ENHANCING STUDENT ASSESSMENT: A DUAL-METHOD APPROACH

Pavlin TSONEV

"Georgi Benkovski" Air Force Academy, Dolna Mitropolia, Bulgaria (ptsonev@yahoo.com)

DOI: 10.19062/2247-3173.2025.26.4

Abstract: *The article presents two methods for assessing students' achievements in the National External Assessment in Mathematics after the 7th grade in Bulgaria. Some advantages of analyzing the assessment using Item Responsibility Theory are shown.*

Keywords: *national external assessment, item responsibility theory, difficulty*

1. INTRODUCTION

Each stage of the Bulgarian education system concludes with a National External Assessment (NEA): administered at the end of IV-th grade, VII-th grade, X-th grade, and XII-th grade. This paper addresses the results in mathematics from NEAs conducted after VII-th grade.

Since the beginning of 2024, the exam has been structured into two distinct parts:

- The first part, lasting 60 minutes, contains 20 test multiple choice questions with a choice of answer A, B, C, D, which are scored with 2, 3 or 4 points. They contribute to a total of 65 points.
- The second part, lasting 90 minutes, contains 3 open-ended mathematical problems. Each problem includes two to four independent sub-tasks, requiring students to provide detailed written solutions. They contribute to a total of 35 points.

The maximum score for the two parts is 100.

The scoring of the tasks is predetermined by the authors of the test. They take into account the perceived difficulty of the items. Usually, these tasks are piloted in advance, and it is expected that the easier task will be evaluated with fewer points, and the more challenging ones are awarded higher point values. Table 1 shows the results after the external assessment in 2024 for the multiple-choice questions (MCQ) tasks, arranged according to the difficulty coefficient obtained from the Classical Test Theory (CTT). As it can be seen, some easy tasks (with numbers 4, 10 and 12) are evaluated with 3 points, while task # 5 is of optimal difficulty and is evaluated with 2 points. At the same time, task # 14 (evaluated with 4 points) is easier than task # 3 (evaluated with 3 points).

22.

A) …

B) Factor the polynomial $M = ax^2 - bx + 45$, where the coefficient a is the smallest value of the expression $(x+3)^2 + 1$ and $b = \frac{|-13|+(-13)^0}{(-1)^{2022}}$.

FIG.1 Condition of task 22 of NEA – 2022

For the evaluation of the extended constructed response (ECR) tasks, the authors of the test develop a scoring rubric, assigning a specific number of points to particular actions or components within a solution. These points are usually multiples of 0.50, and in some tasks, they are multiples of 0.25. This is done for finer differentiation among students' performance, since the points obtained from the NEA are used for admission into profile and professional oriented classes in secondary high schools. Fig. 1 shows the prompt for task # 22 from the NEA – 2022 [2], while table 2 outlines the corresponding scoring scheme. Such a defragmentation of points is associated with a detailed review of each student's solution. In the article [3], it is shown that this scoring is not very good practice, since a small percentage of students receive intermediate results. In addition, this method of scoring tasks is not scientifically justified. In educational measurement, the most widely accepted approaches are Classical Test Theory (CTT) and, more recently, Item Response Theory (IRT) [1].

2. MAIN RESULTS

To explore this further, the author of the report conducted an experiment with 487 students completing grade VII in 2024. In this experiment, their achievements were assessed using two different evaluation methods.

Table 1. Assessment of tasks with IE from NEA – 202 4 years.

Points	#	Difficulty	Interpretation
2	13	0.86	very easy
3	4	0.81	easy
3	10	0.81	easy
3	12	0.79	easy
2	5	0.75	easy
3	1	0.70	optimal
2	2	0.69	optimal
3	17	0.68	optimal
4	14	0.65	optimal
3	15	0.64	optimal
3	18	0.63	optimal
3	11	0.60	optimal
4	16	0.60	optimal
4	20	0.56	optimal
3	3	0.55	optimal
4	7	0.55	optimal
4	9	0.52	optimal
4	8	0.44	optimal
4	6	0.40	difficult
4	19	0.39	difficult

The first method, referred to by the author as *traditional method*, follows the scoring approach currently used in Bulgaria's National External Assessment (NEA). The maximum sum of points for one student is 100 points.

The second method, referred to *as experimental method,* is based on the following scoring scheme:

- Each MCQ task is scored with 1 point for a correct answer and 0 points in other cases.

