School Achievements Monitoring Toolkit: Assessment Framework

Elena Kardanova, Peter Nezhnov
Center for International Cooperation in Education Development
(Russian Federation)
e_kardanova@mail.ru

The presentation describes the assessment framework of the School Achievements Monitoring Toolkit (SAM) that is being developed by the Center for International Cooperation in Education Development (CICED, Russian Federation).

The purpose of SAM is assessment of subject competences of primary school students in mathematics, language and science. The particular feature of SAM is that its assessment model is based on Vygotsky's theory and is designed to evaluate examinees' subject competences on three basic levels: formal, reflexive and functional. Such presentation of assessment results opens a way to deeper interpretation of learning outputs.

Test items are developed for each subject content area in accordance with three levels indicated above. Each block of three items works as a detector that characterizes the level (quality) of mastering the relevant part of the learning program. Examples of items blocks will be demonstrated.

The results of SAM pilot testing will be presented including test- and item analysis and evidence of SAM validity. Different approaches to test results scaling and presentation are discussed.

SAM is supposed to be used in different countries. So questions of test translation and adaptation, as well as item bias are under consideration. The results of corresponding research will be produced.

Key words: Assessment, Measurement, IRT Modeling

Introduction: Overview of SAM Assessment

School Achievements Monitoring Toolkit (SAM) is the attempt to develop an instrument of school achievements assessment through their measurement and qualitative (structural) characteristic. The object of assessment includes subject competences of primary school students, that reflect how well they acquire basic school subjects such as mathematics, native language, science.

The SAM toolkit consists of a set of tools for monitoring on the class/school level the academic subject-matter competences of primary school students in such areas as: mathematics, science, native language. The tools include subject tests in each area, questionnaires for collecting context information and recommendations on the test results interpretation and usage. Also SAM has a recording system, based on a measuring technique with the help of which examinees scores and different report forms are generated. Usage of IRT allows to put test results from different assessments to the same fixed scale that gives an opportunity to compare students achievements, and in the course of time in particular.

It is assumed that the findings of the assessment will be primarily used for optimization of educational process. In other words the basic users of the toolkit SAM are teachers, methodologists, school administration and educationalists including local education management departments.

International monitoring studies, such as TIMSS, PISA, etc., have prompted SAM development. These studies have formed modern vision of the objectives of school education and developed advanced patterns of educational assessment through measurement. At the same time these findings appeared to be incomplete for those teachers who are concerned about quality of acquiring the syllabus. In order to make up for this deficit a group of Russian scientists have developed a test which incorporates the mechanism of diagnosing the quality of syllabus acquisition.

The theoretical base of the toolkit developed is laid by the theory of cultural development of a child, outlined in L.S. Vygotsky's works and further developed by his descendents – D.B. Elkonin, P.J. Galperin, V.V. Davydov, etc. This theory suggests that learning as a necessary prerequisite of a child's psychic development, involves acquisition of sign structures, which crystallize all the basic landmarks of generalized action patterns: a) external characteristics of classes of object situations and corresponding actions; b) understanding of relevant relations within this class of situations, which define direction and limits of possible transformations; c) the essence of the action pattern, i.e. contexts of its meaningful applications.

These three types of landmarks are featured in the cultural action pattern simultaneously. Still when adopting the pattern the role of the cornerstone is taken on first by the external characteristics of the object situation, then the understanding of the relevant relations within it, and finally the corresponding sense field. These three types of action orientation serve as markers of cultural action patterns.

In the first case the pattern is generalized to the minimum and incorporates a limited number of typical situations and corresponding action patterns. The second case implies revealing of the essential link which makes up the basis of the action pattern. This offers a possibility to solve every problem within the given class and corresponding to the given pattern. Finally, the third case, the psychology of which is not yet completely studied [Vygotsky 1982; Galperin 1998; Davydov 1996; Piaget 1969; Nezhnov 2007, 2009; Elkonin 1989], features the action pattern as characterized by functionality, i.e. the possibility of being employed in various contexts.

