HNUE JOURNAL OF SCIENCE
Educational Sciences 2024, Volume 69, Issue 5B, pp. 72-82
This paper is available online at http://hnuejs.edu.vn/es
DOI: 10.18173/2354-1075.2024-0136

ADAPTATION OF 6TH GRADE STUDENTS TO MATHEMATICS ASSESSMENT STANDARDS

Chu Cam Tho[1], Nguyen Thi Quynh Anh[2], Vu Truong An[1], Luu Thanh Ha[3],
Duong Thi Thu Huong[1], Dang Xuan Cuong[1] and Nguyen Viet Dung[2]

[1]*Vietnam Institute of Educational Sciences, Hanoi city, Vietnam*
[2]*Adaptive Learning Global Education Joint Stock Company, Hanoi city, Vietnam*
[3]*Victoria Thang Long Primary and Secondary School, Hanoi city, Vietnam*
[*]Corresponding author: Nguyen Thi Quynh Anh, e-mail: anhntq@aeglobal.edu.vn

Received April 19, 2024. Revised October 10, 2024. Accepted December 27, 2024.

Abstract. This paper examines how the 2018 general education curriculum aligns with competency-based assessment standards, with a particular focus on 6th-grade mathematics. The research encompasses the design and implementation of a computer-based end-of-year mathematics exam, created between March 1 and May 15, 2023, and administered to over 3,850 students via the AEGlobal educational platform. The study underscores the importance of a competency-based assessment framework in education, exemplifies this through the development of a mathematics assessment tool, and offers insights into student adaptability to the new standards. Furthermore, it contributes to enhancing teaching and learning quality across the general education system.

Keywords: assessment, competency, end-of-year examination, Mathematics, standard.

1. Introduction

The policy of using standards for assessment is adopted by many countries to ensure transparency and maintain sustainable quality in education. The term "standard" holds various meanings depending on the context. According to Maxwell [1], the term “standards” encompasses at least five distinct interpretations, categorized here as: (1) standards as moral or ethical imperatives (what someone should do); (2) standards as legal or regulatory requirements (what someone must do); (3) standards as target benchmarks (expected practice or performance); (4) standards as arbiters of quality (relative success or merit); and (5) standards as milestones (progressive or developmental targets). The first three types are considered desirable, necessary, or appropriate, while the last two represent outcome levels. The first type is often conveyed through guidelines or professional codes, the second through performance requirements implying the possibility of failure (e.g., requirements for program approval or certificate awarding), and the third through statements detailing expected outcomes.

Maxwell [2] extended his research on "standards" aiming to offer a clearer description of their structure. He identified four main characteristics of the assessed construct: (1) learning versus performance; (2) development (time-extensive, assessing interim progress) versus achievement (time-limited, assessing degree of success); (3) criterion-referencing versus norm-referencing; and (4) quality (how well) versus quantity (how much). The choice of constructs determines how we represent and express relevant standards. Regarding the purposes of standards, these may include: (1) setting targets for student learning; (2) showing students their progress; (3) promoting consistency in judging achievement or progress; (4) setting qualification requirements (certification); (5) interpreting test performances; (6) setting benchmarks for system monitoring; and (7) ensuring accountability for schools and systems. These purposes are often used in combination. Standards can be implemented in three specific ways: content standards, performance standards (focusing on merit or proficiency), and developmental standards. Content standards act as a "road map" for schools and teachers, outlining the overall knowledge structure for each domain and providing a framework for planning and delivering the curriculum. In the United States, "standards" often refer to "content standards". Performance standards, by contrast, define the levels of knowledge and skills students must attain and serve as a basis for measuring outcomes and applying sanctions when necessary. For example, in Georgia, performance standards go beyond content standards by detailing specific knowledge and skills expected of students, along with assessment tools like tests and student work samples. These standards often use methods like "cut-scores" to define achievement levels. However, a limitation of performance standards is their failure to clearly show how students progress over time, particularly when identical labels (e.g., A–E) and generic descriptors are applied repeatedly. Developmental standards, on the other hand, provide progressive labels and descriptions, illustrating the journey of learning and milestones achieved.

