

**Sofia University "St. Kliment Ohridski"
Faculty of Mathematics and Informatics**



Pavlin Ivanov Tsonev

**METHODS FOR ANALYSIS OF RESULTS OF
STANDARDIZED STUDENT ASSESSMENTS IN
MATHEMATICS**

**Abstract
of the PhD Thesis**

**for the award of educational and scientific degree "Doctor"
in the field of higher education 1. Pedagogical sciences,
Professional field 1.3 "Pedagogy of the education in"**

**PhD program
"Methodology of teaching mathematics and informatics"-
Methodology of teaching mathematics**

**Scientific supervisor
Prof. Kiril Bankov**

Sofia, 2025

Table of Contents

Introduction	3
CHAPTER ONE. Analysis of the National External Assessments (NEA) in mathematics after grade VII using Classical Test Theory (CTT)	7
Classical test theory – basics	7
Analysis of tasks from NEA VII class.....	7
Summary and conclusions	9
CHAPTER TWO. Analysis of NEAs in mathematics after grade VII using Item Response Theory (IRT).....	12
Basic principles and models	12
Choosing an appropriate model for the analysis of the multiple-choice tasks	16
Comparison of the results of the multiple-choice tasks for the period 2020-2023.	22
Scoring the free-response tasks	23
CHAPTER THREE. An experiment for validation of a test for NEA in mathematics after grade VII.....	29
Task statement.....	29
Analysis of the results from the experiment	33
Comparison of traditional and experimental scoring methods on different indicators	36
Implications.....	38
Conclusion.....	39
Contributions of the dissertation	41
Directions for future development.....	42
References	43
Publications of the author related to the topic of the dissertation.....	45
Declaration for originality	46
Acknowledgments.....	47

Introduction

Relevance of the topic

Student assessment is an essential element in the process of education. Assessment in Bulgaria is formative and summative; the latter can be temporary, annual, or upon completion of a given stage of education. The last type of assessment was introduced in the normative documents of the Ministry of Education and Science (MES) in 2007 and is called the National External Assessment (NEA) when students complete grades 4, 7 and 10, and the State Matriculation Examination (SME) when they finish 12th grade. The Bulgarian language and literature (BEL) is compulsory subject for all NEAs and SMEs. Mathematics is compulsory for NEAs, while it is optional for SMEs. For the period from 2007 to 2024, NEA after grade VII contained a different number of tasks of the three types - multiple choice questions (MCQ), short constructed response (SCR) and extended constructed response (ECR). Table 1 shows the number of assignments by type for each year. As can be seen, only the first three years the tasks were only MCQ. In the following years, they were divided into two modules, with the first module mainly containing the MCQ tasks, and the second - constructed response (CR) tasks.

Years	Total tasks	First		Second module	
		MCQ	SCR	SCR	ECR
2007-2009	50	50	There are no modules		
2010-2011	30	25	-	3	2
2012-2018	24	16	4	2	2
2019	25	17	3	2	3
2020-2023	23	18	2	-	3
2024	23	20	-	-	3

Table 1. Number of tasks by type for the period 2006-2024.

Every year, after the completion of the external evaluations, rankings and analyzes of the achieved results are made. Often these analyzes are based on average scores. These scores are frequently compared between the main subjects BEL and Mathematics. Additionally, on the basis of these scores, conclusions are made about the level of success among the different administrative regions of the country, between the settlements in them, between the different schools in these settlements. The results of NEA after grade VII are used for acceptance in profile-oriented and profession-oriented classes in secondary high schools. To what extent this is expedient can be traced in the habilitation work (Bankov, 2012). This issue is not the main topic of the present dissertation, although it will be partly discussed.

In some cases, the analysis of the results of external evaluations is conducted by using scientific test theories. Most often, the well-known Classical Test Theory (CTT) is applied for this purpose - for example (Danchev, Plamen; Bankov, Kiril; Stoimenova, Vessela; Atanassov, Dimitar, 2013), (Djalev, 2013), , (Djalev, 2014), etc. Thorndike 's book from 1904 (Thorndike, 1904) is considered to be the beginning of the CTT worldwide. In Bulgaria, this theory is discussed by several authors - (Bizhkov, 1996), (Stoyanova, 1996), (Bankov, 2012).

The use of CTT, despite its excellent qualities, has certain limitations. At the end of the 20th century, a new theory based on probabilistic methods appeared in the scientific literature. It is known as Item Response Theory (IRT). In 1960 G. Rasch described it for the first time. In Bulgaria IRT was described in 1996 by F. Stoyanova as part of the book (Bizhkov, 1996). Probabilistic modeling is also observed in large-scale international studies, such as TIMSS (Trends in Mathematics and Science Study) and PISA (Programme for International Students Assessment). More about the advantages of IRT, about its principles and its development over time, as well and some of the weaknesses of CTT is discussed in the second chapter of the dissertation.

From the very beginning of the introduction of NEA at the end of grade VII, a few questions arose, such as:

- How can we interpret students' results?
- Do the students in grade VII meet the state educational requirements according to these results?
- Do the tasks that are given to NEAs meet any preliminary specifications?
- Is the data obtained from NEAs used to draw conclusions for improving mathematics education?
- Is it acceptable to use the NEAs results for entrance in profile-oriented and profession-oriented secondary high schools?

In the search for answers to some of these questions, the present dissertation analyzes the results of the students of NEAs after the 7th grade, conducted in the period 2020-2023 in four administrative regions of Bulgaria - Sofia-city, Plovdiv, Pleven and Razgrad. A comparison is made between the conclusions obtained according to the two test theories. Their advantages and disadvantages are discussed. The conclusions drawn were used to conduct a pedagogical

experiment with students graduating grade VII in 2024, which is described in the third chapter. It explores the hypothesis that probabilistic modeling can be used to more precise investigation of the qualities of the test items and can introduce simpler and more objective assessment of students. The methodology described in the book (Ivanov, 2006) was used in the development of this work.

The questions raised above, along with other issues related to NEA are regularly discussed at the Spring Conferences of the Union of Bulgarian Mathematicians, in which the author of the dissertation regularly takes part. One of the first such discussions, after the start of the large-scale national external evaluations, took place in 2010 (Bankov K., Vitanov T., 2010). The test materials for NEA are assigned by the Ministry of Education and Science to the Center for Assessment in Preschool and School Education (CAPSE, 2024), and teachers from various schools from all over the country are involved in their preparation. Since 2010, the author of this work has regularly participated as a member of the expert commissions for compiling the test tasks in mathematics at this center. Also, from the outset of the implementation of the NEA, he has been a member of the regional commissions for their inspection and evaluation.

Purpose and tasks of the study

The main goal of the dissertation is to apply different methods for analysis of the results of standardized assessments of students in mathematics (NEA) and to evaluate the strengths and weaknesses of each method. Additionally, to make recommendations for their use based on the analysis provided in the study.

To achieve this goal, the following *main tasks are completed in the dissertation*:

- 1) An analysis of the scientific and theoretical aspects related to the dissertation's subject is conducted, outlining the problem area of the subject under consideration;
- 2) An analysis is made of the study documentation related to mathematics education in the junior high school stage for the acquisition of basic education;
- 3) An analysis of the test tasks by the NEA for the period 2020-2023 is made using the CTT;
- 4) An analysis of the test tasks from the NEA for the period 2020-2023 is made using IRT;
- 5) The results obtained by the different methods are compared ;

- 6) A pedagogical experiment is being carried out with students who are about to finish grade VII in 2024, incorporating the conclusions drawn from the research;
- 7) Suggestions for calculating the students' competition score are made.

The object of the study is the tests for NEA and the methods used for analyzing their results. The theoretical analysis was carried out in Chapters 1 and 2 on the NEAs held in the period 2020-2023. The empirical research was carried out with students in grade VII from various schools in several towns in Bulgaria in the period April-June 2024 and is detailed in Chapter 3.

The subject of the research is the implementation of different models for analysis of the results of NEA, identifying appropriate scales for assessment of student achievements, depending on the results obtained from these models.

Research methods

The methods applied in the dissertation are document research (including test tasks and their classification); methods for analyzing test results using CTT and IRT; discussions of the analysis obtained from the comparison of these methods; statistical processing of the data from the conducted NEAs; graphical representation of the results.

Research hypothesis

The correct use of CTT combined with modern methods of analysis of test results (IRT) gives a more objective, more accurate and reliable picture of the mathematics achievements of the studied population.

Structure of the dissertation

The dissertation contains an introduction, three chapters, conclusion, references and appendices. The main part of the dissertation consists of 170 pages, which include 105 figures and 104 tables. There are 8 appendices. The references contain 35 titles. Figures in the dissertation are marked in the middle below the figure itself, using the notation **Fig. C.#.**, where **C** is the number of the chapter in which the given figure is, and **#** is the number of the figure itself in this chapter. The tables are marked under the table itself with the notation **Table C.#.**, where **C** is the chapter number and **#** is the number of the corresponding table in the chapter. References is in APA style.

CHAPTER ONE. Analysis of the National External Assessments (NEA) in mathematics after grade VII using Classical Test Theory (CTT)

Classical test theory – basics

The first section of this chapter gives a brief background on CTT, along with some historical notes related to it and an explanation of the concepts used in this theory. After that, the test specification of NEAs in mathematics after class VII is presented. It contains two main directions - substantive and informative. The content direction is described by the different topics of the learning content, and the cognitive direction - by the mental activities that are necessary to solve the tasks. The authors of the test tasks should compose to encompass a variety of content and cognitive areas.

The content area items measured were determined based on the mathematics curriculum included in the fifth-, sixth-, and seventh-grade curricula. These areas are divided into 27 topics, which are grouped into four areas: 1) Numbers, algebra; 2) 2D and 3D shapes. Measurement; 3) Elements of Probability and Statistics; 4) Logical knowledge. Modeling.

In addition to the content area, each test task is also classified according to another parameter - cognitive area (Bankov, 2012). Cognitive domains describe the types of mental activity that are required to solve the test items. According to Bloom's taxonomy, these cognitive levels are six: Knowledge, Comprehension, Application, Analysis, Synthesis and Evaluation.

Analysis of tasks from NEA VII class

In the next four sections NEA in mathematics after grade VII for 2020, 2021, 2022 and 2023, respectively, are discussed in detail. Data on the results of students from 4 regions of Bulgaria were used: Sofia-city, Plovdiv, Pleven and Razgrad. For each of these four years:

- the main statistical characteristics of MCQ tasks obtained according to CTT are given;
- the distribution of students according to their raw scores is graphically displayed, along with several options for transforming these scores into grades according to the six-point marking system generally accepted in Bulgaria;
- by Pearson's χ^2 test, and the hypothesis H_0 was rejected, i.e. it cannot be claimed that the general set of students' scores is normally distributed;
- MCQ tasks are ordered according to coefficients of difficulty and discriminative power, and tables characterizing the distractors of each task are presented;

- Based on this analysis, some specific tasks from each of the studied NEAs were examined, showing some unsatisfactory characteristics according to the above mentioned indicators. Attention was paid to the content area of each task; the cognitive level it covers; the reasons behind the poor performance indicators that were identified; options for improving these indicators were given; the performance of students from different administrative areas was analyzed;
- the reliability and validity of the tests are discussed;

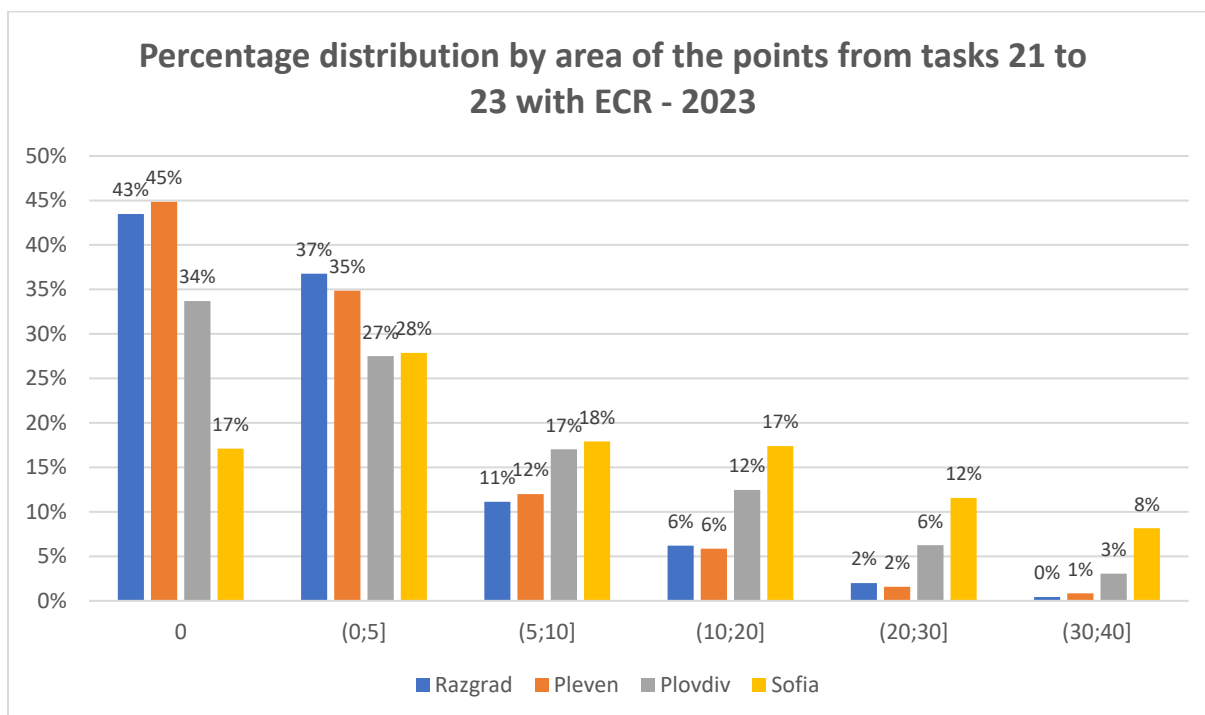


Fig. 1.32. Percentage distribution by area of the points from tasks 21 to 23 with ECR - 2023

- it was concluded that the CR tasks do not significantly contribute to increasing the scores of the students obtained from the MCQ tasks. This applies to almost all areas, except for Sofia-city. One of the reasons for this is that in smaller towns the competition for admission to high schools is reduced to a minimum, while in the city of Sofia there is still significant competition. This conclusion is supported by charts such as the one for 2023 in Figure 1.32, in which the horizontal axis represents the points received by the ECR and the vertical axis indicates the percentage of students who received those points.

