Romanian Reports in Physics **72**, 903 (2020)

COMPARATIVE ANALYSIS OF PHYSICS ITEMS IN 6TH GRADE NATIONAL EVALUATION TESTS

G. DELIU[1,2], S. STEFAN[1,*]

[1] University of Bucharest, Faculty of Physics, P.O.Box MG – 11, Bucharest-Magurele, Romania
[2] "Emil Racovita" Highschool, Brasov, Romania, Email: *gabinan_bv@yahoo.com*
[*] Corresponding authors: *sabina_stefan@yahoo.com*

Received July 15, 2019

Abstract. The aim of the current research is to determine the internal structure of the Physics items present in the 6th grade Romanian National Evaluation Tests. The study also focuses on the efficiency of the Physics item group in the evaluation of logical-scientific thinking competence which represents the construct evaluated by these tests. The analysis of the items includes the 2016–2019 tests. The inclusion of the items at the different levels of the construct scale measured by each test is statistically supported by the Categorical Principal Components Analysis and the "Wright Map" tools. The items calibration is performed within the framework of the Item Response Theory. The results show that the items have content validity and have a high discriminatory power. The small number of Physics items in the National Evaluation Tests, as well as the fact that the internal structure of these items is almost identical in all the tests applied over the four years covered by the analysis, determined their low effectiveness in the competence's evaluation. This study proposes a useful tool for Physics teachers to interpret the results obtained by their students in the National Evaluation Tests, as it complements the image of these items, given by the National Centre for Assessment and Examination in Evaluator's Guidebook, with their strengths, but also their limitations.

Key words: items, evaluation tests, Item Response Theory, content validity, items discrimination.

1. INTRODUCTION

Any evaluation, implicitly the one done with the help of tests, aims at a measurement as accurate as possible of the examined students' competence level. Beyond the measurements' difficulty created by the abstract nature of the measured competences/constructs, there appears another issue determined by the existence of measurement errors. One can mention the existence of systematic errors due to the application of tests that are not valid, or different random errors such as: a variation of the performance level due to an indisposition, the absence of motivation, incorrect interpretation of a question, various deficiencies of items and applied tests, etc. One of the major preoccupations of constructors of tests is the limitation of measurement errors. Systematic errors are controlled through the verification of test validity, whereas random errors are controlled through the verification of test reliability [1]. The validity can be controlled even from the stage of the construction of a test by using the “Construct Map” tool [2]. However, this is not enough. The validity of a test needs to be proved statistically with the help of specific techniques like the items-respondents/students map known as the “Wright-Map” and the Principal Components Analysis. These techniques can be applied after the constructed test was administered to a group of students, as they are based on the students’ answer patterns to test items.

The reliability of a test is influenced by the difficulty and discrimination characteristics of the items [3]. These characteristics can be determined both in the Classical Test Theory (CTT) and in the Item Response Theory (IRT) [4]. This pair of theories presents strengths, but also limitations [5]. However, thanks to the mathematic rigor and the independence of item characteristics from the evaluated student group characteristics, the usage of IRT modeling is preferred for many analyses.

The usage of tests in the assessment of disciplinary competences, developed in students by studying a subject, is an ongoing practice in Romania. Besides these locally applied tests, a national evaluation of inter- and transdisciplinary competences is conducted, although these competences are reflected poorly in the national *curriculum*. The tests that make this type of assessment are the tests from the Programme for International Student Assessment (PISA) applied on a representative sample of schools, and the National Evaluation Tests on the 6th grade that are applied to all the schools in the country, since 2014.

National Evaluation Tests on the 6th grade applied in Mathematics and Natural Sciences target the determination of the development level of the transdisciplinary competence logical-scientific thinking built in students through the study of the subject group: Mathematics, Physics, and Biology [6]. Annual reports by the National Assessment and Examination Center (CNEE) [7, 8] on the results obtained by the students in these tests are complemented with studies related to the unidimensionality of items from the tests applied in 2016 and 2017 and studies in which the characteristics of items considered problematic were analyzed as they affected the equivalence of parallel tests, administered in the same year, in the mentioned time period [9, 10].

This research focuses exclusively on the Physics items from the National Evaluation Tests in Mathematics and Natural Sciences applied to 6th grade students in the years from 2016 to 2019.

The aim of the research is to determine the internal structure of these items, as well as to analyze how efficiently the Physics items group contributes to the assessment of the transdisciplinary logical-scientific thinking competence, measured by these tests, of 6th grade students.

The paper is structured as follows. In Section 2, the methodology of the conducted research, with an eye to reaching the set goal, is presented with its stages and sequences. The results of the scientific research and their interpretation are presented in Section 3. The conclusions on the importance of the research for Physics teachers, as well as teachers from the rest of the scientific fields, end the paper.

2. METHODOLOGY OF RESEARCH

2.1. USED METHODS

The conducted research with the mentioned goal was realized in multiple stages.

In the *first stage* the test validity was verified, respectively the belonging of the Physics items to the levels of the construct measured by each test. This was done in three sequences:

1. The analysis of item statements and their inclusion in one of the three levels of the logical-scientific thinking competence/construct evaluated by these tests: conceptual and procedural knowledge, intuitive and deductive logical-scientific thinking, respectively abstract logical-scientific thinking [6].

