Croatian Journal of Education
Vol.20; No.3/2018, pages: 789-824
Original research paper
Paper submitted: 6th May 2017
Paper accepted: 19th April 2018
https://doi.org/10.15516/cje.v20i3.2791

Development and Validation of New Objective School Achievement Tests in the STEM Field for Primary School Students

Dubravka Glasnović Gracin[1], Toni Babarović[2], Ivan Dević[2] and Josip Burušić[2]
[1]*University of Zagreb, Faculty of Teacher Education*
[2]*Ivo Pilar Institute of Social Sciences in Zagreb*

Abstract

Measuring student achievement is one of the key issues when researching quality and efficiency of an educational system. Several existing studies tried to express student achievement in mathematics or science through separate tests within the particular domain. The aim of this paper is to consider the possibility of measuring knowledge in the STEM area through integrated tests and to present the characteristics of the new tests developed particularly for this purpose.

In this paper, a complete psychometric analysis of the newly developed tests is provided. The paper outlines the steps undertaken in the process of determining content requirements within each knowledge test, preliminary validations of initial test versions, as well as the results obtained in the main study. The main study encompassed 586 grade 4 students, 580 grade 5 students and 632 grade 6 students. In every test, the unidimensional structure was obtained using confirmatory and exploratory factor analyses. Acceptable reliability was obtained for all three tests ($\alpha_{4th} = .78$; $\alpha_{5th} = .70$; $\alpha_{6th} = .79$). The correlations between total test scores and achievement in STEM school subjects were moderate to high. Therefore, all three newly developed tests represent a one-dimensional, reliable, discriminative and valid measure of integrated students' knowledge in the STEM area.

Key words: *measuring knowledge; STEM achievement; test development.*

Introduction

The acronym STEM (*science, technology, engineering, and mathematics*), coined by the American National Science Foundation (NSF), is today widely accepted and used for a

particular area of knowledge and practice (Dugger, 2010). The term *Science* (S) refers to "the study of the natural world, including the laws of nature associated with physics, chemistry and biology, and the treatment or application of facts, principles, concepts, or conventions associated with these disciplines" (Honey, Pearson, & Schweingruber, 2014, p. 14). *Technology* (T) comprises an entire system of people and knowledge, organizations, processes and devices that are involved in designing and operating with technological artefacts created by humans in order to satisfy their needs. *Engineering* (E) refers to the knowledge about the creation of artefacts and about processes of designing and solving problems. One branch of engineering refers to the laws of nature, while others encompass time, money, environmental regulations, available materials, etc. The letter M in the STEM acronym means *Mathematics* – the study about patterns and relationships among quantities, numbers and space objects (Honey et al., 2014).

In the past ten years, the STEM phenomenon was screened from different research perspectives. Some studies were focused on the deficit of STEM experts, consequences of this problem for national economies and its effects on the contemporary and future labour market (European Commission, 2004; Osborne & Dillon, 2008; UNESCO, 2010). Other studies covered issues related to social and gender differences in STEM achievement and self-efficacy beliefs (Anderson & Kim, 2006; Ceci, Williams, & Barnett, 2009; Eccles, 2007, 2009), and the importance of teachers and parents as role models (Adelman, 1999a, 1999b, 2006; Aschbacher, Li, & Roth, 2010; Hagedorn & DuBray, 2010). Other important issues pertain to the effectiveness of certain programs and interventions aimed at promoting STEM careers among students (ASPIRES, 2013; Sjøberg & Schreiner, 2010) and finding theoretical explanations why students continue or leave the STEM field of education (Adelman 1999a, 1999b, 2006; Anderson & Kim, 2006; Hagedorn & DuBray, 2010).

In the last few years, one of the main challenges refers to the appropriate approach to STEM education (European Commission, 2012; Honey et al., 2014). This issue also refers to the question of defining and measuring students' STEM achievement. Xie, Fang, and Shauman (2015) discuss the term *STEM education* with emphasis on the logical and conceptual relations among different STEM areas. Such an approach provides the concept of STEM education as a whole and not as a set of separate disciplines. Following this idea, Honey et al. (2014) developed a research framework for integrated STEM education in order to investigate its positive outcomes. Similarly, Kennedy and Odel (2014) emphasize STEM education that should encompass integration of STEM disciplines, promoting scientific inquiry, project-based learning, engineering design process and mathematical rigor. The authors also involve collaborative approaches to learning, meetings of STEM educators and their students with the broader STEM community, promoting formal and informal learning experiences and integration of technology and engineering into the science and mathematics curriculum.

However, the idea of integrating STEM school subjects may be of a complex and sensitive nature and opens many issues in contemporary educational discussions (Kelley & Knowles, 2016). According to the meta-analysis conducted by Hurley (2001) about the effects of integration on student learning in mathematics and science, the author found fewer positive benefits of integration for mathematics' outcomes compared to the outcomes in science. UNESCO (2015) noted other problems and challenges related to integrated STEM education, such as insufficiently prepared teachers, strong traditional disciplinary boundaries and low status of integrated learning areas compared to single subjects. Besides, STEM educational content becomes rapidly outdated and the level of detail within each discipline could become uncontrollable. Therefore, defining the basics of STEM education is a big challenge for educational experts.

Despite these discussions, knowledge in STEM disciplines is recognized as an important factor in economic development of modern societies, and students' achievement in the STEM area serves as the key predictor of a life where mathematics, science and technology play an important role (Organisation for Economic Co-operation and Development [OECD], 2003). These issues are recognized within large-scale international studies for student assessment such as PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study). The predominant approach of these studies is that all students should receive an education which provides for the acquisition of basic competences for life. These basic competences surely include knowledge in the areas of science and mathematics (Mullis & Martin, 2013; OECD, 2003), and the theoretical framework of international large-scale studies influenced the changes of many national science or mathematics curricula (e.g. Kultusministerkonferenz [KMK], 2003). Also, they influenced the measurement and operationalization of student achievement within national educational contexts.

Analyses show that the large-scale studies such as TIMSS and PISA also influenced the structure and requirements of many national tests of student achievement (Volante, 2016). For example, in England in 2009 the educational authorities recommended that national tests, where possible, should reflect the test items given in a large-scale study in which this country participates (Thomas, Gana, & Muñoz-Chereau, 2016).

The above-mentioned international projects point towards the need for different conceptualization and operationalization of students' achievement. Not just traditionally, as knowledge of separate school subjects, but also as knowledge within integrated areas, such as STEM. The research literature provides some attempts and debates on integrating STEM school subjects. Dugger (2010) considered various possibilities of integrating STEM school subjects. One approach is the integration of one STEM discipline into the other three disciplines. For example, the author suggested the integration of engineering into science, mathematics and technology

courses. The other possibility is to integrate all STEM disciplines and to teach them as one integrated school subject. Another solution is proposed by Kennedy and Odel (2014), who suggest the need of incorporating technology and engineering into mathematics and science. Such integration would promote scientific inquiry and engineering design process. The proposed approach also requires a certain level of curricular and pedagogical coherence through all STEM disciplines (Xie et al., 2015). Similarly, Kelley and Knowles (2016) presented a conceptual framework for integrated STEM education. It positions STEM education as the interplay of mathematical thinking, engineering design, technological literacy, and science inquiry.

Nevertheless, the conceptualization and operationalization of student achievement in STEM areas were mostly focused on the achievement in separated subjects, such as Mathematics and Science. Less attention was paid to measuring achievement as integrated content and knowledge acquired in more than one discipline (Honey et al., 2014). The existing assessment frameworks emphasize the need of connecting the basic ideas and concepts from different STEM disciplines (National Research Council, 2014), but the literature review shows that such integrated knowledge tests barely exist.

This finding was the starting point for this study. The aim of the study was to consider the possibility of conceptualization, development and assessment of integrated STEM achievement tests for primary school students in Croatia. For this purpose, three knowledge tests in the STEM area for 4th, 5th and 6th grade students were developed and assessed on the pupils' samples. This was followed by a comprehensive psychometric validation of the newly developed tests. The results will provide new knowledge about designing and measuring STEM school achievement. Such studies, which encompass designing, assessment and validation of integrated STEM tests, have not yet been conducted in Croatia, and could have significant implications for measurement and conceptualization of STEM achievement in Croatian compulsory education.

Methods

Participants

The sample in the pilot study consisted of 118 students, including 4th grade (N = 41), 5th grade (N = 46) and 6th grade students (N = 31) from one primary school in Zagreb, which was similar in demographic and other student characteristics to the primary schools included in the main study. Two classes were randomly selected within each generation.

The participants in the main study were 1798 primary school students from 16 schools in Zagreb and its surroundings, attending grades 4, 5 and 6 (age 10 – 12). Within each school, two classes of students within one generation were randomly sampled and joined the survey. In the total sample, students were equally represented by gender (49.8% of girls) and by grade ($N_{4\text{th grade}} = 586$, $N_{5\text{th grade}} = 580$, $N_{6\text{th grade}} = 632$).

Measures and Instruments

Integrated Tests of STEM School Knowledge

For the purpose of measuring STEM school knowledge in grades 4 to 6, a separate test for each grade was developed. The development of the tests included several steps.

The first step in test development was the analysis of current curricular documents related to STEM subjects in the fourth, fifth and sixth grade (Ministarstvo znanosti, obrazovanja i športa (MZOS), 2006, 2010). In the fourth grade, it included school subjects Mathematics and Science, and in the fifth and sixth grades Mathematics, Biology, Geography and Technical culture. Thus, catalogues of STEM knowledge for grades 4, 5 and 6 (Table 1, Table 2 and Table 3) were obtained. Since the students were tested at the beginning of the second semester, the topics that should have been acquired by the end of the first semester were used in the development of test items. The catalogues showed that the students' STEM knowledge and competences significantly increase in the period from grade 4 to grade 6. Therefore, each test mainly required knowledge of the subject matter learned in school in the time interval of the past 12 to 18 months.

Table 1

The catalogue of STEM school content (up to the end of the first semester of grade 4)

Subject	Content related to STEM	Item number:
MATH	Arithmetic (grades 1-4): Natural numbers up to a million and related arithmetic operations, Presenting data	1, 2, 5, 9, 10, 12, 16, 20
	Geometry (grades 1-4): 2D and 3D spatial ability (point, segment, line, ray, basic plane and solid geometric shapes, parallel and intersecting lines, perpendicular lines, circle, angle types)	4, 7, 13, 18
	Measurement (grades 1-4): Measuring and estimating length, weight, time and volume of a liquid	3, 5, 6, 8, 14, 15, 19
SCI	(grade 2) Hour, day, month, year, timeline	3, 6
	(grade 3) Cardinal directions, standpoint, horizon, town plan, map; Experiment, three forms of water	7, 11
	(grade 4) Nature, conditions for life: Sun, water, air, soil; Experiments	2, 9, 17, 19, 20

In the second step, a decision on the general test structure for a particular grade was made. It was based on the analysis of curricular documents and the planned test structure (Institut für Didaktik der Mathematik (IDM), 2007; Sullivan, Clarke, & Clarke, 2013). The test structure is shown in Table 4. The usual activities in STEM tasks are calculation, interpretation of a given formula, image or a graph, representation, explanation, and reasoning (according to IDM, 2007). The dimension of complexity refers to cognitive levels of reproduction, making connections, and reflection (Smith & Stein, 1998). The tests include all three cognitive levels, matched to the age of the students in each part of the research. Reproduction (K1) refers to the direct application

of basic terms, rules, procedures, or representations. Making connections (K2) is needed when the task is of a more complex nature and requires the connection of more concepts, rules, procedures or representations, or when it is necessary to link different actions to the whole to solve the problem. A distinction is made between simple and more complex connecting. Reflection (K3) refers to reflection on relationships that are not directly visible from the given facts. Considering the limits of 45 minutes for solving the test, all the items required short and objective answers, either in the form of multiple choice questions or in the form of a short open-ended question in which the respondent would write in the correct answer.