According to the OECD [3], the term "standard" carries different meanings across countries and contexts. In some cases, it is used interchangeably with "objectives" or as criteria for determining whether a specific achievement is adequate or describes the level of progress in a particular domain. A standard can be understood as "defining what someone should know and be able to do to be considered competent in a specific area”. Standards may describe and communicate desired outcomes, quality learning, or best practices. Additionally, they can serve as a benchmark or measurement tool, aiding decision-making by highlighting the gap between actual performance and the minimum level of competence required.

In mathematics, some commonly used assessment standards at the secondary level include the Common Core State Standards for Mathematics (CCSS) [4] and Singapore’s Mathematics Syllabus [5]. Regarding the CCSS, Confrey [4] highlights the importance of developing "sequenced obstacles and challenges for students absent the insights about meaning that derive from careful study of learning, would be unfortunate and unwise”. Recognizing this, the creation of these standards began with research-based learning progressions, detailing how students’ mathematical knowledge, skills, and understanding evolve over time. In Singapore’s Mathematics Syllabus, the primary goals of mathematics education are to: (1) acquire and apply mathematical concepts and skills; (2) develop cognitive and metacognitive skills through a mathematical approach to problem-solving; and (3) cultivate positive attitudes toward mathematics.

In Vietnam, the General Education Curriculum (GEC) introduced the term "standard", particularly "knowledge and skills standards", notably in the GEC 2006. In the GEC 2018 [6], the term "required learning outcomes" is used to refer to the standards that must be "achieved" or "completed", explicitly defining the objectives or expected outcomes for each subject, educational activity, academic year, and educational level. The GEC 2018 specifies the required attainments for each competency and general capacity as standards for each educational level, representing the specific requirements students must meet by the end of a grade or class.

In mathematics, the curriculum defines the required learning outcomes in terms of content appropriate to the class level and skills relevant to the grade. The GCE 2018 outlines five key competency components for students: mathematical thinking and reasoning, mathematical problem-solving, mathematical modeling, mathematical communication, and the ability to use tools and resources for learning mathematics. For 6th-grade mathematics, the required learning outcomes are detailed on pages 46 to 55 of the mathematics curriculum document. These include natural numbers, integers, fractions, decimals, plane shapes in real-life contexts, symmetry in nature, basic geometric shapes, data collection and organization, data analysis, and introductory probability concepts. According to the Ministry of Education and Training's guidelines in Circular 22/2021 [7], regular assessments are mandatory for lower-secondary students to fulfill curriculum requirements. This necessitates the creation of suitable assessment tools aligned with these standards. The academic years 2021-2022 and 2022-2023 marked the initial implementation of the new curriculum for 6th-grade students, during which the standardized assessment process remained under development. Therefore, it is crucial to examine students' adaptation to the curriculum standards through summative tests. This adaptation involves evaluating students' adherence to mathematics standards and assessing how effectively the assessment tools measure their competencies.

This study was conducted to meet curriculum requirements by designing an end-of-year mathematics exam for 6th-grade students. Between March 1 and May 15, 2023, the research team developed a matrix and created the 6th-grade mathematics year-end exam. The test was administered to over 3,850 6th-grade students in a district within Vietnam's northern midland region. All participants completed the computer-based experimental test using the AEGlobal educational ecosystem (els.aeglobal.edu.vn). The purpose of this experiment was to determine the test's scope and evaluate students' adaptability to the constructed standards. Additionally, the study aimed to propose adjustments to the exam questions, moving towards standardizing the test to better align with the competency assessment goals outlined in the 2018 curriculum.

2. Content

2.1. Method

2.1.1. Designing end-of-year assessment standards for 6th-grade mathematics

In alignment with the requirements of the 2018 General Education Curriculum and its corresponding learning objectives, researchers developed an end-of-year assessment standard for 6th-grade mathematics. This standard outlines levels of proficiency and their corresponding behavioral indicators. The specific details of the end-of-year assessment standard for 6th-grade mathematics are presented in Table 1.

