) empirical - external characteristics of the situation and appropriate action;

2) theoretical - the concept of significant relation, which provides the basis for action mode in a given situation;

3) functional - understanding of the field of possibilities of action method with its limits.

These three types of orientation points are transferred to a child as a cultural pattern. But during its assimilation the role of a real support for the action for the first time is played by external characteristics. After that they are joined by the notion of a significant relation, and finally - an appropriate functional field. These three types of orientation mark three levels of assimilation of the cultural mode of action, which briefly can be defined as formal, reflective and functional (Figure 1.1).

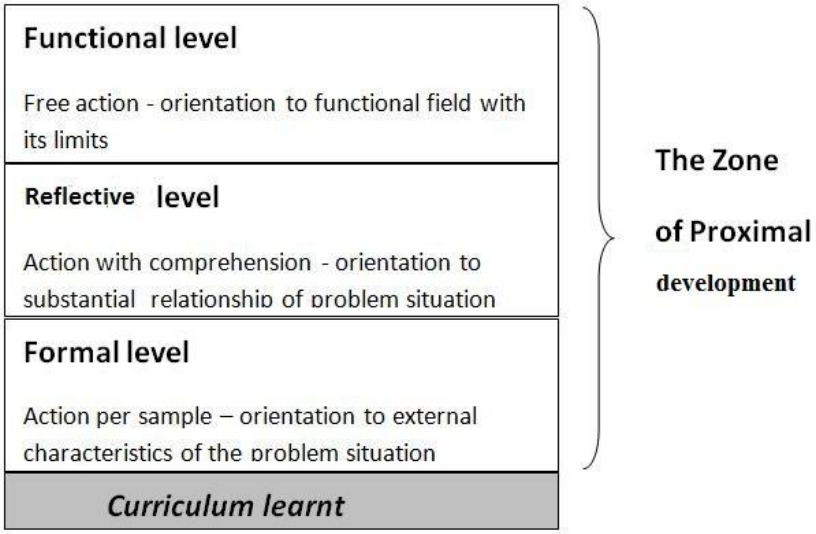


Figure 1.1. Levels of assimilation of the cultural pattern

In case, when the first type of orientation is basic, the measure of generalization is minimum and covers a narrow range of common situations and the corresponding schemes (algorithms) of actions. The second option gives the principal opportunity to solve the full range of tasks corresponding to a given method. Finally, in case of the third option, the method is characterized by the functionality, i.e. by adaptability to different contexts.

It is important to note that each type of orientation is realized through its psychological mechanism which causes a tangible distinctions between them. Thus, the orientation of the first level is based on the direct associative connections. In the language of the teacher, this means rote exterior features standard applications and algorithms for their solution. The orientation of the second level is based on the mental structure that captures the essential relation of situation. In psychology, such structures are termed as "gestalt" and in such cases teachers say, that the child began to understand the subject content, and has learnt to extract significant out of the conditions of the problem. Finally, the orientation of the third level is based on a complex structure, we denote it by the term "functional field", which keeps the area of possibilities of this action method, with its limits. About the student, which demonstrates this level, the teacher can say that he is fluent at subject content, i.e. he can apply it flexibly, intelligently, in accordance with the situation.

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.

LEVEL 1 (FORMAL)

The generic criterion of achieving this level is the ability to act being guided by external parameters of the problem situation and pattern of operation. For example, recognize a problem as referring to a certain class (type) based on particular attributes, and perform a relevant procedure presented as a standard operational scheme or rule; or construct a sequence of operations directly following the statement of problem.

Assimilation of the pattern of operation at the first level is demonstrated by the ability to solve problems whose description either clearly enables to refer them to a certain class with a well-developed solution procedure (standard problems) or directly guides the student to the correct scheme of operation, i.e., contains straightforward prompting messages. In such problems, conceptual understanding required for solution is linked to external parameters of the situation and needs no compulsory statement (identification).

Therefore, at the given level of pattern assimilation, the problem is solved in a purely external context – through direct correlation of the problem description with the description of relevant action as a sequence of operations (Figure 1.2).

Figure 1.2. Orientation of the action at the formal level

LEVEL 2 (REFLECTIVE)

The generic criterion of achieving this level is the ability to act on the basis of substantial analysis of the problem situation, i.e., to develop a conceptual understanding helping to find a principal way to solution.

Assimilation of the pattern of operation at the second level is demonstrated by the ability to solve problems for which no ready-made (standard) patterns of operation can be directly found and applied without conceptual understanding (i.e., without understanding the substance of the objective situation). In particular, this requirement is met in: a) problems where description of conditions hinders their classification; b) problems where the application of a standard pattern requires transformation of conditions; c) problems described in an abstract form that excludes solution through manipulation of explicit data; d) problems suggesting reversion of standard patters of operation (movement from the outcome to conditions), etc. In such problems, solution cannot be found through a direct matching of conditions and ready-made operating pattern but is sought through conceptual understanding, i.e., is based on the interpretation ("understanding") of the situation (Figure 1.3).

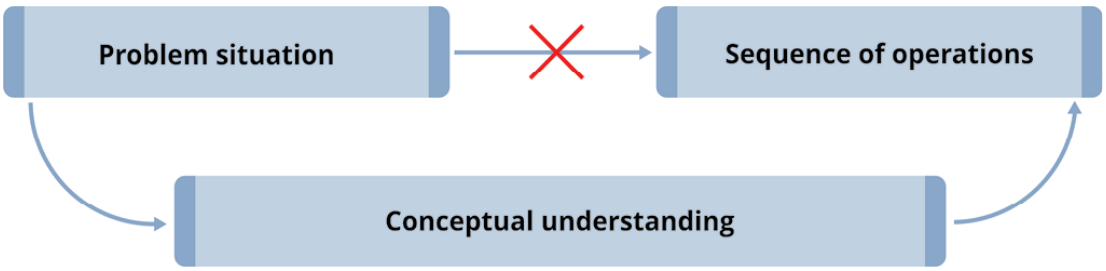


Figure 1.3. Orientation of the action at the reflective level

One can say that finding a solution requires correlation and coordination of two contexts: external (description of the problem situation and pattern of operation) and internal (conceptual understanding of the problem and approach to solution). In this context, the "external" and "internal" should be understood in the logical sense.

LEVEL 3 (FUNCTIONAL)

The generic criterion of achieving this level is the ability to orientate in the domain of possible ways of implementing the general pattern, see its limits, and go beyond them if necessary (Figure 1.4).

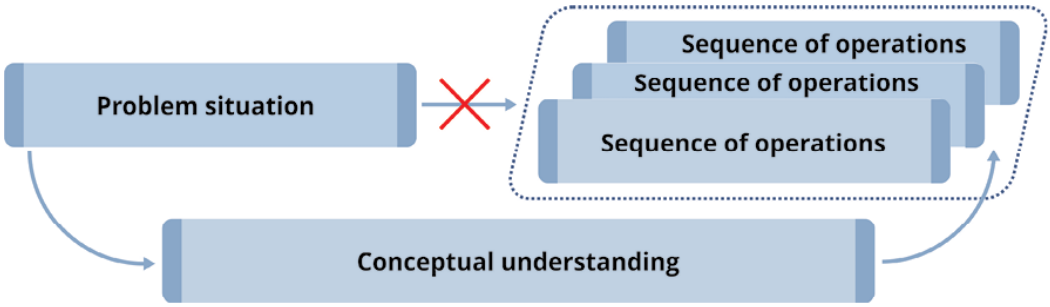


Figure 1.4. Orientation of the action at the functional level

Assimilation of the pattern of operation at the third level is demonstrated by the ability to solve problems:: on reconstruction and probing ways of operation to select the one meeting the specified criterion; on analysis of specified patterns (formulas) of operation to select the adequate one; assuming detection of the limits of the main pattern and use of additional resources; where the main pattern acts as a means to support a higher level operation; assuming the need to adapt the pattern of operation to specific features of the object.

The above typology of problems serves as guidance for designing the blocks of test items. Each block includes level 1, 2 and 3 items referring to one and the same curriculum area. As an illustration, below is a 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 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.*

Test versions are compiled of such blocks consisting of items from the main subject content areas. Accordingly, the test can be viewed as a system of three subtests with each representing a set of items of the same level referring to different content areas (Figure 1.5).

Figure 1.5. Test structure

As a result, the test has two functions: a) measurement of integral educational success (45 test items, covering the main sections of the subject) and b) the diagnosis of the content assimilation (15 blocks).

SAM tests include items of different types: completion ones with a brief answer, multiple-choice ones with a choice from 4-5 offered options, matching items, items that require constructions, etc. Test items are presented in blocks. The sequence of blocks makes 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 single 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.

2. MATH TEST

Tests in two subjects – mathematics and the Russian language – were developed under the SAM model. The tests have a common structure. This paper presents only the description of a math test (Chapter 2). Test specification is given in Annex 1; examples of item blocks – in Annex 2.

2.1. TEST CONTENT

The math test is intended for the assessment of subject competencies of primary school students to evaluate the level of assimilation of the mathematics subject content. In accordance with the theoretical model, students' competencies are to be evaluated at three basic levels: formal, reflexive and functional. The test is targeted for primary school graduates (grades 4 and 5) but can be also adapted for grade 3 students.

The test is built upon the math content included in the majority of existing primary school curricula. The content was selected in accordance with Federal State Standard of Primary Education (MOED Order # 373 of October 6, 2009 *On Approving and Enactment of the Federal State Standard of Primary Education*).

Subject content of the test is divided in five areas.