2. Statistical analysis of item load in the competence/construct evaluated by these tests with the help of the Principal Components Analysis (PCA). By applying this statistical technique, it was intended the verification of the fact that the obtained data through the quantification of the students' answers to items have a high variance in the main component supposed competence/construct evaluated by the analyzed tests, thus bringing statistical support to the validity of these tests. Because the used data had a dichotomous nature, a particular case of PCA was applied, specific to the categorical data, named Categorical Principal Components Analysis (CATPCA) [11].

3. The qualitative analysis of validity through the conducting and interpreting of Wright-Map. The items-respondents map (here items-students) is a graphic representation that allows the comparison of estimated difficulty of items with the evaluated students' competence levels. In the way that it is constructed, Wright-Map facilitates the fast, visual verification of items' difficulty and their belonging to the scale levels of the evaluated competence/construct, therefore it permits a qualitative analysis of the analyzed tests validity [12].

Starting from the fact that the reliability of a test is influenced by the item characteristics that make up the test, in the *second stage* of the research these characteristics were determined. The IRT modeling was applied as the validity analyses conducted in the previous stage support statistically the unidimensionality of the items. The logistic unidimensional model with two parameters was used. The model allowed the determination of difficulty and discrimination characteristics of items. Based on this, item characteristic curves and informational curves were raised. This made possible the identification of relevance intervals in which the items allow the determination of the students' competence levels with minimum errors.

2.2. DATA

The data based on which the conducting of statistical analyses was done, were constituted of answer patterns of students that took the National Evaluation Tests at the sixth-grade level in Mathematics and Sciences during the period 2016–2019. The structure of the 7204 answer patterns gathered throughout the four years is as follows: 883, respectively 853 answer patterns of students who took Test 1, respectively Test 2 in 2016; 1104 answer patterns associated with Test 1, respectively 1053 answer patterns associated with Test 2 in 2017; 950, respectively 949 answer patterns of students who took Test 1, respectively Test 2 in 2018, and 720, respectively 692 answer patterns collected in 2019 associated with Test 1, respectively Test 2. All subjects are 6th students and between 12 and 13 years old.

It is worth mentioning that, during the period 2016–2019 to which the analysis refers, all National Evaluation Tests in Mathematics and Natural Sciences for grade six were made up of 5 Mathematics items, 5 Physics items, 5 Biology items. Regarding the Physics items, two items had closed answers with answer choices and three items had open answers. Although, in this analysis, in order to eliminate the subjectivity factor of the teachers as evaluators in the allocation of an answer code to the items with open answers, all items were treated as dichotomous, attributing a score of 0 for an incorrect or incomplete answer, and 1 for a correct answer.

3. RESULTS AND DISCUSSIONS

3.1. VALIDITY ANALYSIS

3.1.1. Construct Map

The validity foundations of a test are laid since the test construction stage. In this regard, the graphic tool „Construct Map" is used. The making of this map involves the following stages: identifying the test evaluated competence/construct, constructing of a measurable, observable behaviors scale with different load levels in the test evaluated competence/construct and the populating the behaviors scale with items (Fig. 1).

The tests analyzed in this paper have not been constructed by the research team but by the CNEE. This has pushed the research approach in the opposite direction than the one used by the constructor of tests (Fig. 1). Thus, starting from the tasks from the items' statements, the behaviors that should be manifested by students who solve the items were identified.

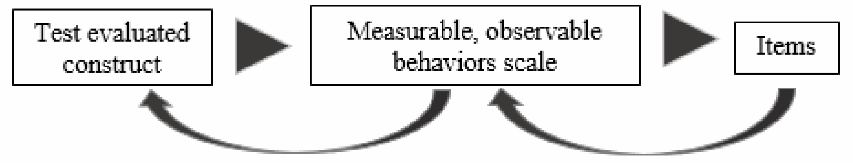

Fig. 1 – Algorithm of Construct Map creation.

These behaviors of cognitive nature were then ranked and grouped, generating a scale of three loading levels in the competence/construct *logical-scientific thinking* evaluated by these tests: *conceptual and procedural knowledge, intuitive and deductive logical-scientific thinking,* and respectively *abstract logical-scientific thinking.*

The statements of all items were analyzed, those in Mathematics and those in Physics and Biology as well. The items associated with these subjects had the same position in the tests each year of the analyzed period. The Physics items had the greatest stability. Their structures were quasi-identical from one year to another, having differences only in the context in which the task was placed. This affirmation is supported by the information presented in Table 1, in which the data and the requirement of Item 4, from the analyzed tests, are presented.

Table 1

Item 4 structure from the analyzed tests

Test/year	Item 4	
	Data	Requirement
1_2016	$d = 21$ km, $v = 5$ m/s	$\Delta t = ?$ (min)
2_2016	$d = 30$ km, $v = 5$ m/s	$\Delta t = ?$ (min)
1_2017	$d = 1.5$ km, $v = 0.5$ m/s	$\Delta t = ?$ (min)
2_2017	$d = 6$ km, $v = 0.4$ m/s	$\Delta t = ?$ (min)
1_2018	$d = 24$ km, $\Delta t = 80$ min	$v = ?$ (m/s)
2_2018	$d = 2.25$ km, $\Delta t = 75$ min	$v = ?$ (m/s)
1_2019	$d = 1.2$ m, $v = 16$ m/s	$\Delta t = ?$ (ms)
2_2019	$d = 5$ cm, $v = 10$ cm/s	$\Delta t = ?$ (ms)

This quasi-identical structure of Physics items from the applied tests in the period 2017–2019 made the framing of these items on the levels of the competence/ construct scale to be the same for every test (Fig. 2).