Table 1. Assessment standards for 6th-grade mathematics

Level	Description	Behavior indicators
Level 1 (L1)	- Recognize, recall, or describe learned content, applying it directly to solve familiar problems. - Corresponding activities at this level include identification, comparison, indication, listing, and more.	Identify integers within given numbers; Apply addition rules directly to add two fractions; Enumerate rays in a given diagram; etc.
Level 2 (L2)	- Connect and organize learned content to address problems with similar contexts. - Activities corresponding to this level involve interpretation, summarization,	Represent numbers on the number line; Transform expressions during mathematical operations; Differentiate between sets of

	retelling, rephrasing, and providing examples based on personal understanding.	numbers; Draw figures based on descriptions; etc.
Level 3 (L3)	- Apply acquired knowledge to solve new problems or provide reasoned responses in both academic and real-life situations. - Activities at this advanced level include model construction, presentation, experimentation, classification, application of principles (laws, theorems, propositions), role-playing, argumentation, critique, and drawing conclusions. Additionally, students are encouraged to create innovative products, fostering creativity within their learning and daily lives.	Apply calculation rules to swiftly solve mathematical expressions; Verify a point as the midpoint of a line segment in real-world scenarios; Solve problems by applying experimental probability knowledge; etc.

2.1.2. Designing table of specifications for 6th-grade mathematics end-of-year test

The test blueprint, or table of specifications, outlined below defines the structure of the 6th-grade mathematics end-of-year assessment. It was developed in alignment with the prescribed assessment standards for this grade level. The examination is designed to be completed within a 60-minute timeframe and consists of 30 multiple-choice questions. Educators have used this test blueprint as the basis for constructing the final year-end examination for the 6th-grade mathematics course.

Table 2. Table of specifications for 6th grade mathematics end-of-year test

Theme	**Content**	**L1**	**L2**	**L3**	**Number of items**
Natural numbers, integers	Set of natural numbers and integers	item1			1
	Math operations with integers	item2	item 13, item 22		3
Fractions, decimals	Solving for variables	item 5	item 14, item 23	item 30	4
	- Fundamental properties of fractions - Math operations with fractions and decimals	item 6	item 15	item 25	3
	Mixed numbers, percentages, and related problems.	item 7	item 16, item 24	item 28	4
Basic elements of geometry	The point lies between two points – the midpoint	item 8	item 17	item 29	3
	Ray	item 9	item 18	item 26	3
Angles	Measurement of geometric elements: - Line segment - Angle	item 10	item 19		2

Statistics	Data representation in tables	item 3	item 21		2
	- Picture chart - Column chart - Dual column chart	item 4, item 11	item 20		3
Probability	Experimental probability	item 12		item 27	2
Total number of items		12	12	6	30

2.1.3. 6th-grade Mathematics testing and analysis procedure

In alignment with Baker's framework [8], Item Response Theory (IRT) emerged in the 1970s to address the limitations of Classical Test Theory (CTT). One of CTT's key weaknesses was its reliance on item parameters, such as difficulty and discrimination, being dependent on the specific sample of participants. Additionally, CTT did not sufficiently account for the influence of items on measuring candidates' latent abilities. In contrast, IRT operates on a fundamental assumption: "If one individual possesses a higher ability than another, the likelihood of that person answering a given item correctly should be greater than for the other individual. Similarly, if one item is more difficult than another, the probability of anyone answering it correctly must be lower than the probability of answering the easier item correctly”. IRT ensures that a candidate’s ability estimation is independent of the specific test items or the sample of examinees used. This feature promotes fairness and consistency, allowing different groups of examinees to take different test versions while still achieving comparable results. IRT enables test designers to create precise and reliable assessments by providing detailed insights into both item characteristics (e.g., difficulty and discrimination) and examinee abilities. Furthermore, IRT facilitates test equating, aligning different test forms on a common scale. This ensures that scores from different versions of a test are directly comparable, thereby enhancing the validity of the assessment process [9], [10].

In this study, data was systematically collected following the administration of the test. Student responses were recorded in a binary format, with correct answers scored as 1 and incorrect answers as 0. The data was then analyzed using a two-parameter IRT model, implemented through R software, enabling a detailed assessment of each student's proficiency level. This approach provided valuable insights for both educational evaluation and curriculum development.