Summary and conclusions

In the last section of this chapter, some summaries are made on the MCQ tasks from the 4 studied years. From figure 1.33 it can be seen that the percentage of students who received points in the interval [0;5] changes consistently across the four years – it grows at almost the same rates as the points increase. Then, in the interval [6;16] there are constant values, and at the end of the interval where the largest points are there is a different characteristic for two pairs of years. For 2020 and 2023, the percentage decreases, while for 2021 and 2022, it increases. It cannot be overlooked that 2020 was the year in which **distance learning in an electronic environment** (DLEE) began, and in 2023 students had already finished this mode of learning. For the middle two years – DLEE was the main mode of education.

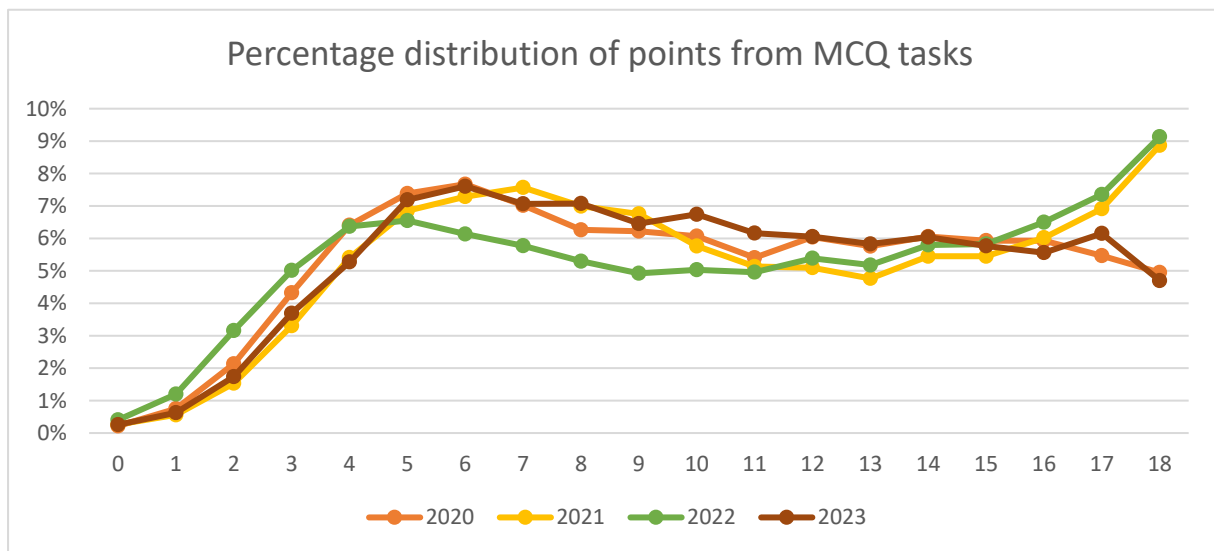


Fig. 1.33. Percentage distribution of points from 2020 to 2023 of MCQ tasks

In this chapter, two options for transformation of points into a six-point system evaluation are proposed - linear and cumulative. For both variants, higher values of poor grades are observed for 2022, which also applies to excellent grades. One of the reasons for this could be the two-year study in an electronic environment, in which the students of this graduating class found themselves. For less prepared students, this training creates problems, while for excellent students it creates an opportunity for better preparation. This phenomenon can be explained by the fact that in online learning, self-training of students is in the foreground, which for the best of them creates opportunities for greater individual progress. For less prepared students, such independent work creates more difficulties. This poor performance of the students from 2022 is also confirmed by the high percentage (68%) of those who received between 0 and 5 points on the CR tasks. This percentage is even higher (82%) for smaller districts, such as Pleven and Razgrad.

In the same section, a comparative table of the MCQ tasks from all four years by content and cognitive area is made. It should be noted that the distribution of these tasks in the two areas was made by the author of the dissertation, and not by the committee that compiled them, because this information is not generally available. Notably, 11 out of all 27 content area topics are either not represented in the MCQ assignments or have only 1 assignment across the four years. On the other hand, a large number of them are present in the CR tasks and it can be assumed that the National External Assessments cover all the content that is outlined in their specifications. According to this table, almost all MCQ tasks are at the cognitive level of Knowledge or Comprehension, with the exception of an Application level task. This is expected for multiple choice tasks.

	Cronbach 's α	SEM
2020	0.8604	1.7476
2021	0.8677	1.7466
2022	0.8867	1.7184
2023	0.8485	1.7762

Table 1.66. Reliability coefficient of test MCQ tasks

At the end of the chapter, based on the coefficient of reliability of the MCQ tasks from the four NEAs in table 1.66, it is concluded that such a test is appropriate:

- to study the mathematics performance of VII students in general. Inferences that are made with a high degree of reliability apply to the entire population;
- to compare these results by different large groups such as: gender, region, schools.

For all years $\alpha < 0.95$ and the standard error (SEM) indicates that the 95 percent confidence interval is 5–6 points long. This gives reason to say that MCQ tasks are not suitable for making decisions related to the ranking of students for admission after VIIth grade. To this end, CR tasks should increase this reliability. Let's see if this is the case for students in general and in individual districts in the country.

Figure 1.36 shows the percentage of students who scored in the interval [0;5]. On average for the country, about 50% of students from the first two years and about 60% from the following two years either did not receive points, or they received a minimum number of ECR tasks. If we look at these data by districts, for the smaller districts of Razgrad and Pleven, these percentages are higher by another 20%. For the Plovdiv region, they are around the average for the country, and for the Sofia-city region, they are lower than the average. A reason for the difference in performance is that for smaller districts, competition for admission after class VII

is either very low or non-existent – while for larger cities, especially Sofia, this competition is very high. Therefore, for a large percentage of students in the country, ECR tasks do not bring added value to MCQ tasks.

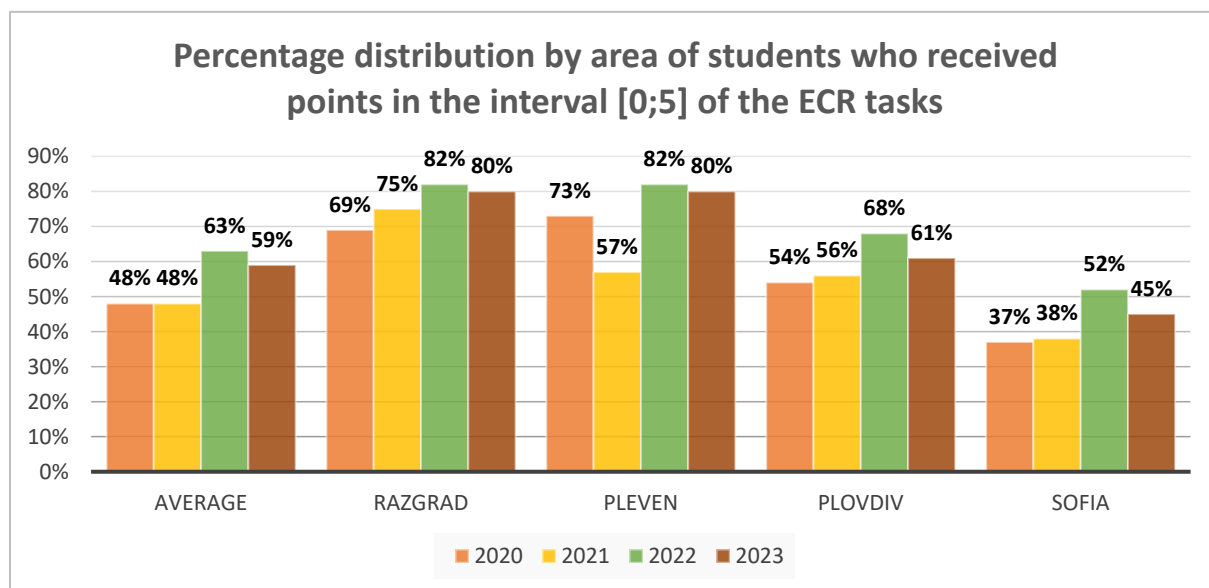


Fig. 1.36. Percentage distribution by area of students who received points in the interval [0;5] of the ECR tasks

Another observation that can be made from Figure 1.36 is the sharp rise in the percentage of students with low results in the second module of NEA for 2022 (from 48% to 63% on average). Considering that the distribution according to the difficulty factor of the MCQ tasks in the four years is not very different, the problem with the display of these percentages can be traced to e-learning or to some large differences in the nature of the CR tasks. It is clear that distance learning in an electronic environment is not suitable in teacher-student communication. This is especially important when the student must describe a solution to a task and the teacher has to comment on that solution, point out weaknesses in it, give supporting questions, etc. The decrease in this percentage in 2023, when students were back in classrooms confirms this conclusion. It will be interesting to follow this trend in the coming years.

The ECR tasks during the studied years have no significant differences. They are of the three main types:

- an algebraic problem that requires solving a linear equation and linear inequality by making several identity transformations.
- a modeling task. In 2020 and 2022 it was of the "movement" type, in 2023 it was of the "work" type. In 2021 an equation had to be modeled and various conclusions drawn needed to be from it in a given chart.

- a geometric task in which, in addition to basic geometric knowledge, additional analytical reasoning must be applied.

CHAPTER TWO. Analysis of NEAs in mathematics after grade VII using Item Response Theory (IRT)

Basic principles and models

The first section of this chapter describes the basic principles of test theory based on probabilistic modeling - Item Response Theory.

Classical test theory is widely used in Bulgaria due to its transparency and accessibility, but it has certain limitations. Here are two of them:

- test scores depend on the tasks selected in the test (their difficulty);
- determining the difficulty of the tasks depends on the specific examinees;

A new theory based on probabilistic modeling has been developing in the scientific literature for decades. It is known as *Item Response Theory (IRT)*. In probabilistic modeling theory different models have been developed. One of the most widely used models is described by G. Rasch. His paper (Rasch, 1980) was first published in 1960. This model is also called *the one-parameter logistic model* or *Rasch-model* and can be seen as a special case of A. Birnbaum's model described in 1968 in (Birnbaum, 1968) and subsequently developed by Hambleton and Swaminathan in the book (Hambleton, R., Swaminathan, H., Rogers, H., 1991). It is described in Bulgarian literature by several authors, for example by F. Stoyanova in the book (Bizhkov, 1996), (Bankov, 2002), (Djalev, 2014) and (Djalev, 2013). The use of probabilistic models is popular in national studies of some countries and especially in international ones (TIMSS, PISA, PIRLS, etc.). In recent years there have been publications in Bulgaria as well - (Bankov, 2012), (Alashka, 2016) and (Alashka, 2017), (Tsonev, 2023-a), (Tsonev, 2024-a) and (Tsonev, 2024-b).

Some of the benefits of IRT are:

- the measure of parameter measurement (especially in the Rasch model) enables the analysis of qualitative data using quantitative methods;
- the assessment of the task difficulty does not depend on the choice of the examinees from whom the data was obtained;

- the assessment of the examinee's ability level does not depend on the set of tasks used for testing them;
- the presence of incomplete data is not a critical flaw in the test study.

Three of the basic assumptions in IRT are (Kim, 2007):

- 1) there are hidden (also called *latent*) personality parameters that are inaccessible to direct observation;
- 2) there are available (called *indicators*) observable parameters that are related to the latent parameters. The importance of these indicators can be used to judge the importance of the latent parameters;
- 3) the estimation of the latent parameter must be one-dimensional. In particular, this means that the test must measure knowledge in only one clearly defined subject area. If the one-dimensionality condition is not met, it is necessary to redo the test, separating the tasks that violate its homogeneity.

