The development of the toolkit for assessment of subject competences of primary school students

Elena Kardanova,
Center for International Cooperation in Education Development of the Academy of National Economy under the Government of the Russian Federation
(Russian Federation)
e_kardanova@mail.ru

The presentation describes the toolkit for assessment of subject competences of primary school students that is being developed by the group of the Russian specialists under the leadership of P. Nezhnov and B. Elconin. This work was influenced by the international comparative monitoring studies (especially PISA) and represents the next logical step on the way to the new understanding of educational goals.

The purpose of the toolkit is to assess primary school students' mathematics, science and reading literacy. The assessment model is based on the Vygotsky theory and is aimed to evaluate examinees on three basic levels of formation of each type of literacy: formal, reflexive and functional.

The toolkit consists of subject tests and context questionnaires. Test items are designed for each unit of school curriculum in accordance with the above levels. Questionnaires include items aimed at the analysis of teaching conditions and characteristics of educational process.

In order to validate the toolkit, IRT modeling is applied (both unidimensional and multidimensional models). It provides an opportunity to build a bank of reliable and valid items.

Complementary scaling models based on IRT are employed for evaluation of each examinee on the above levels and building his profile. The assessment of student achievement based on three indicators gives an opportunity to assure a new interpretation of educational process and its outcomes.

Key words: Measurement, IRT Modeling, Multidimensional Rasch Analysis, Reliability, Validity

Introduction

An effective management in education requires regular and reliable information on educational institutions' performance, outcomes of educational programs and education quality at the different stages of learning. Usually such information is taken from traditional pedagogic reporting. Additionally monitoring studies are used for objective assessing students' achievements and depicting the factors that cause these results.

In order to develop an adequate modernization strategy that relies on monitoring data, it is necessary not only to receive quantitative characteristics of learning outputs but to assess quality levels of learning materials' acquisition, "strategy" of students training in various education systems and to compare efficiency and reliability of educational programs. This means that information on quality of learning outputs is necessary. And it is the most crucial thing for the initial stage of education, when the background for the further learning success is being formed.

Up to now there are several well-known international tools to assess the primary and secondary students' achievements (PIRLS, TIMSS, PISA, etc.). These studies provided us with new ways of thinking about educational goals and gave us a strong political stimulus to raise the quality of education.

The next logical step on the way of new understanding of educational goals is to develop the toolkit that provides both quantitative and qualitative characteristics of students' achievements. First of all, it is important to evaluate competence achievements on the overall subject-matter material to assure a relevant decision making aimed at adjustment and improvement of the education system. Such tool for monitoring achievements of school students (MASS), namely, academic and subject-specific competences of primary school students, is being developed in Russia under the auspices of the Center of International Cooperation in Education Development of the Academy of National Economy under the Government of the Russian Federation.

This paper aims to serve two primary objectives: (1) to present the toolkit for assessment of subject competences of primary school students, (2) to produce evidence of its validity.

Theoretical background

The MASS assessment model has been developed by the group of the Russian psychologists on the basis of the Vygotsky theory (Vygotsky, 1982), the famous Russian psychologist. According to L.Vygotsky, the growth (maturation) of a child is a specific process defined as "a cultural development". This "cultural development" has genetic and functional aspects. A concept of "functional development" relates to the school educational process, process of learning.

In accordance with this theory learning is a necessary condition of a child's psychological development, it is a process of transmitting sign structures (notions, schemes, principles, algorithms, samples of behavior) from an adult to a child. But transmitting the sign structure to the child is only the beginning of the educational process. Then the process of active reconstruction and assimilation of a thinking pattern takes place. The sign structures, acquired by a child, make a zone of proximal development (ZPD) or in other words they launch a process of functional development. The main point of it is reconstruction and assimilation of a cultural means of action, i.e. respective system of abilities, which is a content of a certain subject competence. The child functional development is the internal spontaneous constituent of educational process and the results of this process (intermediate and final) are extremely important for a teacher.