Items 3 and 8 were framed on the fundamental level of the logical-scientific thinking competence/construct, level named conceptual and procedural knowledge. Item 3 is an answer choice type whose solving meant the identification of the length measuring tool (Tests 1 and 2 from 2016), respectively the correlation of the length, area, volume measurement unit's symbol with the measure units of these quantities (tests from 2017, 2018, 2019). The solving of Item 8 meant the identification of some points in a graph, the reading of the temperature coordinate directly from the axes (tests from 2017, 2018, 2019), or through interpolation (tests from 2016) and the conducting of subtraction in order to find out by how much a temperature at a moment t_1 is higher/lower than another at a moment t_2. Item 12 required the representation of a simple, electric circuit, using simple graphic symbols of the circuit elements. Although, the representation with the help of symbols is a first step towards abstract thinking, linking symbols to real objects (lightbulb, wiring, battery/generator) allowed the framing of this item in the thinking area that operates with elements belonging to the real space, so the *intuitive and deductive logical-scientific thinking.*

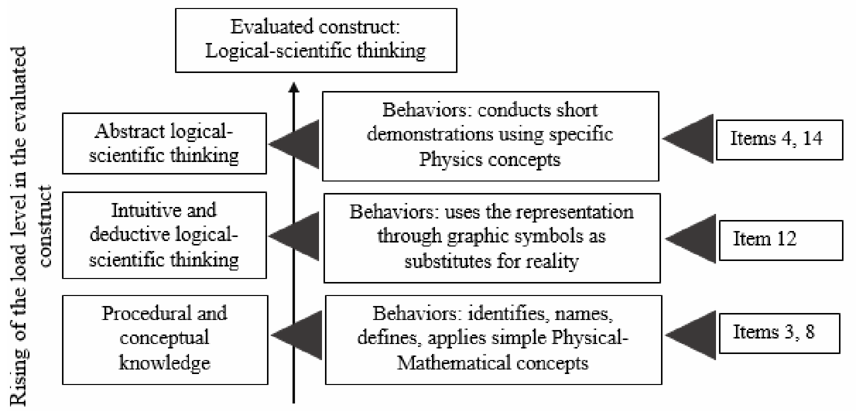
Evaluated construct:
Logical-scientific thinking
Rising of the load level in the evaluated construct
Abstract logical-scientific thinking
Behaviors: conducts short demonstrations using specific Physics concepts
Items 4, 14
Intuitive and deductive logical-scientific thinking
Behaviors: uses the representation through graphic symbols as substitutes for reality
Item 12
Procedural and conceptual knowledge
Behaviors: identifies, names, defines, applies simple Physical-Mathematical concepts
Items 3, 8

Fig. 2 – The framing of the Physics items from the National Assessment Tests in 6th grade Sciences on the levels of the evaluated construct scale.

Regarding Items 4 and 14, we can discuss the way students operate with abstract concepts like speed and density, their competence of putting together short demonstrations, starting from the given problem data in order to solve a task. These items are framed on the superior level of the construct scale, respectively the *abstract logical-scientific thinking* level.

As a wrap up to the first sequence of the first analysis stage, it can be said that Physics items from the analyzed tests have content validity, being framed in one of the three levels of the *logical-scientific thinking competence/construct* (Fig. 2). This was statistically supported by the results obtained in the following two sequences of analysis.

3.1.2. Categorical Principal Component Analysis

For all the analyzed tests it was chosen the extraction of two principal components, as it is important that the consistency of scales and the way in which the components load the test items be identified as precisely as possible.

In the case of Test 2 from 2018, for example, the convergence of the model with two, *a priori* defined, components was obtained only after a single iteration, the result being that the first principal component can explain about 21.21% of the variance of all items (Mathematics, Physics, Biology) (Eigenvalue = 3.18; Cronbach's Alpha = 0.73), while the second component had a reduced consistency (Cronbach's Alpha = 0.29), explaining 9.19% of the item variance (Eigenvalue = 1.55). Therefore, the model with two principal components can consistently explain (Cronbach's Alpha = 0.83) about 30.40% of item variance (Eigenvalue = 4.56). For Test 2 from 2019, the first components can explain 21.92% of the variance of all items (Eigenvalue = 3.07; Cronbach's Alpha = 0.72), while the second component can explain 12.08% of the item variance (Eigenvalue = 1.55; Cronbach's Alpha = 0.44). Therefore, the two-dimensional model can consistently explain (Cronbach's Alpha = 0.85) about 34.01% of the item variance (Eigenvalue = 4.76).

Table 2 specifies the proportions from the total variance with which the components contributed to the item variance belonging to the two tests used for exemplification. These values are called coordinates of item load in the two components.