The question of the one-dimensionality of the 18 test MCQ tasks from the National External Assessments, considered in the present dissertation, is demonstrated on the NEA in Mathematics held in 2020. Factor analysis and SPSS software were used. It turns out that there are two components in the test. The first component is essentially "strong", i.e. the test is primarily one-dimensional (assessing one construct). However, three tasks assess a different aspect related to that construct, which we will refer to as a "second construct" (second dimension). The greatest contribution to this second construct is given by tasks numbered 7, 15 and 18. These are examined in more detail in the dissertation and reasons are provided to explain their deviation from the general behavior of the test.

In contrast to the scoring of test tasks accepted in Bulgaria, in order for them to be examined with IRT, they must be scored according to the type of tasks like this:

- in the MCQ tasks - with 1 point for a correctly specified answer or with 0 points for a wrong answer, unspecified answer or for more than one answer specified;
- in the CR tasks– with an integer number of points in the interval $[0, n]$. This number n is equal to the number of *important steps* in the solution of the problem. The value of n is determined by the authors of the problem and indicates how important these steps are in solving it. This idea is described by (Bankov, 2023). Moreover, if the task is of True-False type, then $n = 1$. Most often $n = 2$ or $n = 3$.

For brevity, below a task in which $n = 1$ we will call *the 0-1 task* if $n = 2$ we will call *the 0-1-2 task* and so on.

To explain the Rasch model, let's consider tasks of the type 0-1. The main idea of the model is that the ability of a student to solve a given task can be measured by an abstract quantity called *ability*, which is denoted by θ . On the other hand, each task can be assigned a so-called *characteristic function*, whose graph is called *a characteristic curve*. This function describes the probability that the task will be correctly solved depending on the same abstract quantity θ , i.e. this function expresses the probability that a student with a given ability will solve the task correctly. The value of θ , at which the probability of correctly solving the task is $1/2$ is called *difficulty* of the task and is denoted by b .

The assumption made in the Rasch model is that θ a student's ability to solve a task and the difficulty b of that task can be placed on the same scale and measured with the same units, called *logits*. This scale is also called *ability scale*. Student success is a function of difference $\theta - b$. If this difference is positive and large, it means that there is a high probability that the student will solve the corresponding task correctly. If this difference is negative and large in modulus, then there is a high probability that the student will not correctly solve the given task.

If the task is 0-1, its characteristic function has the form

$$P(\theta) = \frac{e^{\theta-b}}{1 + e^{\theta-b}},$$

where b is a parameter called task *difficulty*.

The graph of such a curve at $b = 1$ is shown in Figure 2.1. The coordinates of the point of the characteristic curve with the ordinate $P = 0.5$ (in this case point A) show that a student with the corresponding ability (in this case $\theta = 1$) has a 50% probability of getting a correct response to this task. This value of θ (in the case $b = 1$) is the difficulty of the task. Usually, the difficulty of the tasks is in interval $(-2; 2)$. If b it is close to -2 , then the task is easy, and if it is close to 2 , then the task is difficult.

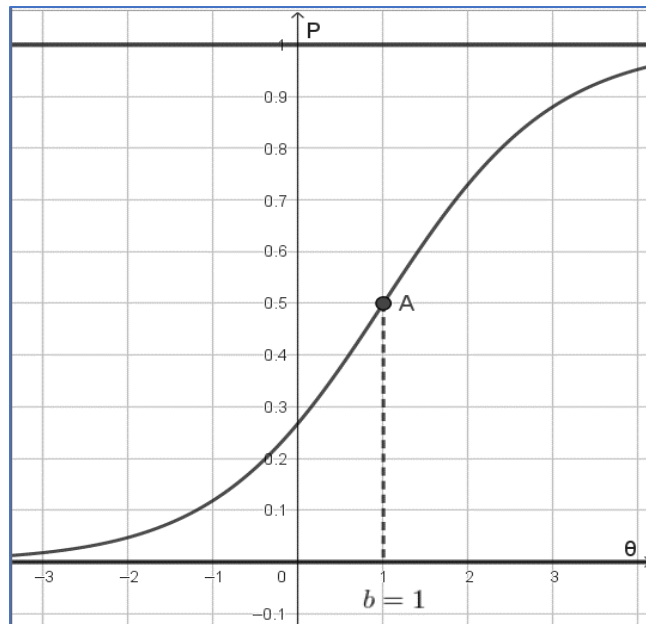


Fig. 2.1. Characteristic curve of the Rasch model of task 0-1

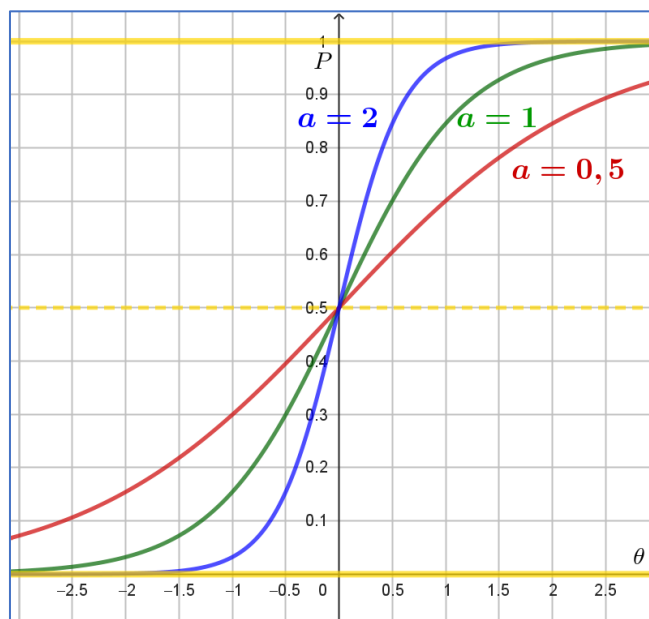


Fig. 2.4. Examples of a 2PL model with three parameter options a

The model proposed by Rasch was subsequently developed further in the book (Birnbaum, 1968) a *two-parameter* and a *three-parameter* model were introduced. In the first model, in addition to the parameter b , a new parameter is introduced a . The characteristic function of each task has the form:

$$P(\theta) = \frac{e^{1.7a(\theta-b)}}{1 + e^{1.7a(\theta-b)'}}$$

where the parameter is the a *discrimination* task sense. It sets the slope of the corresponding characteristic curve at the difficulty point b of the task. The "steeper" this curve is, the better

the task differentiates strong from weak students who have abilities close to b . Most often, the interval in which it accepts values a is $(0;2)$. We will refer to this model as the *2PL model*. In a sense, it is a generalization of the one-parameter Rasch model, which we will also call *the 1PL model*.

Figure 2.4 shows the characteristic curves of tasks with difficulty $b = 0$, which have discrimination $a = 0.5$, $a = 1$ and $a = 2$.

In the case of the three-parameter model (*3PL-model*), a parameter is also introduced c , which is the item *pseudo-guessing* parameter. The characteristic function of this model is:

$$P(\theta) = c + (1 - c) \frac{e^{1.7a(\theta-b)}}{1 + e^{1.7a(\theta-b)}}$$

Graphically, the parameter value c depicts a horizontal asymptote $P = c$ of the characteristic curve at $\theta \rightarrow -\infty$.

Figure 2.5 shows the graphs of characteristic curves of three tasks with difficulty $b = 1$, discrimination $a = 1$ and different inference parameters $c = 0$, $c = 0.25$ и $c = 0.5$.

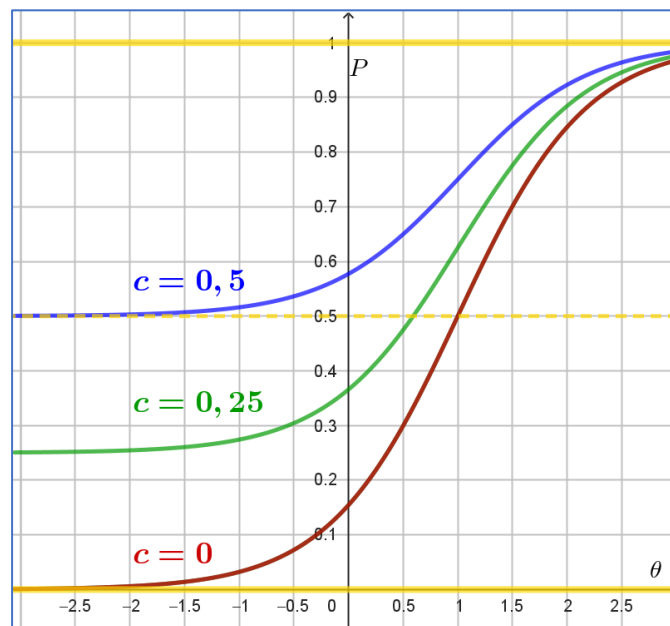


Fig. 2.5. Examples of a 3PL model with three parameter options c

Choosing an appropriate model for the analysis of the multiple-choice tasks

Using the Rasch model to study a given test assumes that the tasks included in it are of equal discrimination and the probability of getting the correct response is minimal. Through this model, it can be determined which tasks do not agree well enough with the empirical data that is obtained after the test is conducted. Its premise is that it is not the model that must match

the empirical data, but the data that must match the model (Kim, 2007). It is generally recommended that problem tasks not be included in subsequent tests until they have been reworked.

For the 2- and 3-parameter models, the approach is different. Through them, a model is sought that best corresponds to the studied test.

In the second section of this chapter, for each MCQ task and for each NEA of the four years examined using the jMetrik application (Meyer, 2024) the obtained characteristic functions were analyzed according to the three IRT models.

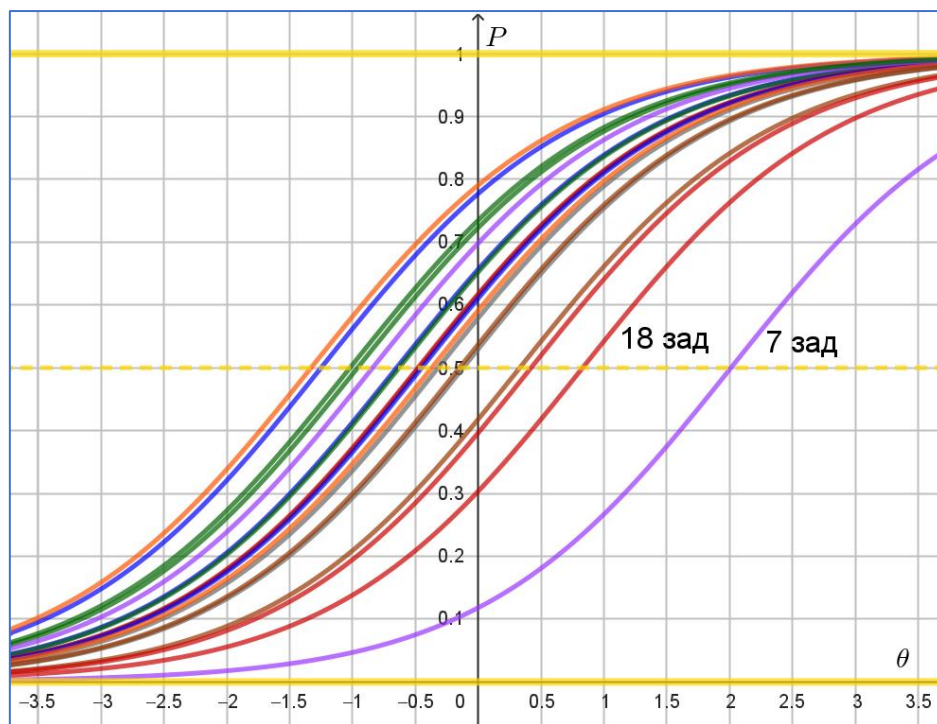


Fig. 2.6. Characteristic curves of tasks from 2020 - 1 PL

In the first point of this section, the characteristics of the MCQ tasks from the NEA conducted in 2020 are examined. For the present study, the results of the NEA conducted at the end of grade VII in mathematics with 18,836 students from the investigated four administrative regions of Bulgaria were analyzed. These regions are Sofia-city, Plovdiv, Pleven, and Razgrad, with the exam having taken place in 2020. The data for the multiple-choice question tasks, which are the first 18 in this test, were processed. The jMetrik software used measures the ability scale so that the ability distribution of all students has a mean of 0 and a standard deviation of 1. Graphs were drawn with the GeoGebra application (GeoGebra, 2024).

Figure 2.6 shows the characteristic curves of all 18 MCQ tasks according to the one-parameter Rasch model (1PL). It shows that these tasks cover the range $(-3.5; 3.5)$ of ability

level θ . The curves of some tasks overlap, for example 1 and 3; 9, 11 and 12; 15 and 16. In this regard, some of them, for example 3, 11, 12 and 16, can be removed from the test without significantly impairing its measurement qualities. At the same time, missing tasks are observed in some difficulty intervals (we look at the level $P = 0.5$). Through further processing, one could look for tasks with difficulty in the intervals $(-0.12; 0.33)$, $(0.42; 0.83)$ and $(0.83; 2.01)$. Obviously, the most difficult tasks are 7 and 18, which is also confirmed by the Classical Test Theory. These tasks are discussed in detail in the first chapter of the dissertation.