According to the MASS assessment model (Nezhnov, 2009; Nezhnov, Froumin, Khasan, Elconin, 2009), the process of cultural development has three key measuring points linked with three possible options of retaining means of action. In the organized educational process these three points mark three levels of mastering means of action:

- 1 level – reproductive – retention of external characteristics of a cultural sample of action (algorithms, rules, forms of action etc.);
- 2 level – reflexive – retention of essential fundamental of generalized means of action;
- 3 level – constructive or functional – retention of possibilities of means of action.

These three levels compose a basic taxonomy of educational targets, which has a psychological background, i.e. it indicates the cultural-psychological structures which are crucial for competence developing from immature stage to mature one. In this taxonomy a level designates a dominant type of a cultural sample retention by a child with possibilities of thinking and acting in the result.

The MASS toolkit is an attempt to develop a measurement instrument on the basis of the described taxonomy. Thus the purpose of the MASS toolkit is to assess the primary school students on three basic levels of formation of each type of literacy: formal, reflexive and functional.

Method

MASS description

The MASS toolkit consists of a set of tools for monitoring on the national/regional level the academic subject-matter competences of primary school students in such areas as: mathematics, science, native language, reading literature texts and reading informative texts (Nezhnov, Froumin, Khasan, Elconin, 2009). The tools include subject tests in each area, questionnaires for collecting context information and recommendations on the test results interpretation and usage.

Each subject test consists of three subscales corresponding to different literacy levels. Each test item belongs to the only one subscale. Test items are designed for each unit of school curriculum in accordance with the literacy levels. Each test represents a set of three items related to the same content at different levels.

Items of MASS tests (on mathematics, science and language) have different formats: multiple choice (with one or two correct options from four or five proposed), opened with a short answer and a free response. The reading tests have more complicated structure. They consist of testlets (bundles of items that share a common stimulus, namely a reading comprehension passage). The items of these tests have similar formats as described above. Most of the items are scored dichotomously, but there are items scored polytomously.

The questionnaires' aim is to collect context information on the factors that influence the students' achievements. Four types of questionnaires are being developed: (1) *a questionnaire for students* (includes questions about student's family; attitude to school, teaching and learning, etc.); (2) *a questionnaire for school administration* (includes questions about school resources, conditions of learning and a system of teaching, etc.); (3) *a questionnaire for parents* (includes questions related to parents'

input in school life, assessment of teachers' work and work of school administration, parents' time devoted to their children home assignment, etc); (4) *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.).

Set of guidelines and methodic recommendations are suggested for different participants of educational process and test results users. This set contains guidelines for education managers of different levels (municipal, regional, federal); methodic recommendations for teacher training; methodic recommendations for school administration; methodic recommendations for a teacher.

It is anticipated that all items in the MASS tests will be scored by automated scoring system. So a computer-aided system for automatically processing the monitoring results is being developed. It includes a software that provides MASS data treatment and students' measurement. It also scales with an opportunity of automatic generalization of different reports and presents the results using different types of analysis. Multidimensional IRT models are used for students' assessment. Each examinee will get three test scores in accordance with three basic levels mentioned above.

On the basis of the test results a student profile can be composed, as well as a class (or any sample of students) profile for each subject. Such presentation of the test data gives a structural vision of the competence being formed and opens a way to the profound interpretation of learning outputs. Additionally it is expected that the integral index of common educational literacy of the primary school students will be developed.

Measurement model

All subscales of the MASS 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. Firstly, we can ignore multidimensionality of the test and apply a unidimensional model. Secondly, we can recognize multidimensionality and apply a unidimensional model to each dimension consecutively. And thirdly, we can apply multidimensional models.

In the unidimensional approach the student's raw test score is treated as the sufficient statistics for estimation of his ability. So the unidimensional approach allows us to get a composite score, i.e. a single estimate of student achievement and its associated standard error. Furthermore, the reliability of these students' estimates is the highest in comparing with other approaches. But a disadvantage of the unidimensional approach is the loss of information about students' achievement at different levels.