Table 2

Coordinates of item load in the two components

Item	Test 2 from 2018		Test 2 from 2019	
	Component 1	Component 2	Component 1	Component 2
1	0.372	–0.371	0.334	–0.417
2	0.352	–0.306	0.425	–0.410
3	0.387	–0.458	0.434	–0.438
4	0.616	0.458	0.403	0.485
5	0.382	–0.079	0.248	0.217
6	0.559	–0.040	0.650	0.031
7	0.460	0.527	0.409	0.433
8	0.497	–0.303	0.462	–0.443
9	0.400	0.001	0.345	0.448
10	0.390	–0.128	0.485	0.036
11	0.547	0.235	0.559	0.252
12	0.449	–0.044	0.473	–0.055
13	0.416	–0.127	0.581	–0.190
14	0.501	0.441	0.473	0.413
15	0.485	–0.259	0.392	–0.193

Each item is graphically represented in the space of the two components, also called dimensions, by a vector whose orientation is given by the coordinates of item load in each component (Fig. 3).

The most important component is component 1. It can be observed, not only from Table 2 but from Fig. 3 as well, that all items, implicitly the Physics items, present high loads in this component. Regarding component 2, it plays a differentiation role between the items. There are items that have a very low load in this component, close to the value 0 (Item 12, for example). There are items that have a great load in component 2 (Items 4 and 14, for example) but also items that have negative loads (Items 3 and 8, for example), approximately equal in absolute value to the high, positive loads in this component. This leads to the separation of items into three clusters visible in all analyzed tests (Fig. 3), statistically supporting the affirmation according to which the three levels of the competence/construct measured by these tests stand behind the analyzed items. This construct is responsible for the variability of students' answers, the data, and the division of the vectors associated with the items in the three clusters.

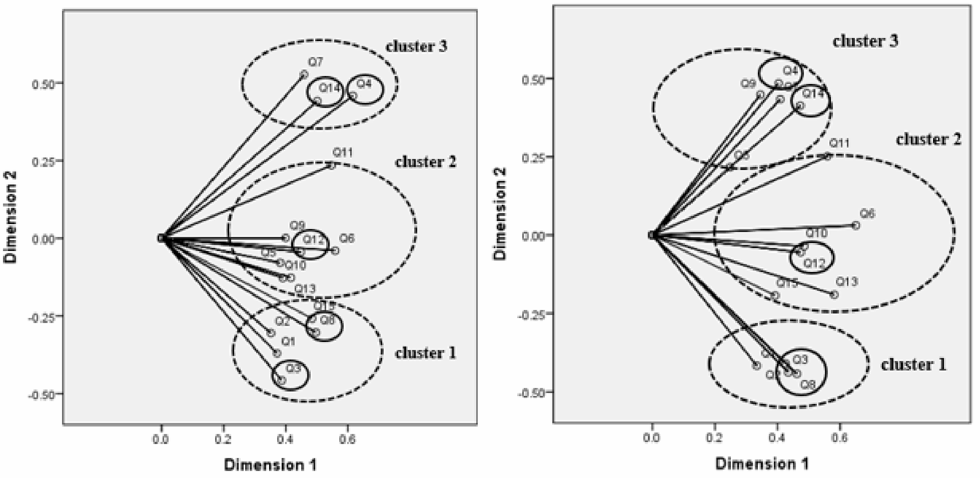

Fig. 3 – The representations of the vector associates with Test 2 items-2018 (left) and Test 2-2019 (right) in the space of the two dimensions.

Regarding the Physics items, in cluster 1 are included Items 3 and 8, placed in the anterior stage of analysis, on the conceptual and procedural knowledge level. The Item 12 part of cluster 2 was included in the intuitive and deductive logical-scientific thinking level. Items 4 and 14 are part of cluster 3, being included in the abstract logical-scientific thinking level, in the anterior stage of the analysis.

By comparison with Mathematics items, respectively Biology items from the analyzed tests, this positioning of the vectors associated with the Physics items in the three cluster is stable, being found in all of these tests.

An exception is represented by the tests from 2016 in which Items 3 and 8 migrated in cluster 2 as a consequence of that year's item requirements being slightly different than the requirements of the same items from the tests applied during the period from 2017 to 2019 (Fig. 4).

For the tests from 2016, the requirement of Item 3 was that of identifying the measurement tool of length and not identifying the length measurement unit symbol, like in the tests from the rest of the years. Item 8, in the tests from 2016, was formulated differently as the reading of values from the axes of coordinates was made through interpolation and not through direct reading from the axes of coordinates, like in the tests from 2017 to 2019.

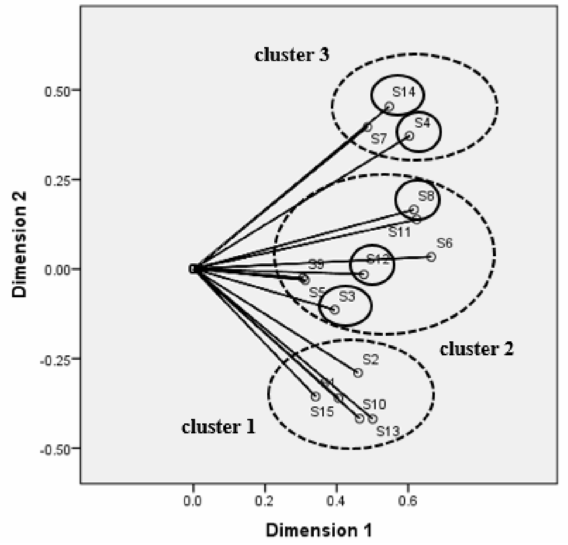

Fig. 4 – The migration of vectors associated with Items 3 and 8 in cluster 3 from Test 1 applied in 2016.