Table 2.3 gives the parameter values according to the three parameterization models. Impressive are two tasks numbered 15 and 9, which have high values of the guessing parameter, respectively $c = 0.37$ and $c = 0.29$. These tasks did not excel on any of the Classical Test Theory metrics and are discussed in this chapter.

No	1PL	2PL		3PL		
	b- param	a- param	b- param	a- param	b- param	c- param
1	-0.65	2.66	-0.44	2.96	-0.23	0.10
2	0.33	1.24	0.25	2.38	0.68	0.19
3	-0.63	2.01	-0.45	2.58	-0.13	0.16
4	0.42	1.12	0.35	2.72	0.81	0.22
5	-1.25	2.29	-0.80	2.25	-0.72	0.04
6	-1.33	1.89	-0.89	2.03	-0.68	0.14
7	2.01	1.12	1.74	2.94	1.51	0.07
8	-0.32	1.21	-0.29	1.98	0.33	0.26
9	-0.47	1.44	-0.39	2.73	0.25	0.29
10	-0.96	2.10	-0.64	2.93	-0.22	0.24
11	-0.48	1.80	-0.36	2.98	0.12	0.24
12	-0.45	2.07	-0.34	2.41	-0.12	0.10
13	-0.37	1.77	-0.29	2.93	0.15	0.22
14	-0.84	1.97	-0.58	2.21	-0.35	0.12
15	-0.12	0.80	-0.13	2.97	0.85	0.37
16	-0.15	1.70	-0.15	2.66	0.24	0.18
17	-1.02	1.27	-0.84	1.60	-0.29	0.26
18	0.83	0.57	1.22	2.57	1.48	0.25

Table 2.3. Parameter values according to the three models of the 2020 tasks.

The next stage of the test analysis examines how well the three parameterization models (1PL, 2PL and 3PL) fit the empirical data. We will apply a graphical method that is described on pages 66-67 in the book (Hambelton, R., Swaminathan, H., Rogers, H., 1991). For each of the three models, each of the students receives a value of the parameter θ (ability), which is a number in the interval $(-3;3)$. This interval is divided into 12 equal parts. If we denote the middle

of each subinterval by θ_i ($i = 1, 2, \dots, 12$), then, of all students with ability θ in this subinterval, we assume that they have the average value θ_i . Then, for each task and for each subinterval, the ratio of the number of students r_i who answered the task correctly to the number of all students is found m_i in this subinterval. This is how the magnitude is obtained

$$p(\theta_i) = \frac{r_i}{m_i},$$

which is the experimental value of the probability of giving a correct response to this task to an able student θ_i .

In the following task, we will demonstrate the idea of examining the correspondence of theoretical curves and empirical data. Figure 2.7 shows the condition of Task 1. The characteristic curves for this task according to the three models are displayed in Figure 2.8, labeled as 1PL, 2PL, and 3PL. The empirical data for each model are represented by the points A, B, C, ..., which have coordinates respectively $A(\theta_1, p(\theta_1))$, $B(\theta_2, p(\theta_2))$, $C(\theta_3, p(\theta_3))$, ... These graphs are also called *residual graphs*. A model has a good fit to the data if the residual plots are close to the corresponding task characteristic curves.

1. The value of the expression $200 - 20 \cdot \left(-2 \frac{1}{2}\right)$ is:

A) -450 B) -72 C) 208 +D) 250

Fig. 2.7. 2020 Assignment 1 Condition

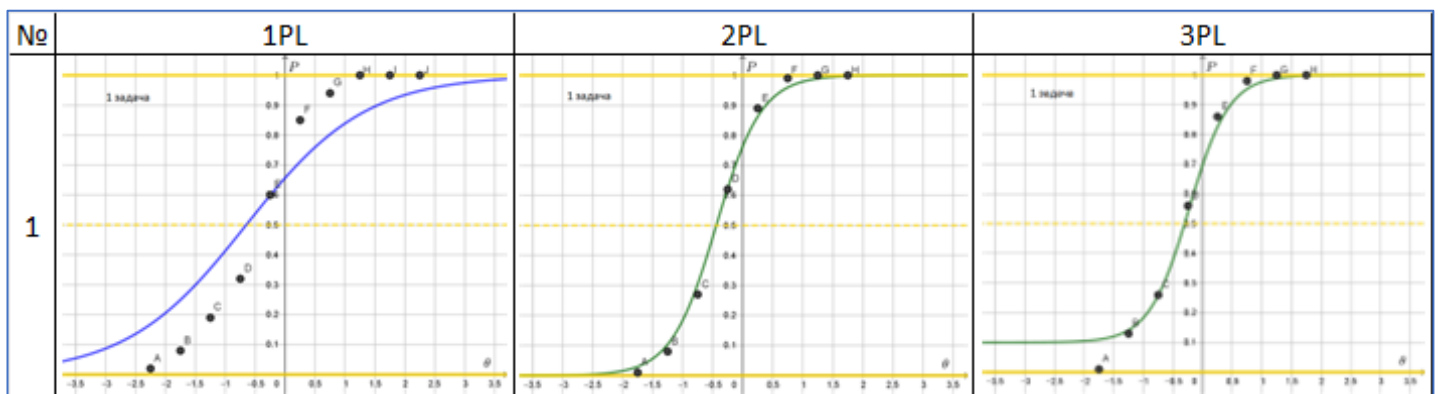


Fig. 2.8. Plot of the residuals and characteristic curve of task 1 of 2020.

It can be said about this task that:

- it is of optimal difficulty – the graphs cross the horizontal line $P = 0.5$ at a point with the abscissa around -0.5 ;

- the graphs of the two- and three-parameter model are straighter, i.e. it has better discriminative power. For both models, the value of the discrimination parameter is $a = 2.66$ (for 2PL) and $a = 2.96$ (for 3PL);
- points corresponding to students with abilities less than -2.5 and greater than are missing 2.5 for the 1PL model, and for the other two models the range of abilities is even narrower $(-2; 2)$;
- the 1PL model has large differences between theoretical and empirical results;
- the 3PL model gives a high guessing rate that does not correlate well with low-ability students (point A is far from the curve);
- the 2PL model agrees to the greatest extent with the empirical data;
- the task is from the content area "Operations with rational numbers" and such tasks are present in all NEAs after grade VII. Similar to this task is task 1 of 2021 in Figure 1.19, and the analysis of both tasks from a CTT perspective is identical.

According to this scheme, four more tasks with numbers 7, 9, 15 and 18 were considered.

Next, the coefficients of difficulty and discriminative power of the MCQ tasks obtained from the two theories are compared. The correlation coefficient was found to be sufficiently high in both cases 0.98981 for the difficulty coefficient and 0.919809 for the discrimination coefficient.

At the end of researching this point in section 1, the so-called *item-person-map* diagram is constructed. For this purpose, it is established that the difficulty of the tasks (b - parameter) and the abilities of the students (θ) can be located on the same scale, referred to as *the Ability scale*. Again, we use the results of the 2PL-model. In Figure 2.15, the highlighted row, denoted by θ/b , is the interval $(-2; 2)$ from the ability scale. The interval is divided into subintervals of length 0.2. Below this row are written the numbers of the 18 tasks, with each task placed where its difficulty parameter falls b . Above the marked line is presented the percentage of students with ability θ in the corresponding interval of length 0.2, as:

- each “#” symbol corresponds to 1% of students;
- at each symbol “.” corresponds to less than 1% of students.

- the most difficult tasks are numbered 18 and 7 and they are located relatively far from the main group of tasks.

According to this procedure, in the remaining three points of this section, an analysis of 10 more tasks from the next three investigated years was made. From the corresponding graphs:

- conclusions were drawn regarding the homogeneity of the tasks in each test;
- by comparing the empirical and theoretical data in the three parameterization models of the results of the four investigated years, it was concluded that the two-parameter model has a certain advantage;
- conclusions were made regarding the coefficients of difficulty and discrimination of the tasks according to the CTT and the two-parameter IRT model, and a high degree of agreement between the two theories was established.;
- via *item-person-map* diagram shows that the difficulty of the tasks correlates well with the students' abilities.

It was found that for most of the MCQ tasks from the four NEAs, there is a correspondence between the coefficients of difficulty and of discrimination according to the two theories - CTT and IRT. However some tasks show certain differences. Additionally, probabilistic modeling found tasks with a high guessing ratio. For these tasks, reasons are sought and identified that allow students to select the correct answer, without making the necessary reasoning or calculations that lead to it. Thus, the main hypothesis in the dissertation work is confirmed: the combination of the two theories in the analysis of the test results gives a more objective, more accurate and reliable picture.

Comparison of the results of the multiple-choice tasks for the period 2020-2023.

In the third section of this chapter, the results of all MCQ tasks from the studied period 2020-2023 are compared. The methodology of this comparison, described in the first point of this section, is based on the idea of using several common tasks (*anchor* tasks) administered in two different tests to two different groups of students and comparing the results between the groups. Due to the lack of common tasks in the years, pairs of tasks with similar characteristics in terms of difficulty and discrimination were identified.

In the second and third points of the section, the procedure for calculating the results of the years 2020 and 2022 relative to 2021 is applied. As a conclusion of this section, the tendency of declining the achievements of the students throughout the studied period has been

established. A finding that somewhat diverges from the official statistics. The reason for this discrepancy is that it only compares the average values of the results of NEAs in mathematics.

The applied methodology, in addition to this comparison, can also be used to create banks of multiple-choice tasks that are "sized" on a single scale and grouped by a difficulty factor, which is obtained objectively, and not according to the opinion of the authors of the test tasks. Often the author has one idea of the difficulty, while the objective reality points to another.

Scoring the free-response tasks

In the fourth section, the CR tasks are discussed. Scoring free-response tasks is often associated with certain difficulties. Two of them are:

- How to determine the maximum number of points for a task?
- How to distribute these points on the solution of the task?

In an effort to make the assessment as objective as possible, points are often awarded in increments of not only whole numbers, but also multiples of 0.5, even 0.25.

Is this correct??

In the present study, one perspective on this question is given, which is based on probabilistic modeling methods. The tasks given in 2023, as well as the results of all 21,060 seventh grade students from the four administrative districts of Sofia-city, Plovdiv, Pleven and Razgrad, were examined. The author's report on this topic at the 53rd Spring Conference of the Union of Mathematicians in Bulgaria was used (Tsonev, 2024-b).

Rasch model, if a task is scored with 0, 1 and 2 points, three characteristic curves are drawn, which are described by the following three functions:

$$P(X = 0|\theta) = \frac{1}{1 + e^{\theta-b-t_1} + e^{2\theta-2b-t_1-t_2}}$$

$$P(X = 1|\theta) = \frac{e^{\theta-b-t_1}}{1 + e^{\theta-b-t_1} + e^{2\theta-2b-t_1-t_2}}$$

$$P(X = 2|\theta) = \frac{e^{2\theta-2b-t_1-t_2}}{1 + e^{\theta-b-t_1} + e^{2\theta-2b-t_1-t_2}}$$

The first of them shows the probability of getting a wrong answer ($X = 0$), the second shows the probability of getting a partially correct response ($X = 1$), and the third – the probability of getting a correct response ($X = 2$).

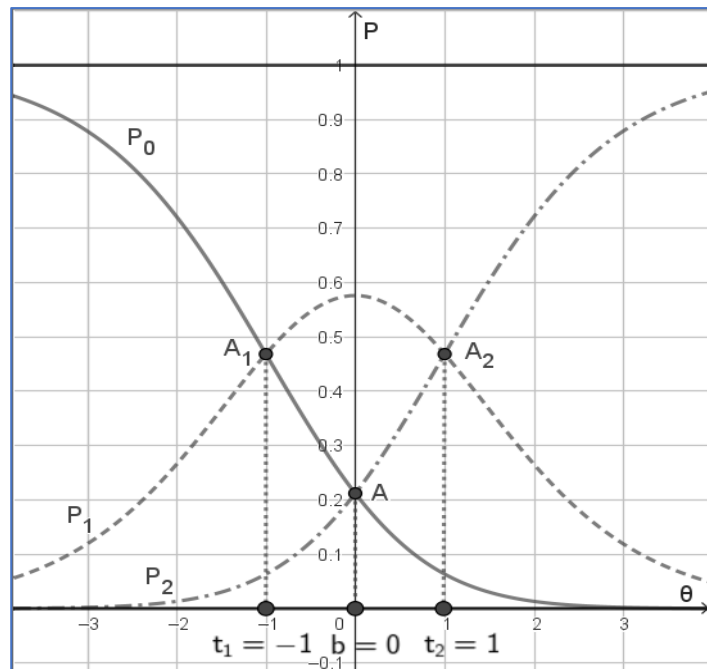


Fig. 2.2. Characteristic curves of Rasch model of task 0-1-2

Figure 2.2 shows an example where the plots P_0 , P_1 and P_2 correspond to the probability of getting 0, 1 and 2 points. The difficulty of such a task is obtained from the abscissa of the intersection point A of P_0 and P_2 (in the case $b = 0$). Of interest are the abscissas of t_1 and t_2 the intersection points A_1 and A_2 respectively of P_0 , P_1 and P_1 , P_2 (in the case $t_1 = -1$ of and $t_2 = 1$). The conclusion that can be drawn from such a task is that students with abilities less than $t_1 = -1$ are most likely to receive 0 points, those with abilities between $t_1 = -1$ and $t_2 = 1$ are most likely to receive 1 point, and those with abilities $t_2 = 1$ above likely to get 2 points.