The consecutive approach implies that raw scores on each dimension are treated as different sufficient statistics and modeled independently as unidimensional constructs. The advantage of this approach is that it produces ability estimates and their standard errors for each dimension. But if a number of items per each dimension is small, the standard errors of students' estimates are essentially large, especially in comparison with the unidimensional approach. This can be explained by the fact that the consecutive approach ignores the possible interrelation of different variables.

Under the multidimensional approach the raw scores on each dimension are treated as distinct information about each student, yet by incorporating the correlation between the latent variables. Due to this the loss in reliability is less than in the unidimensional approach.

At the stage of MASS validization all three approaches were applied. Members of the Rasch family of item response models were employed. The Unidimensional Rasch Models (Wright, Mok, 2000), namely simple Rasch model for dichotomous items, and the Partial Credit Model for polytomous items were employed. The Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM) (Briggs, Wilson, 2003) was applied for modeling tests on mathematics, science and language. The Rasch Testlet Model (Wang, Wilson, 2005) was applied for reading tests.

MRCMLM is a generalized the Rasch type item response model, within the framework of which many existing IRT models can be considered as its special cases, and multidimensional versions of these models can be constructed. For the purposes of our study MRCMLM was adjusted to a three-dimensional Rasch model for dichotomous items, and to a three-dimensional Partial Credit Model for polytomous items. In this three-dimensional model each item only loads on one dimension, which is referred to as between-item multidimensional model (Adams, Wilson, Wang, 1997).

Analysis

For the purpose of MASS validization it is necessary to analyze all three approaches and choose the best one the for test data modeling.

Unidimensional and multidimensional analyses were conducted with *ConQuest* (http://assess.com). The items parameters and population means and variances are estimated by the marginal maximum likelihood technique. There was the constraint for each distribution of abilities to have a mean of 0 and standard deviation of 1. Standard errors and fit statistics are produced for each parameter estimated.

Three other indices are relevant for our study. Firstly, the reliability index was computed. Secondly, goodness of fit of the model was evaluated using the deviance index. It is known that, for two models, one a special case of the other, the difference in deviances has approximately a chi-squared distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models. Thirdly, the correlation between various dimensions in multidimensional and consecutive approaches and between different approaches was analyzed.

Results

The results of the math test analysis are presented at this paper. The math test is being developed by the group of specialists which includes Gorbov S. (the head), Efremova Y.,Ostroverh O., Sviridova O., Zaslavsky V.

It should be noted that validization study of the MASS tests included different kinds of analysis, but only study of the model selection is presented here.

Participants

Data for this study were collected during MASS pilot testing in Krasnoyarsk region of the Russian Federation. The sampling procedure includes 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 418.

Instrument

Five content areas were included in the test of mathematical literacy. They are: numbers and calculations; value measurement; mathematical regularities; dependence between values; geometry elements. For each content area test items in accordance with the three described levels of literacy were developed. Three items related to the same content at different levels form a set. The test contains 15 sets of items and the total number of items is equal to 45.

The test is assumed to be a multidimensional. Items of each literacy level form a subscale. So there are three subscales, 15 items for each one. Each item belongs to only one subscale (dimension). All items were scored dichotomously.

Unidimensional approach

The results of scaling the MASS mathematical data using a unidimensional model are shown in the Table 1. The upper portion of this table presents a summary of the model statistics, and the lower part presents a summary of the item parameter estimates.

The analysis of the table reveals that the reliability is quite high. The test is appeared to be difficult for students: item difficulty average is much more than 0. Item difficulties are widely distributed around the mean point. Fit index averages are close to their expected value 1, but unweighted indices are much more widely dispersed that the weighted indices. It is known, that the unweighted fit statistics are more sensitive to unexpectedly large residuals. This happens if a person with low ability answers a difficult item correctly.

Consecutive approach

The results of the test data scaling using the consecutive approach are shown in the Table 2. The table presents only a summary of the model statistics for each dimension (subscale).