SAM toolkit is principally characterized by developing tasks of three different levels clustered into groups (blocks) when developing tests for each school subject area [Nezhnov, Kardanova, Elkonin 2011]. Each of these clusters functions as a detector of how well a certain subject area has been acquired (this is done through identifying the most difficult task a student managed to complete).

When developing each block a developer employs a system of indicators (a typology of tasks), which reflect the generalized criteria of action pattern acquisition outlined above.

Thus, an indicator of pattern acquisition on the first level implies completing those tasks, in which the link between the task and the action pattern as its finding is transparent. This group includes tasks in which the description of the problem situation makes it apparent that they refer to a certain class with a known solution (so called typical tasks).

An indicator of action pattern acquisition on the second level implies solving problems in which one cannot employ a typical procedure, thus it is crucial to identify the relevant relation which determines the framework of searching for a solution which can be used to build up a solution to the given task. This group includes tasks which avoid direct links between the description of the problem situation and the solution sought: tasks offering an indirect description, abstract tasks, tasks with a description offered in a various visual forms (for instance, tasks involving verbal and non-verbal description – graphs, charts, etc.) and others.

An indicator of action pattern acquisition on the third level implies completing those tasks which need referring to the bank of possibilities of an action pattern. This group is believed to include those tasks which need amending the relevant relation of the task and identifying an array of possible actions to choose a solution which answers to a certain context-determined requirement.

Specific examples of such blocks of tasks within different subject areas will be featured further in the paper.

Thus, a test in any subject features a number of task blocks of the first, second and third levels in the defined subject areas. In accordance with this, a test can be viewed as incorporating three subtests. Each subtest features a set of tasks of one group tapping into various subject areas. When processing test results separately for each level one can draw a chart to outline subject area acquisition pattern in a certain subject by a class or a large group of students (see Figure 1).

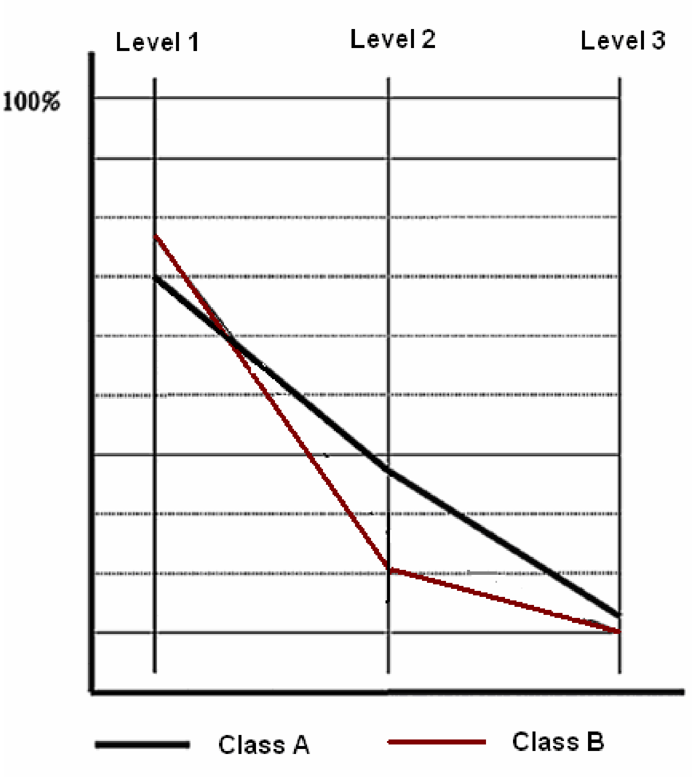

Figure 1. Profiles of two classes

Thus structural approach to featuring test results offers wider possibilities of content-wise interpretation of test findings and qualitative characterization of the competence assessed. The chart (Figure 1) suggests that group A exhibits better understanding of the material learnt than group B (see different results for level 2 against similar indicators for level 1). These preliminary findings can also be used to identify those subject areas which haven't been fully acquired by each of the groups on the second level. As regards the third level scale, it reflects the trend which is believed to be in its formation in primary school given age development descriptions. The test findings can be also placed on an integral scale thus obtaining quantitative characterization of each student's competence level in a particular subject area and ranging the test performance of a student, a class, etc.