Problem tasks are those in which $t_1 > t_2$. Such an example is given in figure 2.3. As can be seen, the graph P_1 is located "below" the graphs of P_0 and P_2 . This means that there is no range of ability where the student is most likely to score 1 point. So, such a task is more suitable to be of the type 0-1.

Tasks with more intermediate points are similarly described, with more bell-type functions.

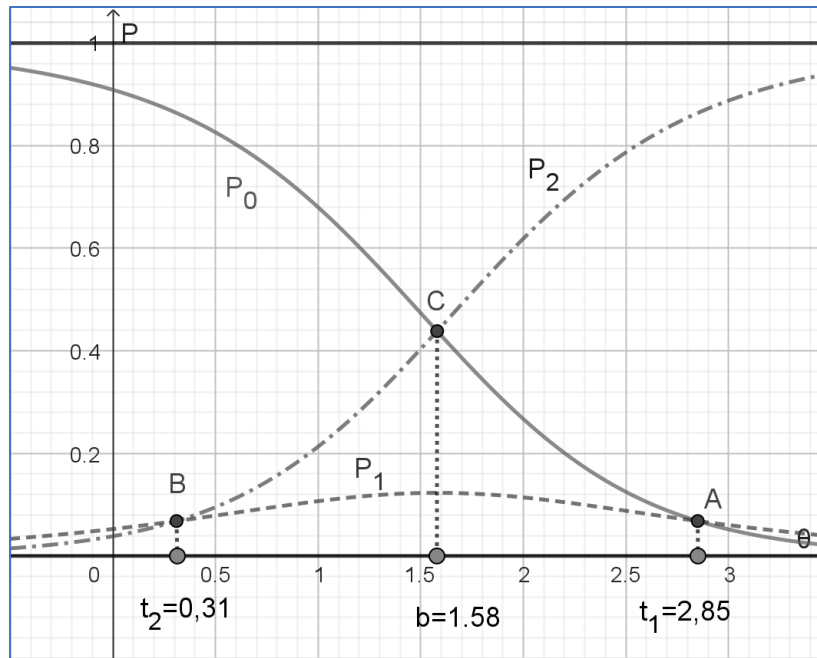


Fig. 2.3. Characteristic curves of Rasch model of task 0-1-2 at $t_1 > t_2$

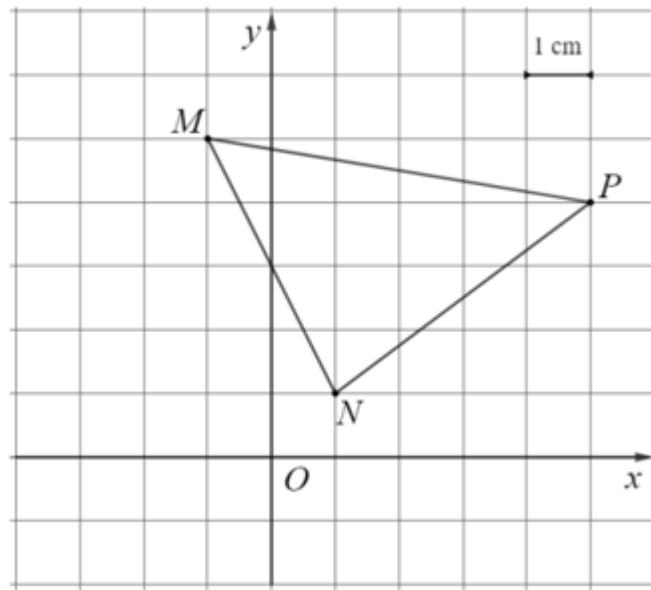
To implement the probabilistic method, the scoring of the 2023 free-response tasks, which was used by the NEA assessors, needs to be changed. Two types of such changes can be distinguished.

The first type is for tasks numbered 19 A), 19 B) and 20 C), which are evaluated respectively with 4, 3 and 4 points for a correct response and 0 points for a wrong or unspecified answer, the scoring is changed to 1 point for correct response and 0 otherwise. So, they become so-called 0-1 tasks.

The second type is in the remaining tasks where the scoring is very fragmented. Such assessment is not applicable to research with IRT methods. For this reason, it is necessary that the number of points in their evaluation be a whole number and be relatively small. In (Bankov, 2023) is analyzed the way of scoring the tasks that assess the student's partial progress in solving them. The concept of an important step is introduced. The purpose of important steps is to distinguish meaningful moments in solving a given task. The author of the task determines these moments.

An example of an extended free-response task that is divided into many important steps is Problem 20 A), the condition of which is in Figure 2.52.

20. In the Oxy coordinate system with a unit segment of 1 cm, the points M, N and P are given.



A) Write the coordinates of the points M, N and P.

B) Write the coordinates of a point Q symmetrical to the point N with respect to the coordinate origin.

Fig. 2.52. In the vocabulary of task 20 A) and B) from 2023.

Points	0	0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.5	2.75	3
%	34.7%	0%	1.6%	0%	3.9%	0%	2.6%	0%	4.1%	0%	6.4%	0%	46.6%

Table 2.19. Evaluation of task 20 A) according to the MES

Grading for this assignment is in increments of 0.25 units. Table 2.19 gives the points and the percentage of students who received these points. As can be seen, the percentage of students who received some points is close to 0%. This gives rise to distinguishing 3 important steps that allow a change in the scoring system. These steps are:

- 1 or 2 correct coordinates are obtained
- 3 or 4 correct coordinates are obtained
- 5 or 6 correct coordinates are obtained

For each step completed, 1 point is obtained and thus the task becomes 0-1-2-3 and it can be investigated with IRT.

The Rasch model of all tasks from 2023 has been applied to the so-accepted scoring and the standard 0-1 scoring of the MCQ tasks.

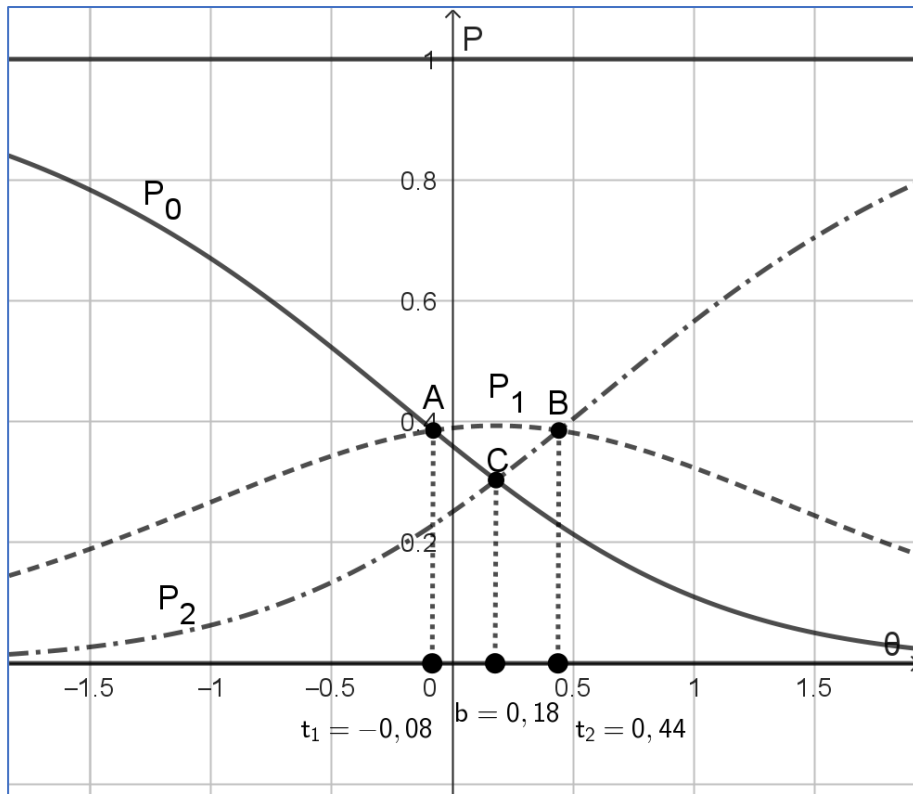


Fig. 2.53. Characteristic curves of task 20 B)

In some of the tasks, a good distribution of the characteristic curves is obtained (this is obtained at $t_1 < t_2$). Such is task 20 B), whose characteristic curves are shown in figure 2.53. It can be said that students with abilities less than -0.08 are most likely to receive 0 points, those with abilities between -0.08 and 0.44 are most likely to receive 1 point, and those with abilities above 0.44 are most likely to receive 2 points. This task has an optimal difficulty ($b = 0.18$) and shows good characteristics regarding the stratification of students across the three assessment levels.

In the task 20 A) discussed above, it turns out that even rounded scoring is superfluous. According to its characteristic curves in Figure 2.56, it can be said that students are most likely to have 0 or 3 points, and intermediate points are unlikely. In other words, if a student can find the coordinates of one point, he can find the coordinates of all three points of the task condition. This task is an example of a situation where the points of intersection of the curves P_0 , P_1 , P_2 and P_3 are arranged in reverse order, i.e. $t_1 > t_2 > t_3$. The conclusion that can be drawn is that it should be scored with 0-1 points. In addition, the difficulty of the task is $b = -0.85$, which

means that it is one of the easy tasks in the test and breaking it up in scoring into 13 parts (according to the instructions of the MES in table 2.19) is unjustified.

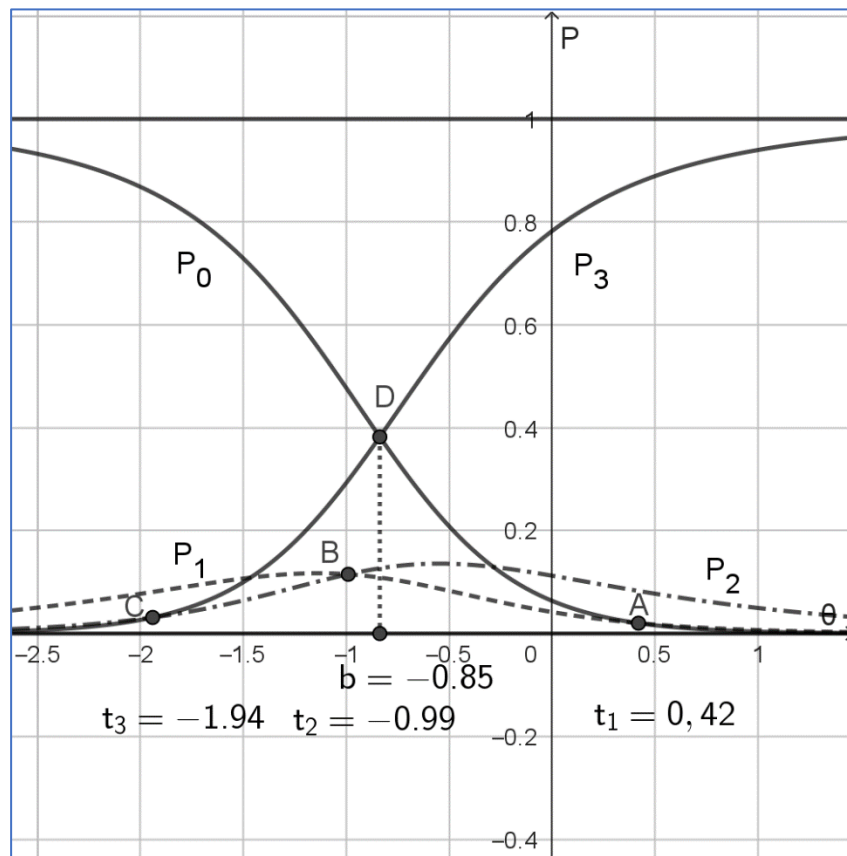


Fig. 2.56. Characteristic curves of task 20 A)

All free-response tasks were examined in a similar manner. Considering the graphs of their characteristic curves, the number of important steps in their solution has been corrected. It turns out that at most 2 important steps stand out. On this basis, the sample scoring guide shown in Table 2.21 can be applied.