2.2. Results and discussion

2.2.1. Test and items analysis

**** Perspectives on standards***

The results obtained after analyzing the test responses through the two-parameter Item Response Theory (IRT) model are described in Table 3.

Table 3. Estimation results of item parameters

Item	Test participants (N)	Correct (M)	Difficulty (b)	Discrimination (a)
1	3850	0.576104	-0.37345	0.953891
2	3850	0.506234	-0.03784	0.923675
3	3850	0.827013	-1.8628	1.027036
4	3850	0.93039	-3.27774	1.373329
5	3850	0.705455	-1.29838	1.665046
6	3850	0.942597	-4.34573	2.174028

7	3850	0.891429	-2.95141	1.67348
8	3850	0.786234	-1.78081	1.465231
9	3850	0.322338	0.839705	0.799427
10	3850	0.838961	-2.254	1.511779
11	3850	0.90961	-3.42724	1.913152
12	3850	0.269351	1.241435	1.150287
13	3850	0.927532	-3.94124	2.108116
14	3850	0.122857	1.971823	-0.12909
15	3850	0.917143	-3.82093	2.182141
16	3850	0.55013	-0.21885	0.577859
17	3850	0.72961	-1.27307	1.216282
18	3850	0.252987	1.122498	0.408507
19	3850	0.209351	1.445262	0.655182
20	3850	0.871429	-2.38619	1.232576
21	3850	0.927792	-3.51107	1.67785
22	3850	0.222078	1.319784	0.498934
23	3850	0.178442	1.534973	-0.1594
24	3850	0.607792	-0.5776	1.216404
25	3850	0.352208	0.665356	0.659159
26	3850	0.234545	1.182825	-0.00515
27	3850	0.762597	-1.72409	1.697101
28	3850	0.383896	0.474761	0.120561
29	3850	0.474805	0.112121	1.070785
30	3850	0.482338	0.074984	1.00445

Questions with higher difficulty levels correspond to lower probabilities of candidates answering them correctly. Baker [8] classifies question difficulty into five levels, corresponding to specific difficulty parameters, as shown in Table 4.

Table 4. Item difficulty indices

Item difficulty	**Range**
Very difficult	
Difficult	
Moderate	
Easy	
Very easy	

Additionally, the item discrimination level indicates its ability to differentiate among candidates. Typically, a positive discrimination level is expected. However, in cases where questions are flawed or incorrectly designed, the discrimination level may be negative. Items with higher positive discrimination values result in greater differences in the probabilities of correct answers between high- and low-ability candidates. In other words, items with higher discrimination levels more effectively distinguish between candidates. Baker categorizes discrimination levels into five groups: very good, good, moderate, poor, and very poor (Table 5).

Table 5. Item discrimination indices

Item discrimination	Range
Defective	
Very poor	
Poor	
Moderate	
Good	
Very good	

To evaluate item difficulty and discrimination levels, we conducted a cross-reference using R software, following the classification recommendations outlined here. Among the items, 14 exhibited acceptable difficulty levels across three categories (L1, L2, and L3) with moderate discrimination levels (starting from 0.65). These items are numbered 1, 2, 3, 5, 8, 9, 12, 17, 19, 24, 25, 27, 29, and 30. Items that showed very poor discrimination levels or had negative discrimination values rendering them ineffective for distinguishing among candidates were identified as defective. These defective items, numbered 14, 16, 18, 22, 23, 26, and 28, require further review and potential modification or replacement. The remaining items, categorized as very easy, should also be adjusted for better alignment with the assessment objectives.

**** Item characteristic curve***

In visually evaluating item quality, characteristic curves, and information function graphs can be used. According to Lam [11], an item's characteristic curve depicts the relationship between the probability of a correct response and the candidate's ability level. The figure below shows characteristic curves for select items (items 3, 5, and 26) from the exam.