Table 2

The catalogue of STEM school content (up to the end of the first semester of grade 5)

Subject	Content related to STEM	Item number:
MATH	Arithmetic (grades 4 and 5): Natural numbers and related arithmetic operations, Estimating quantities; Presenting data	1, 9, 12, 17, 19, 20
	Geometry (grades 4 and 5): 2D spatial ability; Triangle, rectangle, square, parallelogram, rhombus (types of triangles, square as a type of rectangle, square, rectangle and rhombus as a type of parallelogram); 3D spatial ability; Cube and cuboid	2, 6, 11
	Measurement (grades 4 and 5): Circumference of a triangle and types of triangles, circumference of a rectangle, square, parallelogram and rhombus; Area of a square and rectangle; Volume of a cube; Estimations	3, 4, 5, 18
SCI4	(grade 4) Plants, animals, grass, forest, sea	16
BIO5	(grade 5) Natural sciences (microscope, magnifier, sample, characteristics of living organisms, microworld); From cell to organism (cell, parts of a cell, cell division)	17
	Health	13
	Animals (basic structures in animals, invertebrates and vertebrates, how animals move, carnivores, herbivores and omnivores, animal reproduction)	14
GEO	(grade 5) Shape and size of the Earth, gravitation, globe, Equator, North and South pole, planet Earth in space	7, 9
	Geographic grid, geographic map, map scale and map types, orientation on the map	11, 12
	Motions of the Earth, the four seasons	8, 10
	Orientation, cardinal directions, compass	11
TECH	(grade 5) Drawings nets of 3-D objects, dimension lines drawing	2, 5, 11
	Rectangular projection	6, 15

Table 3

The catalogue of STEM school content (up to the end of the first semester of grade 6)

Subject	Content related to STEM	Item number:
MATH	Arithmetic (grades 5 and 6): Decimal numbers and related arithmetic operations; Fractions and operations; Estimation of rational numbers; Presenting data	1, 2, 3, 4, 8, 17, 19
	Geometry (grades 5 and 6): 2D and 3D spatial ability; Sets of points; Adjacent and vertical angles; Perpendicular bisector, Line symmetry	6
	Measurement (grades 5 and 6): Measuring angles; perimeter and area	5, 11, 14, 20
BIO	(grade 5) Plants (organs of flowering plant, what plants need to grow); Root, stem, leaf, flower and fruit – structure and role; Nutrition	12, 13
	(grade 6) Living beings, habitat, life conditions	9
	Forest plants and fungi; Forest animals (adaptations of animals, ecological relationships)	12, 13
	Energy and forms of energy; The Sun's energy, energy transformations, fossil fuels, renewable and non-renewable energy resources, circulation of substances in nature	15, 16
GEO	(grade 5) Earth relief and structure, forces inside Earth	-
	Bodies of water	14
	Weather and climate	8, 10, 11, 18
	Natural resources and preservation of environment (resources, ores, renewable and non-renewable energy resources)	15, 16
	(grade 6) Population (population change, population density, people differ)	20
TECH	(grade 5) Work and energy (force, work, energy; basic forms of energy)	15, 16
	(grade 6) Metric scales; Measuring and contouring in the processing of wood, plastic and rubber materials	1, 7

Table 4
Type and structure of test items

Dimension	Question	Details	Proportion in the test items (%)		
			Grade 4	Grade 5	Grade 6
Activities	What should be done in a particular task?	Presentation (A1)	15 %	5 %	5 %
		Calculation (A2)	60 %	45 %	35 %
		Interpretation (A3)	45 %	65 %	60 %
		Argumentation (A4)	5 %	0 %	0 %
		Factual knowledge (F)	0 %	5 %	15 %
Complexity level	What is the complexity level of a particular task?	Reproduction (K1)	35 %	50 %	50 %
		Making connections (K2)	60 %	40 %	30 %
		Reflection (K3)	5 %	10 %	20 %
Context	What is the context in the task?	No context (C1)	20 %	30 %	30 %
		Realistic (C2)	70 %	40 %	40 %
		Authentic (C3)	10 %	30 %	30 %
Answer	What answer form does the task require?	Multiple choice (MC)	65 %	65 %	75 %
		Short answer (SA)	35 %	35 %	25 %

During the third step, the principle of the integration of content that would be represented in each test was developed. Integration was implemented on several levels: (1) content integration among different school subjects; (2) integration of different concepts within a single school subject; (3) content-context connections (connecting content of one school subject with the context of another subject); (4) mixed tasks within the test as a whole (a test consists of several separate items from different STEM subjects). The integration of content was obtained by textual analysis of similarity and correlation between STEM subjects curricula. The curriculum catalogues served as a basis for linking content in STEM tasks. Also, the framework for integrated STEM education (Kelley & Knowles, 2016) has been consulted with requirements such as technical-logical literacy, mathematical thinking, and engineering design. The authors have also tried to connect the test items through STEM context, not only through STEM content.

In the fourth step, items for each test were produced, considering the varying requirements within the structure shown in Table 4. Although the intention was to put emphasis on integration within a particular task, that was not obtained in every item because it was also necessary to examine some basic subject knowledge in these early grades of STEM education. This step also included the logical aspect, content, language, style, and formatting aspect of designing test items.

In the fifth step, appropriateness and quality of the created items were considered. At this stage, requirements and content of each item were discussed with teachers from

different STEM areas. Since in the previous step many more items were designed than needed, at this stage, the appropriate items were selected and the need for modification of content was discussed. Here is an example of an integrated task for subjects Mathematics and Science in the fourth grade: *"A pot of water is put on a stove top. If you heat it for the next 25° C, the water will reach the boiling point. What is the temperature of the water in the pot?"* In order to solve this computation task (Mathematics), a student should know the concepts of boiling point and its temperature (Science).

In the sixth step, an initial version of the tests was made.

In the seventh step, the final test was checked, instructions for students were written, images were graphically completed and the test was prepared for the pilot study. After the pilot study, the observed problems were discussed and the tests were prepared for the main study.

The final versions of each of the three tests contain 20 items that are considered to be a reliable, discriminate and valid measure of student knowledge in the STEM area. Dominating contents of each task are presented in Tables 6, 7 and 8, while the integration dimension can be seen in the right columns of Tables 1, 2 and 3. In each item, there is only one correct answer and these answers are encoded with 1 or 0. The total score on the test is the sum of the correct answers ranging from 0 to 20 whereby a higher score indicates better knowledge in the STEM area.

School Achievement in the STEM Area Represented by School Grades

The study also included student school grades in school subjects related to STEM (subject Mathematics in the 4th grade, and Biology, Mathematics, Technical culture and Geography in grades 5 and 6). Grades in school subject Social science (in Croatian: Priroda i društvo) were not included because of saturation with non-STEM content (social part of the school subject). School subject grades were collected directly from the school administration based on the existing school records on student achievement in the school year 2014/15 and school year 2015/16. Within a particular school grade, the average STEM achievement was calculated as the arithmetic mean of the grades in STEM-related school subjects.

Procedure

Data were collected within regular school classes. Prior to testing, parents were thoroughly informed about the aims and features of the study and a written consent for participation in the study was obtained from the parents/guardians. The consent return rate was 89.8%.

In the pilot study, initial versions of the tests were applied to students in the 4th, 5th and 6th grade. Students were tested within regular school classes, and testing lasted for 45 minutes. Initial tests were revised based on the results of the psychometric analysis. Already in the initial version of the tests, appropriate reliability ($\alpha > .70$) and a one-

factor structure were obtained for all three tests. The distribution of the total score was nearly normal, with an appropriate variability of the overall test result. It has been shown that only a small number of items do not improve the reliability of the test with low loadings on the general factor ($r < .30$), low variability, or being too easy or too difficult for the target population ($p > .80$, or $p < .20$). In some multiple-choice items, poor distractors, with less than 5 % of retained answers, were identified. Based on the pilot study analysis, only items with appropriate variability and items that improve the overall reliability and validity were included in the final tests. Some items, which were out of the acceptable difficulty range, were revised to be easier or more difficult. The strategy was to adapt the content of the items or to choose better distractors. Items not related to the total test score (not improving validity or reliability) were revised and modified or new items with the similar testing objective were designed.

In the main study students from grades 4, 5 and 6 were tested. They were assessed during regular school classes and they had 45 minutes to solve the test. Students were informed about the study and the test in advance, but they were not pre-trained for the test since the focus was on basic STEM knowledge and on connections among STEM disciplines.

Data Analysis and Statistics

Statistical analyses were made via SPSS 23 and AMOS 7.0. Scores on the developed STEM knowledge test in grades 4, 5 and 6 were calculated and normality of distribution was checked. Descriptive statistics were used for the validation of newly developed tests and examination of psychometric properties. Internal consistency was checked using Cronbach alpha coefficients. The Principal component analysis and Confirmatory factor analysis were used to examine structural validity of the tests. The external validity of the tests was examined by analyzing Pearson correlations of the total test scores and earlier school achievement in the STEM area.

Results

The collected data were analysed separately for the fourth, fifth and sixth grade. The total score was calculated by a total number of correctly answered items, with the theoretical range from 0 to 20 points. The average test score in the 4th grade was $M = 11.81$ (SD = 4.32), in the 5th grade $M = 9.34$ (SD = 3.66) and in the 6th grade $M = 10.93$ (SD = 3.97). The average scores demonstrate appropriate difficulty level of all three tests. Average item difficulty is appropriate for all three tests, $p = .42$ to $p = .61$ (Table 5).

Table 5
Test results of the 4th, 5th and 6th grade students

	Grade 4	Grade 5	Grade 6
N	586	580	632
M	11.81	9.34	10.93
Mdn	12.0	9.0	11.0
SD	4.32	3.66	3.97
p	.61	.42	.58
Skewness	-0.24	0.08	0.06
Std. error of skewness	0.10	0.10	0.10
Kurtosis	-0.69	-0.63	-0.62
Std. error of kurtosis	0.20	0.20	0.19

The distribution of test results in the 4th grade is nearly normal although slightly negatively skewed (Figure 1). The distributions in the 5th and the 6th grade are nearly normal. All three distributions are slightly platykurtic, with higher tails and fewer results grouped in the middle of the distribution compared to a normal distribution. The distributions indicate good discriminating power of all three tests. Discriminating power is a property of a psychometric test to be able to distinguish between individuals with higher and lower levels of knowledge through the whole range of results (Nunnally & Bernstein, 1994).

The lowest and the highest scores in the 4th grade test were $X_{min} = 0$ and $X_{max} = 20$, in the 5th grade test $X_{min} = 1$ and $X_{max} = 19$ and in the 6th grade test $X_{min} = 1$ and $X_{max} = 20$. The ranges mostly cover the theoretical range and also indicate good discriminating power of the tests. The Ferguson δ (delta) coefficient (Krković, 1978) also indicates good test discriminating power. Ferguson's Delta is the ratio of the observed between-person differences to the maximum number possible, and ranges from 0 - no differences are observed, to 1 - all possible between-person discriminations are made. Ferguson's Delta for all three tests show that they are very discriminative ($\delta_{4th} = .98$, $\delta_{5th} = .97$, $\delta_{6th} = .98$).

Item difficulty in the 4th grade test was adequate and varies from $p = .41$ for the most difficult item to $p = .91$ for the easiest (Table 6). Three items were too easy for the target population with low variability and $p > .80$. For example, in the easiest item in this test (the 2nd item in the test) the respondents had to select an appropriate calculating operation and to display the presented text with mathematical symbols. This task belongs to the complexity level 1 – reproductions and direct application of definitions and rules, and the percentage of students that correctly answered this item was 91%. With respect to content, 90% of the fourth-grade items refer to Mathematics and 45% of them to Science (Table 6). In about a half of items content or context integration was accomplished.