The analysis of the table reveals the substantial reduction in reliability for separate subscales in comparison with the unidimensional approach: .08 for the subscale 1, .19 for the subscale 2 and .57 for

the subscale 3. The standard errors of students' measurement by each subscale are extremely high. It means that separate subscales cannot be considered as independent measurements. So we can conclude that the consecutive approach is unacceptable for the MASS math data. This result is expected due to small number of items in each subscale and in light of the fact that all items were scored dichotomously.

Table 1
Summary of Unidimensional Model Scaling

Model Summary			
Number of parameters	Deviance	Reliabilty	Standard error mean of students' estimation
46	14919.96	.86	.45

Item Summary				
	Item difficulty (logits)	Standard error	Unweighted fit statistics	Weighted fit statistics
Mean	1.53	.18	1.01	1.00
(SD)	1.9	.1	.44	.1

Table 2
Summary of Consecutive Model Scaling

Model Summary		
	Reliabilty	Standard error mean of students' estimation
Dimension 1	.78	.65
Dimension 2	.67	.78
Dimension 3	.29	.92

Multidimensional approach

The results of the test data scaling using the multidimensional approach are shown in the Table 3. The upper portion of this table presents a summary of the model statistics for each dimension (subscale), and the lower part presents a summary of the item parameter estimates (only overall).

This proves that the multidimensional approach provides an improvement in reliability compared to the consecutive approach. Under the multidimensional approach the reliability for each dimension comes closer to the unidimensional reliability estimate.

Table 3
Summary of Multidimensional Model Scaling

Model Summary				
	Number of parameters	Deviance	Reliabilty	Standard error mean of students' estimation
	51	14891.1		
Dimension 1			.83	.48
Dimension 2			.80	.53
Dimension 3			.61	.54

Item Summary

	Item difficulty (logits)	Standard error	Unweighted fit statistics	Weighted fit statistics
Mean	1.63	.18	.96	1.00
(SD)	1.87	.1	.3	.1

Model comparison

The multidimensional approach is hierarchically related to the unidimensional approach. So the model fit can be compared to the change in the deviance value. As the Tables 1 and 3 indicate, the difference in deviance between the two models is 28.86. This difference is approximately distributed as a chi-square with 5 degrees (the difference in the number of parameters estimated) of freedom. This suggests that the multidimensional model fits the data much better than the unidimensional model. Additionally the comparison of these two models was implemented by means of Akaike's Information Criterion (AIC), which is a transformation of the Deviance index. For the unidimensional model AIC index is equal to 15011.96, for the multidimensional model it is 14993.1. Thus the multidimensional model provides the best explanation of the data.

In order to illustrate that the multidimensional approach is not always better than the unidimensional approach, we divided the math test into different subscales by another way, for example, using content areas. The five content areas and their notations were: numbers and calculations (C, 6 items); value measurement (M, 18 items); mathematical regularities (R, 6 items); dependence between values (D, 12 items); geometry elements (G, 3 items). The Table 4 presents the results of models' comparison between the unidimensional model and the multidimensional models. The first multidimensional model is a five-dimensional model where each content area is represented as a different dimension (M_D_C_R_G). The second model is a three-dimensional model where items of three content areas (numbers and calculations, mathematical regularities and geometry elements) were combined because of small number of items in each area (M_D_CRG). The last raw of the table presents the results for the multidimensional model above examined.

The analysis of the table reveals that all multidimensional models fit data better than the unidimensional model, but only the last model does it significantly better. This provides statistical support for the use of the three-dimensional model where different dimensions are based on literacy levels.

Table 4
Comparison of Multidimensional Models to the Unidimensional Model

Model	Number of parameters	Deviance	Ch-square	df
Unidimensional	46	14919.96		
Multidimensional M_D_C_R_G	60	14893.32	26.64	14
Multidimensional M_D_CRG	51	14918.7	1.26	5
Multidimensional (literacy levels)	51	14891.1	28.86	5

Correlation between dimensions

The Table 5 presents the correlation between variables under the multidimensional model. Note that the correlation produced in the ConQuest analysis is not the raw correlation between the students' ability estimates. These correlations are corrected for error, so they are relatively free of measurement noise. We see that the correlations among the three dimensions are reasonably high, indicating that they all measure student "math literacy". The Figure 1 displays the relationship between dimension 2 and dimension 3, for example.