This idea of evaluating free-response tasks could be implemented in future NEAs. So the examiners will give a whole number of points for each important step (no more than 2-3 per task). These steps are predetermined by the test authors. After that, a team of experts will analyze the obtained results and according to the obtained difficulty coefficients, coefficients will be given for each task to obtain the required amount of a finite number of points (in this case 100). Modern technologies allow data processing to take place within a few hours. The work of the experts should also not take more than 1 working day.

In the next chapter, an experiment is described in which assessment is done in the traditional way and in the way described in this section of the dissertation, on a sample topic for NEA after grade VII in mathematics.

No	Completed activity	Suggestion
19A	received 3	1
19B	received 10	1
20A	obtained correct coordinates of at least 1 point	1
20B	<ul style="list-style-type: none"> • one correct Q coordinate is obtained • a second correct Q coordinate is obtained 	<ul style="list-style-type: none"> • 1 • 1
20C	received $S = 11$	1
21A	found an expansion and solved the resulting equation	1
21B	roots 3 and 9 obtained	1
21C	<ul style="list-style-type: none"> • wrong inequality, but correct conclusions about the resulting roots in the interval obtained by the inequality • correct conclusions that 3, 6, and 9 are solutions to the inequality 	<ul style="list-style-type: none"> • 1 • 1
22A	<ul style="list-style-type: none"> • the unknowns x and $x - 2$ are introduced • received number of workers 8 and 6 	<ul style="list-style-type: none"> • 1 • 1
22B	found 5 and 4 days	1
23A	<ul style="list-style-type: none"> • plausible drawing and found angle 40° or 100° • found all three angles 40°, 40° and 100° 	<ul style="list-style-type: none"> • 1 • 1
23B	<ul style="list-style-type: none"> • established an equality that proves a rhombus • the rhombus proof is complete 	<ul style="list-style-type: none"> • 1 • 1
23C	proved $AL = BQ$	1
23D	proved $AP > PQ$	1

Table 2.21. Sample scoring of the 2023 free-response tasks.

CHAPTER THREE. An experiment for validation of a test for NEA in mathematics after grade VII

Task statement

The first section presents the conceptual model of **the pedagogical experiment**, which involved 487 students from different types of schools across 9 regional cities of the country. These students were about to graduate from grade VII in 2024. In connection with the questions raised in the previous two chapters, the following **goal** was formulated: to apply and compare the methods of CTT and IRT for the analysis of test tasks for the formation of the test score of the students at the NEA in mathematics after grade VII. To achieve this goal, **research tasks** have been set:

- to prepare a research toolkit including appropriate measurement tools;
- to explore the possibilities of cooperation with teachers in the country;
- to create an organization to collect the necessary resources that will be used in the implementation of the experiment;
- to collect, analyze, and summarize the empirical material to draw relevant conclusions about the students' scoring methods.

A hypothesis of the experiment was formulated:

The use of probabilistic modeling offers a more precise evaluation of the characteristics of the test tasks and helps to introduce a simpler and more objective assessment of students in NEAs in mathematics after grade VII.

This experiment went through three main stages - **preliminary, main** and **final**.

In **the preliminary** stage, an average of 5 MCQ tasks were selected from the four NEAs held in the period 2020-2023. The criteria for selecting these tasks were as follows:

- represent different content areas;
- cover as much of the V-VII curriculum as possible;
- vary in difficulty level – easy, optimal and difficult;
- demonstrate relatively poor characteristics when tested with CTT and IRT.

For tasks with CR, due to the changes in the type of NEA, only those with extended CR were considered. The following aspects were studied for them:

- what types of such tasks have been included in past NEAs;
- how the tasks are worded (e.g. with a few bullet points, without drawings, etc.);
- how their assessment was carried out by the assessment committees;
- what characteristics they exhibited in the IRT analysis, as described in the second chapter.

A study conducted to explore the possibilities of colleagues in the country to conduct a trial NEA with their students. In addition to this, it had to be determined how much of the study material would be covered by mid-May to ensure that unlearned study content was not included.

The main stage of the experiment went through the following steps:

- 20 MCQ tasks and 3 ECR tasks were created based on the new model of the NEA for 2024, considering the conclusions obtained in the preliminary stage. Some of the MCQ assignments were almost identical to those given over the years, while other assignments were modified to address issues identified in earlier chapters of the dissertation. The goal was to track how much these changes affected the measurement qualities of the problem tasks – whether they led to improvements or not. Some modifications involved rewording of the tasks, while others changes in the selection of distractors;
- The trial NEA was divided into two parts:
 - Part one, in which students work for 75 minutes, contains tasks with only 4 possible answers, only one of which is correct;
 - The second part, in which the working time is 90 minutes, consists of 3 ECR tasks, to which the students must write the complete solutions with the necessary justifications and drawings;
- MCQ tasks, according to the 2024 model, are scored with 2, 3 or 4 points, and their total must amount to 65 points. Several options exist for distributing the number of tasks to meet this requirement. A variant with 3 easy tasks (worth 2 points each), 9 optimally difficult tasks (3 points each) and 8 difficult tasks (4 points each) was chosen for the experiment;
- The difficulty of the created MCQ tasks was adjusted according to the difficulty of previous years' tasks that served as a model. In this sense, it can be assumed that they are approved;
- There are three tasks with extended CR, one is algebraic, the other is applied, and the third is geometric. Each of them includes two or three sub-items;
- Two versions of the assessment guide have been prepared for the ECR tasks:
 - *traditional variant* - which is accepted for verification in the current NEAs. With it, each stage of the solution to the task is evaluated with 1 point, it also allows for partial scoring, with points awarded as fractions - 0.5 or 0.25. Thus, the points for each of these tasks is 12 points for tasks #21 and #23 and 11 points for task #22. Their total is 35 points;
 - *experimental variant* - each sub-item of each task is treated as a separate task and is evaluated with an integer number of points based on the important steps in its

solution, as described in the second chapter. This approach allows for the application of the IRT methods. Thus, 0-1 task is #23 A), 0-1-2 tasks are #21 A), #21 B), #21 C), #22 A), #22 B) and #23 C), and 0-1-2- 3 task is #23 B).

- A form is provided for each student to mark the correct response to each MCQ task, and the examiner must enter the points obtained for correct responses to these tasks. The ECR tasks are described on separate sheets, and the obtained points for each sub-item are entered by the examiner on the same form;
- The trial exam was held with students from different schools in the country - primary, secondary and mathematics high schools in the period April 27 - June 1, 2024;

Type of school	City	Number	Total
MG	Pleven	32	80
MG	Montana	25	
MG	Razgrad	23	
SU	Pleven	32	90
SU	Sofia	19	
SU	Plovdiv	39	
OU	Blagoevgrad	38	317
OU	Pleven	117	
OU	Dobrich	32	
OU	Montana	67	
OU	Noisy	63	
Total number		487	

Table 3.1. Number of students participating in the experiment

- The number of students according to the school in which they study is given in table 3.1 . As can be seen, most students are from secondary or primary schools, and there are also from mathematics high schools, where it is assumed that the preparation in mathematics is at a higher level than other schools. Students from 9 administrative regions of Bulgaria took part;
- Traditional grading was done by the students' teachers, after which the students' works were sent to the dissertation author for experimental grading. For this purpose, these works were either scanned or sent by courier.

The final stage of the study includes:

- Analysis of the tasks after the implementation and evaluation of the trial NEA;
- Comparison of traditional and experimental scoring methods by various indicators;

- Proposals for the method of scoring in future NEAs.

Analysis of the results from the experiment

In the second section, an analysis of the obtained results is made. In the first point of this section, the MCQ tasks by means of CTT are analyzed. It has been established that the tasks of the trial NEA correspond to the main characteristics of those of the NEA conducted in the four years studied.

Points	No	Difficulty	rpbis	Interpretation
2	13	0.86	0.36	many easy
3	4	0.81	0.41	easy
3	10	0.81	0.26	easy
3	12	0.79	0.46	easy
2	5	0.75	0.35	easy
3	1	0.7	0.4	optimal
2	2	0.69	0.38	optimal
3	17	0.68	0.41	optimal
4	14	0.65	0.4	optimal
3	15	0.64	0.37	optimal
3	18	0.63	0.55	optimal
3	11	0.6	0.39	optimal
4	16	0.6	0.45	optimal
4	20	0.56	0.16	optimal
3	3	0.55	0.4	optimal
4	7	0.55	0.45	optimal
4	9	0.52	0.54	optimal
4	8	0.44	0.46	optimal
4	6	0.4	0.25	difficult
4	19	0.39	0.42	difficult

Table 3.3. Ordinance of MCQ tasks from 2024 according to their difficulty

No	Difficulty	rpbis	Interpretation
20	0.56	0.16	low
6	0.4	0.25	average
10	0.81	0.26	average
5	0.75	0.35	good
13	0.86	0.36	good
15	0.64	0.37	good
2	0.69	0.38	good
11	0.6	0.39	good
1	0.7	0.4	good
3	0.55	0.4	good
14	0.65	0.4	good
17	0.68	0.41	many good
4	0.81	0.41	many good
19	0.39	0.42	many good
16	0.6	0.45	many good
7	0.55	0.45	many good
12	0.79	0.46	many good
8	0.44	0.46	many good
9	0.52	0.54	many good
18	0.63	0.55	many good

Table 3.4. Regulation of MCQ tasks from 2024 according to their discrimination

Tables 3.3 and 3.4 give the numbers of the tasks, arranged according to the difficulty coefficients and discriminative power (rpbis), respectively. As evident, there is a variety in the difficulty of the tasks. There is one very easy task, 4 easy, 13 optimal and 2 more difficult, which is advantageous if the NEA is intended as a normative test. When composing the test tasks, the easier tasks were evaluated with 2 points, the difficult ones with 3 points, and the most difficult ones with 4 points.

According to Table 3.3, there is a slight discrepancy among the easier tasks, but it can be said that it is not significant. The discrimination of the tasks, except for #20, is at an acceptable level. This assignment was given without being tested. An experimental component was embedded in its condition – the correct response is an elementary sum of the two numbers

given in the prompt, making it easy to guess. We will see later that this is confirmed with IRT methods. Tasks 6 and 10, although tested, showed lower discriminative power.

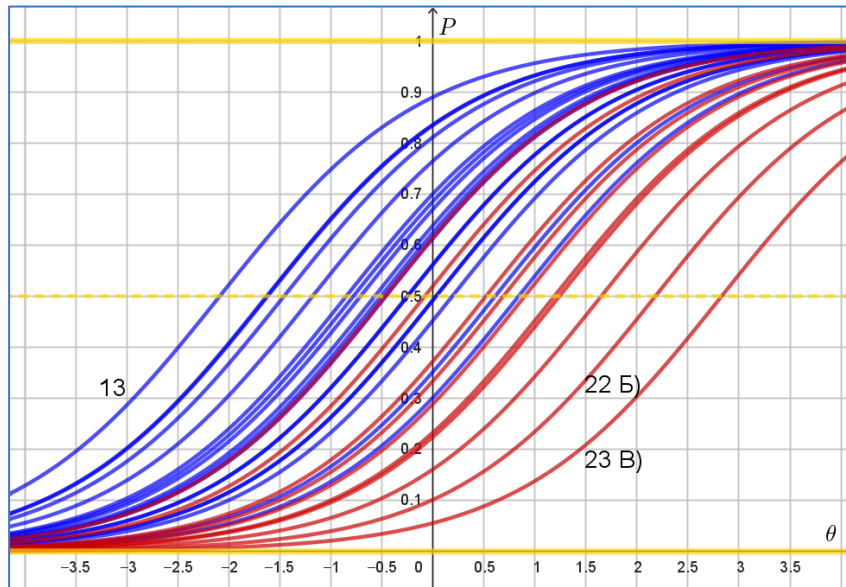


Fig. 3.3. 2024 task characteristic curves.

In the second part of this section, the Rasch model is applied to all tasks, and each sub-item of the ECR tasks is considered as an independent unit. The characteristic curves (Figure 3.3) of these tasks span a consistently wide range of difficulty.

Number of students	.				.	#	.	#											
	#	.	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
θ/b	-2.4	-2.0	-1.6	-1.2	-0.8	-0.4	0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0	4.4	4.8
Task No	13	4	12	5	1	2	17	11	3	23A)	6	19	22A)	21C)	22B)	23C)			
		10				14	16	7	8	23B)									
						15	20	9	21B)										
						21A)													
						18													

Fig. 3.4. Chart of 2024 tasks and respondents.

According to the diagram in figure 3.4, the test composed of the presented tasks appears well balanced for the respective students, because the two distributions peak at close values and cover almost the same range of ability/difficulty. However, for better prepared students with achievement level above, 2.8 there are no tasks that reach such a high level of difficulty.