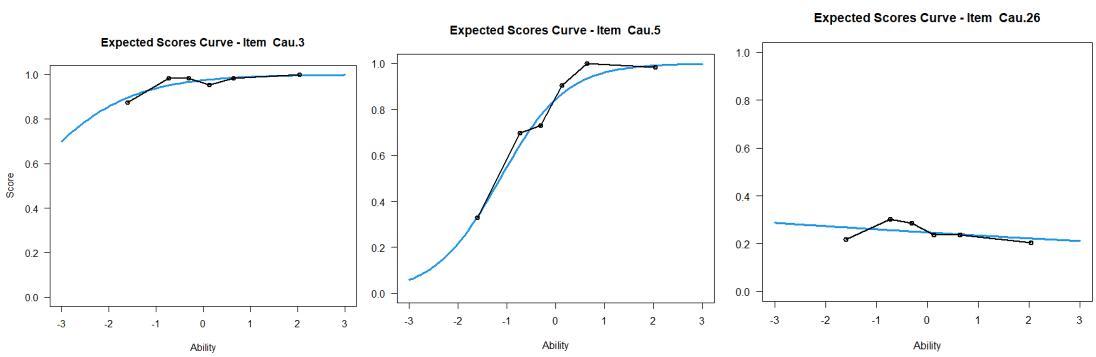


Figure 1. Item characteristic curve of select items

The steepness of an item's characteristic curve reflects its discrimination power: the steeper the curve, the greater the discrimination. Among the three curves, item 5 demonstrates the highest discrimination. Item 3 shows moderate discrimination, effectively distinguishing only among candidates with lower ability levels (ranging from -3 to 0). Interestingly, the characteristic curve of item 26 suggests that candidates with lower ability levels have a higher probability of answering correctly than those with higher abilities, indicating a design flaw in the item. When comparing these results with the discrimination levels in Table 1, item 3 is classified as having low difficulty and moderate discrimination, item 5 as having low difficulty and high discrimination, and item 26 as exhibiting negative discrimination.

**** Test information function***

The test information function is a critical tool for evaluating multiple-choice tests. The item information function indicates the extent to which each item contributes to measuring a candidate's ability. The figure below illustrates the test information function of select items from the designed exam.

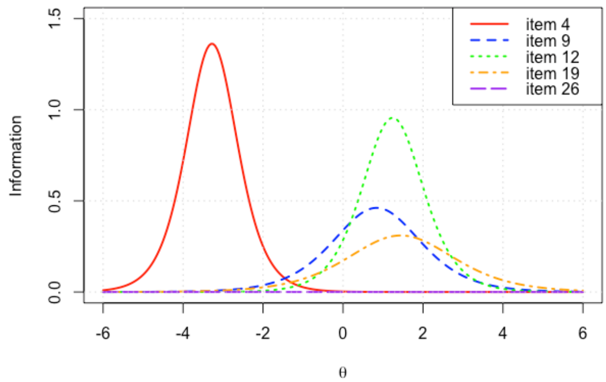


Figure 2. Item information function of select items

Item 4 provides significant information for assessing candidates with low abilities but contributes virtually no information for those with high abilities, indicating that it is a very easy item. Within the same ability range, item 12 offers more discriminating information than item 19, although its assessment range is narrower. Item 26 provides almost no information for evaluating candidates' abilities.

The total test information function, derived from all items in the test, represents the test's overall informational value. A multiple-choice test can be designed to measure a specific ability range with the highest possible precision. Depending on the test's assessment objectives, the desired shape of the test information function also referred to as the target information function is determined. Suitable items are then selected or modified to ensure that the test information function aligns with the desired pattern. Figure 3 illustrates the graph of the test information function.