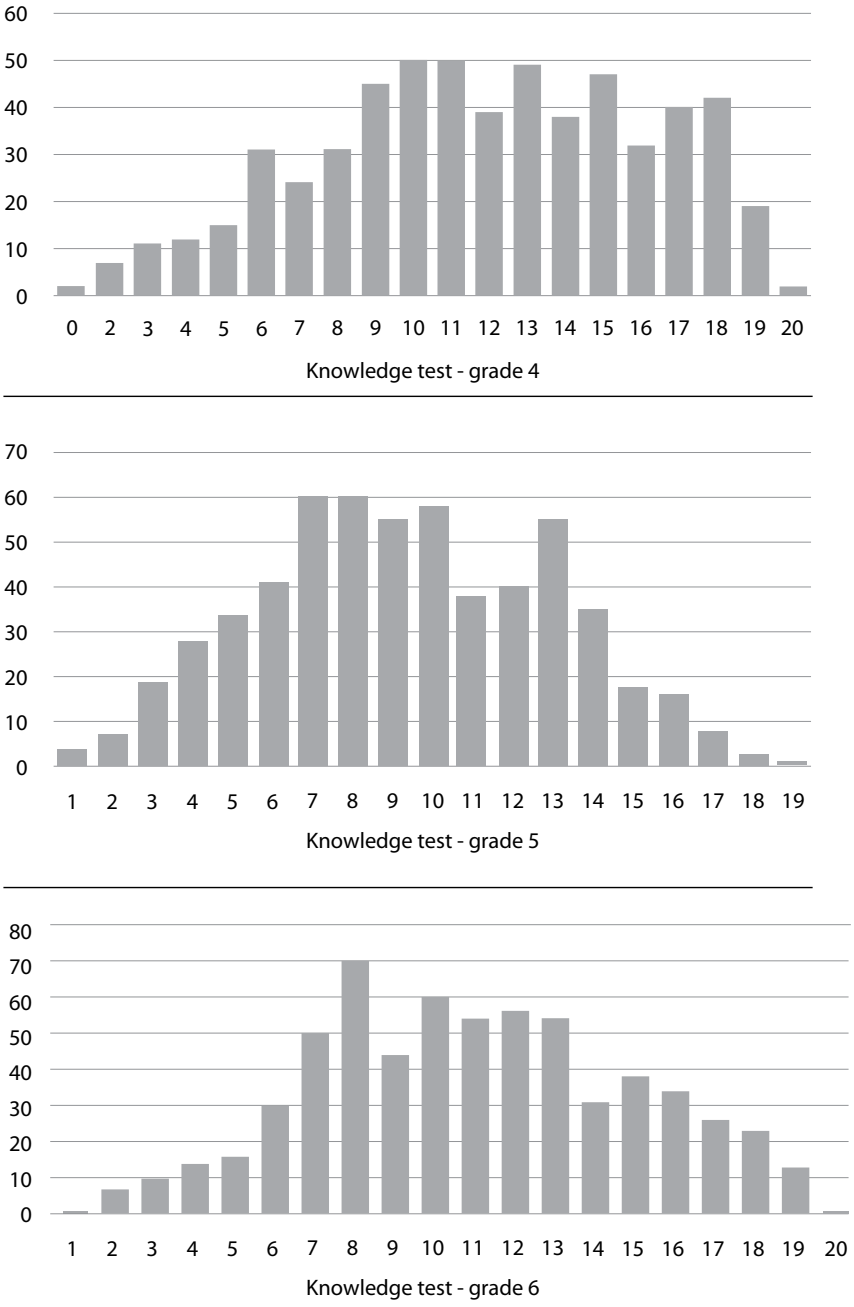

Figure 1. Distribution of students' scores on STEM knowledge test in grades 4, 5 and 6

Table 6
Average item difficulty and item content structure in the 4th grade test

Item	p	SD	K	A	C	Content	Answer	Subject
1	.78	.42	K2	A2	C1	Arithmetic	MC	Math
2	.91	.28	K1	A1	C2	Arithmetic, animals	MC	Math+Sci
3	.53	.50	K2	A2	C2	Measuring time	MC	Math+Sci
4	.63	.48	K1	A3	C1	Length estimation	MC	Math
5	.74	.44	K2	A3,A2	C2	Measuring weight, arithm.	SA	Math+Math
6	.52	.50	K2	A2	C3	Measuring time	SA	Math+Sci
7	.81	.39	K1	A3	C2	2D spat. ability, Standpoint	MC	Math+Sci
8	.56	.50	K2	A2	C3	Measuring liquid	MC	Math
9	.52	.50	K2	A2	C2	Water and measuring temp.	SA	Sci+Math
10	.43	.50	K2	A2	C2	Arithmetic	SA	Math
11	.56	.50	K1	A3	C2	Reading a map	MC	Sci
12	.59	.49	K1	A2	C2	Quantity	MC	Math
13	.51	.50	K2	A3,A2	C1	3D spatial ability	MC	Math
14	.47	.50	K1	A3	C2	Measuring liquid	MC	Math
15	.41	.49	K2	A2	C2	Measurement units	SA	Math
16	.47	.50	K2	A3,A2	C2	Presenting data, arithm.	SA	Math+Math
17	.73	.45	K3	A4	C2	Temperature	MC	Sci
18	.75	.44	K1	A3,A1	C1	2D spatial ability	MC	Math
19	.45	.50	K2	A2	C2	Measuring temperature	SA	Math+Sci
20	.84	.37	K2	A3,A1	C2	Presenting data, animals	MC	Math+Sci

Legend: K1-reproduction, K2-making connections, K3-reflection; A1-presentation, A2-calculation, A3-interpretation, A4-argumentation, F-factual knowledge; C1-no context, C2-realistic context, C3-authentic context, MC-multiple choice; SA-short answer.

Item difficulty in the 5th grade test was appropriate for most items (Table 7). It ranges from $p = .19$ to $p = .85$. Only two items are somewhat inappropriate with lower variability. Item difficulty of the third item in the test was $p = .85$ and this was the easiest test item. This item refers to geometry and measurement with a complexity level 1 (K1). The most difficult item for students in this test ($p = .19$) was item number 4, even though it was set at the complexity level 1 (K1). With respect to content, 60% of the fifth-grade items refer to Mathematics, 30% to Geography, 20% to Technical culture and 20% to Biology (Table 7). In 40% of items content or context integration was accomplished.

Table 7
Average item difficulty and item content structure in the 5th grade test

Item	p	SD	K	A	C	Content	Answer	Subject
1	.59	.49	K1	A2	C1	Arithmetic	MC	Math
2	.68	.47	K1	A3,A1	C1	2D and 3D spatial ability	MC	Tech+Math
3	.85	.36	K1	A3	C2	Perimeter of a plane figure	SA	Math
4	.19	.39	K1	A2	C2	Measuring volume	MC	Math
5	.51	.50	K2	A3,A2	C1	Measuring area	SA	Math+Tech
6	.46	.50	K1	A3	C2	Light and spatial ability	MC	Tech+Math
7	.43	.50	K3	A3	C3	Earth	MC	Geo
8	.47	.50	K2	A3	C3	Earth – rotation	MC	Geo
9	.45	.50	K1	A3	C3	Geography and percentages	MC	Geo+Math
10	.42	.49	K1	F	C3	Earth – rotation	SA	Geo
11	.58	.49	K2	A3,A2	C2	Metric scale and map	SA	Geo+Tech
12	.24	.43	K2	A2	C2	Metric scale – computation	MC	Geo+Math
13	.43	.50	K1	A3	C3	Human and health	MC	Bio
14	.71	.45	K1	F	C3	Animals	MC	Bio
15	.42	.49	K1	A3	C1	Ground plan	MC	Tech
16	.65	.48	K3	A3	C2	Plants, data	MC	Bio
17	.23	.42	K2	A2	C2	Nature, arithmetic	SA	Math+Bio
18	.39	.49	K2	A2,A3	C1	Measuring length	SA	Math
19	.47	.50	K2	A2	C1	Numbers, estimation	MC	Math
20	.44	.50	K2	A3,A2	C2	Presenting data, arithmetic	SA	Math+Math

Legend: K1-reproduction, K2-making connections, K3-reflection; A1-presentation, A2-calculation, A3-interpretation, A4-argumentation, F- factual knowledge; C1-no context, C2-realistic context, C3-authentic context, MC-multiple choice; SA-short answer.

Item difficulty for the 6th grade test was appropriate for almost all items with the lowest index of $p = .13$ and the highest of $p = .91$ (Table 8). Only three items were either too easy or too difficult for the respondents. The easiest item, which requires reproduction (K1), was item number 12 and the percentage of students that correctly answered this item was 91%. The most difficult item was number 3, which refers to the cognitively most complex level – reflection (K3). The percentage of students that correctly answered this item was 13%. With respect to the content, 60% of the six-grade items refer to Mathematics, 40% to Geography, 25% to Biology and 15% to Technical culture (Table 8). Content integration was accomplished only in 35% of sixth-grade items, mainly because of the broadness of the subject matter. Some items in the sixth grade contain the integration of three school subjects.

Table 8
Average item difficulty and item content structure in the 6th grade test

Item	*p*	SD	K	A	C	Content	Answer	Subject
1	.44	.50	K1	A2	C2	Measuring units	SA	Math+Tech
2	.64	.48	K1	A3	C1	Arithmetic	MC	Math
3	.13	.34	K3	A3,A1	C1	Fraction representations	MC	Math
4	.51	.50	K2	A2	C1	Arithmetic	SA	Math
5	.79	.41	K1	A3	C1	Plane geometry	MC	Math
6	.60	.49	K1	A3	C1	Orientation on the line	MC	Math
7	.60	.49	K3	A3	C3	Scales and proportions	MC	Tech
8	.48	.50	K3	A3	C3	Climate	MC	Geo+Math
9	.60	.49	K1	A3	C2	Photosynthesis	MC	Bio
10	.63	.48	K1	A3	C2	Meteorology	MC	Geo
11	.52	.50	K2	A2	C3	Altitude and temperature	SA	Geo+Math
12	.91	.29	K1	F	C2	Animals	MC	Bio
13	.86	.34	K1	A3	C2	Animals	MC	Bio
14	.47	.50	K2	A2	C3	Sea	SA	Geo+Math
15	.63	.48	K1	F	C1	Energy	MC	Tech+Bio+Geo
16	.73	.44	K1	F	C3	Energy	MC	Tech+Bio+Geo
17	.37	.48	K2	A2	C2	Arithmetic	SA	Math
18	.73	.44	K3	A3	C2	Climate	MC	Geo
19	.38	.49	K2	A3,A2	C2	Presenting data, arithmetic	MC	Math
20	.39	.49	K2	A3,A2	C3	Population density	MC	Geo+Math

Legend: K1-reproduction, K2-making connections, K3-reflection; A1-presentation, A2-calculation, A3-interpretation, A4-argumentation, F- factual knowledge; C1-no context, C2-realistic context, C3-authentic context, MC-multiple choice; SA-short answer.

Average test difficulty index for the 5th grade is lower in comparison to the ones in the other two grades (Table 5). Results presented in Tables 6 and 8 show that the test for the 4th grade does not have low difficulty items and the 6th grade test has only one such item, while the 5th grade test contains as much as three items with the difficulty index lower than 0.25 (Table 7). One item refers to K1 and two items refer to connections (K2): more complex use of computations using the contexts of two

new school subjects, Geography and Biology. Also, the fifth-grade test contains only two items with $p > .7$.

Furthermore, the acceptable reliability was obtained for all three tests. Cronbach alphas were $\alpha_{4th} = .78$, $\alpha_{5th} = .70$, and $\alpha_{6th} = .79$, respectively by grades. Items within the tests generally correlated moderately to highly with the total test scores. Based on the reliability indices we can conclude that each of the three tests measures a common construct.

The Principle component analysis was applied to examine structural validity of the tests. As expected, the unidimensional structure was obtained for all three tests, with items saturated moderately to highly with general factor, indicating the existence of one underlying latent construct. The decision for retaining only the first component was based on the Cattell Scree test (Figure 2), Velicer's MAP test and the Parallel analysis[1] conducted for all three data sets.

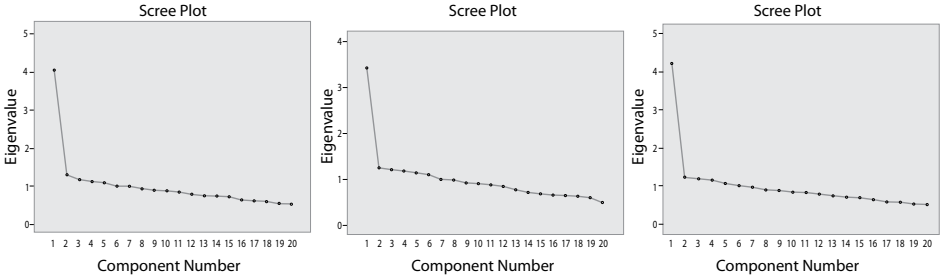

Figure 2. Scree plots from the principal components analysis of the test items in grades 4, 5 and 6

Confirmatory factor analysis also confirmed the unidimensional structure of all three tests. Five indicators of model fit were used: *the Chi-Square Test* (χ^2), *Relative Chi-Square Test* (χ^2/df), *Comparative Fit index* (CFI), *Tucker-Lewis index* (TLI) and *Root Mean Square Error of Approximation (RMSEA)* (Browne & Cudeck, 1993). Relative Chi Square values between 1 and 5 indicate good model fit with values closer to 1 indicating better fit (Mulaik et al., 1989). Hu and Bentler (1999) proposed using the CFI and TLI indices equal or higher than .95 and RMSEA index equal or lower than .06 to indicate good model fit.