"NUMBERS AND CALCULATIONS". This content area is relevant to the formal aspect of the concept of natural numbers (positional representation of numbers, standard algorithms of operations with numbers, order of operations, properties of operations). It also includes materials related to representation of numbers on the coordinate line. The latter is important for the comprehension of real numbers and assimilation of coordinate method.

"MEASUREMENT OF VALUES". This content area includes materials related to direct and indirect measurement operations, and also incorporates geometric measurements.

As to the applied aspect of this content area closely related to specific practical measurements and their representation as diagrams and charts (data analysis), it can rather be relevant to the environmental studies.

"REGULARITIES". This content area is related to construction of numerical and geometric sequences and other structured objects, and measurement of their quantitative parameters. This area is highly important for the development of mathematical thinking (first of all, algorithmic and combinatory).

"DEPENDENCES". This content area is related to identification and description of the mathematical structure of relations between values usually represented in test items.

"ELEMENTS OF GEOMETRY". This content area covers geometric materials related to identification of spatial forms and relative position of objects.

The content framework of the math test can be presented as a matrix (Table 2.1.) that includes:

- Subject content areas;
- Mathematical tools (concepts, principles, formulas, algorithms, etc.) providing the orientation for mathematical operations.

Content areas	**Orientation tools for mathematical operations**
Numbers and Calculations	■ *Sequence of natural numbers* ■ *Number line* ■ *Positional principle* ■ *Properties of arithmetic operations* ■ *Order of operations*

Measurement of Values	■ *Relationship between the number, value and unit* ■ *Whole-part relationship* ■ *Formula of rectangle area*
Regularities	■ *"Induction step"* ■ *Recurrence (periodicity)*
Dependences	■ *Relationship between like values (equality, inequality, multiplicity, difference, "whole-part")* ■ *Direct proportion between values* ■ *Derived values: velocity, labor productivity, etc.* ■ *Relationship between units*
Elements of Geometry	■ *Form and other properties of figures (main types of geometrical figures)* ■ *Spatial relationship between figures* ■ *Symmetry*

Table 2.1. Content of the math test

2.1.2 DESCRIPTION OF TEST ITEMS

General description of test items corresponding to the first level of subject content assimilation can be specified for math content areas.

To assess the assimilation of level 1 in the **"NUMBERS AND CALCULATIONS"** content area, the test toolkit uses items related to arithmetic operations and application of some standard techniques used in calculations, e.g., estimated result or rounding. Level 2 items are focused on identification and consideration of the structure of a multiple number or expression rather than calculations. And, finally, in level 3 items an abstract expression should be specified to satisfy a certain condition.

Test items on **"MEASUREMENT OF VALUES"** are related to simple measurements. In case of direct measurements, the result is being achieved either through direct plotting of units (length or square area measurements) or using well-known instruments (e.g., a ruler or clock). Anyway, items of this level require no transformation of objects being measured. Items assuming indirect measurement may require simple calculations by the known formulas (e.g., that of rectangle area). Level 2 is tested with such items where basic operation algorithms cannot be applied at once, and it is needed either to reduce the situation to a standard one by transforming the objects to be measured (in case of direct measurement) or outline a plan of calculations with due regard for "barriers" implied in problem statement. Level 3 items require going beyond the limits of the operating algorithm, i.e., identifying its potential and involving an additional intellectual resource.

Level 1 test items on **"REGULARITIES"** include sequences with an easily identified "step" where the number of elements in a structured object is found through a direct calculation (e.g., if a structured object consists of few elements). Level 2 includes items where direct counting of elements in a structured object is difficult, for example, if an object consists of a large number of elements. "Inverse" items can also be used where an object should be identified based on the specified pattern of its structure. Testing of level 3 involves items on finding a formula for an "irregular" object, i.e., a formula where, in addition to variations, one needs to consider certain irregularities in the object.

Level 1 items on **"DEPENDENCES"** include standard text problems containing a few easily identified relationships. Level 2 items include text problems with "latent" relationships whose identification requires construction of a model or additional reasoning. Level 3 can be assessed with items requiring a mathematical operation involving several relationships.

Level 1 items on **"ELEMENTS OF GEOMETRY"** require knowledge of figures with easily recognized forms and positions. In level 2 items, figures and their position do not

correspond to their standard visual images. Another type of items implies consideration of idealized properties of geometric figures contradictory to their image (e.g., line of infinity). Achievement of level 3 is demonstrated by performing items that require reconstructing a range of options for the recognition of a complex geometric figure.

Sample items for each content area are given in Annex 2. Tentative ratio between items of different levels and content is presented below (Table 2.2).

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

Table 2.2. Tentative ratio between different items in the test book

THE FOLLOWING TYPES OF ITEMS ARE USED IN THE TEST:

- completion items with a brief answer;
- multiple-choice items with a choice from 4-5 offered options;
- items requiring constructions.

The majority of items (about 80%) are the completion ones with a brief answer.

3. PSYCHOMETRIC PARAMETERS OF THE TEST

The math test under the SAM model was piloted using paper- and computer-based forms in 2010-2012 in Russian regions and the Republic of Kazakhstan. The pilot involved about 10 000 4-5-grade students from general education schools. Testing of 4-grade students took place in spring, and that of 5-grade students was conducted in autumn.

The pilot testing was undertaken in several stages. The first stage (2010-2011) had a goal of determining statistical characteristics of items, assessing their quality, and identifying deficient items (e.g., items with low discrimination or items with non-functioning distractors, etc.)—that is, those that were to be re-worked or removed; also assessing the test's reliability and measurement error. The second stage (2011-2012) was designed to verify the properties of the items as well as the overall test after its re-work, to substantiate the test quality as a measuring tool. The sample of the first stage totaled around 2000 students, and of the second stage around 8000 students. The results of the first stage of the pilot testing are summarized below.

The results of piloting were analyzed both in the framework of the classical test theory, and the IRT.

In the classical analysis (where item difficulty[11] and discriminativity[2] are considered the main parameters), all math test items of the SAM model demonstrated high performance. Classical reliability index (Chronbach's Alfa) was above 0.9 in all versions and both forms of testing.

Below are the average values and standard deviations of difficulty and discrimination indices for items of various levels in one of the test forms (Table 3.1). (Test form No.1 of the paper-based test is used as an example here and elsewhere.)

Item levels	Number of items	Item difficulty		Discrimination index	
		Average value	Standard deviation	Average value	Standard deviation
Level 1 items	*15*	*0,76*	*0,15*	*0,39*	*0,16*
Level 2 items	*15*	*0,43*	*0,14*	*0,57*	*0,13*
Level 3 items	*15*	*0,19*	*0,10*	*0,37*	*0,15*
Total test	***45***	***0,46***	***0,27***	***0,44***	***0,17***

Table 3.1. Characteristics of math items referred to different levels

Almost all items demonstrate high discriminativity (high average values). Some of level 1 items show comparatively low discriminativity with difficulty indices being equal to 0.8 and higher (which means that over 80% of test takers performed the task correctly). Relatively low discriminativity of these items is attributed to their simplicity for the students.

According to the theoretical model of SAM, three items (levels 1, 2 and 3) within a block should form a hierarchy in terms of difficulty, which is the case for almost all blocks of items (Table 4.2).

Content areas	Level 1	Level 2	Level 3
Numbers and Calculations	*0,75*	*0,47*	*0,19*
Measurement of Values	*0,73*	*0,37*	*0,27*
Regularities	*0,79*	*0,46*	*0,16*
Dependences between values	*0,86*	*0,44*	*0,19*
Elements of Geometry	*0,72*	*0,46*	*0,06*

[1] *Item difficulty is defined as a sharepart (percent) of test takers who performed the item correctly.*

[2] *Discriminativity designates the distinctive character of an item, i.e., its ability to distinguish test takers with different ability scorelevel.*

Table 3.2. Item difficulty by different levels

Table 3.3 presents average values and standard deviations of item difficulty and discriminativity depending on the content area. Comparison between Tables 3.2 and 3.3 shows that item difficulty is weakly dependent on the content area but is mainly defined by the item level (although items from the "Elements of Geometry" content area generally appeared more difficult for the students).

Content area	**Number of items**	**Item difficulty**		**Discrimination index**	
		Average value	**Standard deviation**	**Average value**	**Standard deviation**
Numbers and Calculations	*12*	*0,47*	*0,27*	*0,47*	*0,17*
Measurement of Values	*15*	*0,46*	*0,23*	*0,49*	*0,17*
Regularities	*6*	*0,47*	*0,30*	*0,41*	*0,18*
Dependences between values	*6*	*0,50*	*0,31*	*0,44*	*0,11*
Elements of Geometry	*6*	*0,41*	*0,36*	*0,31*	*0,16*

Table 3.3. Characteristics of items from major content areas of the math test

A more thorough analysis of math items from the SAM test, as well as their properties (for all versions and both forms of testing) using the Item Response Theory allows the following conclusions:

■ The test can be considered as essentially unidimensional;

■ The test is optimal in terms of difficulty and well-centered in respect to the population of test takers;

■ The absolute majority of items demonstrate good psychometric parameters and fit the model of measurement (G. Rasch dichotomous model);

■ Therefore, the SAM math test can be considered as a high quality tool to assess mathematical competency of primary school students.

4. METHODOLOGY OF TEST TAKER ESTIMATION

4.1. ESTIMATION MODEL

The SAM toolkit is intended to ensure compatibility of test results obtained from different samples, at different time and from partially different tests. Therefore, a modern IRT approach was selected as a basis for estimation model. Test scores obtained using the IRT are plotted on the metric scale thus enabling to compare test results achieved by different groups of students, and use a wide range of mathematical statistics methods to study and verify various hypotheses. In addition, the metric nature of the scale allows (provided that some additional conditions are observed) equating test results obtained from different test forms and at different time.