SAM Mathematics Framework

An example to follow looks into the test in mathematics, that has been developed by a group of researchers under S. Gorbov.

Five content areas are included in the test in mathematics. They are: numbers and calculations; value measurement; mathematical regularities; dependence between values; geometry elements. Each part outlines mathematical tools (notions, principles, algorithms), which are to be comprehended to enable understanding and fulfilling of appropriate mathematical operations. Besides, each part features indictors of subject acquisition for each subject area. Let us look into "Numbers and calculations" part to see how it works.

In order to determine how well students have managed to acquire the material on the first level it is necessary to construct tasks which imply performing a mathematical operation and standardized operations, used when performing calculations such as result assessment or approximating. The tasks of the second level are built upon identifying and considering multidigit number or expression structure rather than mere calculations. Analogously this level incorporates those tasks which require that a student works out his/her own strategy of calculating. Finally, the third level features those tasks which need an abstract expression specified given certain description of the problem situation. This can be exemplified with the case to follow.

What number shall we get if 10472 is divided by 34?
Solution: __________

Petya who is absent-minded copied a sum in which he was meant to multiply two digits. The first factor was copied correctly, it was a 7, the second factor was copied inaccurately as Petya switched around the figures in this factor. Thus the solution came out wrong, it turned out to be 147. What would the response be if Petya copied both factors accurately?
Solution: __________

What would be the greatest possible solution if we substitute letters in the equation AB5+BC2 with numbers (different letters should relate to different numbers)?
Solution: __________

In the example all three tasks relate to "Numbers and calculations" part. The first task of the block implies mere use of the calculation rule (algorithm) (successful completion about 70%). The second task implies analyzing the mistaken mathematical operation (taking in consideration the position principle) and working out a program of its amendment (34%). Finally, the third task implies "amending" the position principle to define some particular solutions to the equation, which would answer to the requirement of getting the highest possible solution (15%).

Different versions of a test are built of such blocks. Math test includes 15 three-level blocks, 45 items in total. Each block relates to the same content at different levels. In accordance with it, test can be considered as consisting from three subscales. Each subscale represents a set of items of the same level but different contents areas.

The most of test items have an opened format with a short answer in the numeric or verbal form. Some items have multiple choice format (with one correct option from four or five proposed) or other formats (matching, required construction etc.). The test time is 90 minutes. All items are scored dichotomously: examinee gets 1 point for correct answer and 0 points otherwise.

SAM Questionnaires

The questionnaires aim to collect context-determined information about the factors that influence students' achievements. They are offered to students as well as to their classroom teachers. These questionnaires stand out due to their focus on identifying those factors that directly influence the qualitative characteristics of a child's development and thus the test results. A questionnaire for students includes questions about a student's family; his/her attitude to school, teaching and learning; school activities; classroom; relations with other students, etc. A questionnaire for teachers includes questions related to availability and quality of training programs, organization of educational process, availability of modern educational technologies and their efficiency, etc.

Six factors are chosen as the most important ones that influence students' achievements: psychological ambience at school (friendly to hostile); extracurricular activities at school (full to scanty); syllabus (thorough to basic); communication style of a teacher with students (democratic to authoritarian); communication between students (informative to cliche); parents - school relations (close to sporadic).

SAM Assessment Design

Measurement model

As described above, all tests can be considered as consisting of three subtests. Each subtest represents a set of items of the same level but different content areas.

All subtests of the SAM tests measure related (but supposedly different), latent examinees' characteristics. So, the tests in MASS are assumed to be multidimensional. There are three approaches in the item response modeling to such kind of tests: unidimensional, consecutive and multidimensional. All these approaches were analyzed and compared in [Kardanova, 2010]. It was shown that the consecutive approach was unacceptable for the SAM tests data because of extremely high standard errors of students' measurement by each subscale separately. The unidimensional model as well as the multidimensional model based on literacy levels adequately account for the SAM test data, although more complex multidimensional model provides slightly better explanation. However the correlation between variables under the multidimensional model is very high (0.75 and higher) and additional research was shown that tests could be considered as essential unidimensional ones. And what is more, using multidimensional models has technical difficulties connected with need of different scales equating, that is hard to implement because scales don't have common items. As a result, multidimensional model was used for SAM validization rather than for students scaling.