The obtained fit indices indicated a very good fit of the specified unidimensional model to the data in the 4th grade (χ^2 (170) = 239; $p < .05$; χ^2/df = 1.14; CFI = .951; TLI = .945; RMSEA = .026), in the 5th grade (χ^2 (170) = 313.15; $p < .05$; χ^2/df = 1.84; CFI = .941; TLI = .935; RMSEA = .035) and in the 6th grade (χ^2 (170) = 248.8; $p < .05$; χ^2/df = 1.46; CFI = .938; TLI = .931; RMSEA = .027). The confirmatory factor analysis confirmed the unidimensional test structure, indicating that these tests can be used as a measure of integrated knowledge in the STEM area.

[1] Results of MAP test and Parallel analysis can be provided on request.

Table 9
Correlation of total test score in grades 4-6 and prior STEM school achievement

	STEM test grade 4	STEM test grade 5	STEM test grade 6
STEM school achievement 2014/2015	.53	.53	.59
STEM school achievement 2015/2016	.52	.62	.59
Mathematics 2014/2015	.53	.53	.54
Mathematics 2015/2016	.52	.62	.61
Geography 2014/2015	-	-	.51
Geography 2015/2016	-	.54	.54
Biology 2014/2015	-	.43	.47
Biology 2015/2016	-	.53	.46
Technical culture 2014/2015	-	-	.42
Technical culture 2015/2016	-	.44	.38

* All correlation coefficients are significant at $p < .01$ level

The external validity of the test was examined by analyzing the correlations of total test scores and earlier school achievement in the STEM area (Table 9). Since the tests were developed as an integrated measure of knowledge in the STEM field, moderate correlations with average grades in the STEM school subjects were expected, and obtained (ranging from $r = .52$ to $r = .62$). The total test scores had the highest correlations with Math grades and somewhat lower correlation with grades in other STEM school subjects (Geography, Biology, Technical culture).

Discussion and Conclusion

The research results presented in this paper show that achievement in the STEM area can be measured by a single measure, despite the fact that these contents are taught within several separate school subjects. The developed tests have proven to be reliable and valid measures of knowledge in the STEM area at primary school level. The unidimensional structure has been proved using the exploratory and confirmatory factor analysis. Nearly normal distribution of test scores was obtained in all three tests with adequately large variability. Average test difficulty indices indicated appropriate difficulty of all three tests. The total test scores were in correlation with school grades in STEM school subjects indicating good construct validity of the tests. Therefore, the tests are measuring what they are intended for - integrated knowledge in the STEM area. These findings encourage the conceptualization and development of integrated testing tools in the STEM. It is especially important, because, according to Honey et al. (2014), such tests are still rare compared to tests that measure knowledge in separate STEM disciplines.

During the development of these knowledge tests, our attention was, besides to test content, also devoted to other components of item development (Sullivan et al., 2013). The students' cognitive activities during task solving were taken into account, as well as the type of questions, and the context of each item. The proportions of these components varied between tests, and were different in the fourth grade test compared to tests for the fifth and sixth grades. For example, in the fifth and the sixth grade tests half of the items required direct application of basic knowledge (K1), while the share of such tasks in the fourth grade was considerably lower (35%). The reason for this is that teaching in the fourth grade is in the form of class teaching and the subject covered by the test items was Mathematics with smaller portion of Science. Whereas in the fifth and sixth grades students learn a broader spectrum of content in four different school subjects with much wider (and shallower) scope of information. That was the reason for the greater proportion of integration in the fourth grade.

Also, given that the content of STEM knowledge in the fourth grade is restricted, there is a growing need for using higher cognitive processes to manipulate with wider range of STEM information presented in the higher grades. Additionally, in the fourth grade, 60% of test items required the knowledge of making connections (K2), while in the fifth and sixth grade making connections was represented in 40% and 30% of the items, respectively. However, as the proportion of items that require making connections (K2) decreases from the 4th to the 6th grade, the proportion of items that require deeper thinking and reflection (K3) increases through the educational level. This finding is in accordance with students' cognitive development.

The analysis of task requirements and item difficulty shows that cognitive complexity of items is mostly consistent with the difficulty of the item. For example, many items assigned with the lowest cognitive level (K1) can be characterized as the easy ones compared to the average test difficulty. However, it should be noted that this is not true for all test items. The item difficulty is influenced also by other factors and not just by cognitive complexity of the tasks. For example, in the fifth grade test, the easiest and the most difficult items are both categorized at the lowest cognitive level (K1). The reason why students have solved the task poorly, when direct application of basic knowledge was needed, is probably due to low representation of specific item content in the current curriculum (MZOS, 2006). This item refers to measurement and space geometry, and it is a fact that in school practice many teachers skip this content at the end of the fourth grade of primary school (Glasnović Gracin, 2016). Also, the lower test difficulty index for the 5th grade lies in the higher proportion of more difficult items in comparison to the other two grades. The reason may be the concept of Measurement present in many items, which is important for STEM competence, but is not highlighted as a domain in the current curriculum in Croatia (MZOS, 2006). Further, the finding that the level of cognitive complexity of the item does not correspond directly to item difficulty is in accordance with the Austrian Mathematical Standards for Education (IDM, 2007) where detailed examples of cognitive levels K1, K2 and K3 are given.

In all grades, items that are characterized as the best regarding test reliability are items at the K2 cognitive level – making connections. In this type of items, to solve the task correctly, student should link different concepts from the STEM area. This finding indicates that *making connections* is a specific cognitive activity needed for successful integration of STEM subject areas. It should be the basic cognitive activity targeted throughout item development in order to make an appropriate test of integrated STEM knowledge. Making connections in many aspects (concepts, contexts, activities, etc.) is also described in literature as a specificity of integration within STEM area (Honey et al., 2014). Furthermore, the items that increase reliability of all three tests require *calculation activities* in a given STEM situation (temperature measurement, sea salinity, etc.). It is widely expected, because calculation as a mathematics application is traditionally an important activity within the STEM area (UNESCO, 2010), and mastering arithmetic operations is one of the most important predictors of STEM achievement (Nakakoji & Wilson, 2014). Furthermore, although tests predominately have multiple-choice questions, it should be noted that the items which mostly contribute to the reliability and validity of the tests require short open-ended answers. Correspondingly, all items that do not contribute to the reliability of the tests are multiple-choice items. These findings can be explained by the random guess response strategy some students had used.

Insights into items that undermine the reliability of the test show that two out of three such items require some kind of estimation from students. Although estimation, together with measurement, is part of STEM activity (Honey et al., 2014), estimation by itself is not emphasized in the current Primary School Curriculum in Croatia (MZOS, 2006). It is also true for the activity of image interpretation and choosing different strategies, which are also not emphasized in the current curriculum (Glasnović Gracin, 2011).

Analyses of the students' activities needed to resolve the tasks show that calculation and interpretation prevail. At the same time, the proportion of calculation tasks decreases from the fourth to the sixth grade, while the proportion of interpretation tasks increases. In all three tests the contextual tasks prevail, which is in line with the description of the framework for an integrated STEM education (Kelley & Knowles, 2016). In Croatian math textbooks, on the other hand, symbolic tasks with a minimal share of the context dominate (Glasnović Gracin, 2011).

Finally, the presented newly developed tests can serve as an efficient, reliable and valid measure of STEM knowledge. For this reason, they could be used not only for research purposes but also in assessing STEM knowledge of primary school children in Croatia. A possible drawback of these tests can be found in few items that do not contribute to the reliability and validity of the tests, or in the few items with an inappropriate difficulty level. However, these items do not disrupt the overall validity of the tests and can be easily modified in future test versions.

The presentation of test items and its development within this paper also have some limitations. We could not publish complete tests or some test items, as they are

being applied throughout a longitudinal study. Thus, the items of all three tests should remain unknown to the potential participants, and cannot be publicly available. In addition, the possibility of wider application of these tests after the completion of the planned study also limits the public availability of all its items[2]. Nonetheless, the development and testing of the metric characteristics of integrated STEM knowledge tests was a great challenge. Such tests are relatively new in the educational context, both in research area and in school practice. Therefore, the authors believe that the presented results will be helpful in promoting STEM education and in improving STEM knowledge assessment.

Acknowledgement

This research was supported by grants from the Croatian Science Foundation through project IP-09-2014-9250 "STEM career aspirations during primary schooling: A cohort-sequential longitudinal study of relations between achievement, self-competence beliefs, and career interests (JOBSTEM)".

References

Adelman, C. (1999a). *Answers in the Tool Box: Academic Intensity. Attendance Patterns, and Bachelor's Degree Attainment.* Washington, DC: US Department of Education.

Adelman, C. (1999b). *Women and men of the engineering path: A model for analysis of undergraduate careers*. Washington, DC: US Department of Education.

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college.* Washington, DC: US Department of Education.

Anderson, E., & Kim, D. (2006). *Increasing the success of minority students in science and technology*. Washington, DC: American Council on Education.

Aschbacher, P. R., Li, E., & Roth, E. J. (2010). Is science me? High school students' identities, participation and aspirations in science, engineering, and medicine. *Journal of Research in Science Teaching, 47*, 564-582.

ASPIRES (2013). *ASPIRES - Young people's science and career aspirations, age 10-14.* London: Department of Education and Professional Studies, King's College London.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions, 154*, 136-136.

Ceci, S., Williams, W., & Barnett, S. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, *135*(2), 218-26. https://doi.org/10.1037/a0014412

[2] This is similar to other studies with non-publicly available test tasks, such as PISA, with the goal to protect the integrity of the developed tests (OECD, 2016).

Dugger, E. W. (2010). *Evolution of STEM in the United States*. Paper presented at the 6th Biennial International Conference on Technology Education Research in Australia. Retrieved from http://citeseerx.ist.psu.edu

Eccles, J. S. (2007). Where are all the women? Gender differences in participation in physical science and engineering. In J. S. Ceci, & W. M. Williams (Eds.), *Why aren't more women in science? Top researchers debate evidence* (pp. 199-210). Washington: American Psychological Association.

Eccles, J. S. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist*, *44*(2), 78–89.

European Commission (2004). *Europe needs more scientists!* Brussels: European Commission, Directorate - General for Research, High Level Group on Human Resources for Science and Technology in Europe. Retrieved from https://ec.europa.eu/research/conferences/2004/sciprof/pdf/conference_review_en.pdf

European Commission (2012). *Rethinking Education: Investing in skills for better socio-economic outcomes.* Brussels: European Commission. Retrieved from www.cedefop.europa.eu/files/com669_en.pdf

European Commission (2013). *Addressing Low Achievement in Mathematics and Science*. Retrieved from http://ec.europa.eu/dgs/education_culture/repository/education/policy/strategic-framework/archive/documents/wg-mst-final-report_en.pdf

Glasnović Gracin, D. (2011). *Requirements in mathematics textbooks and PISA assessment. (Doctoral dissertation).* Klagenfurt: University of Klagenfurt.

Glasnović Gracin, D. (2016). Više prostora za geometriju prostora! [More space for space geometry!]. *Matematika i škola*, *85*(17), 194-195.

Hagedorn, L. S., & DuBray, D. (2010). Math and science success and nonsuccess: Journeys within the community college. *Journal of Women and Minorities in Science and Engineering*, *16*(1), 31-50. https://doi.org/10.1615/JWomenMinorScienEng.v16.i1.30

Hair, J. F., Tatham, R. L., Anderson, R. E., & Black, W. (2010). *Multivariate data analysis*. Upper Saddle River, NJ: Pearson Prentice Hall.