SAM tests consist of three level items, and the set of items referring to the same level within a single test can be viewed as an independent subtest. Therefore, tests taken as a whole can originally be considered as multidimensional ones.

Three approaches can be used in modeling of such tests [17]:

■ Ignore multidimensionality, and use a unidimensional model;

■ Consider items within the same level as a separate subtest, and successively apply unidimensional model to each subtest;

■ Use multidimensional models.

The first approach evidently possesses obvious advantages: it assumes that each test taker gets one test score with a minimal measurement error since the whole variety of items is viewed as a single test. In addition, the approach is preferable due to its simplicity, ease of interpretation, and opportunity to address all specific testing challenges. On the deficiency side is the loss of information on the student's achievements in individual (theoretical) levels; however, establishing benchmarks based on these levels enables to eliminate this drawback.

In view of the above, a unidimensional dichotomous Rasch model [7] was selected as a testing model. Applicability of the model was justified in a special study [18].

4.2. SCALING OF TEST RESULTS

To establish a universal scale, a study was performed on a specially designed sample, representative in respect to the general population of potential test takers (primary school graduates). The sample was stratified on the basis of two parameters: school location (city / village) and school type (general educational institution / magnet school – gymnasium, lyceum, etc.). The sample size made up about 1000 persons.

Estimates of academic achievements were obtained for all participants from the basic sample using the selected testing model (Rasch model). The estimates – integral scores of test takes on the logit scale – are unsuitable for reporting to test takers and stakeholders since they contain fractional and negative values. The estimates are traditionally transformed from the logit scale to another, more convenient one. To this end, a 1000-point scale was selected, which is widely used in the world testing practice, in particular, in international monitoring assessments.

Transition to the 1000-point scale was performed using a linear transformation that maintains the scale metricity and does not distort intervals between the objects. Scores of all test takers from the basic sample were transformed from the logit scale to the 1000-point scale with the average value about 500 and standard deviation of 50. Reduction of the average value to 500 points allowed the alignment of the basic sample approximately in the center of the scale. Standard deviation of 50 points enabled, on the one hand, to adequately stretch the basic sample (approximately in the range from 350 to 650 points), and, on the other hand, to leave sufficient "gaps" in the upper and lower parts of the scale to plot the results of subsequent testing (inter alia, of other age groups).

In the future all scores obtained as a result of the practical application of the SAM toolkit will be plotted on the same scale thus allowing a comparison of each test taker's achievements in time, i.e., perform achievement monitoring. Of course, average values and standard deviation for other samples can be different.

4.3. BENCHMARKS AND PROFICIENCY LEVELS

For the purpose of test score interpretation based on the three-level assessment model, benchmarks were established to divide all test takers in 4 groups according to proficiency levels. Benchmarks were established following a procedure described below.

A benchmark score (in logits) is established for each subtest where the probability of completing an "averaged" subtest item equals about 50%, which sets the lower limit of the respective level. All test takers with the score below this limit are considered as those who failed to achieve the given level, as well as all subsequent ones.

Hence, first of all, a benchmark corresponding to a 50-percent probability of completing an "averaged" level 1 item is established. This value sets up the lower limit of the first proficiency level. All test takers whose score is below this benchmark are considered as those who failed to achieve the first (and, therefore, the second and third) level. Further, a benchmark is established corresponding to a 50-percent probability of completing an "averaged" level 2 item. All test takers whose score is below this benchmark but above the limit of the first level are considered as those who failed to achieve the second (and consequently the third) level but satisfy the conditions of achieving the first level. The benchmark for the third level is established in a similar way.

In the course of this procedure three values in logits are obtained that set up the benchmarks for each of the three proficiency levels. The most convenient way to establish benchmarks is to use characteristic curves for individual subtests that are later transferred to the 1000-point scale.

Figure 5.1 illustrates establishment of benchmarks in math. Benchmarks plotted on the X-axis are transferred to the 1000-point scale through the same linear transformation used for the test takers scores. As a result, the following benchmarks are obtained: transition from level 0 to level 1 – 430 points; from level 1 to level 2 – 500 points; from level 2 to level 3 – 570 points.

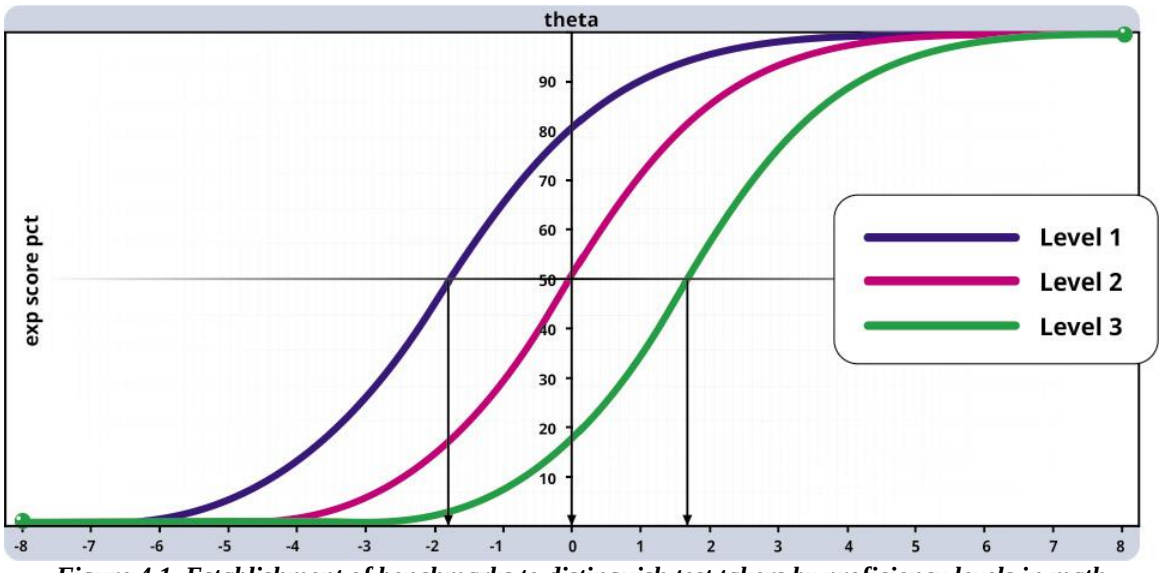


Figure 4.1. Establishment of benchmarks to distinguish test takers by proficiency levels in math

Hence, there are 4 proficiency levels corresponding to the following substantive criteria:

Zero level – even the first level of orientation has not been assimilated: the student completes less than 50% of level 1 items. And, as is seen from Figure 4.1, the probability of completing level 2 and 3 items is almost zero.

First level – only the first level of orientation is assimilated: the student completes at least 50% of level 1 items but less than 50% of level 2 items, the probability of completing level 3 items is very low.

Second level – the second orientation level has been assimilated: the student completes at least 50% of level 2 items but less than 50% of level 3 items, and, as is seen from Figure, he/she is able to perform over 80% of level 1 items but less than 50% of level 3 items.

Third level – the third orientation level has been assimilated: the student completes at least 50% of level 3 items, and he/she is very likely to perform any level 1 item and at least 80% of level 2 items.

Figure 4.2 shows a variable map for one of the math test forms.