Thus, unidimensional approach is applied for test data modeling and students scaling, more exactly one parameter dichotomous Rasch model or OPLM (One Parameter Logistic Model) [Rasch models…,1995]. All analyses were conducted with Winsteps or with OPLM software.

Calibrating test items and evaluating fit of the model

Firstly at the stage of pilot testing all test items were calibrated and their quality was confirmed. At the stage of item parameter estimation all omitted and not-reached items were treated as missing data. However, at the stage of generating students test scores these items were treated as incorrect responses.

After the calibration was completed, the fit of Rasch model was evaluated. For this purpose two approaches were used: analysis of item fit statistics available in Winsteps and comparing theoretical and empirical item characteristics curves (ICC). Figure 2 shows an example of ICC plot for one of math items generated by Winsteps. In this plot, the horizontal axis represents the profiency scale, and the vertical axis represents the probability of correct response. The theoretical curve based on the estimated item parameters is shown as a red line.

The empirical results are represented by daggers. The empirical results are first obtained by dividing the whole sample into 10 groups of equal size and then by counting the proportion of students in each group responding the item correctly. Additionally the boundaries of confidence intervals for these proportions are shown.

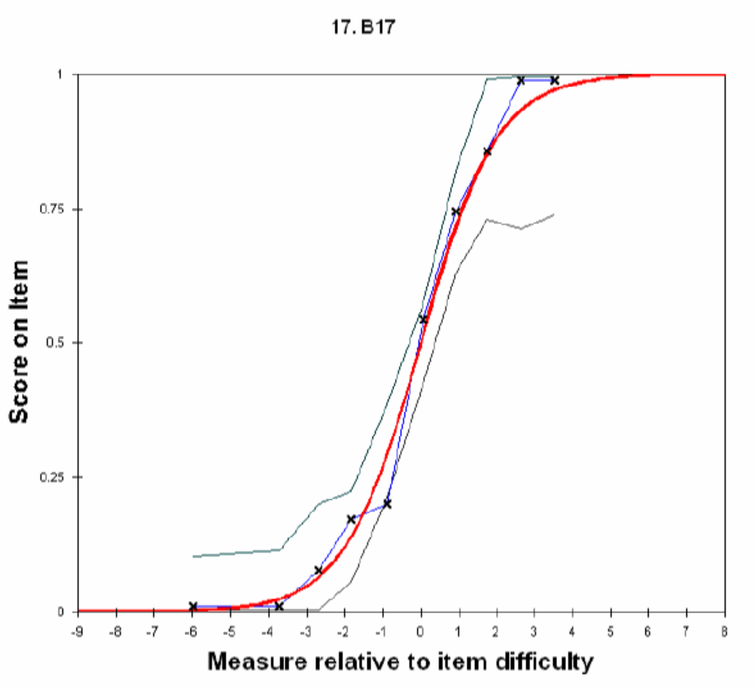

Figure 2. An example of ICC for math test

The final version of tests contains only valid items that fit the model. These tests are used for students assessment.

Estimating students achievement

All test scores are rescaled from a logit scale (that is not suitable for reporting purposes as it containes negative and fractional values) to a scale that has been formed as a result of a special scaling study. This scale has been established by appropriate linear transformation to create a scale with mean of 500 and standard deviation of 50. All test results from further assessments are transformed to this scale by applying the same linear transformation.

The theoretical levels of mastering action patterns determine the benchmarks that are used to describe students' achievement at four different degrees of mastery. There are four degrees of mastery identified:

0 degree: students belonging to this cluster can complete fewer than 50% of first level items. The probability of their completing items of the second and third levels is verging 0.