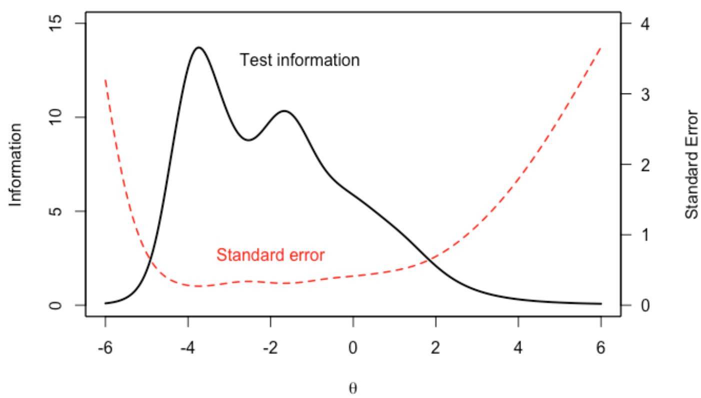


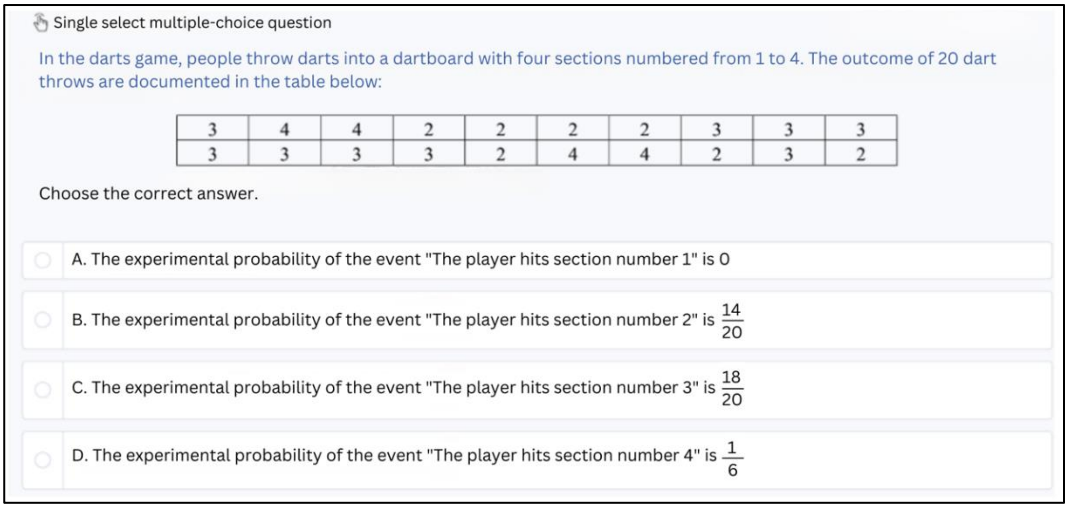
Figure 3. Test information curve

The information function curve for this test highlights a focus on assessing candidates with low abilities (ranging from -4 to nearly 2) while providing limited information for distinguishing candidates in the high-ability group. The lower portion of the information function graph represents the standard error of ability measurement: where the information value is high, the standard error is small. Thus, the test effectively evaluates the extent to which it meets the requirements specified in the 6th-grade curriculum. However, to improve discrimination and the assessment of high-ability students, the research team proposes specific modifications to the test, as outlined in Section 3.3.

2.3. Proposed adjustments to the multiple-choice test

Based on the earlier analysis, the research team suggests several modifications to enhance the test's quality and alignment with the following evaluation objectives:

(1) For items with acceptable levels of difficulty and discrimination (items 1, 2, 3, 5, 8, 9, 12, 17, 19, 24, 25, 27, 29, and 30), reconsider assigning appropriate difficulty levels to match the test blueprint or adjust the content accordingly. For example, item 27 was evaluated at level 3 (difficult) in the blueprint but had a difficulty parameter $b = -1.72409$, equivalent to an easy item.



Single select multiple-choice question

In the darts game, people throw darts into a dartboard with four sections numbered from 1 to 4. The outcome of 20 dart throws are documented in the table below:

3	4	4	2	2	2	2	3	3	3
3	3	3	3	2	4	4	2	3	2

Choose the correct answer.

A. The experimental probability of the event "The player hits section number 1" is 0

B. The experimental probability of the event "The player hits section number 2" is $\frac{14}{20}$

C. The experimental probability of the event "The player hits section number 3" is $\frac{18}{20}$

D. The experimental probability of the event "The player hits section number 4" is $\frac{1}{6}$

Figure 4. Item 27's content

Upon reviewing the item content, it was determined that, although students were required to read data from a chart and apply the formula for the experimental probability of an event, the item's design significantly increased the likelihood of selecting the correct answer. The item was presented as a single-select multiple-choice question, with the correct answer being option A, making it easy for students to guess using a trial-and-error approach. Therefore, we propose modifying the question using one of the following methods:

- Change the item format to a fill-in-the-blank question.