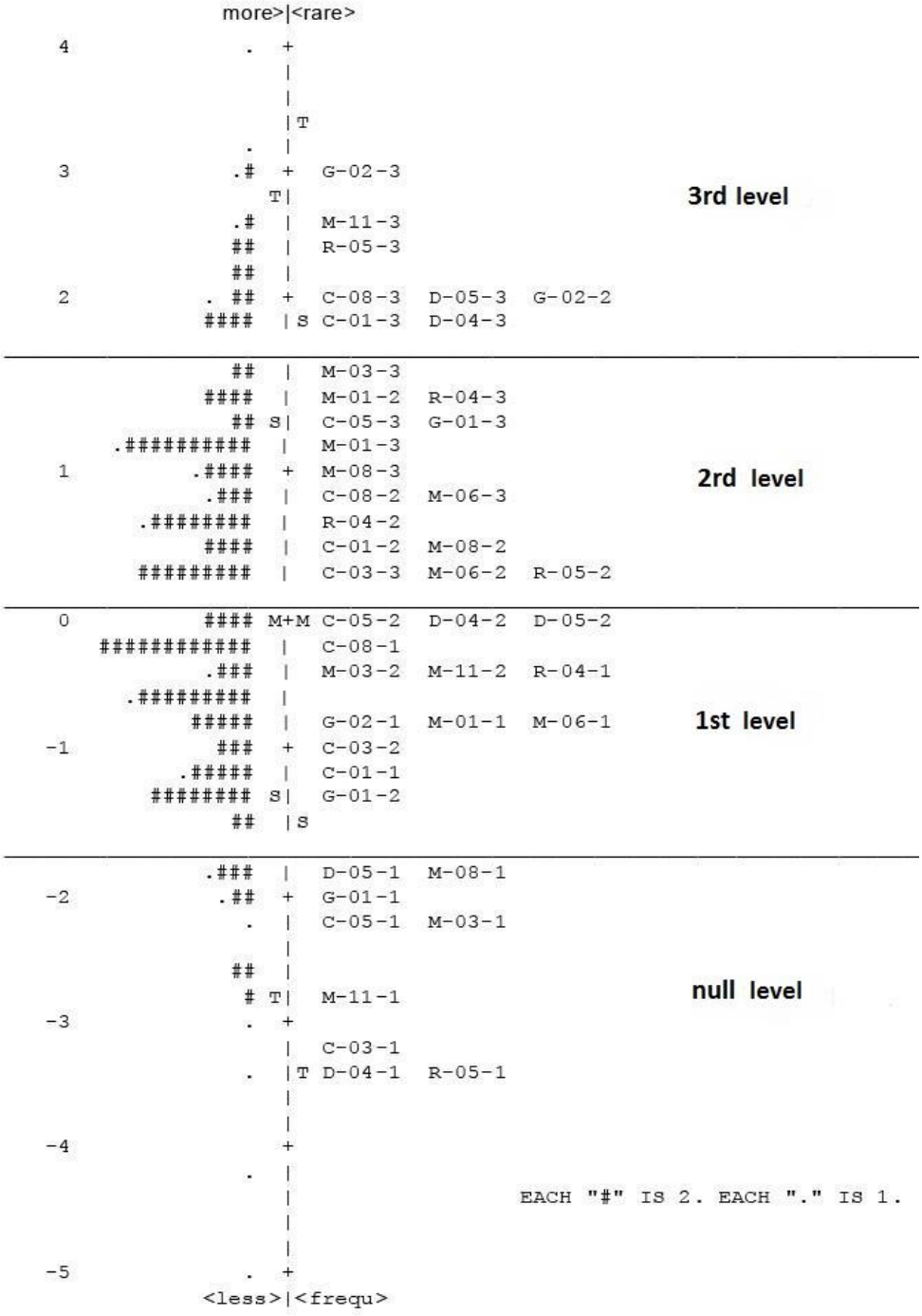


Figure 4.2. A variable map of math test

The map shows relative distribution of test takers and test items plotted on the same metric scale. Test takers are given at the left, and test items – at the right, the left side of the map presents a logit scale. More difficult items and more ability students are located in the upper part of the map; more easy items and low ability students are in the lower part of the map. It is seen from the map that the sample of test takers is well centered in relation to the variety of test items, the items being optimal in terms of difficulty.

Horizontal lines denote boundaries between proficiency levels corresponding to benchmarks. Hence, the majority of test takers are at the first and second levels with just a few students being at the third level. This is consistent with model developers' expectations that by the end of primary school students just start to acquire the ability to solve level 3 problems.

The above differentiation of test takers by levels enables to interpret test results in terms of the level assessment model, i.e., support the integral score with informative interpretation.

4.4. EQUATING THE RESULTS OF SUBSEQUENT TESTING

Under the SAM methodology, the scores of all test takers should be plotted on the common scale created for the basic sample according to procedures described above. At the same time, it is assumed that new items will be added to test forms; therefore, placing the results of subsequent tests to the same scale requires equating. The equating procedure is based on the method of common items: the tests on which the scale was built, and the ones that will be used in the future will contain common items in the amount sufficient to perform reliable equating procedures. Given the specific features of proposed tests (each test consists of blocks of items referring to the same content area), new tests should include at least 2 blocks from the initial test versions.

The common scale was designed using a method of separate calibration with identification of common parameters and reflection of all parameters on the common scale [8].

Figure 5.3 shows characteristic curves[3] of two math tests: a characteristic curve for the basic sample and that for one of the subsequent testing. As is seen from the Figure, both characteristic curves almost coincide, which is indicative of the successful equating procedure: examinee measures obtained from two studies are equivalent.

[3] *Test characteristic curve is the mathematical expectation of the test percent score depending on the students' ability score.*

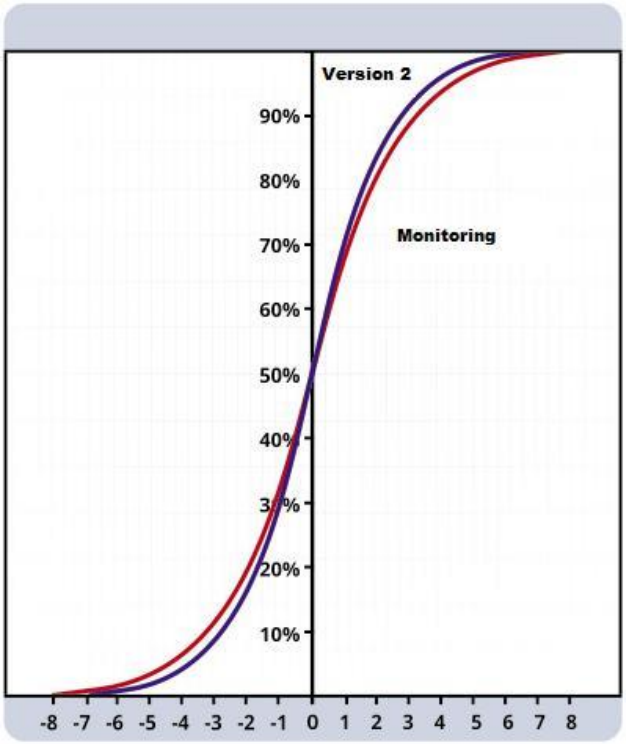


Figure 4.3. Characteristic curves of two math tests

Given that after the equating all parameters (both of items and test takers) are reflected on the common scale created for the basic sample, test takers can be divided in groups using the same benchmarks that were previously established for the basic sample. Therefore, application of the IRT allows presentation of all parameters of test takers and items on the common scale, and the use of common benchmarks to perform comparisons and interpret test results.

In the future, the proposed equating procedure will be applied to all testing data, and the results will be shown on the common scale.

5. OUTCOMES OF SAM APPLICATION

5.1. MAJOR TESTING RESULTS AND FORMS OF PRESENTATION

Theoretical framework and the construct of SAM tests predetermine three key groups of indicators: integral score, proficiency levels, and three-dimensional profiles.

INTEGRAL SCORE

Raw score is the sum of points achieved by a test taker.

Test score is the result of mathematical treatment of the raw score aimed at obtaining estimates on the metric scale common for all test takers irrespective of the time of testing and specific set of items performed. Test results under this indicator are presented on a 1000-point scale designed for each subject test based on special studies with a basic sample.

PROFICIENCY LEVELS

For the purposes of SAM tests, a graduated scale of achievements was developed where each level corresponds to a qualitative characteristic based on theoretically outlined levels of orientation (assimilation of cultural patterns). There are 4 levels of achievement that satisfy the following criteria:

Below first level – a student performs less that 50% of level 1 items;

First level – a student performs at least 50% of level 1 items;

Second level – a student performs at least 50% of level 2 items;

Third level – a student performs at least 50% of level 3 items.

Recall that here we refer to probability estimates. Thus, assigning the zero level to a test taker assumes that the given student is highly unlikely to complete over 50% of level 1 items.

Benchmarks for the math test: transition from level 0 to level 1 – 430 points; from level 1 to level 2 – 500 points; from level 2 to level 3 – 570 points.

THREE-DIMENSIONAL PROFILES

The SAM test enables to obtain a structural characteristic of the assessed competency: its three-dimensional profile. The profile shows the share of material assimilated at each of the three levels, i.e., constituents of the integral score presented in three subscales.

The profile is designed on the basis of raw (percentage) score obtained at each level. It is not the result of measurements but just characterizes the relative share of completed test items at each level. However, the ease of construction and interpretation makes the profile highly useful. The more so that it is the profiles and their changes in the course of monitoring that most consistently and explicitly reflect assimilation of the subject content.

PRESENTATION OF TEST RESULTS

Several types of tables and diagrams were developed for the purpose of test results presentation.

Table 5.1 contains summary individual test results that serve as a basis for deriving all other indicators. Thus, a raw score is given for each student, to which the following indicators are assigned:

a) Achievement profile (raw score distributed in three levels);

b) Test score;

c) Confidence interval where the true scope occurs with the probability of 90%;

d) Proficiency level (projection of the three-level pattern on the common scale).

N	*Name*	*Class*	*School*	*Raw score*				*Test score*	*Proficiency level*	*Confidence interval*
				Total	*Levels*					
					I	*II*	*III*			
1	*A.V. Petrov*	*4A*	*11*	*16*	*9*	*6*	*1*	*472*	*1*	*(455,489)*
2										
3										
…										
	Average values									

Table 5.1. Individual test results

To allow a comparative analysis of students' achievements, the table can be adjusted in accordance with individual test scores both within classes, and within the sample as a whole. In this case, the table will give an insight into the distribution of individual achievements on the common scale.

Table 5.2 presents the average integral and level-specific score for classes and schools, as well as the sample as a whole. And (like in the previous Table) an averaged achievement profile is assigned to each value.

School	*Class*	*Average raw score (%)*				*Average test score*
		Total	*Levels*			
			I	*II*	*III*	
11						
11	*4a*					
11	*4b*					
……..						
Sample average						

Table 5.2. Averaged testing results

Achievement profiles can be presented both as tables and as diagrams. The table form enables to summarize and visualize data on a large number of tested objects (Tables 5.1 and 5.2). However, graphic representation is more illustrative (Figure 5.1).