Honey, M., Pearson, G., & Schweingruber, H. (2014). *STEM Integration in K-12 Education: Status, Prospects, and an Agenda for Research*. Washington, DC: The National Academic Press.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hurley, M. M. (2001). Reviewing Integrated Science and Mathematics: The Search for Evidence and Definitions from New Perspectives. *School Science and Mathematics*, *101*(5), 259-268. https://doi.org/10.1111/j.1949-8594.2001.tb18028.x

IDM - Institut für Didaktik der Mathematik. (2007). *Standards für die mathematischen Fähigkeiten österreichischer Schülerinnen und Schüler am Ende der 8. Schulstufe* [Standards for mathematical competences of Austrian students at the end of the 8th school grade]. Klagenfurt: Alpen-Adria-Universität.

Kelley, T. R., & Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education*, *3*(11), 1-11. https://doi.org/10.1186/s40594-016-0046-z

Kennedy, T., & Odell, M. (2014). Engaging students in STEM education. *Science Education International, 25*(3), 246-258.

KMK - Kultusministerkonferenz (2003). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss*. Beschluss vom 15. 10. 2004. [Educational standards for mathematics at the end of middle grades, Released Oct 15, 2004]. Retrieved from http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2003/2003_12_04-Bildungsstandards-Mathe-Mittleren-SA.pdf

Krković, A. (1978). *Elementi psihologije I* [Elements of psychology 1]. Filozofski fakultet: Zagreb.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*(3), 430. https://doi.org/10.1037/0033-2909.105.3.430

Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *TIMSS 2015 Assessment Frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

MZOS - Ministarstvo znanosti, obrazovanja i športa (2006). *Nastavni plan i program za osnovnu školu* [Teaching plan and program for primary school]. Zagreb: Ministarstvo znanosti, obrazovanja i športa Republike Hrvatske.

MZOS - Ministarstvo znanosti, obrazovanja i športa (2010). *Nacionalni okvirni kurikulum za predškolski odgoj i obrazovanje te opće obvezno i srednješkolsko obrazovanje* [National Framework Curriculum for Preschool Education, General Compulsory and Secondary School Education]. Zagreb: Ministarstvo znanosti, obrazovanja i športa Republike Hrvatske.

MZOS - Ministarstvo znanosti, obrazovanja i športa (2014). *Strategija obrazovanja, znanosti i tehnologije* [The Strategy of Education, Science and Technology]. Zagreb: Ministarstvo znanosti, obrazovanja i športa Republike Hrvatske. Retrieved from http://www.azoo.hr/images/AZOO/Cjelovit_sadrzaj_Strategije_obrazovanja_znanosti_i_tehnologije.pdf

Nakakoji, Y., & Wilson, R. (2014). Maths is a strong predictor of STEM attainment in first year university. *Proceedings of the Australian Conference on Science and Mathematics Education* (pp. 149-155). Sidney: University of Sidney.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. McGraw-Hill: New York.

OECD - Organisation for Economic Co-operation and Development (2003). *The PISA 2003 assessment framework – mathematics, reading, science and problem solving knowledge and skills*. Retrieved from http://www.pisa.oecd.org/dataoecd/46/14/33694881.pdf

OECD - Organisation for Economic Co-operation and Development (2016). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. Retrieved from http://www.mecd.gob.es/dctm/inee/internacional/pisa-2015-frameworks.pdf?documentId=0901e72b820fee48

Osborne, J., & Dillon, J. (2008). *Science Education in Europe: Critical Reflections*. London: The Nuffield Foundation.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education. Washington, DC: The National Academies Press.

Sjøberg, S., & Schreiner, C. (2010). *The ROSE project: An overview and key findings*. Oslo: University of Oslo.

Smith, M. P., & Stein, M. K. (1998). Selecting and creating mathematical tasks: From research to practice. *Mathematics Teaching in the Middle School*, *3*(5), 344-350.

Sullivan, P., Clarke, D. M., & Clarke, B. A. (2013). *Teaching with tasks for effective mathematics learning*. New York (NY): Springer. https://doi.org/10.1007/978-1-4614-4681-1

Thomas, S., Gana, Y., & Munoz Chereau, B. (2016). The Intersection of International Achievement Testing and Educational Policy Development in England. In L. Volante (Ed.), *The Intersection of International Achievement Testing and Educational Policy: Global Perspectives on Large-Scale Reform.* New York, NY: Routledge.

UNESCO. (2010). *Engineering: Issues, Challenges, and Opportunities for Development*. Paris: UNESCO. Retrieved from http://unesdoc.unesco.org/images/0018/001897/189753e.pdf

UNESCO. (2015). *STEM education and the curriculum: Issues, tensions and challenges*. International STEM High-level Policy Forum on "Evidence-based Science Education in Developing Countries", Kuala Lumpur, 26-27 May 2015. Paris: UNESCO.

Volante, L. (Ed.) (2016). *The Intersection of International Achievement Testing and Educational Policy: Global Perspectives on Large-Scale Reform.* New York, NY: Routledge.

Xie, Y., Fang, M., & Shauman, K. (2015). Stem Education. *Annual Review of Sociology*, 41, 331-357. https://doi.org/10.1146/annurev-soc-071312-145659

Dubravka Glasnović Gracin
University of Zagreb, Faculty of Teacher Education
Savska cesta 77, 10 000 Zagreb, Croatia
dubravka.glasnovic@ufzg.hr

Toni Babarović
Ivo Pilar Institute of Social Sciences
Marulićev trg 19/I, 10 000 Zagreb, Croatia
toni.babarović@pilar.hr

Ivan Dević
Ivo Pilar Institute of Social Sciences
Marulićev trg 19/I, 10 000 Zagreb, Croatia
ivan.devic@pilar.hr

Josip Burušić
Ivo Pilar Institute of Social Sciences
Marulićev trg 19/I, 10 000 Zagreb, Croatia
josip.burusic@pilar.hr

Razvoj i validacija novih objektivnih testova znanja unutar STEM područja za učenike osnovnih škola

Sažetak

Mjerenje učeničkog postignuća jedno je od središnjih pitanja u istraživanjima kvalitete i učinkovitosti obrazovanja. U području prirodoslovlja i matematike nekoliko je dosadašnjih studija nastojalo iskazati postignuće učenika jedinstvenim pokazateljem znanja unutar tog područja. Cilj je ovog rada razmotriti mogućnost mjerenja znanja učenika u STEM području te prikazati obilježja novih testova razvijenih posebno s tom svrhom.

U radu je prikazana cjelovita psihometrijska analiza novokonstruiranih testova te su prikazani koraci i pristupi u određivanju sadržajnih zahtjeva testova, zatim ishodi preliminarnih validacija inicijalnih inačica testova, kao i rezultati do kojih se došlo u glavnom dijelu istraživanja. U glavnom istraživanju sudjelovalo je 586 učenika četvrtog razreda, 580 učenika petog razreda i 632 učenika šestog razreda.

Postupcima eksploratorne i konfirmatorne faktorske analize potvrđena je jasna jednofaktorska struktura korištenih testova koja je praćena solidnim pokazateljima unutarnje konzistencije ($\alpha_{4.r}$ = ,78; $\alpha_{5.r}$ = ,70; $\alpha_{6.r}$ = ,79). Ukupan rezultat u testu umjereno je do visoko povezan sa školskim postignućem u STEM školskim predmetima. Zaključeno je da sva tri novokonstruirana testa predstavljaju jednodimenzionalnu, a ipak pouzdanu, osjetljivu i valjanu mjeru znanja učenika u STEM području.

Ključne riječi: *mjerenje znanja; razvoj testa; STEM postignuće*

Uvod

Američka Nacionalna zaklada za znanost (NSF, National Science Foundation) oblikovala je akronim STEM (*science, technology, engineering and mathematics*) koji je široko prihvaćen i danas se univerzalno koristi kako bi se njime označilo posebno područje znanja i spoznaja (Dugger, 2010). Termin *Science* (S) odnosi se na izučavanje prirodnog svijeta, uključujući primjenu, ideje i zakone fizike, kemije, biologije i srodnih znanosti (Honey, Pearson, i Schweingruber, 2014). *Technology* (T) obuhvaća čitav sustav ljudi i organizacija, znanja, procesa i uređaja koji sudjeluju u kreiranju

i operiranju s tehnološkim artefaktima koje je čovjek stvorio kako bi zadovoljio svoje potrebe. Termin *Engeneering* (E), inženjerstvo, odnosi se na znanje o kreiranju artefakata, kao i na procese dizajniranja i rješavanja raznih problema. Jedna grana inženjerstva odnosi se na prirodne zakonitosti, a druge obuhvaćaju vrijeme, novac, dostupne materijale, regulacije okoliša i sl. Slovo M u akronimu STEM odnosi se na matematiku (*Mathematics*) kao znanost o uzorcima i odnosima među količinama, brojevima i prostornim objektima (Honey i sur., 2014).

Fenomen STEM-a u posljednjem je desetljeću sagledavan iz različitih perspektiva. Pritom su razmatrane posljedice deficita STEM stručnjaka na nacionalne ekonomije i njihov utjecaj na postojeće i buduće tržište rada (European commission, 2004; Osborne i Dillon, 2008; UNESCO, 2010), kao i pitanja vezana uz socijalne i rodne razlike i uvjerenja (Anderson i Kim, 2006; Ceci, Williams, i Barnett, 2009; Eccles, 2007, 2009). Također, istraživanja vezana uz STEM problematiku usmjeravaju se prema važnosti roditelja i nastavnika kao modela za STEM zanimanja (Adelman 1999a, 1999b, 2006; Aschbacher, Li, i Roth, 2010; Hagedorn i DuBray, 2010). Ostale važne studije odnose se i na uspješnost određenih programa i intervencija usmjerenih na promicanje STEM zanimanja među učenicima (ASPIRES, 2013; Sjøberg i Schreiner, 2010) i na pronalaženje prikladnih teorijskih objašnjenja u vezi s tim zašto učenici ustraju ili napuštaju STEM obrazovna područja (Adelman 1999a, 1999b, 2006; Anderson i Kim, 2006; Hagedorn i DuBray, 2010).

Posljednjih godina jedan od zanimljivijih izazova vezanih uz STEM odnosi se na primjerenu vrstu nastave u tom području (European Commission, 2012; Honey i sur., 2014), što kao posljedicu ima i pitanje operacionalizacije i iskazivanja učeničkih postignuća. Xie, Fang, i Shauman (2015) govore o terminu STEM obrazovanja (eng. *STEM education*) u kojem je naglasak na logičkoj i konceptualnoj vezi među različitim STEM područjima. Takvim se pristupom na STEM obrazovanje gleda kao na cjelinu, a ne kao na skup odvojenih disciplina. S tom svrhom Honey i sur. (2014) razvili su istraživački okvir s ciljem proučavanja pozitivnih ishoda integriranog STEM obrazovanja, u što su uključeni razni parametri poput učeničkog interesa, motivacije i STEM postignuća. Kennedy i Odel (2014) ističu da kvalitetno STEM obrazovanje treba obuhvatiti integraciju STEM područja i projektnu nastavu, poticati znanstveni pristup, inženjerski dizajn i matematičku strogoću. Također, autori smatraju da treba uključivati suradničke pristupe učenju uz povezivanje učenika i nastavnika s ekspertima unutar STEM-a, poticati iskustva formalnog i neformalnog učenja te implementirati prikladne tehnologije za unapređenje učenja.

No, ideja integriranja STEM predmeta u školskoj praksi vrlo je kompleksna, osjetljiva i otvara za sobom brojna pitanja u suvremenim obrazovnim raspravama (Kelley i Knowles, 2016). Prema metaanalizi koju je provela Hurley (2001) o istraživanjima i projektima vezanima uz integraciju nastave matematike i prirodnih predmeta, pokazalo se da je dobrobit takve integracije za ishode u matematici mnogo manja u usporedbi s ishodima u drugim prirodnim predmetima. UNESCO (2015) upozorava

na još neke probleme vezane uz integrirano STEM obrazovanje, poput nepripremljenih nastavnika za ovakav interdisciplinarni pristup, tradicionalno snažnih granica između zasebnih predmeta i loš status integriranih u odnosu na samostalne predmete. Uz to, STEM sadržaji i zahtjevi brzo zastarijevaju, a svaka STEM disciplina za sebe je toliko razvijena da sadrži veoma visok stupanj detalja. Stoga je velik izazov definirati sadržajne i druge osnove unutar STEM obrazovanja.