The CATPCA results exemplified on the test items from 2018 and 2019 appear in all the tests from the analyzed period, with the aforementioned exception, and statistically support the construct validity of these tests.

3.1.3. Wright Map

Previous sequences of the analysis had the role of showing that the items are, indeed, correlated with the levels of the construct scale, but only predictions were made regarding the level of item difficulty. Considering the way in which the construct scale was built, it was expected that the items associated with the inferior scale level, items conceptual and procedural knowledge, have reduced difficulties, items associated with the intuitive and deductive logical-scientific thinking level have medium difficulties, while items associated with the superior level of the construct scale, abstract logical-scientific thinking level, have high difficulties.

The way of operating of this sequence of analysis is represented in Test 2 from 2018 and 2019.

As can be seen from Fig. 5 and 6, Physics items manifest a level of difficulty corresponding to the construct scale level anticipated in the previous sequences of analysis.

Items 3 and 8 have negative Logits, within the range of (–3.5:–2) which means that the items are very easy. Item 12 has Logits close to the value 0 which means that it is an item of medium difficulty, while items 4 and 14 have positive Logits, very high, within the range of (+1.5:+3.5) which means that the items have a great difficulty.

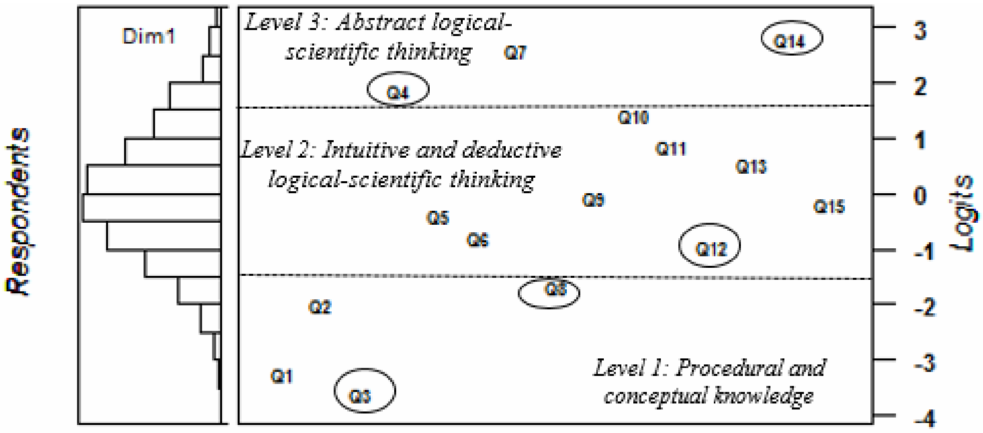

Fig. 5 – Items-student map of Test 2 applied in 2018.

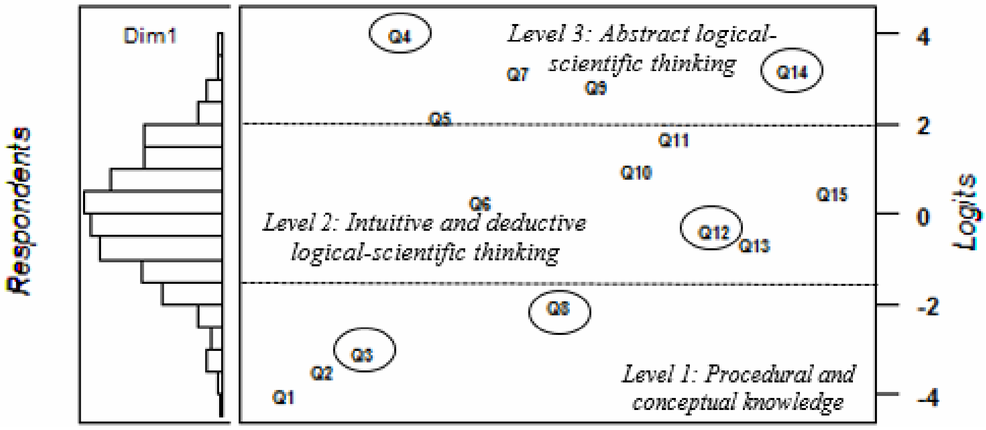

Fig. 6 – Items-student map of Test 2 applied in 2019.

In all analyzed tests the Physics items have the same behavior. An exception is Item 8 from the 2016 tests whose Logits has a slightly positive value which means that its level of difficulty is medium (Fig. 7). This behavior is explicable because the reading through interpolation of data from the axes of coordinated is, at the age level of sixth grade students, a more difficult task than the direct reading of data from the axes of coordinates.

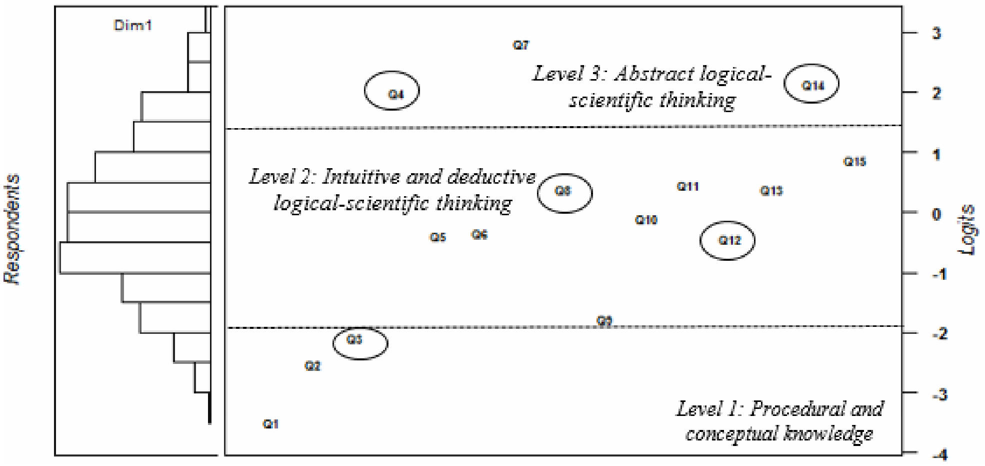

Fig. 7 – Items-student map of Test 1 applied in 2016.