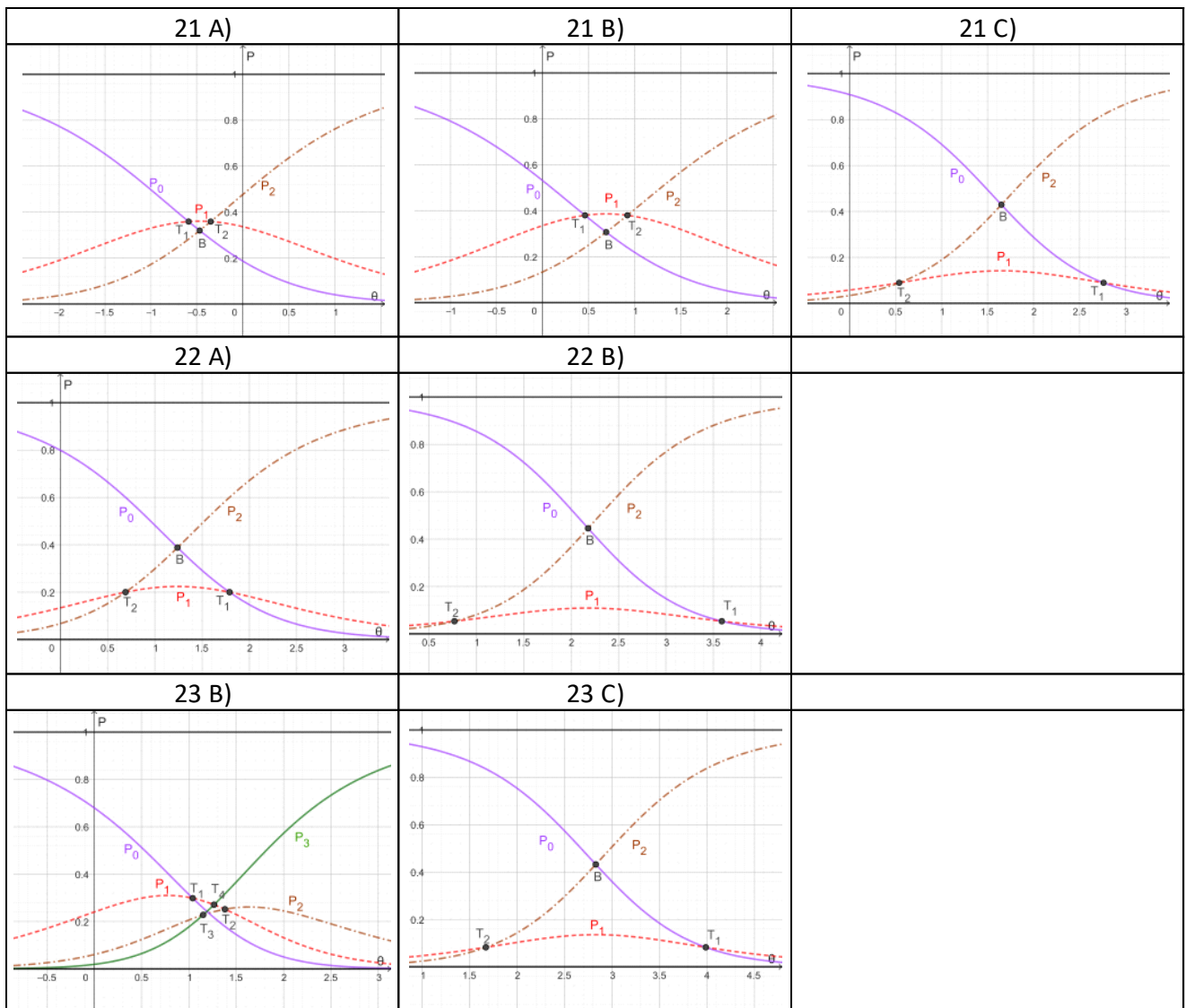


Fig. 3.5. 2024 CR task characteristic curve graphs.

In Figure 3.5 it can be observed that only in tasks 21 A) and 21 B) students are most likely to receive 0, 1 or 2 as predicted in the assessment. In these tasks, the abscissas of the points T_1 are smaller than those of T_2 . In contrast, in tasks 21 C), 22 A), 22 B) and 23 C), students are more likely to get 0 or 2 points, i.e. either solve the task correctly or not solve it at all. In these tasks, the abscissas of points T_1 are greater than those of T_2 , suggesting that scoring could be simplified to a 0-1 scale, or adjusted to allow 1 point under less stringent conditions. Task 23 B) is scored with 0-1-2-3 point scale, but as can be seen from the graph, the probability of having students with a score 2 points is low. It is more likely to have students who achieve 0, 1 or 3 points, suggesting it may be more fitting as a 0-1-2 task. These tasks were created keeping in mind the scoring of the NEAs conducted during the studied four-year period. However, only 2 (possibly 3) of them have the possibility to obtain intermediate points. For the other tasks, if the students have started the task correctly, they manage to finish it successfully. One of the

reasons can be found in the fact that not all students work on the second module. Most who write solutions to the ECR tasks are more prepared and manage to solve them to the end. They might make a small mistake, but it shouldn't affect their points significantly.

Comparison of traditional and experimental scoring methods on different indicators

Recall that the two assessment methods we are considering here are:

- "traditional" - each MCQ task is scored 2, 3 or 4 points, for a total of 65, and CR tasks are scored 11 or 12 points - a total of 35 points. Thus, the maximum score from both parts of the NEA is **100 points**;
- "experimental" - each task with MCQ is scored with 1 point, and each sub-item of the CR tasks is given a maximum of 1, 2 or 3 points. So the maximum score is **36 points**.

To compare the results obtained through traditional and experimental methods, one can use various approaches. One of them is described in (Bankov, 2012). For this purpose, we will equate the score obtained from **the experimental (X)** assessment (ranging from 0 to 36 points) to those from **the traditional (Y)** assessment (ranging from 0 to 100 points). This is done with the formula:

$$X_i^* = \frac{\sigma_Y}{\sigma_X} (X_i - \mu_X) + \mu_Y,$$

where

- X_i are the points of the i-th student according to the experimental assessment;
- X_i^* are those points equated to traditional scoring;
- μ_X and μ_Y are the mean values of students' raw scores obtained from the experimental and traditional assessments, respectively;
- σ_X and σ_Y are their respective standard deviations.

This gives the formula:

$$X_i^* = 2.76(X_i - 17.63) + 49.$$

After that, the correlation coefficient between **the scores obtained (X*)** and those from the traditional assessment (Y) was obtained. The value of this coefficient is **0,99192**, which means

that the two evaluations are identical. For comparison, table 3.14 shows the scores of the first 20 students on the three scales. Apparently, in most cases the differences are 0 or 1 points, although some larger differences of up to 10 points are also observed.

No	X	X*	Y
1	34	94	94
2	27	75	80
3	32	89	88
4	26	72	66
5	27	75	79
6	32	89	89
7	26	72	65
8	33	92	93
9	21	58	61
10	33	92	98.5
11	35	97	94
12	34	94	95
13	24	67	77
14	28	78	82
15	30	83	85
16	36	100	94
17	36	100	100
18	21	58	64
19	34	94	96
20	9	25	24

Table 3.14. Scores of the top 20 students on three rating scales

One of the reasons for the significant score discrepancies is that each evaluator may not equally evaluate the student's progress on the tasks with CR. This is not an unusual situation at the NEA, and for this reason there is an arbitrator who determines which of the two assessors' assessments is closer to the real performance. Let's recall that 1 point according to the experimental assessment corresponds approximately to 3 points according to the traditional one. Evaluators who have traditionally graded their students sometimes tend to award 1-2 "bonus" points to students who normally do well in class, but in this case have made mistakes. These errors, however, in the experimental assessment are penalized as provided in the assessment manual.

Another situation where larger point differences exist, they between X^* and Y can be illustrated with student number 13. He made 1 error on task 2, where the sum of his raw score on the MCQ tasks is $X_{13} = 19$, giving an equal score $X_{13}^* = 53$, while the traditional assessment score is $Y_{13} = 63$. Thus, the difference between the two evaluations in his overall score is also 10 points.

Although, cases of larger differences are few and do not significantly affect the overall statistics, as evidenced by the high correlation coefficient between the X^* and Y scores.

Implications

The fourth section discusses the implications of the conducted pedagogical experiment. Its results can be classified according to the three main directions of research:

1. Regarding the multiple-choice questions:
 - a. an important element in their composition is not to allow the correct response to be obtained through elementary arithmetic operations of the numbers in the condition - tasks 12 and 20;
 - b. the drawings to the geometric tasks should not give too much information about obtaining the answer, especially in cases where it is necessary to apply some properties of the objects - tasks 17 and 19;
 - c. if the drawing of the geometric task is sufficiently suggestive, to choose answers that could not be found easily - task 19;
 - d. concepts that are rarely used in grade VII often do not lead to good discrimination of students - *scale* in task 4 and *prime numbers* in task 6.
2. Regarding the 2023 extended free-response tasks.
 - a. a positive trend is to divide them into several sub-items, and the author's opinion is that each sub-item can support the decisions of the following ones, and not necessarily be independent of them. In this way, less prepared students have a greater opportunity to get points in the second part of the NEA - a problem that is seen every year;
 - b. each sub-item to be evaluated by separating several important steps in its solution. These steps should not be more than 2-3. Thus, it is possible to conduct a study using probabilistic modeling and to more accurately compare the results of different NEAs;
 - c. to allow relatively easy obtaining of 1 point, which is comparable to solving a MCQ task.
3. Regarding score formation:

- a. the scoring of 2, 3 or 4 points for the MCQ tasks is not very justified. Obviously, even if they are well known and tried and tested, easier tasks do not always correspond to fewer points;
- b. theory recommends that MCQ tasks be scored 0-1, regardless of their difficulty. This coefficient shows in fact what proportion of students indicated the correct response, so it is expected that if a student is well prepared, he will be able to indicate more correct responses;
- c. the evaluation of the individual sub-items of the tasks with extended free response should be an integer number of points 0-1-2 or 0-1-2-3, which correspond to the important steps in their solution;
- d. the high value of the correlation coefficient between the two types of evaluation obtained in point 3.3 of this chapter shows the identity of the two methods of evaluation. Additionally, the experimental method is amenable to analysis using the methods of modern test theory;
- e. if a unified scale is needed, the obtained points can be presented in percentages from 0 to 100.

Recommendations for future developments:

- The use of the common tasks from the trial NEA of 2024 and the past NEA can serve as "anchor" tasks for a more objective comparison of the student results across the four studied years. This method was discussed in point 3 of the second chapter;
- A system for comparing these results can be established through similar exemplary NEAs to be held in specially selected schools in the country.

It can be recommended that the authors of future NEAs in mathematics try out the tasks using the ideas of this experiment. In this way, some inadvertent errors in both the conditions and responses to the MCQ tasks would be avoided. Also, it can be seen how well the important steps in solving the extended CR tasks separate the solution, so that students are able to get the intermediate points of these tasks with good probability.

Conclusion

This research is the result of the author's long-term participation in committees for the development and validation of various exams and assessments at both regional and national

level. The focus of the dissertation work is the National External Assessments in Mathematics, administered to students completing grade VII . The study addresses long-standing questions, surrounding these assessments, including: how to interpret the results achieved by the students; to what extent the students have mastered the learning material; do the given tasks correspond to the set preliminary requirements regarding the content and knowledge area; what the quality of these tasks is in terms of theories of multiple-choice or free-response test tasks. Furthermore, the research explores methods to improve tasks that exhibit specific issues, evaluates the justification for using NEAs as criteria for high school admissions, considers alternative frameworks for task evaluation, and examines the impact of e-distance learning on student achievement.

The main goal of the dissertation is to apply different methods for the analysis of results of standardized assessments of students in mathematics (NEA) and to analyze the strengths and limitations of each method. Using this analysis, recommendations for their effective use is provided. The main hypothesis is that the correct use of the Classical theory of tests, combined with the modern Item responsibility theory, yields a more objective, more accurate and reliable assessment of the mathematics achievements of students of grade VII. This hypothesis was tested in the dissertation.

To substantiate this thesis, the author has researched and analyzed publications related to the application Classical Test Theory and Item Response Theory in both national and international studies of student achievement. Moreover various software tools for processing test results were examined.

The following **conclusions** can be drawn from the observations and analyses:

Chapter One:

- All topics from the content framework of the test specification are represented;
- The MCQ tasks are mainly at the cognitive level of Knowledge and Comprehension
- Difficulty and discriminative power are well measured, with a few minor exceptions;
- Tasks with CS do not provide added value for a large part of students in the small and medium-sized regions of the country. However, this trend does not hold true for Sofia-city and, to some extent, for Plovdiv;
- The COVID epidemic affected the results of students, especially those who studied at DLEE and graduated in 2022 grade VII;

Chapter Two:

- IRT models showed that the two-parameter model provided the best fit to the empirical data. The results of this model confirm the conclusions obtained with CTT;
- Applying the three-parameter model makes it possible to detect tasks that have a high guessing ratio. For these tasks, it is advisable to look for a better option or not to include them in subsequent evaluations;
- The difficulty of the MCQ tasks is well matched to students' abilities, with no tasks specifically designed for very strong and very weak students;
- Comparing students' achievements using anchor-tasks shows a decline over the four-year period studied, and this decrease turns out to be statistically significant;
- Scoring tasks in the conventional manner does not always correspond to their difficulty. There are many examples in which a difficult task is awarded a few points, while an easy task receives more points;
- An assessment approach based on student outcomes is proposed, which can be implemented in future NEAs;

Chapter Three:

- Several factors have been identified that can negatively impact the qualities of tasks with IS. Methods for detecting such tasks have been proposed, as well as options for improving their qualities;
- It is advisable to score MCQ tasks with 1 point for a correct response and 0 points for all other responses;
- Assessment of CR tasks does not need to be fragmented as is done in current NEAs. It is enough that their sub-items are evaluated with 1, 2 to 3 points, which correspond to the important steps in solving them;
- If a unified scale of 0 to 100 is needed, students' raw scores obtained from the above two assessments can be presented on a percentage scale.