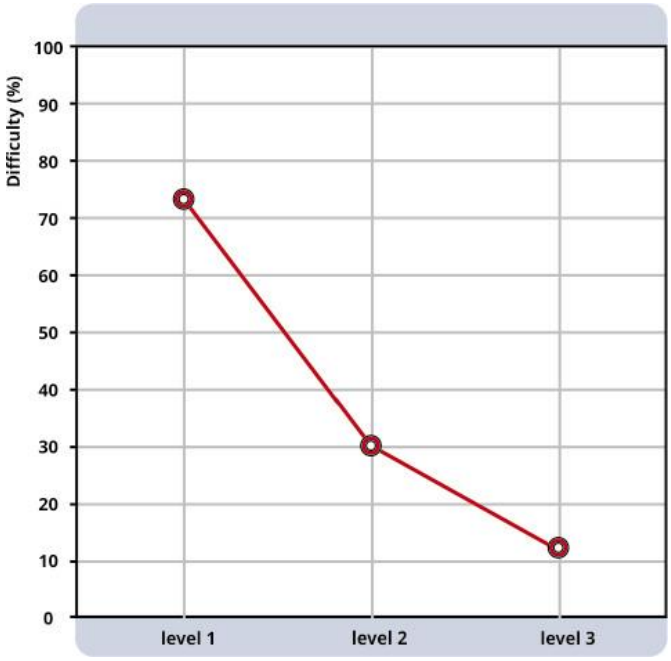


Figure 5.1. Average math profile for a sample of students

The profile is relatively simple to interpret. If we assume that the test covers major content areas in the proper proportion, then level 1 scale informs that three fourths of this content has been formally assimilated; level 2 scale reports that about one third has been assimilated reflectively (with comprehension); and the last scale informs that the functional level of assimilation covers just a modest part of the content.

Experience shows that profiles obtained from testing of large samples of students (e.g., from two different regions) demonstrate highly similar configurations, and differ mostly in the height of location on the coordinate grid. Differences in the form of profiles appear at the level of schools and classes. And here the structural presentation of results provides additional grounds for their comparative evaluation.

Thus, for example, the diagram on Figure 5.2 enables to state that class *4a* is highly competitive with class *4b* in solving standard problems but demonstrates a somewhat weaker comprehension of the assimilated material (see the difference in level 2 with equal performance in level 1). Differences between the classes become even more evident in level 3.

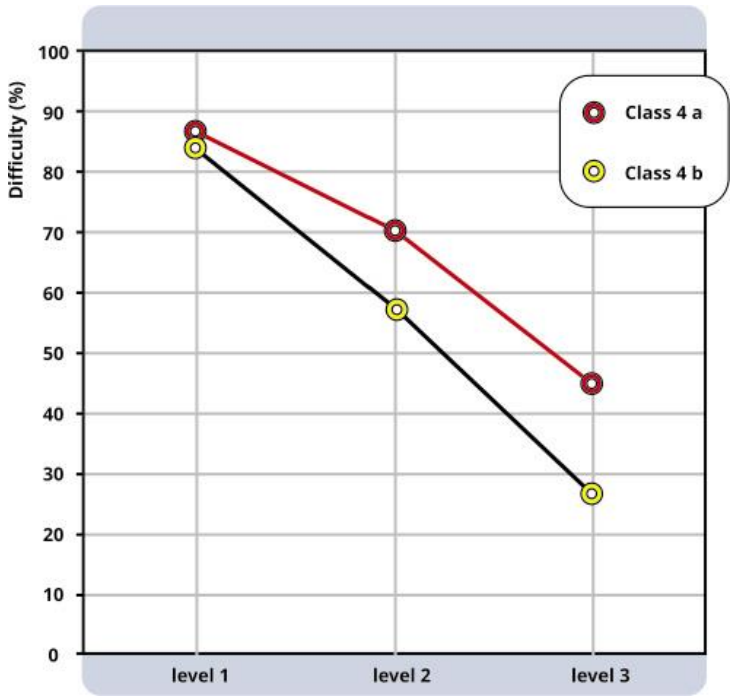


Figure 5.2. Profiles of two classes from the same school

Comparison of profiles is especially significant when we are dealing with classes that have close average score, i.e., generally demonstrate similar testing results (Figure 5.3).

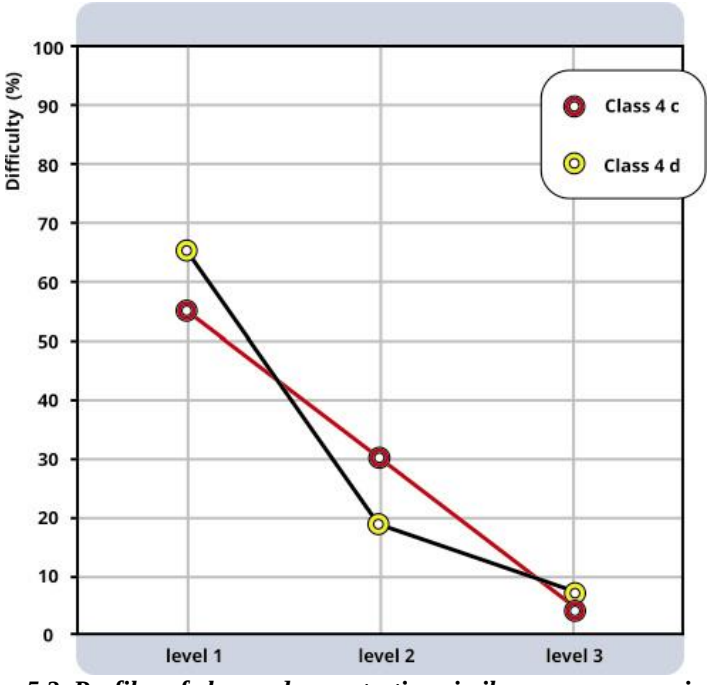


Figure 5.3. Profiles of classes demonstrating similar average score in math

As is seen from the diagram, class *4 c* ranks below class *4 d* in the comprehension of subject material (scale 2), and grades up in the total outcome due to a better assimilation of curriculum at level 1 (level 1 scale). If the classes were initially equal in terms of ability, one can assume that class *4 c* mostly focused on solving the large number of simple problems, and in class *4d* special attention was given to assimilation of main concepts, including modeling of respective mathematical relationships.

In addition to the individual and group results, substantive information is provided by the percentage distribution of a sample by grades of achievement (Table 5.3). These data are required to teachers to select the educational strategy adequate to the specific group of students.

School	Level 0	Level 1	Level 2	Level 3	Total
Secondary General School No.1	*20*	*54*	*25*	*1*	*100*
4a	*15*	*52*	*30*	*3*	*100*
4b	*20*	*56*	*24*	*0*	*100*
4c	*25*	*54*	*21*	*0*	*100*
Secondary General School No.2	*31*	*48*	*21*	*0*	*100*
4a	*21*	*50*	*29*	*0*	*100*
4b	*28*	*56*	*16*	*0*	*100*
4c	*42*	*39*	*19*	*0*	*100*
…..					

Table 5.3. Distribution of test takers by proficiency levels (%)

A string diagram is highly suitable for graphic representation of these data. Figure 5.4 illustrates a distribution of test takers by proficiency levels in different schools from the same region. The horizontal axis shows the share of students at each level, and the vertical axis indicates the schools. Schools are arranged in order of decreasing average total test scores.

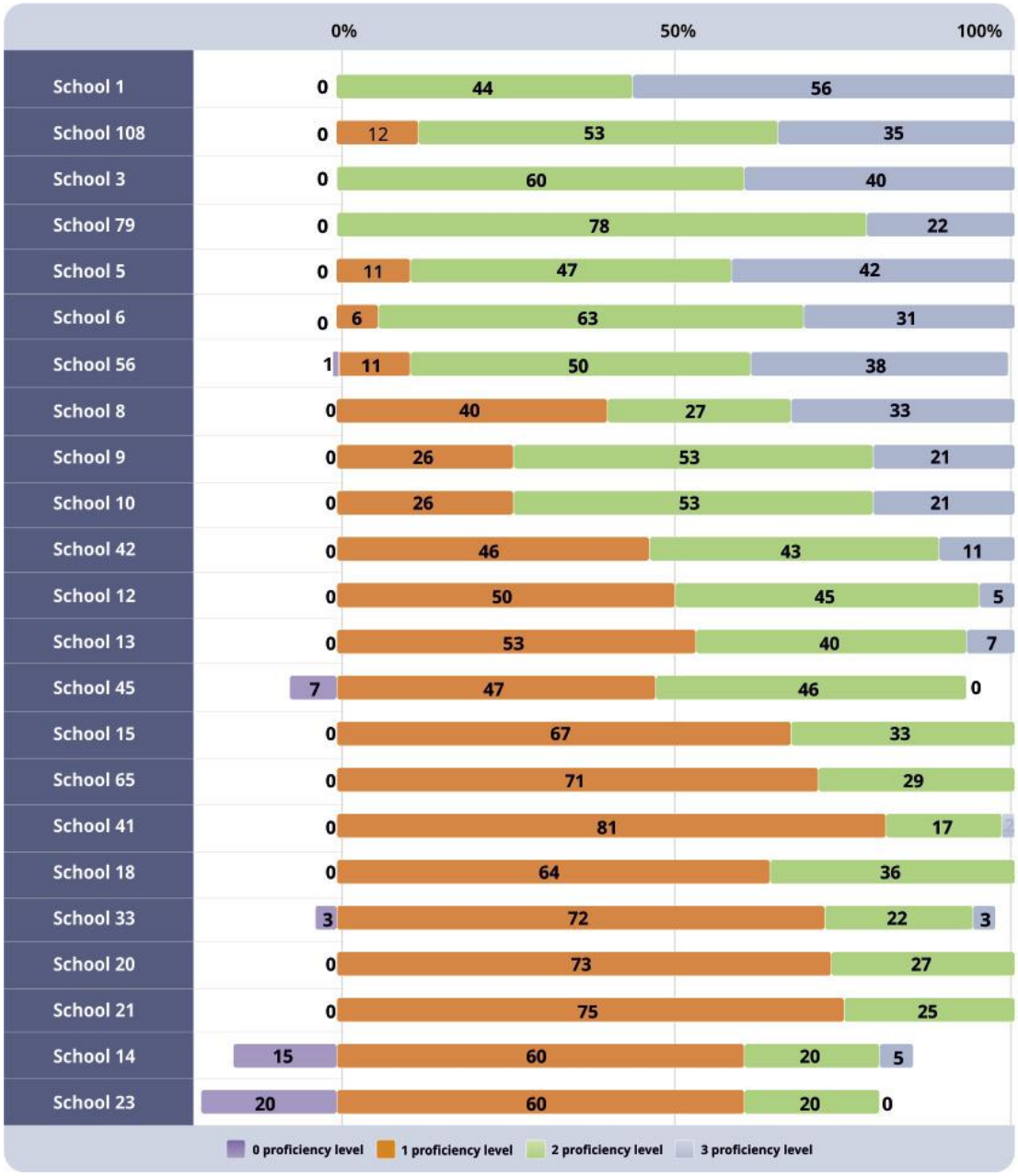


Figure 5.4. Distribution of students from different schools proficiency levels

Similar diagrams can be designed for different classes. Note that bar diagrams can be also used here. Thus, Figure 5.5 shows distribution of students by proficiency levels in four classes of the same school (7 classes took part in testing from this school).