Table 5
Correlations between Dimensions

	Dimension 1	Dimension 2	Dimension 3
Dimension 1	1		
Dimension 2	.88	1	
Dimension 3	.74	.79	1

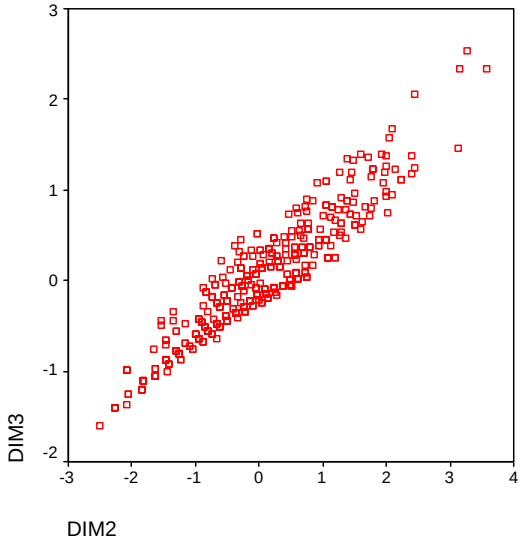

Figure 1. Relation between Student Ability Estimates on Dimension 2 and Dimension 3

Additionally the whole test (45 items) was investigated on dimensionality using principal component analysis of standardized residuals (Smith, 2002). The results indicate that the test can be considered as an essential unidimensional one. This proves that all 45 items measure the single variable – "math literacy".

Discussion

Our analysis suggests that, although the unidimensional model adequately accounts for the MASS test data, slightly more complex multidimensional models provide better explanation. The best model seems to be the three-dimensional model based on literacy levels. This provides statistical support for the theoretical hypothesis that the MASS math test is able to assess primary school students on three basic levels of each type of literacy formation: formal, reflexive and functional. The assessment of student's achievement based on three indicators opens an opportunity of new interpretation of education process and results.

Furthermore the single estimate of students' math literacy can be received using the unidimensional approach.

Many other important aspects of MASS tests investigation remained beyond this paper. They include analysis of test items and subscales; tests forms equating; scaling students' estimates on different dimensions; construction of student's profile, etc. All these results are available from the author on request.

Conclusion

In order to improve school efficiency, monitoring tools based on the theory of learning process should be developed. A monitoring toolkit of classroom subject competences (MASS) of primary school

students on the basis of the Vygotsky theory is being developed in Russia. This tool can be useful for Russia and other countries.

There are two levels of the MASS results use: classroom assessment and large scale assessment. Teachers can use MASS for objective assessing the results of their work, understanding their advantages and deficiencies, improving and developing their teaching practice. Secondly, the MASS results can be used by education authorities in order to improve school efficiency. It is important that MASS 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 measuring.

Acknowledgment

The author would like to thank Ovchinnikov V. for his help in conducting this study.

References

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

Briggs, D.C., Wilson, M. (2003). An Introduction to Multidinemsional Measurement using Rasch Models. *Journal of Applied measurement, 4*(1), 87-100.

Nezhnov, P.G., Froumin, I.D., Hasan, B.I., Elconin, B.D. (Eds.), (2009). *Diagnostics of academic successfulness in primary school.* Moscow: Opened Institute "Developmental Education".

Nezhnov, P. (2009). Toolkit for assessment of subject competences of primary school students. In *Innovation in assessment to meet changing needs.* The 10-th Annual AEA-Europe Conference. Malta, 22

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

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

Wang, W.-C., Wilson, M.(2005). The Rasch Testlet Model. *Applied Psychological Measurement, 29* (2), 126–149.

Wright, B.D., Mok, M. (2000). Rasch Model Overview. *Journal of Applied Measurement. 1*(1), 83-106.