1 degree: students belonging to this cluster can complete at least 50% of first level items, but fewer than 50% of second level items. The probability of their completing items of the third level is very small.

2 degree: students belonging to this cluster can complete at least 50% of second level items, over 80% of first level items, but fewer than 50% of third level items.

3 degree: students belonging to this cluster can complete at least 50% of third level items. At the same time they will most likely complete any first level item and at least 80% of second level items.

Benchmarks establishment is shown in Figure 3. In the plot, the horizontal axis represents the proficiency scale, and the vertical axis represents expected percentage subtest score. Three curves are subtest characteristic curves. They are the theoretical curves based on the estimated item parameters. The benchmarks are indicated at the bottom, on the axis. Then they are transformed from the logit scale to the fixed proficiency scale. As a result we get the following benchmarks: 450 (the boundary between 0 degree and 1 degree); 520 (the boundary between 1 degree and 2 degree) and 590 (the boundary between 2 degree and 3 degree). The 1,2,3 degrees can be interpreted as corresponding to reproductive, reflexive and functional levels of mastering action patterns accordingly. The 0 degree means that even the reproductive level hasn't been attained.

These levels of mastering compose the basic taxonomy of educational goals, which has a psychological background, i.e. it indicates cultural-psychological structures which are crucial for competence development from immature to mature stage.

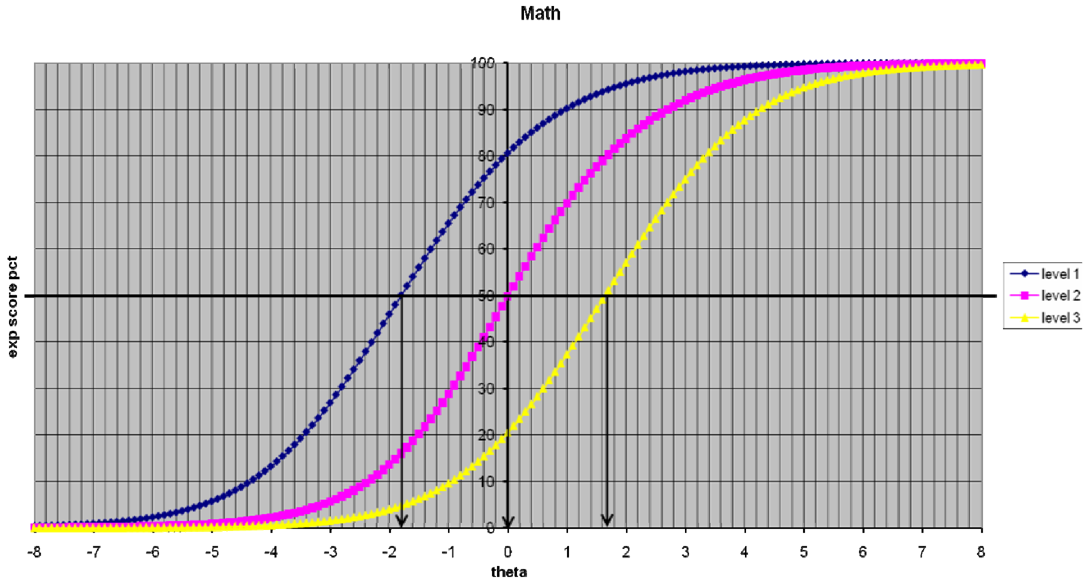

Figure 3. Establishment of benchmarks to divide examinees into different clusters

Item and Test Characteristics: Results of Pilot Testing Analysis

The results of the math test analysis are presented here. The data for this study have been collected in SAM pilot testing in Krasnoyarsk region of the Russian Federation. The sampling procedure included two variables: type of school and school location. All examines were 11-year-old students of the last (fourth) grade of primary school. The total number of participants for this test form was 484.

Table 1 contains summary of classical test statistics.