**Contributions of the dissertation
Scientific and applied contributions**

1. Different IRT models (with one-, two- and three-parameters) were compared with the empirical data obtained from the MCQ tasks in NEAs conducted in the period 2020-2023, encompassing the results of all students from four administrative districts in Bulgaria. It was concluded that the two-parameter model fit these data best.
2. A comparison has been made of different methods for evaluating the CR tasks. A simplified rating scale for CR tasks was developed so that IRT methods could be applied

to study their indicators. This scale is justified on the basis of the results obtained from the conducted NEAs;

3. Student achievements in mathematics in the four studied years were compared, using a method based on the two-parameter IRT model to evaluate different populations through the use of anchor-tasks. A decline in these achievements was observed with each subsequent year;
4. Variants are proposed for converting the scores from the MCQ tasks into the six-point grading system widely used in our country.

Applied Contributions

5. CTT methods were applied to the MCQ tasks from NEAs after grade VII in the studied four-year period;
6. An analysis of the MCQ tasks from NEAs in the studied four years was conducted based on the content and cognitive area;
7. A pedagogical experiment was conducted with 487 students from 9 regional cities in the country, who were set to graduate from the VIIIth grade in 2024;
8. 23 original tasks (by the author of the dissertation) have been compiled, which are in line with those of the NEAs conducted in the studied four-year period;
9. The evaluation of the MCQ tasks from the conducted experiment was compared using the generally accepted method (with 0, 2, 3 or 4 points) and the condition for a true-false task (with 0 or 1 point);
10. The evaluation of the tasks with extended CR based on the generally accepted method was compared with the evaluation, which was carried out by the author according to the number of "important" steps in solving them,;
11. It was found that there was an excellent correspondence of the obtained scores of the students in the two ways described in contributions 9 and 10.

Directions for future development

Some ideas for future developments that emerged during the development of this paper:

- ✓ The use of the common tasks between the 2024 trial NEA and the past NEA in the period 2020-2023 can serve as anchor tasks for a more objective comparison of the results of the students of the four studied years. This method was outlined in point 3 of Chapter Two;
- ✓ Through such exemplary external assessments to be conducted in specially selected schools in the country, a system can be created to compare the results of such

mathematics assessments. This approach can also be experimented in other academic disciplines;

- ✓ IRT test research methods would improve the objectivity of these assessments.

References

- Alashka, R. (2016). Two-Factor Probability Models in Education [In Bulgarian]. *Mechanics, transport, communications*, 14(3/2), VII-7 - VII-12.
- Alashka, R. (2017). *Application of Probabilistic Models for the Analysis of Exam and Test Results, Dissertation for the award of the scientific and educational degree "PhD" [In Bulgarian]*. Sofia: Sofia University "St. Kl. Ohridski".
- Bankov K., Vitanov T. (2010). The External Assessment in Mathematics, Status and Prospects [In Bulgarian]. *Mathematics and mathematics education*. Albena: SMB.
- Bankov, K. (2002). Probabilistic Modeling for Measurement of Students' Achievement [In Bulgarian]. *Mathematics and Informatics*(4), 41-49.
- Bankov, K. (2012). *Introduction to Test Theory [In Bulgarian]*. Sofia: Izkustva.
- Bankov, K. (2012). *Large Scale Studies in Educational Assessment [In Bulgarian]*. Sofia.
- Bankov, K. (2023). Scoring Test Tasks and Using Scales in Standardized Assessments [In Bulgarian]. *Mathematics education 75 years of mission and history*, 73-86.
- Birnbaum, A. (1968). *Some Latent Trait Models and Their Use in Inferring an Examinee's Ability* (Lord, F.M. and Novick, M.R., Eds., *Statistical Theories of Mental Test Scores* ed.). Addison-Wesley: Reading, 397-479.
- Bizhkov G., Kraevski V. (2002). *Methodology and Methods of Pedagogical Research [In Bulgarian]*. Sofia: University Publishing House "St. Kl. Ohridski".
- Bizhkov, G. (1996). *Theory and Methodology of Didactic Tests [In Bulgarian]*. Sofia: Prosveta.
- CAPSE. (2024). Retrieved from Center for Assessment in Pre-school and School Education: <https://www.copuo.bg>
- Chehlarova, T. (2022). National External Assessment in Mathematics in 4th Class in 2022 [In Bulgarian]. *Mathematics & Informatics*, 65(6), 344 – 357.
- Danchev, Plamen; Bankov, Kiril; Stoimenova, Vessela; Atanassov, Dimitar. (2013). *Pilot Implementation of Statistical Models for Estimation of the Value-Added of Bulgarian Schools Using National Student Assessment Data*. Washington, D.C.: World Bank Group. From <http://hdl.handle.net/10986/24491>
- Djalev, L. (2013). *A Comparative Analysis of the Applicability of Two Psychometric Test Theories (on data from the general education test). Dissertation work for the award of the scientific and educational degree "PhD" [In Bulgarian]*. Sofia: NBU.

- Djalev, L. (2014). Applicability of Classical Test Theory and Item Response Theory: a Literature Review [In Bulgarian]. *Bulgarian Journal of Psychology*(1-3), 81-102.
- GeoGebra. (2024). Retrieved from <https://www.geogebra.org>
- Hambelton, R., Swaminathan, H., Rogers, H. (1991). *Fundamentals of Item Response Theory*. SAGE Publications.
- Ivanov, I. (2006). *Pedagogical Diagnostics [In Bulgarian]*. Shumen: University Press "Bishop Konstantin of Preslav".
- Kim, V. S. (2007). *Testing of Educational Achievements [In Russian]*. Ussuriysk: UGPI.
- Lord, F. M., M.R. Novick (ed.). (1968). *Statistical Theories of Mental Test Scores*. Reading/Mass, Addison-Wesley,.
- Meyer, J. P. (2024, 09 16). Retrieved from Psychomeasurement Systems: <https://itemanalysis.com>
- MON-2020. (2024, 09 16). *NEA for students in grade VII during the academic year 2019-2020 [In Bulgarian]*. From MON: https://www.mon.bg/nfs/2020/06/nvo-viikl-math_17062020.pdf
- MON-2021. (2024, 09 16). *NEA for students in grade VII during the academic year 2020-2021 [In Bulgarian]*. From https://www.mon.bg/nfs/2021/06/nvo-math_7kl_18062021.pdf
- MON-2022. (2024, 09 16). *NEA for students in grade VII during the academic year 2021-2022 [In Bulgarian]*. From https://www.mon.bg/nfs/2022/06/7kl_nvo_math_16062022.pdf
- MON-2023. (2024, 09 16). *NEA for students in grade VII during the academic year 2022-2023 [In Bulgarian]*. From https://www.mon.bg/nfs/2023/06/nvo_math_7klas_variant2_16.06.2023.pdf
- Open Data Portal of the Republic of Bulgaria. (2024, 09 16). Retrieved from <https://data.egov.bg/>
- Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright*. Chicago: The University of Chicago Press.
- Stoimenova, E. (2000). *Measurement Quality of Tests [In Bulgarian]*. Sofia: New Bulgarian University Press.
- Stoyanova, F. (1996). *Testology for Teachers [In Bulgarian]*. Sofia: Atika.
- Tsonev, P. (2022). Some Conclusions on the Results of the National External Assessment in Mathematics for 7th Grade [In Bulgarian]. *Mathematics and Informatics*, 65(6), 587-601.

- Tsonev, P. (2023-a). Choosing an Appropriate Model for Investigating the Tasks of School in Mathematics after Grade VII According to IRT Methods [In Bulgarian]. *Mathematics and Informatics*, 66(4), 491-505.
- Tsonev, P. (2023-b). Comparison of the Results Based on the Multiplechoice Items from the National External Assessment in Mathematics After Grade VII Held in 2020, 2021 and 2022. [In Bulgarian]. Annual International Scientific Conference of VVVU “G. Benkovski”, (pp. 353-363). Dolna Mitropolia.
- Tsonev, P. (2024-a). Challenging Mathematic Task-Types for Students Finishing 7th Grade in Bulgarian Schools. *International conference KNOWLEDGE-BASED ORGANIZATION*, 30 (Issue 2), pp. 1-4. Sibiu.
- Tsonev, P. (2024-b). Possible Scoring of the Open-ended Items from the National External Assessment after Grade VII in 2023, based on on Item Response Theory [In Bulgarian]. *Mathematics and Education in Mathematics*, 53, 161-168.
- Tsonev, P. (2024-c). Analysis of Some Multiple Choice Tasks from a Trial National External Assessment Conducted in 2024 [In Bulgarian]. Annual International Scientific Conference of VVVU “G. Benkovski”, D. Mitropolia, 2024, 380–389
- Yankova, T. (2024, 09 16). *Testing for Normality [In Bulgarian]*. From Exercises in Probability and Statistics: https://www.fmi-plovdiv.org/evlm/DBbg/database/probstat/VerStatXPtema21_Testing%20normality%20_BG.pdf

Note:

References include all titles that are cited in the dissertation. Those of them that are not cited in the auto-reference are in italics. The number of cited titles in the abstract is 27.

Publications of the author related to the topic of the dissertation

- Tsonev, P. (2023-b). Comparison of the Results Based on the Multiplechoice Items from the National External Assessment in Mathematics After Grade VII Held in 2020, 2021 and 2022. [In Bulgarian]. Annual International Scientific Conference of VVVU “G. Benkovski”, (pp. 353-363). Dolna Mitropolia.
- Tsonev, P. (2024-a). Challenging Mathematic Task-Types for Students Finishing 7th Grade in Bulgarian Schools. *International conference KNOWLEDGE-BASED ORGANIZATION*, 30 (Issue 2), pp. 1-4. Sibiu.
- Tsonev, P. (2024-b). Possible Scoring of the Open-ended Items from the National External Assessment after Grade VII in 2023, based on on Item Response Theory [In Bulgarian]. *Mathematics and Education in Mathematics*, 53, 161-168.
- Tsonev, P. (2024-c). Analysis of Some Multiple Choice Tasks from a Trial National External Assessment Conducted in 2024 [In Bulgarian]. Annual International Scientific Conference of VVVU “G. Benkovski”, D. Mitropolia, 2024, 380–389

Declaration for originality

by Pavlin Ivanov Tsonev

PhD student in the program

"Methodology of teaching mathematics and informatics" - Methodology of teaching
mathematics

FMI, Sofia University "St. Kliment Ohridski"

In connection with the conduct of the procedure for acquiring the educational and scientific degree "doctor" in the FMI of SU "St. Kliment Ohridski" and defense of the dissertation work presented by me, I declare:

The results and contributions of the dissertation research conducted, presented in my dissertation on the topic "Methods for the analysis of results of standardized assessments of students in mathematics" are original and are not borrowed from research and publications in which I have no participation.

Declarant:

Pavlin Tsonev

Sofia, 2025

Acknowledgments

I would like to express my sincere gratitude to:

- my supervisor Prof. Bankov for his professionalism and patience in addressing my numerous questions. I am grateful for the encouragement he provided in the development of this research work and the publications that accompanied it. Moreover, I thank him for his clear and timely guidance;
- the colleagues from CAPSE and specifically its director Mrs. Neda Kristanova , who assisted me in obtaining the data from the conducted NEAs in mathematics for class VII , without which this research would have been impossible;
- the teachers with whom I collaborated to conduct the pedagogical experiment:
 - from Blagoevgrad – Mediha Topalova;
 - from Dobrich - Milena Avramova;
 - from Montana – Yordanka Elenkova and Stefcho Nakov;
 - from Pleven - Ani Tsoneva, Valeri Naumov, Georgi Melnikliyski , Diana Danova, Dolores Nenova, Elka Linkova, Emilia Melnikliska , Marusya Kehaiova, Milanka Decheva, Petranka Kubratova, Plamen Mitev, Plamen Petranov , Svetla Petrova, Silvia Spiridonova, Teodora Todorova;
 - from Plovdiv – Elena Todorova;
 - from Razgrad – Emilia Stankova;
 - from Sofia - Dimitar Dimitrov;
 - from Shumen – Antoaneta Kovacheva;
- my colleague Vanya Katsarska from Georgi Benkovski Air Force Academy, who gave me invaluable assistance in the English version of the abstract;
- my family and colleagues from Georgi Benkovski Air Force Academy for the understanding and moral support!