- Each ECR task is scored with 0, 1, 2 or 3 points, which correspond to the number of "important" steps in solving it.

Thus, the maximum number of points according to the experimental method is 36 points.

Table 2. Assessment of task 22 of NEA – 2022

Description	Points
$a = 1, b = 14$ and $M = x^2 - 14x + 45 - (x - 9)(x - 5)$	5 points
Finding the minimum value and determining $a = 1$	1 point
Finding $b = 14$	1 point
For factoring $M = x^2 - 14x + 45$	3 points
Method I: Representing $-14x = -9x - 5x$	1 point
For appropriate grouping in pairs	1 point
For factoring out the common term	1 point
Method II: (completing the square)	0.5 points
Representing $-14x = -2.7x - 2.7x$	1 point
For representing $x^2 - 2.7x + 7^2 - 7^2 + 45$	1 point
For writing $(x - 7)^2 - 4$ or $(x - 7)^2 - (2)^2$	0.5 points
For factoring $(x - 7 - 2)(x - 7 + 2)$ or $(x - 9)(x + 5)$	1 point

Table 3. All tasks from NEA – 2024, sorted by difficulty

#	b-param
13	-2.09
4	-1.63
10	-1.62
12	-1.46
5	-1.18
1	-0.85
2	-0.78
17	-0.71
14	-0.57
15	-0.51
21 A)	-0.47
18	-0.46
11	-0.25
16	-0.25

#	b-param
20	-0.08
3	0.03
7	0.03
9	0.19
23 A)	0.52
8	0.61
21 B)	0.69
6	0.84
19	0.90
23 B)	1.19
22 A)	1.24
21 C)	1.65
22 B)	2.18
23 C)	2.83

The use of the experimental assessment method enables all tasks—both MCQs and ECRs—to be analyzed using Item Response Theory (IRT) on a single difficulty scale, based on the b-parameter. This coefficient is usually in the interval (-3;3). Moreover, if a task is easier, the coefficient b is closer to -3, and if it is more difficult, it is closer to 3.

Comparing tables 1 and 3, it can be established that MCQ tasks from # 1 to # 20 are arranged in the same way according to both theories – CTT and IRT. At the same time, the ECR tasks with numbers from 21 to 23 are included in Table 3. As can be seen, the tasks cover a relatively uniform difficulty interval from -2.09 to 2.83. As anticipated, the ECR tasks generally possess higher b-parameter values, reflecting their greater complexity relative to the MCQs.

To compare the scores obtained by the students in the two methods, one can approach it in different ways. One of them is by equating the scores obtained by the experimental method X ∈ [0;36] to those obtained by the traditional method Y ∈ [0;100]. This is done with the formula

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

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

This is how the formula is obtained:

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

The value of the correlation coefficient is 0.99192, which means that the two assessments are identical.

The experimental evaluation method offers several key advantages:

- it enables the application of IRT for analyzing student performance;
- it allows for the comparison of results across different regions and schools;
- it facilitates more reliable comparisons between distinct populations, surpassing the traditional method used in Bulgaria, which relies primarily on average exam scores;
- it supports more accurate comparisons between different populations through so-called “anchor” tasks. This method is shown for example in [4]

3. CONCLUSIONS

In conclusion, the conventional scoring method used in the NEA can be effectively replaced by a simpler experimental approach without compromising the assessment's regulatory function. Moreover, the research on the results of these assessments is subject to analysis using various educational measurement theories, including both Classical Test Theory (CTT) and Item Response Theory (IRT). If necessary, student scores from the experimental method can be easily converted to a unified 0–100 scale by representing the raw scores as percentages, ensuring compatibility with existing reporting standards.

REFERENCES

[1] R. Hambelton, H. Swaminathan, H. Rogers, *Fundamentals of Item Response Theory*., SAGE Publications, 1991;

[2] MON-2022, (07.04.2025) 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;

[3] P. Tsonev, *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, 2024;

[4] P. Tsonev, *The National External Assessment in Mathematics After Grade VII Held In 2022 and 2023: Comparison of the Results Based on the Multiple-Choice Items*, in Journal of Philology and Intercultural Communication, vol. VIII, No. 2, 97-103, 2024, Bucharest, Romania, Military Technical Academy “Ferdinand I” Publishing House ISSN: 2558-8478 ISSN-L: 2558-8478 ISSN Online: 2558-9830.