Unatoč aktualnim raspravama, poznavanje disciplina unutar STEM područja u svakom je slučaju prepoznato kao važan čimbenik u ekonomskom razvoju društva, a učenička postignuća na testovima kao ključni prediktori za pripremljenost mladih osoba za život u suvremenom društvu u kojem matematika, prirodne znanosti i tehnologija imaju važnu ulogu (Organisation for Economic Co-operation and Development [OECD], 2003). Tim su se pitanjem bavili istraživači u sklopu određenih opsežnih međunarodnih projekata usmjerenih na mjerenje postignuća učenika, poput PISA-e (Programme for International Student Assessment) i TIMSS-a (Trends in International Mathematics and Science Study). Unutar tih studija prevladava pristup da svim učenicima treba pružiti obrazovanje koje će im omogućiti stjecanje temeljnih kompetencija za život, pri čemu je naglasak stavljen i na učenička postignuća u područjima prirodoslovlja i matematike (Mullis i Martin, 2013; OECD, 2003). Teorijski okviri tih projekata utjecali su na promjene mnogih suvremenih nacionalnih kurikula iz matematike ili prirodnih predmeta (npr. Kultusministerkonferenz [KMK], 2003). U svjetlu tih međunarodnih studija potrebno je naglasiti i njihov utjecaj na kasnije postupke mjerenja i operacionalizacije postignuća učenika u nacionalnim obrazovnim kontekstima. Analize pokazuju da su velika istraživanja poput TIMSS-a i PISA-e u mnogim zemljama utjecala na strukturu i zahtjeve unutar nacionalnih testova za mjerenje znanja učenika (Volante, 2016). Primjerice, u Engleskoj su 2009. godine obrazovne vlasti preporučile da, gdje god je moguće, nacionalni ispiti trebaju biti povezani sa zadatcima iz velikih međunarodnih ispitivanja u kojima ta zemlja sudjeluje (Thomas, Gana, i Muñoz-Chereau, 2016).

Primjeri spomenutih međunarodnih projekata upućuju na potrebu razmatranja mogućnosti konceptualizacije i operacionalizacije iskazivanja postignuća učenika u pojedinim tematskim područjima, prije nego stalnog iskazivanja postignuća u specifičnim predmetnim područjima. Jedno od područja u kojem je u svakom slučaju potrebno razmotriti mogućnost iskazivanja zajedničkog postignuća jest STEM područje. Neki dosadašnji pokušaji bili su upravo na tom tragu. Dugger (2010) je razmatrao različite mogućnosti integracije STEM školskih predmeta u školskoj praksi. Jedan od načina je integrirati neku od STEM disciplina u preostale tri. Primjerice, autor predlaže da se inženjerstvo može integrirati u nastavu prirodoslovlja, matematike i tehnologije. Drugi je način prožeti sve STEM discipline međusobno i podučavati ih u sklopu jednog integriranog predmeta. Kennedy i Odel (2014) sugeriraju potrebu integriranja tehnologije i inženjerstva u nastavu matematike i prirodnih predmeta,

što će promovirati znanstveni pristup i proces inženjerskog dizajna. Takav pristup zahtijeva određeni stupanj kurikulne i pedagoške usklađenosti u sva STEM područja (Xie i sur., 2015). Konceptualni okvir za integrirano STEM obrazovanje prikazan je u radu Kelley i Knowles (2016), u kojem autori prikazuju STEM obrazovanje kao integraciju matematičkog mišljenja, inženjerskog dizajna, tehnološke pismenosti i znanstvenog pristupa.

Dosadašnja razmatranja operacionalizacije postignuća u pojedinim STEM predmetima ukazala su na jedan važan problem. Naime, konceptualizacija i iskazivanje postignuća u STEM predmetima primarno su do sada bili usmjereni na postignuća učenika u pojedinačnim predmetima. Mnogo se manje pažnja usmjeravala na međupredmetna znanja, a mjerenje očekivanih kompetencija učenika u integraciji znanja stečenih u više disciplina vrlo je malo prisutno (Honey i sur., 2014). To je, osim u pristupu, vidljivo i u izvedbi pojedinačnih testova ili drugih mjernih postupaka za mjerenje postignuća učenika. Predloženi okviri za testiranje obuhvaćaju naglasak na osnovnim idejama svake STEM discipline, primjeni i povezivanju koncepata unutar STEM-a (National Research Council, 2014), ali u literaturi implementacije takvih testova gotovo da i nema.

U ovom smo istraživanju krenuli upravo od toga. Osnovni cilj ovog istraživanja jest razmotriti mogućnost konceptualizacije, izrade i provedbe mjerenja integriranog STEM postignuća učenika osnovnih škola u Hrvatskoj. S tom svrhom pošlo se od potrebe sastavljanja testova postignuća u STEM području primjerenog učenicima četvrtog, petog i šestog razreda, a zatim njihove primjene. Drugi, jednako važan dio ovog istraživanja, odnosi se na sveobuhvatnu psihometrijsku validaciju konstruiranih testova. U Hrvatskoj još nije bilo studija koje obuhvaćaju sastavljanje, provedbu i provjeru valjanosti integriranih testova STEM znanja, tako da ovdje prikazani testovi mogu imati i značajne praktične implikacije za mjerenje i konceptualizaciju STEM postignuća tijekom osnovne škole.

Metode

Sudionici

U predistraživanju se koristio prigodni uzorak od 118 učenika 4. razreda (N = 41), 5. razreda (N = 46) i učenika 6. razreda (N = 31), odabranih iz jedne zagrebačke osnovne škole koja svojim demografskim i drugim obilježjima u značajnom dijelu slijedi obilježja drugih osnovnih škola iz glavnog istraživanja. U uzorak su uključena po dva cijela razreda iz svake generacije.

U glavnom istraživanju sudjelovalo je 1798 učenika četvrtog, petog i šestog razreda iz 16 osnovnih škola u Gradu Zagrebu i okolici, u dobi od 10 do 12 godina. Odabir sudionika unutar svake od 16 škola učinjen je po slučaju na način da su odabrana 2 razredna odjela jedne generacije učenika unutar pojedine škole. U ukupnom uzorku učenici su podjednako zastupljeni po spolu (49,8 % djevojčica) i s obzirom na razred koji pohađaju ($N^{4.razred}$ = 586, $N^{5.razred}$ = 580, $N^{6.razred}$ = 632).

Mjere i instrumenti

Integrirani testovi STEM školskog postignuća

Za potrebe mjerenja STEM školskog postignuća u 4., 5. i 6. razredu za svaki je razred konstruiran zaseban test. Konstrukcija testova obuhvaćala je nekoliko koraka.

Prvi korak u konstrukciji testova sastojao se od analize važećih kurikulskih dokumenata koji se odnose na Nastavni plan i program STEM predmeta za četvrti, peti i šesti razred (Ministarstvo znanosti, obrazovanja i športa [MZOS], 2006, 2010). U četvrtom razredu i prije njega to su školski predmeti Matematika i Priroda i društvo, a u petom i šestom razredu Matematika, Priroda, Geografija i Tehnička kultura. Tako su dobiveni katalozi STEM znanja za 4., 5. i 6. razred (Tablica 1, Tablica 2 i Tablica 3). Pri konstrukciji testova, s obzirom na primjenu testova početkom drugog polugodišta, koristile su se nastavne teme i postignuća koja su prema programu trebala biti usvojena do kraja prvog polugodišta. Katalozi su pokazali da učenička znanja i kompetencije u STEM području znatno rastu u razdoblju od 4. do 6. razreda. Stoga je svaki test primarno obuhvatio sadržaje koji su se učili u školi u proteklih 12 do 18 mjeseci, do trenutka pisanja testa.

Tablica 1, 2 i 3

U drugom koraku, na temelju analize kurikulskih dokumenata i literature o strukturi testova (Institut für Didaktik der Mathematik [IDM], 2007; Sullivan, Clarke i Clarke, 2013), donesena je odluka o općoj strukturi testova za pojedini razred. Struktura za izradu testova vidljiva je u tablici 4. Uobičajene aktivnosti u STEM zadatcima su računanje, interpretiranje dane formule, slike ili grafa, prikazivanje, objašnjavanje i procjenjivanje (prema IDM, 2007). Dimenzija kompleksnosti odnosi se na kognitivne razine reprodukcije, povezivanja i refleksije (Smith i Stein, 1998). Testovi znanja u okviru JOBSTEM projekta obuhvaćaju sve tri kognitivne razine, usklađeno s dobi učenika u pojedinom dijelu istraživanja. *Reprodukcija (K1)* se odnosi na direktnu primjenu osnovnih pojmova, pravila, postupaka ili prikaza. *Uspostavljanje povezivanja (K2)* potrebno je kada je zadatak kompleksnije prirode pa zahtijeva povezivanje više pojmova, poučaka, postupaka ili prikaza, ili pak treba povezati različite radnje u cjelinu kako bi se problem riješio. Pritom se povezivanja mogu još dodatno razdvojiti na jednostavnija i složenija povezivanja. *Refleksija (K3)* se odnosi na promišljanje o odnosima koji nisu neposredno vidljivi iz danih matematičkih činjenica. S obzirom na duljinu pisanja testa od 45 minuta svi zadatci su tražili kratke i objektivne odgovore, bilo u obliku zadataka višestrukog izbora bilo u obliku otvorenog pitanja gdje se točan odgovor upisuje na crtu.

Tablica 4

U trećem koraku razmotrena je i donesena odluka o načelu integriranosti sadržaja koji će biti zastupljeni u pojedinom testu. Integriranost se u ovim testovima implementira na više razina: (1) povezivanje sadržaja različitih školskih predmeta; (2) povezivanje sadržaja istog školskog predmeta, ali njegovih različitih tema; (3)

sadržajno-kontekstna povezivanja (povezivanje sadržaja jednog predmeta samo s kontekstom iz drugog predmeta); (4) mješovitost zadataka na razini testa (unutar jednog testa pojavljuju se i pojedinačni zadatci iz različitih STEM predmeta). Pritom smo se koristili tekstualnom analizom sličnosti i korelacija među kurikulima STEM predmeta. Katalozi su poslužili kao baza za povezivanje sadržaja u STEM zadatcima, a u zadatke je uklopljen i okvir za integrirano STEM učenje (Kelley i Knowles, 2016) sa zahtjevima poput tehnološke pismenosti, matematičkog mišljenja i inženjerskog dizajna. Osim u sadržaju pri sastavljanju testa povezanost se nastojala dobiti i putem STEM konteksta.

U četvrtom koraku producirani su pojedini zadatci za svaki od testova. Pritom se vodilo računa o variranju zahtjeva unutar strukture prikazane u tablici 4. Iako je naglasak bio na integriranosti unutar pojedinog zadatka, ona nije uvijek bila moguća jer neka osnovna predmetna znanja na ranim stupnjevima obrazovanja čine temelj STEM edukacije. Taj je korak uključio i logičku, sadržajnu, jezičnu, stilsku i oblikovnu stranu formuliranja zadataka.

U petom koraku pristupilo se promišljanju o primjerenosti i kvaliteti kreiranih zadataka. U toj su se fazi provodile rasprave članova tima s nastavnicima iz različitih STEM područja o zahtjevima i sadržajima u pojedinom zadatku. S obzirom na to da je u prethodnom koraku osmišljeno mnogo više zadataka nego je potrebno, u ovoj se fazi pristupilo njihovu odabiru i potrebnoj modifikaciji sadržaja. Primjer integriranog zadatka koji povezuje sadržaje Prirode i društva s Matematikom u 4. razredu je: *"U posudi se nalazi voda. Ako se temperatura vode uveća za 25 °C, postići će se vrelište. Kolika je temperatura vode u toj posudi?"* Kako bi učenik uspješno riješio taj zadatak računanja (Matematika), treba poznavati pojmove vrelišta i temperature vrelišta (Priroda i društvo).

U šestom koraku izrađena je inicijalna inačica testova za učenike.

U sedmom koraku provedena je završna provjera zahtjeva u testovima, napisane su upute za učenike, slike su grafički dovršene te je napravljen prijelom testa za predtestiranje, a nakon toga i za glavno testiranje.