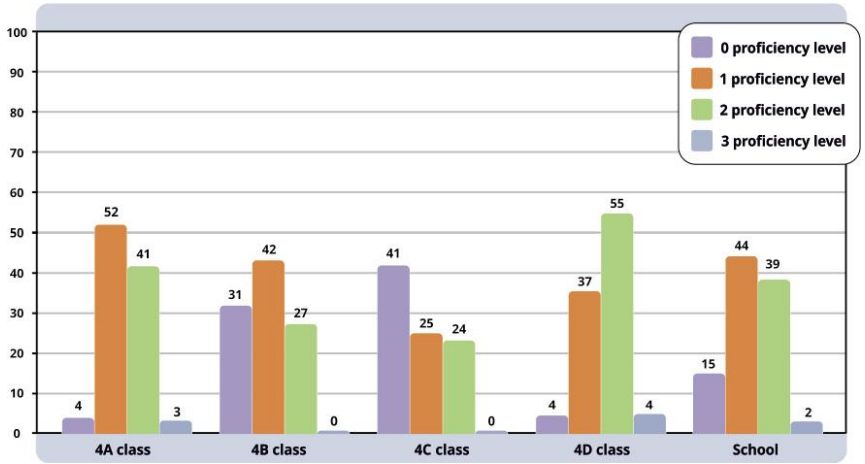


Figure 5.5. Distribution of students by proficiency levels

When comparing distributions by proficiency levels, one can see that level 1 dominates in class *4a*, which corresponds to the first (formal) level of subject assimilation. While in class *4d* level 2 is dominating, i.e., over one half of students have assimilated the subject at the second (reflexive) level. And, finally, in class *4c* 41% of students are at level 0, which means that they failed to assimilate even the first (formal) level. Knowledge of these relationships is highly relevant to the selection of strategies for the work with one or another class.

Separate diagrams can be made for each level to illustrate the share of students referred to the given level in the region as a whole (red bar on the right) and in each school (or each class). Figure 6.6 gives an example of such a diagram (for level 1).

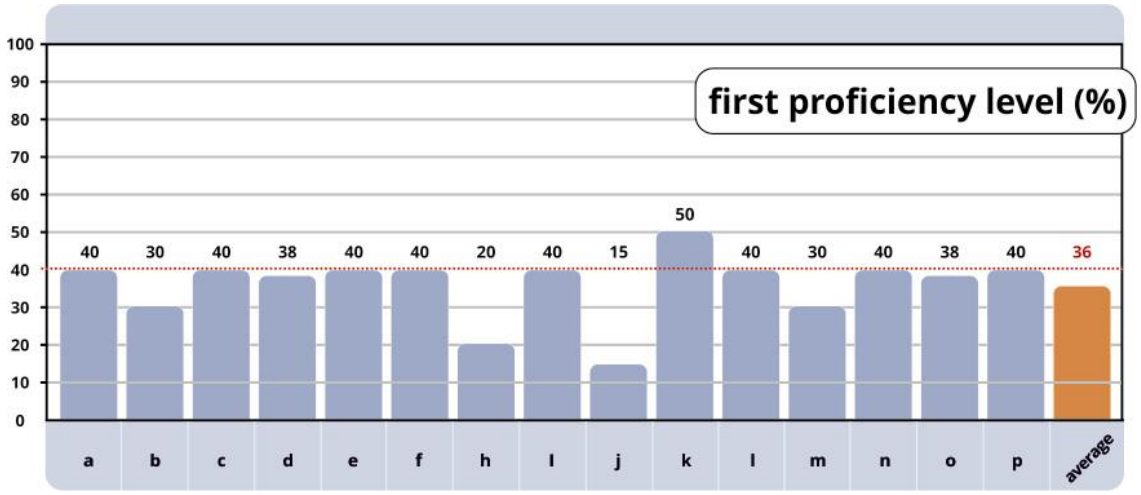


Figure 5.6. Data on the first proficiency level

5.2. DATA INTERPRETATION AND EVALUATION

Basic interpretation of individual testing results follows from the accepted multi-level pattern, and it is essentially included in the SAM indices. That is to say, the level that a child can reach reveals the leading type of his (or her) orientation within the curriculum. At the same time, such categorization of testing results lets the teacher see the next closest development zone for a specific child within the framework of the given curriculum area and also take this into account while doing the teacher's work.

When performing testing data interpretation and evaluation, one should always remember that full assimilation of the cultural content by a child (i. e., functional development, according to L. S. Vygotsky) is a long process doing beyond the time frame of studying the given curriculum. This consideration is strongly supported by the results of the same subject tests carried out by students of 4, 6, 8 and 10 grades. Below are the results of SAM math test (Figure 5.7).

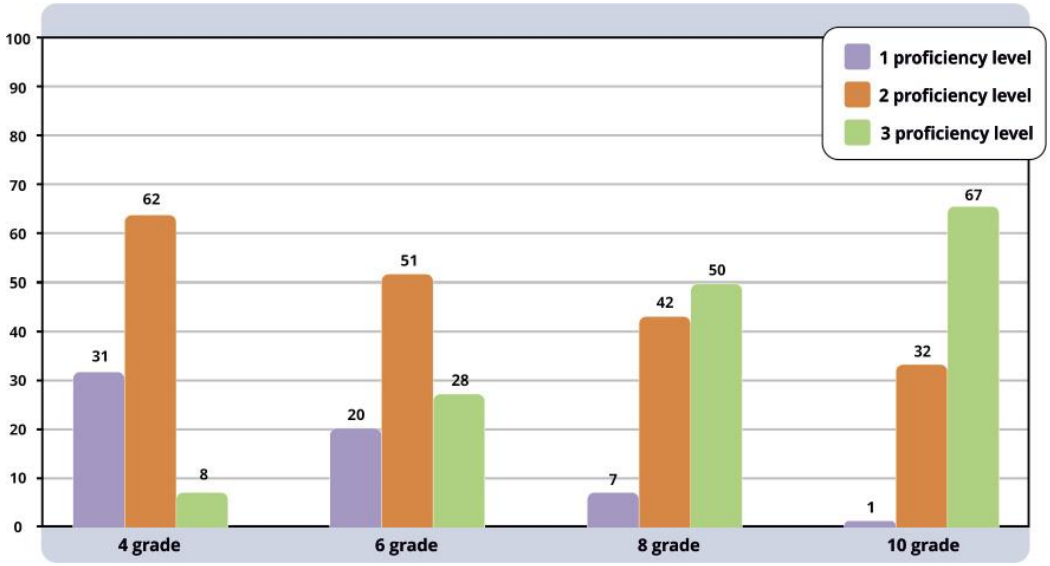


Figure 5.7. Distribution of students of different age by proficiency levels

The results of this study are in agreement with the cultural development theory (as divided into periods by D. B. Elkonin), and they give the reason to believe that the age-related norm by the end of the elementary school corresponds to the acquisition of the curriculum at the second level while the achievement of the functional level should be expected rather in the middle school. Beside the age-related norm that has the cultural and psychological basis, statistical norms (as arrived at over a representative student sample) provide an important assessment reference:

а) statistically average norm, which is an average test completion indicator over the whole student sample;

б) socio-cultural norm, or a realistic "tomorrow" norm, which is the test completion indicator for the students in the leading group.

Thus the following normative basis that includes four reference items can be provided for the assessment of the testing results within the SAM framework:

- Absolute (ideal) cultural norm – 100% test completion; Age-related norm by the end of the elementary school – demonstrating the achievement of the reflexive level; Statistically average norm – average test score over the whole sample;
- Socio-cultural norm – average test score over the group of leading schools.

5.3. TARGETING OF TESTING DATA

The list of main indicators obtained from SAM tests includes the following:

■ Matrix of primary estimates for each test item — *student*

■ Raw score — *student / class / school / district*

■ Achievement profile — *student / class / school / district*

■ Test score — *student / class / school / district*

■ Proficiency levels — *student*

■ Distribution of students (%) by proficiency levels: *class / school / district*

The above data contain a considerable amount of information that, in its entirety, can be of interest, above all, to educational researchers. As to various categories of practitioners, it makes sense to select specific parts of testing data that can be used in the context of their activities.