- Retain the single-select format, but change the position of the correct answer and adjust the event described in the correct answer.

(2) For items assessed as very easy (items 4, 6, 7, 10, 11, 13, 15, 20, and 21), adjustments should be made to increase their difficulty and better align with the test blueprint. For instance, item 6, as illustrated below, was evaluated as very easy. Teachers can modify this question by altering the fractions in the calculation to make it more challenging.

Figure 5. Item 6's content

(3) For items with poor discrimination or design flaws, consider replacements. For instance, item 14 showed negative discrimination.

Figure 6. Item 14's content

Upon closely examining the item content, the research team identified potential challenges for students as follows:

- Content Knowledge: The item required the application of algebraic transformations, specifically the transposition rule. However, this content is only introduced in a general form in the 7th-grade curriculum, making it potentially unfamiliar to 6th-grade students.

- Questioning Technique: The item did not directly ask for the value of xx, which may have caused students who did not read the question carefully to select misleading answers. To enhance the quality of the test and ensure alignment with assessment standards and evaluation objectives, this item should be replaced.

3. Conclusions

To effectively assess students' adaptability to the curriculum, developing standardized assessment tools and tests is essential. These tools, when designed in accordance with established standards and analyzed using scientific methods, provide valuable insights into student performance and adaptability. The data gathered offers critical feedback on teaching methodologies and test design effectiveness, enabling educators to align their strategies with students' actual needs.

The findings from this research indicate that students successfully adapted to both the computer-based testing format and the curriculum's required learning outcomes, as outlined in the GCE 2018 framework. However, a significant challenge persists in differentiating between high-achieving students, a common issue in Vietnam's educational system, where high examination scores are frequently awarded.

Additionally, item analysis revealed a need for test design adjustments, such as rephrasing questions, modifying question types, and aligning assessments more closely with intended teaching and learning outcomes. These findings emphasize the importance of standardizing question design and development to ensure accurate evaluation of student competence. Moving forward, assessments must be continuously refined to better reflect students' true capabilities, rather than relying solely on teacher experience or conventional practices.

REFERENCES

[1] Graham SM, (2002). Are Core Learning Outcomes 'Standards'?. *Queensland School Curriculum Council*.

[2] Graham SM, (2009). Defining Standards for the 21st Century, In: Claire WS and Joy C. *Educational Assessment in the 21st Century: Connecting Theory and Practice*. Springer Netherlands, p. 263-286. doi: 10.1007/978-1-4020-9964-9_14.

[3] OECD, (2013). *Learning Standards, Teaching Standards and Standards for School Principals: A Comparative Study*. doi: 10.1787/5k3tsjqtp90v-en.

[4] Council of Chief State School Officers, (2023), *Mathematics Standards Common Core State Standards Initiative*. https://www.thecorestandards.org/Math/.

[5] Singapore Ministry of Education, (2013). *Mathematics Syllabus - Primary One to Six*. https://www.moe.gov.sg/-/media/files/primary/mathematics_syllabus_primary_1_to_6.pdf.

[6] Vietnam Ministry of Education and Training, (2018). *Circular 32/2018/TT-BGDDT – 2018 General Education Curriculum* (in Vietnamese).

[7] Vietnam Ministry of Education and Training, (2021). *Circular 22/2021/TT-BGDDT – Guidelines for assessing high school students* (in Vietnamese).

[8] Frank BB, (2001). *The Basics of Item Response Theory* (Second Edition). https://eric.ed.gov/?id=ED458219.

[9] Margaret W, Hak PT & Tsung-Hau J, (2016). *Educational Measurement for Applied Researchers*. Springer Singapore. doi: 10.1007/978-981-10-3302-5.

[10] Gábor S, (2008). *Applying Item Response Theory in Language Test Item Bank Building*. Peter Lang. doi: 10.3726/978-3-653-01167-8.

[11] Thiep LQ, (2010). *Measurement in Education - Theory and Application*. Vietnam National University Press (in Vietnamese).