3.2. ANALYSIS OF TEST RELIABILITY. IRT MODELING

The test reliability being influenced by the difficulty and discrimination levels of items, in this stage of analysis, these item characteristics were determined.

Because Physics items present big similarities from one test to another, with the aforementioned exceptions, in this section only the item characteristics from Test 2, applied in 2019, will be presented and analyzed. The results were obtained following the application of IRT modeling with two parameters.

As mentioned in the previous stages of the analysis, Item 3 is very easy. This is statistically supported by the IRT model used, which estimates a very low value of the difficulty parameter of this item, b = –2.02. As can be observed from the characteristic curve of the item (Fig. 8), students with very low competence levels (θ = –2.02) manifest a probability of answering correctly of 50%. The item has a very high discrimination level, a = 1.70. For this reason, the characteristic curve has a very accentuated slope which leads to a fast change of the probability of answering correctly to the item from a region to another of the θ competence scale.

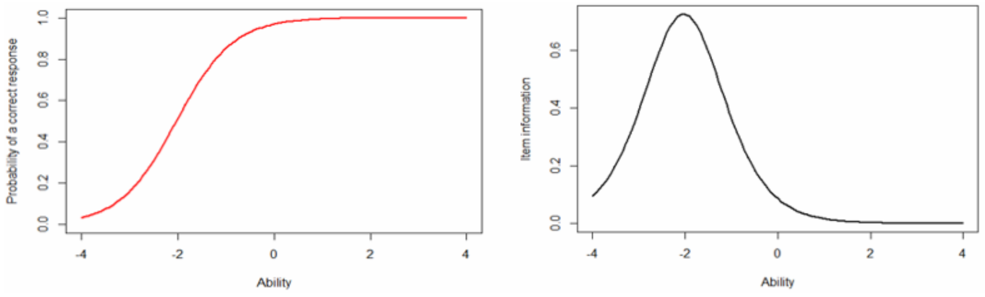

Fig. 8 – Characteristic curve (left) and information curve (right) of Item 3.

For students whose competence level θ ≥ –1, the probability of answering correctly to the item is higher than 80%. This is visible in the respondent histogram from Fig. 9 which presents that almost all students who take the test, excepting those whose competence levels are very low (Logits ≤ –2), solve the item, that is, they associate the symbol of the length measurement unit with the measurement unit of this quantity. The very high power of the item discrimination determines a very high value of the item's informational curve maximum and its narrow relevance interval, centered on the value θ = –2.02 (Fig. 8). This means that the item can make estimations of the students' competence level, with a very low level of error, but for the students with very low competences (Fig. 9).

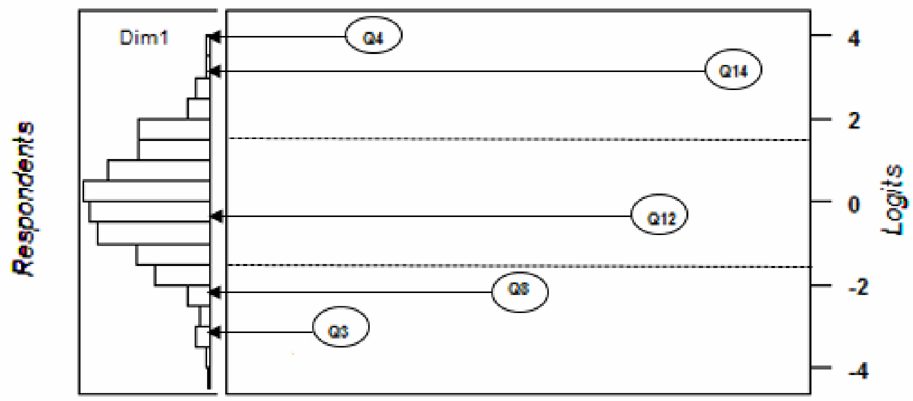

Fig. 9 – Physics items correlation-respondents. Test 2 from 2019.

Item 8 has a low level of difficulty, *b* = –1.55, which places it the area of easy items, having a high discrimination power *a* = 1.29 (Fig. 10). This means that the item allows the determination of students' competence level, with a low level of error but in its relevance interval. In the case of this item, the relevance interval is centered on the value θ = –1.55, therefore still in the area of low competence levels, but not as low as in the case of Item 3 (Fig. 9).