Završna inačica testa sadrži 20 zadataka po svakom testu za koje se smatralo da bi trebali predstavljati pouzdanu, osjetljivu i valjanu mjeru znanja učenika iz STEM područja. Dominirajuće sadržajne komponente u pojedinom zadatku vidljive su u tablicama 6, 7 i 8, a integriranost unutar pojedinog zadatka završnih inačica testova prikazana je u desnim stupcima tablica 1, 2 i 3. U svakom zadatku samo je jedno točno rješenje pa su odgovori kodirani s 1 ili 0. Ukupan rezultat na testu čini zbroj točnih odgovora ispitanika s mogućim rasponom od 0 do 20, pri čemu veći rezultat ukazuje na bolje znanje u STEM području.

Školsko postignuće učenika u STEM području iskazano školskim ocjenama

U istraživanju su se koristile i školske ocjene učenika iz školskih predmeta koji pripadaju STEM području, a to su Matematika u 4. razredu, Priroda, Matematika, Tehnička kultura i Geografija u 5. i 6. razredu. U sadržaju predmeta Priroda i društvo u

velikom dijelu zastupljene su i društvene teme koje ne pripadaju STEM području, tako da se ocjene iz tog predmeta nisu koristile. Ocjene iz školskih predmeta prikupljene su izravno od škola na temelju postojeće školske dokumentacije o uspjehu učenika u školskoj godini 2014./15. i 2015./16. Kao pokazatelj školskog postignuća u STEM području koristio se prosjek ocjena iz navedenih školskih predmeta unutar pojedinog razreda.

Postupak

Podatci su prikupljeni testiranjem učenika unutar redovne **školske** nastave. Prije početka provođenja testiranja roditelji su iscrpno informirani o svim ciljevima i obilježjima ovog istraživanja te je pribavljena pisana suglasnost roditelja/skrbnika za sudjelovanje djece u istraživanju. Pristanak za sudjelovanje djece u istraživanju dalo je 89,8 % roditelja.

U predistraživanju su na uzorcima učenika 4., 5. i 6. razreda primijenjene inicijalne inačice testova znanja. Testiranje je provedeno grupno, u sklopu redovite **školske** nastave, u maksimalnom trajanju od 45 minuta. Nakon analize podataka prikupljenih predistraživanjem pristupilo se reviziji inicijalnih testova na temelju psihometrijskih rezultata. Utvrđeno je da testovi imaju zadovoljavajuću pouzdanost već u svojoj inicijalnoj inačici (**α >** ,70) i da imaju jasnu jednofaktorsku strukturu. Distribucija ukupnog uratka u testu bila je približno normalna, sa zadovoljavajućim varijabilitetom ukupnog rezultata testa. Pokazalo se da manji broj zadataka narušava pouzdanost testa, odnosno da ima niske saturacije prvom glavnom komponentom (r< ,30) ili da ima mali varijabilitet, odnosno da su prelagani ili preteški ciljnoj populaciji (p> ,80, odnosno p< ,20). Za određeni broj zadataka višestrukog izbora dobiveno je da pojedini distraktori nisu dovoljno jaki, odnosno da na sebe vežu manje od 5% učeničkih odgovora. Na osnovi provedenih analiza, u fazi izrade završnih inačica testova, odabrani su zadatci koji imaju dovoljno velik varijabilitet te koji doprinose pouzdanosti i faktorskoj valjanosti testa. Određeni broj zadataka koji nije udovoljavao kriteriju zadovoljavajućeg varijabiliteta modificiran je u smjeru njihova otežavanja ili olakšavanja. Strategija je bila usmjerena adaptaciji njihova sadržaja ili odabiru adekvatnijih distraktora. U slučaju da je pojedini zadatak odstupao od predmeta mjerenja cijelog testa (narušava pouzdanost ili faktorsku valjanost), pristupilo se njegovoj sadržajnoj adaptaciji, ili kreiranju novog zadataka koji ima isti ili sličan predmet mjerenja.

Glavno istraživanje znanja za učenike 4., 5. i 6. razreda provedeno je grupno, u sklopu redovne **školske** nastave, a za rješavanje testa učenici su na raspolaganju imali 45 minuta. Učenici su unaprijed bili informirani o istraživanju i testu znanja, ali nisu unaprijed vježbali za taj test jer je fokus bio na osnovnim STEM znanjima koja učenici inače posjeduju, kao I na veze između pojedinih STEM disciplina.

Statistička analiza podataka

Statističke analize napravljene su s pomoću programa SPSS 23 i AMOS 7.0. Izračunat je ukupan rezultat ispitanika u novokonstruiranim testovima te je provjerena

normalnost distribucija. Deskriptivni podatci koristili su se za validaciju testova i provjeru psihometrijskih svojstava testa. Unutarnja valjanosti testa izražena je Cronbachovim alfa koeficijentom. Analiza glavnih komponenata i konfirmatorna faktorska analiza koristile su se za provjeru strukturalne valjanosti testa. Vanjska valjanost testova izražena je Pearsonovim koeficijentom korelacije između ukupnog uratka u testu i postignutog školskog uspjeha u STEM području.

Rezultati

Prikupljeni su podatci analizirani zasebno za četvrti, peti i šesti razred. Rezultat učenika na testu definiran je ukupnim brojem točno riješenih zadataka kojih je bilo ukupno 20. Prosječni uradak u testu znanja u **četvrtom** razredu bio je M = 11,81 (SD = 4,32), u petom razredu M = 9,34 (SD = 3,66) i u **šestom** razredu M = 10,93 (SD = 3,97). Izračunata prosječna rješivost testova ukazuje na težinsku primjernost svih triju testova. Prosječna rješivost cijelog testa prikazana je u tablici 5.

Tablica 5

Distribucija odgovora učenika 4. razreda je približno normalna, iako blago negativno asimetrična (Slika 1). Distribucije rezultata u testu za 5. i 6. razred približno su normalne. Sve tri distribucije blago su platikurtične, s izduženim krajevima distribucije i nešto manjim grupiranjem rezultata oko središnje vrijednosti u usporedbi s normalnom raspodjelom. Dobivene distribucije ukazuju na to da su konstruirani testovi znanja diskriminativni i da dobro razlikuju učenike preko cijelog raspona bruto rezultata.

Slika 1

Raspon rezultata kod učenika četvrtog razreda seže od X_{min} = 0 do X_{max} = 20, kod učenika petog razreda od X_{min} = 1 do X_{max} = 19 te kod učenika šestog razreda između X_{min} = 1 i X_{max} = 20. Navedeni rasponi uglavnom prekrivaju teorijski raspon i govore o dobroj osjetljivosti testa. O dobroj osjetljivost testa, koja je definirana kao mogućnost razlikovanja ispitanika na osnovi njihova ukupnog individualnog rezultata u testu (Nunnally i Bernstein, 1994), dodatno govore broj ostvarenih razlikovanja testom (BOR) i Fergusonov delta-koeficijent (Krković, 1978). BOR je podjednak u testu za četvrti i peti razred te čak još veći za test u šestom razredu ($BOR_{4.r}$ = 160501, $BOR_{5.r}$ = 155408, $BOR_{6.r}$ = 185541), što ukazuje na vrlo velik broj ostvarenih razlikovanja. Fergusonov koeficijent osjetljivosti, koji predstavlja omjer između broja različitih opaženih bruto rezultata (BROR) i maksimalnog broja mogućih različitih rezultata (BRR), vrlo je visok za sva tri testa, što ukazuje na adekvatnu osjetljivost testova znanja ($\delta_{4.r}$ = ,982; $\delta_{5.r}$ = ,970 $\delta_{6.r}$ = ,976).

Zadatci testa za 4. razred u prosjeku su primjerene težine i kreću se u rasponu od p = ,41 za najteži zadatak do p = ,91 za najlakši zadatak (Tablica 6). Tri zadatka su bila

prelagana ciljnoj populaciji, odnosno imaju sužen varijabilitet (p > ,80). Na primjer, najlakši zadatak u ovom testu (drugi zadatak po redu) tiče se odabira primjerene računske operacije i prikazivanja teksta matematičkim simbolima te po zahtjevnosti pripada najnižoj kognitivnoj razini – reprodukcija i direktna primjena definicija ili pravila (K1). Taj zadatak točno je riješilo 91 % ispitanih učenika. Tablica 6 prikazuje da se najveći udio zadataka odnosi na matematičke sadržaje (90 %), a da se 45 % njih odnosi na prirodu. U otprilike polovini zadataka postignuta je sadržajna ili kontekstualna integracija.

Tablica 6

Zadatci u testu za 5. razred uglavnom su primjerene težine (Tablica 7). Prosječna rješivost zadataka u tom testu kreće se od p = ,19 do p = ,85. Svega dva zadatka nisu primjerene težine i imaju sužen varijabilitet. Treći zadatak u testu za 5. razred ima p = ,85 i ujedno je najlakši zadatak u testu. Zadatak sadržajno pripada području geometrije i mjerenja, a s obzirom na zahtjevnost u najnižu kognitivnu razinu (K1). Učenicima je najteži zadatak (p = ,19) u ovom testu zadatak br. 4, unatoč tome što pripada najnižoj kognitivnoj razini kompleksnosti (K1). Radi se zadatku koji se po sadržaju odnosi na geometriju prostora i mjerenje. Tablica 7 prikazuje da se 60 % zadataka za 5. razred odnosi na matematičke sadržaje, 30 % na geografiju, a po 20 % na tehničku kulturu i prirodu. U 40 % zadataka postignuta je sadržajna ili kontekstualna integracija.

Tablica 7

U testu za učenike 6. razreda zadatci su većinom primjereni po težini, a prosječna rješivost zadataka kreće se od p = ,13 do p = ,91 (Tablica 8). Manji broj zadataka, njih svega 3, učenicima je preteško ili prelagano. Najlakši zadatak u ovom testu, koji je i koncipiran na način da od učenika zahtijeva reproduktivno znanje (K1), jest dvanaesti zadatak testa, a točan odgovor odabralo je 91 % učenika. Najteži zadatak u ovom testu po redu je treći zadatak u testu, koji ujedno pripada kognitivno najzahtjevnijim zadatcima jer se od učenika traži najviša kognitivna razina – refleksija. Na taj zadatak točno je odgovorilo 13 % učenika. Tablica 8 prikazuje da se 60 % zadataka odnosi na matematičke sadržaje, 40 % na geografiju, 25 % na prirodu, a 15 % na tehničku kulturu. Zbog širine sadržaja, u samo 35 % zadataka za 6. razred postignuta je sadržajna integracija, pri čemu su u nekim zadatcima integrirani sadržaji čak triju predmeta.

Tablica 8

Prosječna rješivost cijelog testa za 5. razred manja je u odnosu na rješivosti testova u 4. i 6. razredu (Tablica 5). Rezultati prikazani u tablicama 6 i 8 pokazuju da test za 4. razred ne sadrži zadatke s niskom rješivošću, test za 6. tazred ima jedan takav zadatak, a test za 5. razred ima čak tri zadatka s prosječnom rješivošću manjom od

0,25 (Tablica 7). Jedan od tih zadataka odnosi se na K1, a dva na složenija povezivanja (K2) računanja u kontekstima drugih predmeta, novih u petom razredu (geografija i priroda). Također, test petog razreda sadrži samo dva lakša zadatka (p > ,7).

Nadalje, dobivena je zadovoljavajuća pouzdanost zadataka u sva tri testa. Pouzdanost mjerena metodom unutarnje konzistencije (Cronbach α) iznosi $\alpha_{4.r}$ = ,78 za test za četvrti razred, za peti razred $\alpha_{5.r}$ = ,70 i za šesti razred $\alpha_{6.r}$ = ,79. Zadatci unutar testa uglavnom umjereno do visoko koreliraju s ukupnim uratkom i na osnovi dobivenih pouzdanosti možemo zaključiti da svaki od triju testova mjeri jedinstven predmet mjerenja.

Konstruktna valjanost testova utvrđena je faktorskom analizom – metodom analize Glavnih komponenata. Provedene su tri faktorske analize, po jedna za svaki test znanja. Prema očekivanjima, dobivena je jednofaktorska struktura u sva tri testa znanja s umjerenim do visokim saturacijama većine čestica prvom glavnom komponentom, što ukazuje na postojanje jednog latentnog konstrukta u pozadini čestica testa. O opravdanosti zadržavanja samo prve glavne komponente zorno govori Cattelov Scree test (Slika 2) i dodatno provedeni Velicerov MAP test i Paralelna analiza[1] za sva tri seta podataka.