Table 1. *Summary of classical statistics*

Maximum possible raw score	45
Maximum obtained raw score	41
Minimum obtained raw score	0
Average raw score (standard deviation)	19.3 (7.9)
Average item difficulty	0.43
Average discrimination index	0.42
Average point-biserial correlation	0.40
Standard error of measurement	2.6
Reliability (Cronbach's Alpha Coefficient)	0.89

The theoretical model on the base of which the toolkit is being developed implies that items of three levels that relate to the same subject area should feather a proper hierarchy of items completion. So, if subject area is mastered on the second level, it is presumed that it is mastered on the first level too. That means that a student who has completed a second level item, should also be able to complete a first level item in the same block. In other words, the hierarchy of the difficulty levels should be observed inside each block of three items. Almost all items blocks meet this demand in the test analyzed: completing potential decreases gradually from the first level to the third one in each block. This serves as an additional argument in favour of the validity of the instrument. Table 2 contains average difficulty levels and discrimination indeces for each subtest.

Table 2. *Average difficulty levels and discrimination indeces of subtests*

	Average difficulty level	Average discrimination index
Items of the 1-st level	0,70	0,50
Items of the 2-nd level	0,41	0,48
Items of the 3-d level	0,18	0,27
The whole test	0,43	0,42

The results of IRT analysis are not presented here because of limited paper size. Only items exhibiting good measurement properties were selected for final version of tests.

Reporting students achievement

The SAM recording system generates a number of tables and graphs aimed at reporting students achievement and comparing different classes and schools. Figure 4 represents one of them. In the plot, the horizontal axis represents the percentage of students, and the vertical axis represents codes of schools analyzed.

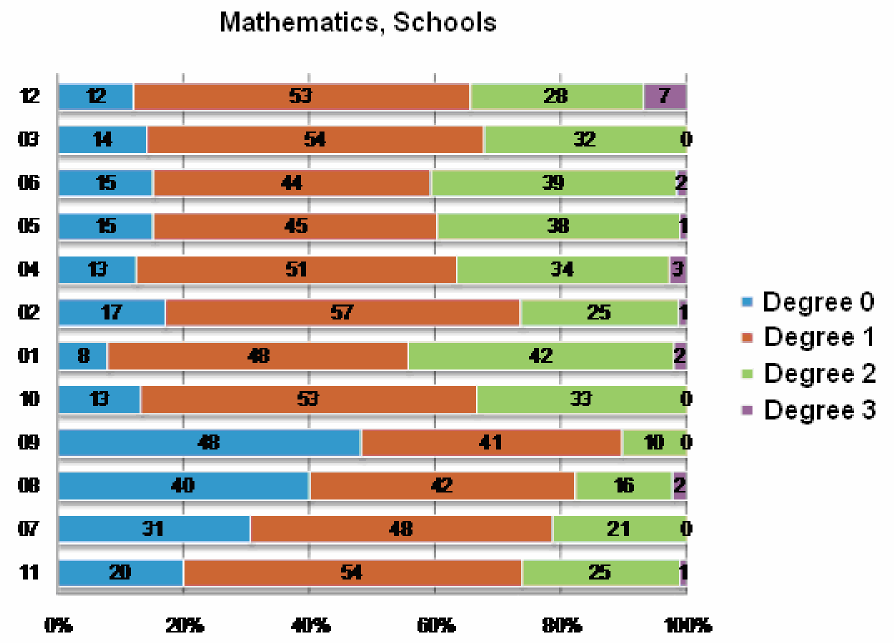

Figure 4. Distribution of students of different schools

We see that a number of students at the 3 degree is very small for all schools. This fact can be explained by lack of functional mastering level for most primary school graduates. To confirm it, an additional research aimed at comparing the test completion by students of different age was conducted.

The participants were students of four grades: 11-year-old students of the last (fourth) grade of primary school; 13-year-old students of the sixth grade; 15-year-old students of the eighth grade and 17-year-old students of the tenth grade. The total number of participants of each age was approximately 100.