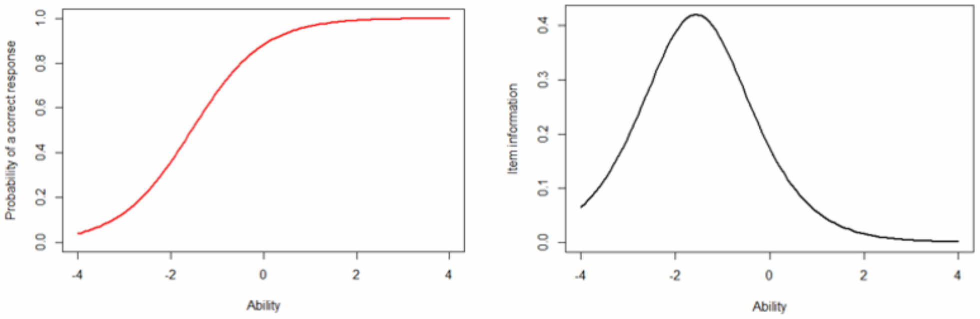

Fig. 10 – Characteristic curve (left) and information curve (right) of Item 8.

Item 12 has a medium difficulty, b = –0.32 and o moderate discrimination, a = 0.96 (Fig. 11). This item is dedicated to students with medium competence levels. On the characteristic curve of the item there can be observed that the probability of answering correctly to the item is equal to 50% for a level θ placed in the area of medium competence levels, θ = –0.32. The moderate discrimination power of the item makes the variation of the probability that the answer to item is correct not too accentuated from a region to another of the θ competence levels. Another effect is a maximum of the informational function of lower value, $I_{\max}$ = 0.25, which means that the estimation of the competence levels from the relevance interval of this item is made with a bigger error than in the case of the other analyzed items.

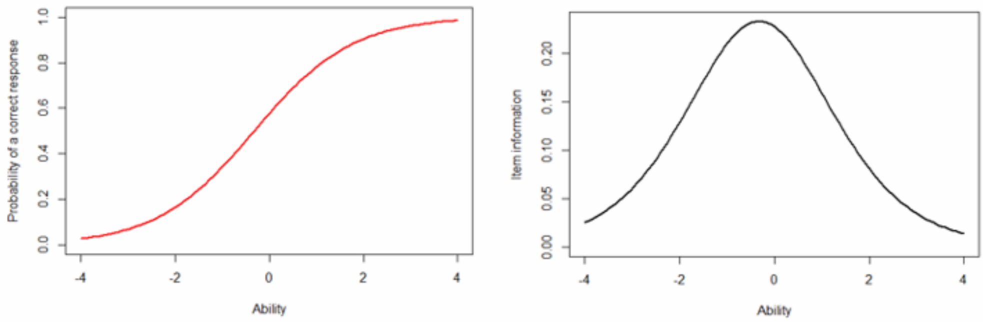

Fig. 11 – Characteristic curve (left) and information curve (right) of Item 12.

Items 4 and 14 with difficulty levels b_4 = 2.16, respectively b_{14} = 1.83 are the most difficult items of the test (Fig. 12 and Fig. 13).

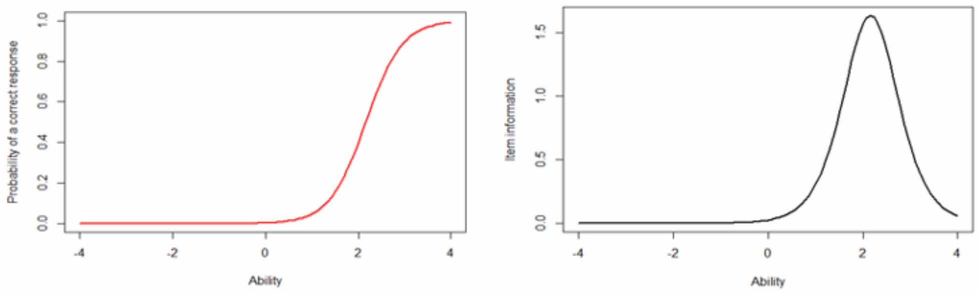

Fig. 12 – Characteristic curve (left) and information curve (right) of Item 4.

Certainly, they address students with very high competence levels, being solved by very few students as it can be observed in the respondent histogram associated with this test (Fig. 9). Their discrimination powers, a_4 = 2.55 and a_{14} = 2.29, are also very high. In Figs. 12 and 13, there can be observed that the probability of a correct answer to these items is null for the low and medium competence levels, manifesting an accentuated increase from a region to another of the θ scale, in the interval (+1.5; +4) of very high competence levels.

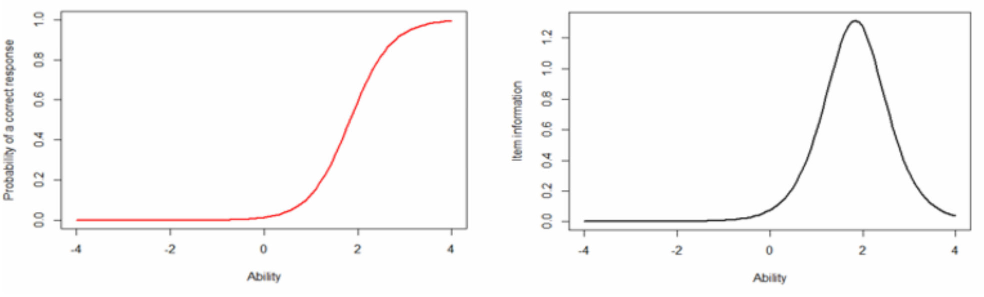

Fig. 13 – Characteristic curve (left) and information curve (right) of Item 14.