Slika 2

Postojanje jednofaktorske strukture dodatno je potvrđeno i konfirmatornom faktorskom analizom. Da bismo provjerili u kojoj se mjeri modeli slažu s podatcima, koristili smo pet indikatora slaganja: *hi-kvadrat test* (χ^2), *relativni hi-kvadrat* **(χ^2/df),** *komparativni indeks slaganja* (CFI), *Tucker-Lewis indeks podudarnosti* (TLI) i indeks *prosječne standardne rezidualne pogreške* (RMSEA) (Browne i Cudeck, 1993). Granični raspon vrijednosti relativnog hi-kvadrata koji pokazuje dobro pristajanje modela podatcima jest između 1 i 5, pri čemu vrijednost bliže 1 upućuje na bolji model (Mulaik i sur., 1989). Prema preporukama Hu i Bentler (1999), dobrim pristajanjem modela smatraju se vrijednosti CFI i TLI indeksa jednake ili veće od ,95 i vrijednost RMSEA indeksa jednaka ili manja od .06.

Postavljeni jednofaktorski modeli dobro pristaju podatcima prikupljenim u četvrtom razredu (χ^2 (170) = 239; p < ,05; χ^2/df = 1,14; CFI = ,951; TLI = ,945; RMSEA = ,026), podatcima u petom razredu (χ^2 (170) = 313,15; p < ,05; χ^2/df = 1,84; CFI = ,941; TLI = ,935; RMSEA = ,035) i podatcima prikupljenim u šestom razredu (χ^2 (170) = 248,8; p < ,05; χ^2/df = 1,46; CFI = ,938; TLI = ,931; RMSEA = ,027). Dakle, provedene konfirmatorne faktorske analize potvrđuju jednofaktorsku strukturu testa, odnosno potvrđuju da testovi mjere jedinstveno znanje iz STEM područja.

Vanjska valjanost testa provjerena je na osnovi povezanosti ukupnog uratka u testu i dosadašnjeg školskog uspjeha u STEM području (Tablica 9). Budući da je

[1] Rezultate MAP testa i Paralelne analize moguće je dobiti na zahtjev.

test konstruiran kao integrirani test koji mjeri uspjeh u STEM području, dobivene su očekivane umjereno visoke korelacije s pokazateljima prosječne ocjene u STEM školskim predmetima (od r = ,52 do r = ,62). Ukupni uradak u svakom od triju testova znanja najviše je povezan s dosadašnjim uspjehom iz Matematike i nešto niže s uspjehom u ostalim školskim predmetima iz STEM područja (Geografija, Priroda, Tehnička kultura).

Tablica 9

Rasprava i zaključak

Rezultati istraživanja prikazanog u ovom radu pokazuju da se postignuća iz područja STEM-a mogu mjeriti jedinstvenom mjerom, unatoč tome što se ti sadržaji u školi poučavaju unutar više zasebnih školskih predmeta. Testovi znanja pokazali su se pouzdanim, valjanim i osjetljivim mjerama znanja u STEM području u osnovnoj školi. Eksploratornom i konfirmatornom faktorskom analizom potvrđena je njihova jednofaktorska struktura. U svim trima konstruiranim testovima znanja dobivena je približno normalna distribucija odgovora sa zadovoljavajućim varijabilitetom ukupnog rezultata. Prosječni indeksi lakoće testova ukazuju na težinsku primjerenost svih triju testova. Ukupan rezultat u testu povezan je s uspjehom u školi u školskim predmetima iz STEM područja, što ukazuje na dobru konstruktnu valjanost i, prema tome, test mjeri upravo ono za što je namijenjen, znanje iz STEM područja. Time je dan doprinos konceptualizaciji i razvoju instrumenta za integrirano testiranje u STEM-u jer, prema Honey i sur. (2014), takvi testovi su još uvijek rijetki u odnosu na testove koji mjere znanje u zasebnim STEM područjima.

Prilikom izrade testova znanja pažnja je, osim sadržajima, posvećena i drugim komponentama konstrukcije zadataka (Sullivan i sur., 2013). Vodilo se računa o aktivnostima koje se od učenika očekuju prilikom rješavanja zadatka, kognitivnim razinama, tipu pitanja i vrsti konteksta u zadatku. Udjeli tih komponenata po pojedinom testu pokazuju da su one različito zastupljene u četvrtom razredu u usporedbi s petim i šestim razredom. Primjerice, u petom i šestom razredu polovina je zadataka zahtijevala znanje izravne primjene osnovnih znanja (K1), a udio je takvih zadataka u četvrtom razredu bio znatno manji (35 %). Razlog leži u tome što četvrti razred pripada razrednoj nastavi i obuhvaćeni predmeti bili su samo Matematika i manji dio Prirode i društva, a u petom i šestom razredu radi se o širem spektru sadržaja iz četiriju predmeta pri čemu se ne ide previše u sadržajnu dubinu. To je razlog i veće integracije STEM sadržaja u četvrtom razredu u odnosu na peti i šesti razred. Nadalje, u četvrtom je razredu čak 60 % zadataka zahtijevalo znanje povezivanja (K2), a u petom i šestom razredu povezivanje je zastupljeno u 40 %, odnosno 30 % zadataka. No, kako se udio povezivanja smanjuje od 4. prema 6. razredu, tako se udio zadataka koji traže dublje promišljanje i refleksiju (K3) povećava prema svakom višem razredu. Taj nalaz u skladu je s kognitivnim razvojem učenika.

Analiza strukture zahtjeva u zadatcima i prosječne rješivosti pokazuje da je kognitivna zahtjevnost u određenom broju zadataka u skladu s težinom zadataka. Primjerice, mnogi zadatci s najnižim kognitivnim stupnjem (K1) zaista se mogu okarakterizirati lakšima prema prosječnoj rješivosti. Ipak, kako na rješivost utječu i drugi faktori, a ne samo kognitivna kompleksnost, treba naglasiti da to ne vrijedi za sve zadatke. Primjerice, u petom razredu i najlakši i najteži zadatak pripadaju najnižem kognitivnom stupnju (K1). Razlog zašto su učenici tako loše riješili zadatak izravne primjene osnovnih znanja vjerojatno je u slaboj zastupljenosti mjerenja i geometrije prostora u trenutno važećem Nastavnom planu i programu (MZOS, 2006), kao i u školskoj praksi u kojoj mnogi učitelji preskaču taj sadržaj na kraju četvrtog razreda osnovne škole (Glasnović Gracin, 2016). Također, slabija prosječna rješivost testa u petom razredu leži u većem udjelu težih zadataka u usporedbi s ostalim dvama razredima. Razlog može biti koncept Mjerenja koji je prisutan u mnogim testnim zadatcima i koji je važan za STEM kompetenciju, ali nije naglašen kao domena u važećem Nastavnom planu i programu (MZOS, 2006) pa su učenici sa zadatcima mjerenja u 5. razredu mogli imati više problema. Nadalje, nalaz da stupanj kompleksnosti zadatka ne odgovara u potpunosti težini zadatka, u skladu je s napomenom iz Austrijskih obrazovnih standarda za matematiku (IDM, 2007) u kojima se detaljno daju primjeri zadataka s kognitivnim razinama K1, K2 i K3.

Zadatci koji su okarakterizirani najboljima u vezi s pouzdanošću testova jesu svi zadatci povezivanja (K2), i to u svim razredima. Oni povezuju različite pojmove iz STEM područja, kao i različite aktivnosti koje se očekuju od učenika u zadatku. Taj nalaz ukazuje na to da je upravo *razina povezivanja* specifična za integraciju STEM predmeta i za sastavljanje zadataka iz tog područja. Povezivanje na mnogim komponentama (pojmovi, kontekst, aktivnosti i sl.) opisano je i u literaturi kao specifičnost integracije unutar STEM-a (Honey i sur., 2014). Također, najbolji zadatci po svim razredima u vezi s pouzdanošću tog testa zahtijevaju *aktivnosti računanja* u nekoj zadanoj STEM situaciji (mjerenje temperature, salinitet mora i sl.). To ne čudi jer je računanje kao primjena matematike tradicionalno važna aktivnost unutar STEM područja (UNESCO, 2010), a ovladavanje aritmetikom jedan je od važnih prediktora za STEM postignuće (Nakakoji i Wilson, 2014). Također, iako svakim testom dominiraju zadatci višestrukog izbora, zanimljivo je primijetiti da zadatci koji najviše doprinose pouzdanosti i valjanosti testova od učenika traže da napiše odgovor na crtu, dakle, zadatci kratkih otvorenih odgovora. Dodatno, treba napomenuti da su svi zadatci koji narušavaju pouzdanost testa zadatci višestrukog izbora, što se može povezati s primjenom strategije pogađanja odgovora kod nekih učenika.

Uvid u zahtjeve zadataka koji najviše narušavaju pouzdanost testa pokazuju da se u dva od tri takva zadatka radi o aktivnosti procjene koja se traži od učenika. Iako je procjena, uz mjerenje, sastavni dio STEM aktivnosti (Honey i sur., 2014), ona nije naglašena u trenutno važećem Nastavnom planu za osnovnu školu (MZOS, 2006). Također, u navedeni se zadatcima traži aktivnost interpretiranja slike i traženja strategija, što također nije naglašeno u tekućem planu (Glasnović Gracin, 2011).

Uvid u aktivnosti koje se očekuju od učenika kako bi uspješno riješio zadatak pokazuje da u zadatcima prevladavaju zadatci računanja i interpretacije. Pritom udio zadataka s računanjem pada od četvrtog prema šestom razredu, a udio interpretacija raste. U STEM zadatcima prevladavaju zadatci s kontekstom, što je u skladu s opisom okvira integriranog STEM obrazovanja (Kelley i Knowles, 2016). U hrvatskim udžbenicima matematike, naprotiv, prevladavaju simbolički zadatci s minimalnim udjelom konteksta (Glasnović Gracin, 2011).

Na kraju možemo zaključiti da ovdje prikazani testovi mjere znanje iz STEM područja vrlo ekonomično, pouzdano i valjano. Zbog toga bi se mogli koristiti ne samo u istraživačke svrhe već bi svoje mjesto mogli naći i u praktičnom mjerenju znanja iz STEM područja u osnovnim školama u Hrvatskoj. Eventualni nedostatak testova je određeni manji broj zadataka koji ne doprinose u većoj mjeri pouzdanosti i valjanosti testova te nekoliko težinski neprimjerenih zadataka koje bi trebalo modificirati.

Prezentacija konstrukcije i validacije testova znanja u ovom radu ima i određenih ograničenja. U sklopu poglavlja o rezultatima prikazana su psihometrijska i sadržajno-strukturna obilježja svih zadataka, ali sami zadatci unutar testova nisu mogli biti prikazani. Zadatci svih triju testova nisu dostupni javno kako bi se mogli upotrijebiti u cijelom longitudinalnom istraživanju tijekom kojega bi ispitanicima zadatci trebali ostati nepoznati do trenutka provedbe ispitivanja. Osim toga, mogućnost šire primjene ovih testova nakon završetka postojeće studije također ograničava javnu dostupnost svih čestica[2]. No, s obzirom da su takvi testovi novost u edukacijskom okruženju, izrada i provjera integriranog testa za mjerenje postignuća u STEM području predstavlja velik izazov pa smatramo da metoda i prikazani rezultati mogu biti od pomoći u promoviranju STEM obrazovanja i kvalitetnijem mjerenju STEM postignuća.

Napomena

Ovaj je rad izrađen u okviru projekta "Profesionalne aspiracije prema STEM zanimanjima tijekom osnovne škole: longitudinalno istraživanje odnosa postignuća, vjerovanja o vlastitim kompetencijama i interesa za zanimanja (JOBSTEM)" koji u potpunosti financira Hrvatska zaklada za znanost. Projekt se vodi pod brojem IP-2014-09-9250.

[2] Slično je i kod drugih studija s ispitnim zadatcima koji nisu javno dostupni, poput PISA-e, kako bi se zaštitio integritet izrađenih testova (OECD, 2016).