For example, *LOCAL EDUCATION AUTHORITIES* could be interested in the following set of indicators:

■ Test score — *school / district*

■ Distribution of students (%) by proficiency levels — *school / district*

The district-average test score gives an indication of the general level of educational outcomes for the school network as a whole. This scale score can be directly compared to similar data on schools from other districts thus allowing a comparative evaluation of the local school system.

School-average test score enables to evaluate the achievements of each school as compared to other schools, both local and those located in other regions.

Distribution of students by proficiency levels at the district level, above all, characterizes, the child population: first, in respect to compliance with the cultural and age norm (which can be accepted as the averaged educational achievements of several leading schools); second, from the viewpoint of qualitative homogeneity or inhomogeneity in terms of academic achievements. Peculiar features of the distribution can, in turn, raise questions concerning the implemented educational policy.

Context specialists can be interested in the analysis of specific context area assimilation at the school and district level. Such data can be useful for defining steps to improve education methods and professional upgrading of teachers.

The set of indicators that is of interest to *SCHOOL WORKERS* overlaps with the previous one but is substantially different:

■ Primary score — *student / class / district*

■ Achievement profile — *student / class / district*

■ Test score — *student / class / district*

■ Proficiency levels — *student*

■ Distribution of students (%) by grades of achievement: *class / district*

SCHOOL ADMINISTRATIONS ARE INTERESTED IN:

■ Test score — *student / class / district*

■ Distribution of students (%) by proficiency levels: *class / district*

Average test score of each class in the given school and district enables to compare classes between themselves in terms of academic achievements, and also to compare school outcomes with those at the district level.

It is also useful for the administration to know individual test scores of the most advanced students to be able to select potential participants of school olympiads.

And, finally, the administration should have a general idea of the students' distribution by proficiency levels, especially as compared to the results of the best schools in the district. This information can be relevant for making administrative decisions concerning the school's educational strategy.

TEACHERS are interested in almost all testing data:

■ Test score — *student / class / district*

These scale indicators help the teacher to evaluate the achievements of his/her students in comparison with that of students from parallel classes, and with the representative sample (the district and higher levels).

■ Achievement profile — *class / district*

Review of structured indicators enables the teacher to evaluate the qualitative specifics of his/her approach as compared to the averaged picture.

The following indicators help the teacher to focus on his/her class, and get a detailed picture of students' learning achievements.

■ Primary score — *student*

These data allows ranking of students in terms of curriculum assimilation success.

The following two indicators describe the qualitative level of each student's competency, and ratio between students with different levels of competency in the given class.

■ Proficiency level — *student*

This indicator helps to define the range of teaching targets for individual students.

■ Distribution of students (%) by proficiency levels— *class*

These data provide an insight on the organization of educational process adequate to the given class.

The last group of indicators enables to evaluate the qualitative level of subject competency of individual students, and see how they handle various test items to select adequate subject material.

■ Achievement profile — *student / class*

■ Matrix of primary estimates for each test item — *student*

5.4. SAM SPECIFIC OPTIONS AND LIMITATIONS

SAM tests differ from other tests in several aspects. First and foremost, the content area of the testing is revealed not via the normative assessment, but by indicating the key orientation means that ensure competent action. This characteristic of the SAM impacts the system of test units which may not fully correspond to the existing curriculum codifiers.

The second SAM characteristic involves the selection of items that gravitate towards the lower limits of the level which we had defined, that is those items which are formally least complex. As a result of this, a certain part of items, which have been traditionally used in education, escape the control area.

The third SAM characteristic lies in the fact that it has a built-in mechanism of the categorization of testing results within the framework of a multi-level pattern following a certain psychological theory. This specific may to some degree prevent SAM implementation in the studies conducted by experts who advocate different theories. Empirical approach is more relevant in such a case for it would aim to achieve objective material that can be independently analyzed by various experts holding different points of view. Tests for schools, however, must provide a frame for initial analysis and test results interpretation within the terms of a taxonomy of educational objectives, which calls for choosing one's theoretical stand. This strategy was implemented in the SAM tests, where the integral score of each student is revealed as a three-dimensional profile showing the correlation of the three options of orientation in the curriculum. This same result as shown in the graded scale will pinpoint the leading type of orientation and indicate the next development zone for a specific child which provides the teacher with an easier choice for his (or her) action strategy.

The value of the diagnostic representation will, of course, depend on the quality of the psychological theory implemented in the tool system. This is why we took the theory of cultural development by L. S. Vygotsky as a basis for our work of developing the SAM—this theory's adequate functioning has been confirmed in many multiple ways.

Literature

Adams, R.J., Wilson, M., Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. Applied Psychological Measurement. 21(1), p. 1-23.

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

Kardanova, E., Nezhnov, P. (2011): School Achievements Monitoring Toolkit: Assessment Framework. The paper presented at the 37-th Annual Conference IAEA 2011. Manila.

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.

Russian editions

Выготский Л.С. Мышление и речь //Собр. соч.: В 6 т., т. 2. М.: Педагогика, 1982.

Выготский Л.С. История развития высших психических функций //Собр. соч.: В 6 т., т. 3. М.: Педагогика, 1983.

Гальперин П.Я. Психология как объективная наука. М., 1998.

Давыдов В.В. Теория развивающего обучения. М.:ИНТОР, 1996.

Диагностика учебной успешности в начальной школе. М.: ОИРО, 2009.

Исаев Е.И. Психологическая характеристика способов планирований у младших школьников // Вопр. психологии. 1983. № 2.

Карданова Е.Ю. Моделирование и параметризация тестов: основы теории и приложения. – М.: Федеральный центр тестирования, 2008, 304 с.

Карданова Е.Ю., Нейман Ю.М. Проблема выравнивания в современной теории тестирования. Вопросы тестирования в образовании, 2003, № 8, с. 21-40.

Магкаев В.Х. Экспериментальное изучение планирующей функции мышления в младшем школьном возрасте // Вопросы психологии, 1974, № 5.

Микулина Г.Г., Савельева О.В. К психологической оценке качества знаний у младших школьников // Психологическая наука и образование. 1997. № 2.

Нежнов П.Г., Медведев А.М. Исследование теоретического анализа у школьников // Вопросы психологии, 1990, № 5.

Нежнов П.Г. Опосредствование и спонтанность в модели «культурного развития» // Вестн.Моск.Ун-та. Серия 14. Психология. 2007, № 1 (Специальный выпуск: 40 лет факультету психологии МГУ).

Нежнов П.Г., Карданова Е.Ю., Эльконин Б.Д. Оценка результатов школьного образования: структурный подход // Вопросы образования, 2011, № 1, стр. 26-43.

Савельева О.В. Психологические критерии качества знаний младших школьников // Автореф… дисс. канд. пед. наук. М., 1989.

Эльконин Б.Д. Введение в психологию развития. М., 1994.

Эльконин Д.Б. Избранные психологические труды. М.: Педагогика, 1989.

ANNEX 1: SPECIFICATION OF THE SAM MATH TEST

1. OBJECTIVE

The test is intended for the assessment of subject competencies of primary school students to evaluate the level of assimilation of the *Mathematics* subject content. The test model is based on the theory of cultural development (L. S. Vygotsky, V. V. Davydov, D.B. Elkonin et al.), and assumes assessment of students' competencies at three basic levels: formal, reflective and functional.

2. TARGET AUDIENCE

The test is targeted for primary school graduates and can be performed by grade 4 and 5 students.

3. CONTENT

The test includes the main math areas from the primary school curricula.

The content was selected in accordance with Federal State Standard of Primary Education (MOED Order # 373 of October 6, 2009 *On Approving and Enactment of the Federal State Standard of Primary Education*).

Subject content of the test is divided in five areas.

"Numbers and Calculations". This content area is relevant to the formal aspect of the concept of natural numbers (positional representation of numbers, standard algorithms of operations with numbers, order of operations, properties of operations). It alsoincludes materials related to representation of numbers on the coordinate line. The latter is important for the understanding of real numbers and assimilation of coordinate method.

"Measurement of Values". This content area includes materials related to direct and indirect measurement operations, and also incorporates geometric measurements.

As to the applied aspect of this content area closely related to specific practical measurements and their representation as diagrams and charts (*data analysis*), it can rather be relevant to the *Outside World* subject.

"Regularities". This content area is related to construction of numerical and geometric sequences and other structured objects, and measurement of their quantitative parameters. This area is highly important for the development of mathematical thinking(first of all, algorithmic and combinatory).

"Dependences". This content area is related to identification and description of the mathematical structure of relations between values usually represented in test items.

"Elements of Geometry". This content area covers geometric materials related to identification of spatial forms and relative position of objects.

The content framework of the math test can be presented as a matrix (Table 3.1.) that includes:

■ Subject content areas (6 areas);

■ Mathematical tools (concepts, principles, formulas, algorithms, etc.) providing the orientation for mathematical operations.

Table 1. Content of the math test

Content areas	**Orientation tools for mathematical operations**
Numbers and Calculations	■ *Sequence of natural numbers* ■ *Number line* ■ *Positional principle* ■ *Properties of arithmetic operations*

	■ *Order of operations*
Measurement of Values	■ *Relationship between the number, value and unit* ■ *Whole-part relationship* ■ *Formula of rectangle area*
Regularities	■ *"Induction step"* ■ *Recurrence (periodicity)*
Dependences	■ *Relationship between like values (equality, inequality, multiplicity,difference, "whole-part")* ■ *Direct proportion between values* ■ *Derived values: velocity, labor productivity, etc.* ■ *Relationship between units*
Elements of Geometry	■ *Form and other properties of figures (main types of geometrical figures)* ■ *Spatial relationship between figures* ■ *Symmetry*

4. PRINCIPLES OF TEST DESIGN

The test was developed using two approaches: norm-referenced and criterion-referenced ones — combined in accordance with the modern test theory — Item Response Theory (hereinafter — IRT), which also allowed a criterion-referenced interpretation of the test score scale.