Thanks to their very high discrimination power, the maximums of the informational functions have very high values. Consequently, these items can be used for the estimation of very high competence levels, with a low error.

Taking into account the high and very high discrimination powers of these five items it can be said, about each item, that they are very efficient in measuring the competence levels in their relevance intervals.

However, analyzing how efficient the Physics item group contributes to the evaluation of transdisciplinary competence, logical-scientific thinking, measured by these tests, it is noted that two of the five analyzed items have very low difficulties being solved by almost every respondent student (Fig. 9). These items can bring relevant information to the teacher as an evaluator only about a very small number of students. This refers to the students whose competence level is smaller or equal with the very low difficulty level of these items. Other two items, respectively 4 and 14, have very high difficulty levels. Just like Items 3 and 8, these items can bring relevant information about a very small number of students. In this case, this refers to the students with very high competence levels, greater or equal to the very high difficulty level of these items. In the area of medium competences, where the biggest number of respondent students exists (Fig. 9), a single Physics item functions, Item 12.

4. CONCLUSIONS

The conducted analyses, based on data obtained through the application of National Evaluation Tests in 6th-grade Sciences in the 2016–2019 period, showed that the tests present construct validity and the Physics items are correlated with the levels of the construct scale. CATPCA and Wright Map techniques statistically support the affiliation of these items to the three-levels of the construct scale:

procedural and conceptual knowledge, intuitive and deductive logical-scientific thinking, and abstract logical-scientific thinking.

IRT modeling shows that the Physics items from the analyzed tests have a superior discrimination power and are very efficient in measuring the θ competence levels positioned strictly in the relevance interval of the items. However, the small number of Physics items from these tests and their distribution on the construct scale do not contribute efficiently to the general picture created around the development level of the logical-scientific thinking in sixth-grade students. These image is created of the level of every test by all the test items.

Moreover, the fact that the analyzed items have a very similar structure in all the tests may lead, in the case of students, to a drop in study motivation and mechanic assimilation of some algorithms of solving, and in the case of teachers, it may lead to the conducting of some learning activities in which students mainly exercise the solving of some items of this type.

The results obtained show that the study is a useful resource not only for the Physics teachers but for all the teachers from the rest of the scientific fields, respectively Mathematics and Biology. They interpret the answers obtained by students in the National Evaluation Tests and complement the information, provided by the National Centre for Assessment and Examinations, regarding these items in the Evaluator's Guidebook with their strengths but also their limitations.

REFERENCES

1. C. Opariuc-Dan, *Statistică aplicată în Ştiinţele Socio-Umane. Analiza relaţiilor şi a diferenţelor dintre variabile*, Arhip-Art, Sibiu, 2011.
2. N. Derbentseva, F. Safayeni and A. J. Canas, *Concept Map: Experiments on Dynamic Thinking*, Journal of Research in Science Teaching **44**, 448–465 (2007).
3. P. V. Engelhardt, *An Introduction to Classical Test Theory as Applied to Conceptual Multiple-choice Tests*, Getting Started in Physics Education Research **2**, 1–40 (2009).
4. L. Ding, R.J. Beichner, *Approaches to data analysis of multiple-choice questions*, Physical Review Special Topics Physics Education Research **5**, 020103 (2009).
5. R. K. Hambleton, R. W. Jones, *Comparison of Classical Theory and Item Response Theory and their applications to test development*, Educational Measurement: Issues and Practice **12**, 38–47 (1993).
6. G. Deliu, C. Miron and C. Opariuc-Dan, *Item Dimensionality Exploration by Means of Construct Map and Categorical Principal Components Analysis*, Journal Baltic of Science Education **18**, 209–225 (2019).
7. CNEE, *Raport Naţional Matematică şi Ştiinţe – Analiza rezultatelor înregistrate la Evaluarea Naţională la finalul clasei a VI-a (ENVI) 2016.* Retrieved April 25, 2019, from http://rocnee.eu/sites/default/files/2017-09/Raport_national_ENVI_2016_Mate_stiinte.pdf
8. CNEE, *Raport Naţional Matematică şi Ştiinţe – Analiza rezultatelor înregistrate la Evaluarea Naţională la finalul clasei a VI-a (ENVI) 2017.* Retrieved April 25, 2019, from http://rocnee.eu/sites/default/files/2017-12/Raport_national_ENVI_2017_Mate_stiinte.pdf

9. G. Deliu, C. Miron and C. Opariuc-Dan, *Characteristics of the Items Administered in the National Evaluation Science Testing 6th grade*, in Proceedings of the 13th International Conference on Virtual Learning, Bucharest, 2018, p.210.
10. G. Deliu, C. Miron and C. Opariuc-Dan*, Exploratory Analysis of the Equivalence of Tests, applied to National Evaluation Tests – Mathematics and Natural Sciences – 6 th grade*, in Proceedings of the 13th International Conference on Virtual Learning, Bucharest, 2018, p.217.
11. C. Opariuc-Dan, *Analiza Componentelor Principale pentru date Categoriale*, Psihologia Resurselor Umane **10**, 103–117 (2012).
12. B. Rittle-Johnson, P.G. Matthews, R. S. Taylor and K.L. McEldoon*, Assessing Knowledge of Mathematical Equivalence: A Construct Modeling Approach*, Journal of Educational Psychology **103**, 85–104 (2011).