Figure 5 represents the results of test completion by the students of four grades analyzed. We see that a number of students at the 3 degree (that corresponds to the highest – functional - level of mastering action pattern in the theoretical model) is increasing with the grade - 8% in the 4-th grade to 67 % in the 10-th grade. A number of students at the 1 degree (that corresponds to the lowest – reproductive - level of mastering action pattern in the theoretical model) is decreasing with the grade - 31% in the 4-th grade to 1% in the 10-th grade. Furthermore, beginning with the 8-th grade (when finishing middle school) the 3 degree dominates. This provides support for the SAM theoretical model and its validity.

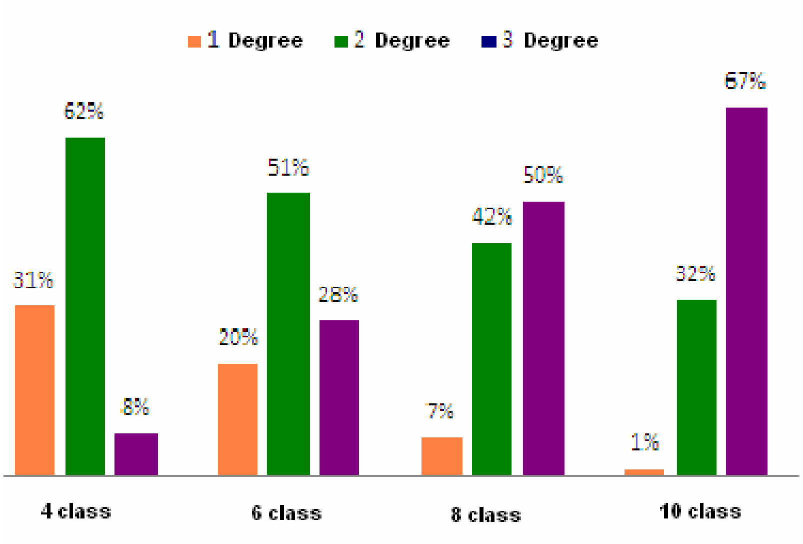

Figure 5**.** Distribution of students of different grades

Discussion

Many other important aspects of SAM research remain beyond this paper. They include DIF analysis; tests adaptation when translating into another language; comparing the results from two forms of test administration (paper and pencil vs. computer-based); test equating, etc. All these directions of research are being pursued.

Conclusion

In order to improve school efficiency, monitoring tools based on the theory of learning process should be developed. The presentation describes the assessment framework of the School Achievements Monitoring Toolkit (SAM) of classroom subject competences of primary school students that is being developed on the basis of the Vygotsky theory by the Center for International Cooperation in Education Development (CICED, Russian Federation). This toolkit can be useful for Russia and other countries.

Teachers can use SAM for objective assessing the results of their work, understanding their advantages and deficiencies, improving and developing their teaching practice. Secondly, the SAM results can be used by education authorities in order to improve school efficiency. It is important that SAM will be supplied with both methodic recommendations on interpretation and use of the test results, and instrument for the test data treatment and students' achievement estimation.

References

Vygotsky, L. (1982): *Thought and language*. Cambridge, Mass: MIT Press

Galperin, P. (1998): *Psihologiya kak ob'ectivnaya nauka*. Moscow

Davydov, V. (1996): *Teoriya razvivayushchego obuchenia*. Moscow

Piaget, G. (1969): *Izbrannye psihologicheskie trudy*. Moscow

Nezhnov, P. (2007): Oposredstvovanie i spontannost' v modeli "kul'turnogo razvitiya". In: *Vestnik Moskovskogo universiteta*. Seria 14. Psihologiya. № 1. P. 133-146

Elkonin, D. (1989): *Izbrannye psihologicheskie trudy*. Moscow

Nezhnov, P. (2009): *Toolkit for assessment of subject competences of primary school students. In Innovation in assessment to meet changing needs*. The paper presented at the 10-th Annual AEA-Europe Conference. Malta

Nezhnov, P., Kardanova, E., Elkonin, B. (2011): Otsenca resultatov shkolnogo obrazovaniya: strukturniy podhod. In: *Voprosy obrazovaniya.* №1. P. 26-43

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

Rasch models. Foundations, Recent Developments, and Applications. (1995)*:* G.Fisher, I.Molenar, Eds. N.-Y.