Norm-referenced approach enables to compare the performance of a student with that of other test takers, as well as with his/her previous performance. To this end, each test taker is assigned an integral test score obtained as a result of mathematical treatment of test results. Test scores of all test takers are shown on a common scale, irrespective of the time of testing and specific set of items done.

The second (criterion-referenced) approach allows a qualitative evaluation of the content area assimilation through indicating the type of orientation in problem solving. To this end, a graduated scale of achievements was developed based on integral scores of test takers and benchmarks classifying all test takers in groups corresponding to different qualitative levels of achievement.

5. TEST STRUCTURE

The test consists of three level items grouped in blocks. In total, there are 45 items and 15 blocks. The test can be viewed as a system of three subtests with each representing a set of items of the same level referring to different content areas. The overall structure of the test is shown in Figure 1.

Figure 1. Test structure

6. ARRANGEMENT OF TEST ITEMS

A block consisting of three items (levels 1, 2 and 3) corresponding to a subject content area is a structural unit of the test. Items are presented in blocks. The sequence of blocks makes no difference.

7. RATIO BETWEEN ITEMS REFERRING TO DIFFERENT CONTENT AREAS AND LEVELS

Table 2. Tentative ratio between different items 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*

8. TYPES OF ITEMS

The following types of items are used in the test:

■ completion items with a brief answer;

■ multiple-choice items with a choice from 4-5 offered options;

■ items requiring constructions.

The majority of items (about 80%) are the completion ones with a brief answer.

9. NUMBER OF VERSIONS

4 versions of the test with similar statistical parameters were developed. All versions include common items enabling to perform equating and plot the scores of all test takers on a single scale. The number of common items in different versions makes up at least 6 (at least two common blocks).

10. SCORING OF ITEMS

Scoring procedure uses a dichotomous approach: students get 1 point for the correct answer and 0 for incorrect (or absent) answer. Therefore, 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 level equals 15.

11. RECOMMENDED DURATION OF TEST TAKING

Recommended duration of test taking is 90 minutes (two 45 minute lessons with a break). Tentative time of performing level 1 items – 1 minute, level 2 – 2 minutes, level 3 – 3 minutes. Testing can be conducted during two days: a half of the test (blocks 1-8) to be completed on the first day, and the second half (blocks 8-15) – on the second day.

12. FORMS OF TEST TAKING

The test can be offered in paper-based and computer-based forms. It is assumed that testing data do not depend on the form of test taking but this issue is still being studied.

13. CONDITIONS OF TEST TAKING

Testing can be performed by a primary school teacher (if the testing takes place in grade 4) or math teacher (is testing takes place in grade 5). The teacher in charge of testing should help the test takers to complete identification information on the front of the test booklet, explain obscure passages in the instructions, keep track of the time and maintain order in the class during the testing.

14. PROCESSING OF TEST RESULTS

The SAM toolkit includes an automated information system (computer module) intended to estimate test takers and automatically generate various reports (tables, charts and diagrams) both on each test taker, and different classes and schools. At his/her wish, the teacher can carry out an independent scoring (manually check the items using the keys). However, computer module would not only simplify the task but also enable to utilize the whole range of reports generated by the system to analyze the test results. In addition, the module can accumulate the data and take account of previous testing data when generating reports on subsequent testing thus allowing a comparison of results.

15. ESTIMATION OF TEST TAKERS

Estimation is performed using a special measuring technique based on the modern test theory IRT. The following parameters are estimated for each student:

■ Integral test score placed on the scale common for all test takers irrespective of the time of testing and specific set of items they have done. Integral test scores are presented on a 1000-point scale;

■ One of the four achievement grades:

■ **ZERO LEVEL:** even the first (formal) level has not been assimilated;

■ **FIRST LEVEL:** only the first (formal) level has been assimilated;

■ **SECOND LEVEL :** the second (reflective) level has been assimilated;

■ **THIRD LEVEL:** the third (functional) level has been assimilated.

16. PRESENTATION OF TEST RESULTS

For the purpose of test results presentation, the system generates two types of tables:

■ Aggregate data on schools and classes;

■ Individual data on test takers.

Automated information system allows presentation of estimation data in the form of charts and diagrams showing:

■ Distribution of students by proficiency levels;

■ Subject success of classes (profiles).

The above tables and illustrations can be supplemented with data obtained from surveys among the test takers, which enables to track the influence of various factors on the quality of students' learning achievements.

ANNEX 2: SAMPLES OF MATH ITEM BLOCKS

NUMBERS AND CALCULATIONS

M-C-03-1-1
M-C-03-2-1 *Find out the largest of the following numbers:* *10073, 1801, 9999, 10110* *Answer: _____*
Key: 10110
Item difficulty: 85%

M-C-03-1-2
M-C-03-2-2 *What number should be multiplied by 152 to obtain a value larger than 1300 but smaller than 1500?* *Answer: _____*
Key: 9
Item difficulty: 61%

M-C-03-1-3
M-C-03-1-3 *Below are three numbers:* *53* *535* *5* *Write them down one by one in such an order that the resultant six-digit number be as small as possible.* *Answer: _____*
Key: 535355
Item difficulty: 25%

MEASUREMENT OF VALUES

M-M-11-1-1
M-M-11-1-1 *The square side equals 3 cm.* *Find out its perimeter.* *Answer: _____ cm*
Key: 12
Item difficulty: 76%

M-M-11-1-2
M-M-11-1-2 *The length of a rectangular sports ground equals 40 m.* *Find out its width if the perimeter equals 120 m.* *Answer: _____ m*
Key: 20

Item difficulty: 41%

M-M-11-1-3

M-M-11-1-3

A square with 8 cm side was cut in two rectangles.
The perimeter of one of them equals 26 cm.
What is the perimeter of the other rectangle?
Answer: _____ cm

Key: 22

Item difficulty: 22%

REGULARITIES

M-R-05-1-1

M-R-05-1-1

Squares and triangles are arranged
in a row following a certain rule:

Continue the row (draw the next three figures).

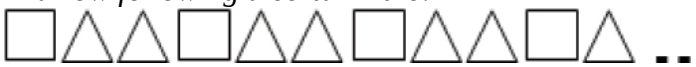
Ключ:

Item difficulty: 81%

M-R-05-1-2

M-R-05-1-2

Squares and triangles are arranged
in a row following a certain rule:



How many triangles in such row if it consists of 30 figures?

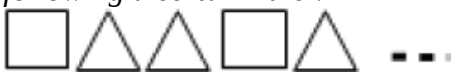
Key: 20

Item difficulty: 47%

M-R-05-1-3

M-R-05-1-3

Squares and triangles are arranged in a row
following a certain rule :



What is the longest row (how many figures are there) can be constructed if there are 7 squares and 8 triangles?
Answer: _____

Key: 13

Item difficulty: 16%

DEPENDENCES

M-D-03-1-1

M-D-03-1-1

Nick is higher than Peter by 15 cm. nick's height is 1 m 60 cm. Find out Peter's height.
Answer: _____

Key: 1 m 45 cm (or 145 cm)

Item difficulty: 70%

M-D-03-1-2

M-D-03-1-2

Last year Alex was lower than Mary by 7 cm.
During the year, Alex has grown up by 9 cm, and Mary – by 4 cm.
Who of them is higher than the other, and by how much?
Answer: _____

Key: Mary is higher by 2 cm

Item difficulty: 42%

M-D-03-1-3

M-D-03-1-3

Mike's height is 1 m 50 cm. Nick's height differs from that of Mike by 5 cm. Victor's height differs from Nick's height by 10 cm.
It is known that a year ago Victor's height was equal to 1 m 48 cm, and now it is lower than 1 m 60 cm. What is Victor's height now?
Answer: _____

Key: 1 m 55 cm (or 155 cm)

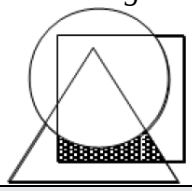
Item difficulty: 24%

ELEMENTS OF GEOMETRY

M-G-01-1-1

M-G-01-1-1

Put a point so that it lies within the square and triangle but outside the circle.



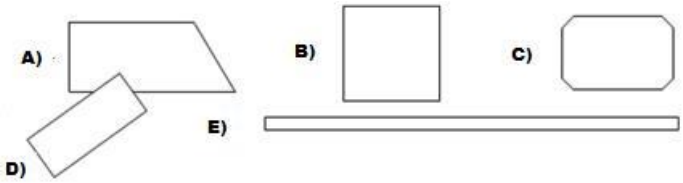
Key: Any point within the dashed area (but not the area itself).

Item difficulty: 82%

M-G-01-1-2

M-G-01-1-2

Which of the figures shown below are rectangles?
Mark all correct answers.



Answer _____

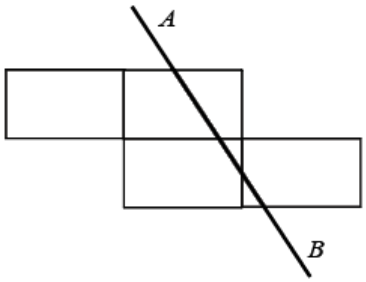
Key: B, D, E

Item difficulty: 63%

M-G-01-1-3

M-G-01-1-3

How many rectangles are crossed by line AB?



Key: 6

Item difficulty: 18%